CAFCA

Chapter 6

Component Compatibility

© M. Zandee 1989,1996 All Rights Reserved.

February 1996

THIS PAGE INTENTIONALLY LEFT BLANK

VI. COMPONENT COMPATIBILITY

BIOGEOGRAPHIC ANALYSIS

AREA-CLADOGRAM

INTRODUCTION.

PURPOSE

In a biogeographic analysis we try to match the distribution of (several groups of) taxa over areas with the phylogenetic history of these (groups of) taxa. This matching will result in a diagram called an area-cladogram expressing the historical relationships among areas in a hierarchical way, in the same manner as a cladogram does for taxa. In case the matching is based on more than one study group of taxa the resulting diagram is called a generalised area-cladogram. The goal of historical biogeography is to document the history of the involvement of geographic areas or biotas in speciation events. By adopting this position one does not rule out the possibility that the 'same area' may have been involved in a variety of speciation events in different episodes (Brooks, 1990). In this respect, the analysis of the historical relations of areas or biotas can be quite different from that of species. Areas that inhabit species may represent different historical origins, i.e. may not be directly related (be a member of the same 'clade'), while species and monophyletic groups have only a singular origin (Cracraft, 1988; Sober, 1988)

To this end the phylogenetic relations (i.e., the in- and exclusion relations among clada) together with the distributional data of the taxa over the areas or biotas involved, are used as basic data to derive an area-data matrix for the areas occupied by the (terminal) taxa. The rules for the derivation of these area-data matrices and the method used in their subsequent analysis is the subject of this chapter.

ASSUMPTIONS

A few important assumptions underlie the derivation of an area-data matrix. In the literature these assumptions come under different numbers, viz. number 0, 1, and 2. The latter two are from Nelson & Platnick (1981; see also Page, 1988 a, b), the first, assumption zero, is from Zandee & Roos (1987; see also Wiley, 1988 a, b). I refer to this literature for an account on all details. All these assumptions regard the way in which problems caused by widespread taxa, or taxa lacking occurrence in one or more areas, or redundancy of distributional information, is handled.

Assumption zero is the least complicated one. It implies that in the event of biogeographic problems the reconstructed phylogeny of the taxa concerned has to be trusted a priori. Only afterward, when the biogeographic analysis is completed and problems as to the interpretations of widespread or missing taxa still abound, doubt can be raised against the phylogeny reconstruction for the taxa involved. Maybe some apparent monophyletic groups are not that monophyletic after all, and breaking them up while accepting alternative solutions

will subsequently dissolve some of the biogeographic stumble blocks that were in the way of a harmonious matching of biogeography with phylogeny.

Assumptions 1 and 2, however, as formulated originally by Nelson & Platnick (1981) in the opinion of Zandee & Roos (1987, but see Page, 1988 b) imply a doubt of the reconstructed phylogeny for taxa from the start onward, especially with regard to the widespread and/or missing taxa. This doubt has to be accounted for in the derivation of an area-data matrix for biogeographic analysis. The implications of assumption 1 are exemplified in the second example.

I refer to Page (1988 a, b) for a discussion of how assumptions 1 and 2 should be strictly implemented according to Nelson & Platnick (1981). In contrast, CAFCA does not manipulate the data matrix to allow for redundancy of distributional information under assumption 1 and 2. In figure 6.1 (after Page, 1988a, but see also fig 2 in Zandee & Roos, 1987) I show how widespread species and missing areas are handled under assumptions 1 and 2. There is one wide spread taxon, T3. The relations of area C with other areas might be correctly indicated by T3 or those of D, but not for both at the same time.

Under assumption 1 additional data are created to account for the placement of D on the branches showing open circles in case C is correct, as well as for C in case D is already correctly placed.

Under assumption 2 additional data are created to account for the placement of C and D, respectively, on the branches showing open as well as closed circles (i.e., all branches). The same type of placements takes place for missing areas.

I consider these manipulations unwarranted as they are a direct violation of Hennig's auxiliary principle, stating that we *should not assume homoplasy beyond necessity*. I think 'homoplasy' **is** assumed when taxon T3 is rejected as an indication of the relations of areas C and D with respect to each other and to the other areas. Moreover, these manipulations are inconsistent from a methodological point of view as we will never consider them to take place in cladistic character analysis with respect to either redundancy showed by character state distributions over taxa, or widespread character states (see also Wiley, 1988 a, b).

For a detailed discussion of an analysis using assumption 2 as implemented in CAFCA the user is referred to Zandee & Roos (1987). Here I mention only briefly their conclusion that this type of analysis is as unwieldy as unnecessary as their method (component compatibility) can resolve the conflicts in the area-data matrix as to widespread and missing taxa quite sufficiently using assumption zero only.

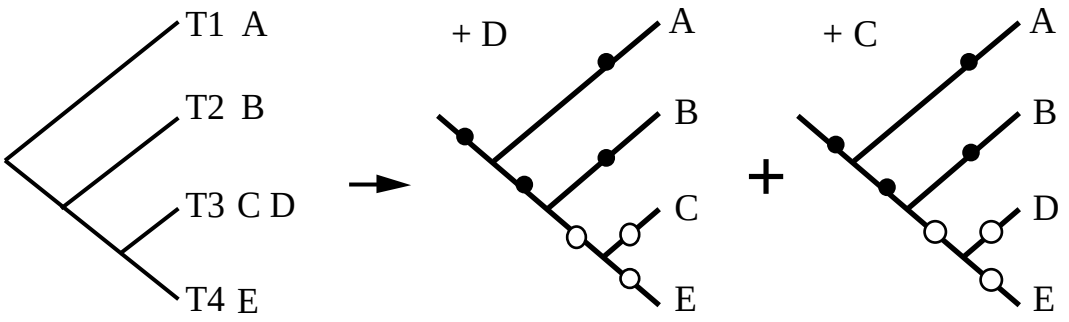


Figure 6.1 The consequences of assumptions 1 and 2 as to widespread species (after Page, 1988; see text for explanation).

BASIC DATA

No matter which assumption obtains, in all cases two items must be available to run a biogeographic analysis.

First comes a table (matrix) that summarises the distribution of terminal taxa over areas. The terminal taxa appear in the rows of the table; the areas occupied by these taxa appear in the columns. For each taxon occurring in a certain area we score a 1 in the appropriate entry in the table, otherwise we score a zero.

A distribution matrix for the taxa in our example PLANT may look like this:

```
1 1 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 1
0 0 0 0 1
```

The taxa are represented by the rows; the columns represent the areas where the taxa can be found. All columns together represent the distributional range for all taxa.

You may also use a transposed distribution matrix, like this

```
Aarea 10000
Barea 11000
Carea 00100
Darea 00010
Earea 00011
```

Areas are now in the rows, and taxa in the columns. In this matrix you may include the names for the areas (instead of offering them in a separate name-file, as is necessary with the upper distribution matrix). CAFCA can recognise the format you are offering (taxa x areas, or areas x taxa), as long as the number of taxa in the distribution matrix coincides with the number of terminal taxa in the cladogram.

Only one type of biogeographic problem occurs in this matrix, viz. widespread species. Taxon 1 occurs in both areas 1 and 2, and taxon 4 occurs in area 4 and 5. As a consequence, areas 2 and 5 have more than taxon from the same monophyletic group. I will deal with missing taxa and redundancy in the example of a generalised analysis.

Second comes a cladogram expressing the phylogenetic relations among the terminal taxa. This cladogram must be available either in parenthesis's notation or as a table with terminal taxa in the rows and an indication of their groupings (clada) in the columns. CAFCA provides these tables as a result from a primary or secondary analysis, so in most cases you do not have to worry about transcribing a diagram into a binary table. In our first example we use a cladogram from our primary analysis on PLANT (see chapter 3) to run a biogeographic analysis. The binary representation of the selected cladogram for PLANT (table 3.6) looks like this:

```
Aus  1 0 0 0 0 0 0 0 1
Bus  0 1 0 0 0 0 0 1 1
Cus  0 0 1 0 0 0 1 1 1
Dus  0 0 0 1 0 1 1 1 1
Eus  0 0 0 0 1 1 1 1 1
```

Taxa are represented by the rows, the columns indicate the grouping patterns present in the cladogram by means of additive binary coding. One of the

methodological problems we must deal with is now immediately clear. The additive coded columns in this matrix, indicating groupings of taxa, can hardly be considered to represent independent data items. The group {Dus, Eus} is not (logically) independent of {Cus, Dus, Eus}. Later on, we will see if and how a coding scheme using a partitioning vector, like we employed for binary coded multi-state characters, may help to solve this problem.

If you want to enter a cladogram, say, from the literature you must prepare a text file (ASCII only) presenting this cladogram either in parentheses format, like this (1,(2,(3,(4,5)))), or as a binary table, like the one shown above. Below, a tree-file as exported by PAUP is shown as another example.

```
#NEXUS

begin trees;  [Treefile saved Thursday, July 1, 1993  6:04 PM]
[!>Heuristic search settings:
>   Addition sequence: simple (reference taxon = Aus)
>   1 tree(s) held at each step during stepwise addition
>   Tree-bisection-reconnection (TBR) branch-swapping performed
>   MULPARS option in effect
>   Steepest descent option not in effect
>   Initial MAXTREES setting = 100
>   Branches having maximum length zero collapsed to yield polytomies
>   Topological constraints not enforced
>   Trees are rooted
>   Total number of rearrangements tried = 968
>   Length of shortest tree found = 18
>   Number of trees retained = 3
>   Time used = 1.12 sec
]
      translate
            1 Aus,
            2 Bus,
            3 Cus,
            4 Dus,
            5 Eus,
            6 Fus,
            7 Gus,
            8 Hus,
            9 Ius,
            10 Anc
            ;
tree PAUP_1 = ((((1,2),((6,7),8)),((3,4),(5,9))),10);
tree PAUP_2 = (((((1,2),8),(6,7)),((3,4),(5,9))),10);
tree PAUP_3 = ((((((1,2),8),6),7),((3,4),(5,9))),10);
end;
```

CAFCA can read such tree-files. On the other hand, a file with contents as shown below suffices to generate the cladogram shown left in binary coding (do **not** forget the closing semi-colon after each cladogram) The text after each slash (/) is just a comment and may be omitted:

```
/ One cladogram in parenthesis
/ format.
(1,(2,(3,(4,5))));
```

EVALUATION

Before we look at the examples we must say something in general about the interpretation of an evaluated area-cladogram. As monophyletic groups cannot be 'real characters' for areas in the same sense as intrinsic characters from morphology, anatomy, etc... are for taxa, some of our usual interpretations of what happens with characters during phylogeny have changed.

As 'real' characters for areas are absent we cannot speak of evolutionary novelties for areas. The analogy of an evolutionary novelty in an area-cladogram is a single origin explanation for the distribution of a monophyletic taxon

over areas. If an area splits in daughter areas, also the taxa (for not just one but generally all groups to be considered) occupying these areas must split in daughter taxa in order to fit a same single origin explanation. This explanation then conforms to the model of vicariance biogeography. A species not responding to a vicariance event will result in a so-called widespread distribution for that species.

All other auxiliary explanations needed to fit the observed distributions in terms of area-cladograms are ad hoc and as such analogies of events of homoplasy. What we call a parallelism when dealing with real character states equates a dispersal event when dealing with areas and monophyletic groups, and in the same manner a reversal equates extinction.

The quality of an area-cladogram is measured by the degree in which it explains the observed distribution of taxa in terms of vicariance events, relative to the number of ad hoc statements. To that end we can use the same criteria as applied in measuring the explanatory power of ordinary cladograms.

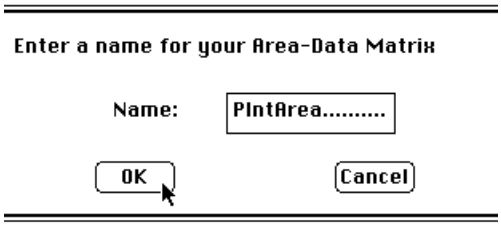
EXAMPLE OF AN ANALYSIS USING ASSUMPTION ZERO.

I now give you a hands-on introduction to a biogeographic analysis. After we have run the analysis and printed the results we will discuss the different items.

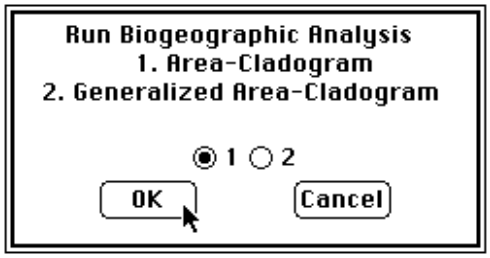
TUTORIAL

1. Select **Biogeographic Analysis** from the **Run** menu.

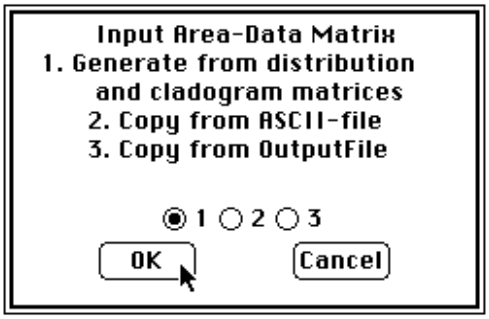
2. In the next dialog, type a name, **PlntArea** for example, for the area-data matrix to be used in this run by CAFCA.



3. Click **OK** for default **1** (**Area-Cladogram**) in the next dialog.

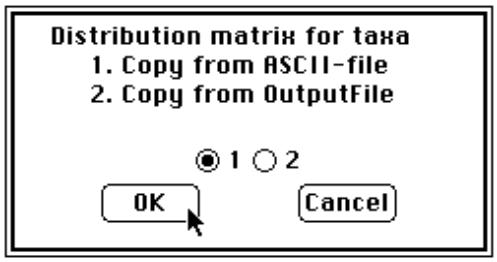


4. Click **OK** for the default button **1** (**Generate from distribution and cladogram matrix**) in the dialog prompting for the source of the area-data matrix .



In a biogeographic analysis we use binary or other representations for both the cladogram and the distribution of taxa. In this first example, CAFCA must read a distribution matrix from an ASCII file as it is not yet present as a saved object in the CAFCA OutputFile system containing the results of the analysis of PLANT.

5. Click **OK** for default value **1** (**Copy from ASCII file**) in the dialog box presenting different sources for a distribution matrix.



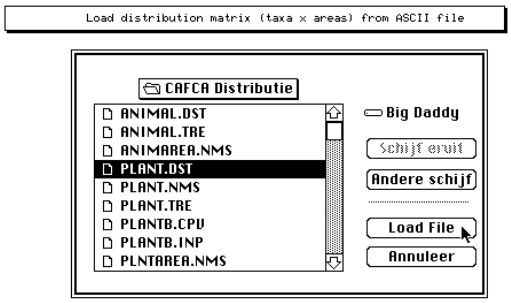
6. Select the appropriate name for a distribution matrix in the next file selector box and click **Load File**. You may use PLANT.DST, which provides no names for the areas,

```
1 1 0 0 0
0 1 0 0 0
0 0 1 0 0
0 0 0 1 1
0 0 0 0 1
```

or the transposed form in the file PLANT.ARS, the one including the names for the areas.

```
Aarea 1 0 0 0 0
Barea 1 1 0 0 0
Carea 0 0 1 0 0
Darea 0 0 0 1 0
Earea 0 0 0 1 1
```

(In the latter case you will not be prompted to provide names for the areas as in step **12**).

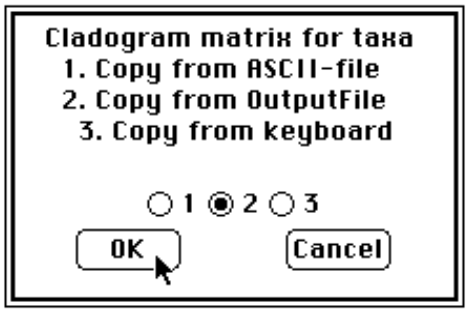


We will use a cladogram from our primary analysis on PLANT to run a biogeographic analysis. The binary representation of the selected cladogram for PLANT (table 3.6) looks like this:

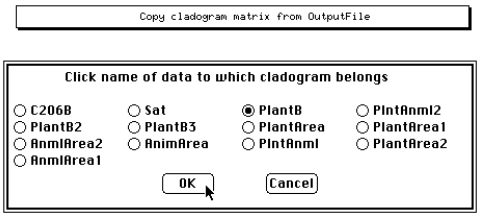
```
Aus  1 0 0 0 0 0 0 0 1
Bus  0 1 0 0 0 0 0 1 1
Cus  0 0 1 0 0 0 1 1 1
Dus  0 0 0 1 0 1 1 1 1
Eus  0 0 0 0 1 1 1 1 1
```

Taxa are represented by the rows, the columns give the grouping patterns present in the cladogram. The program already made such a cladogram matrix for you during the primary analysis. It is present among the objects you saved in an OutputFile after finishing your primary analysis on PLANT, and we can copy it from this OutputFile.

7. Click **2** (**Copy from OutputFile**) as a source for a cladogram matrix in the next dialog.

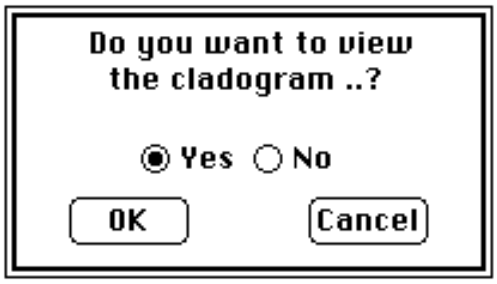


8. In the next dialog box, click a name of the data matrix to which the cladogram belongs, PlantB in this example, or any other name that you used for your data matrix in the first example of a primary analysis.

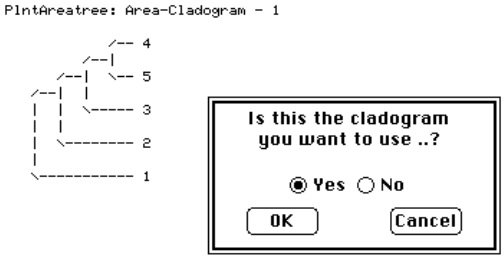


In case you did not save the output of the primary analysis in an OutputFile, as exemplified in chapter 3, you can click **Cancel** and redo step **7.** Click **1** (**Copy from ASCII file**) in dialog step **7** and in the next step (file selector box) select PLANT.TRE from the examples folder on your distribution disk.

9. Click **OK** in the next dialog if you want to see the cladogram.

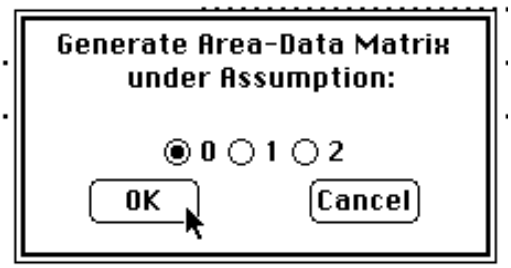


10. Click OK in next dialog if cladogram is indeed correct. If you click **Cancel** step **7** will appear again.

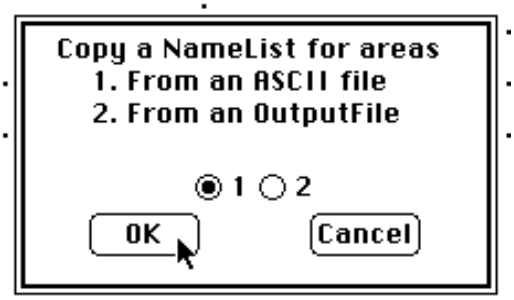


From these two basic pieces of information an area-data matrix will be composed by substituting taxa in the cladogram matrix by corresponding areas from the distribution matrix. The derivation of the area-data matrix is constrained by one of the possible assumptions 0, 1 and 2.

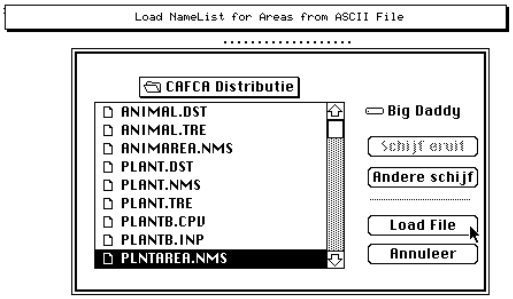
11. Click **OK** for the default value **0** in the **Assumption** dialog.



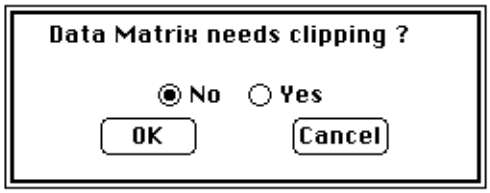
12. Click **OK** for default value **1** (**From an ASCII file**) in the **Copy a namelist for areas** dialog.



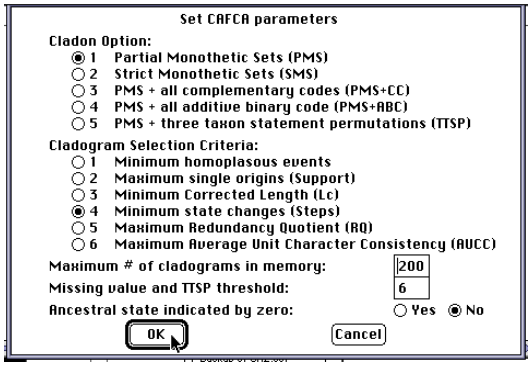
13. Select the appropriate filename for the names of the areas in the next file selector box and click **Load File**.



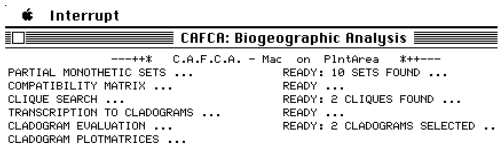
14. Click **OK** for the default **No** in the **Data matrix needs clipping ?** dialog box.



15. Take all defaults in the **Set CAFCA Parameters** dialog box.

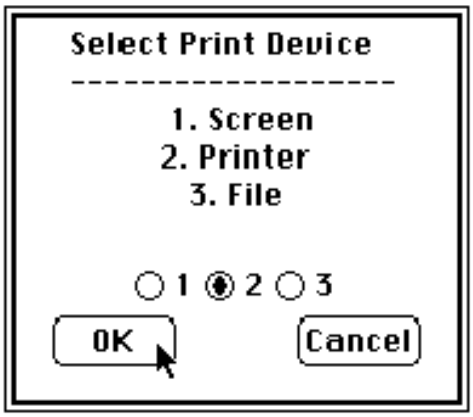


16. The biogeographic analysis now starts running. You can follow its progress on the screen.



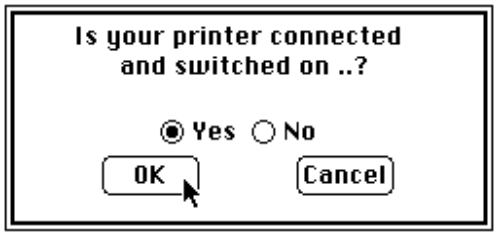
17. After the analysis is finished, as shown by the elapsed time message on the screen, select **All Results** from the **Print** menu.

18. Click **2** for **Printer** in the **Print Device** dialog.



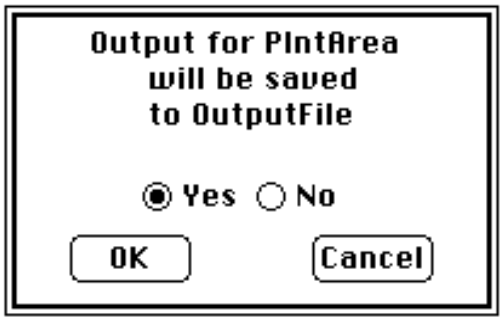
19. Click the printer of your choice in the **Select Printer** dialog box.

20. Click **OK** in next dialog if everything is all right with your printer.

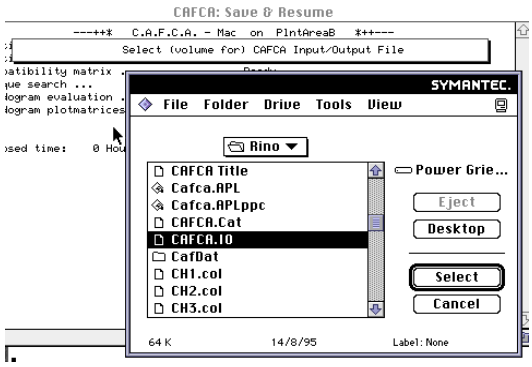


21 After printing you may want to save all your results for later use or inspection. Select **Save & Resume** in the **Output-File** menu to do so.

22. Click **OK** in the **Output for PlntArea will be saved to OutputFile** dialog box if you really want to save your results this way. It may come in handy for our next examples.



23. Select an Outputfile in the following file select box.



After you clicked **Select** CAFCA will start saving your results to the Outputfile.

DISCUSSION OF RESULTS.

I will now present a discussion of the results obtained in this biogeographic analysis. The area-data matrix (areas x clada) for our example is given in table 6.1.

How did this matrix come about? Starting from the binary representation of the cladogram the entries in each column are replaced by corresponding entries from the distribution matrix.

Area–Data Matrix (binary) : PlntArea

	1	2	3	4	5	6	7	8	9
Aarea	1	0	0	0	0	0	0	0	1
Barea	1	1	0	0	0	0	0	1	1
Carea	0	0	1	0	0	0	1	1	1
Darea	0	0	0	1	0	1	1	1	1
Earea	0	0	0	1	1	1	1	1	1

(multi–state) : PlntArea

	1	2	3	4	5	6
Aarea	1	0	0	0	0	1
Barea	1	1	0	0	0	2
Carea	0	0	1	0	0	3
Darea	0	0	0	1	0	4
Earea	0	0	0	1	1	4

Column Partitioning Vector :
1 1 1 1 1 4

Table 6.1 Area-data matrix for a biogeographic analysis.

Take for instance the sixth column in the cladogram matrix.

```
Aus  1 0 0 0 0 0 0 0 1
Bus  0 1 0 0 0 0 0 1 1
Cus  0 0 1 0 0 0 1 1 1
Dus  0 0 0 1 0 1 1 1 1
Eus  0 0 0 0 1 1 1 1 1
```

This column indicates the grouping of taxa 4 and 5. The distribution matrix shows that these taxa (row # 4 and 5) occur in the areas (columns) 4 and 5. Substituting areas for taxa in the cladogram matrix therefore renders an identical sixth column in the binary area-data matrix.

Now take the first column of the cladogram matrix as an example. We see that taxon 1 is indicated. Looking at taxon 1 in the distribution matrix makes clear that this taxon occurs in the areas # 1 and 2. Substituting column 1 in the cladogram matrix from taxon to area indication gives us column 1 as depicted in the binary area-data matrix (table 6.1).

By treating all columns of the cladogram matrix in this manner a binary **area-data matrix** is build, where **rows indicate areas and the columns indicate the monophyletic groups (as depicted in the cladogram) in these areas**.

There is also a multi-state representation of the data matrix, because we can justify a partition vector for the binary columns. The additive coded columns in the binary matrix represent the hierarchical structure of the cladogram, and therefore they are interdependent; they must be seen as the character states of a multi-state character. This character is treated as an ordered character in the analysis. However, as we shall see this does not make the interpretation of state changes in the area-cladogram any easier, especially in the case of many wide-spread taxa or redundancy.

The remainder of the output is very much the same as for a primary character analysis, so the explanation as given in chapter 3 also applies here.

Partial Monothetic Sets of areas in PlntArea

1	1
2	2
3	3
4	4
5	5
6	1 2
7	4 5
8	3 4 5
9	2 3 4 5
10	1 2 3 4 5

Partial Monothetic Sets of Monophyletic Groups (= Components) in PlntArea

1	
2	2
3	3
4	
5	5
6	1
7	4 6
8	7
9	8
10	9

Table 6.2 Components from the area-data matrix PLNTAREA.

The building-blocks for area-cladograms, or components (table 6.2), are derived from the area-data matrix in the same way as clada are from a normal data matrix. However, in a biogeographic analysis the recognition of components is restricted by using the partial definition of monothetic sets only. Components correspond to partial monothetic sets of areas, defined by the unique occurrence of at least one monophyletic group of taxa.

```
Character states on root
(= Start transf. ser.) for PLNTAREA
-----------------------------------
Rownumbers refer to index numbers of cladograms.
Column numbers refer to columns of multi-state data matrix.

      1  2  3  4  5  6
    ---------------------
 1 |  0  0  0  0  0  2
 2 |  1  0  0  0  0  1
```

Table 6.3 Indication of ancestral monophyletic groups present in ancestral area.

Character states on the root (table 6.3) in the case of a biogeographic analysis can be interpreted as ancestral monophyletic groups present in an ancestral area.

Both area-cladograms explain the distributions of taxa equally well considering all but one criterion, RQ, which prefers area-cladogram # 1 (table 6.4). Area-cladogram # 2 has the same topology as the cladogram for taxa we used as a starting point; area-cladogram #1 is different.

```
Selection criteria for cladograms of: PlntAreaÆB
Column numbers refer to numbers of cladograms
-----------------------------------------------
Row 1 : Total number of homoplasous events
Row 2 : Total number of single origins (Support)
Row 3 : Corrected Extra Length (x1000; CEL: Turner + Zandee)
Row 4 : Total number of state changes (S: Steps)
Row 5 : Redundancy Quotient (x1000; RQ: Zandee + Geesink)
Row 6 : Rescaled Redundancy Quotient (x1000; RQc)
Row 7 : Consistency Index (x1000; CI), with autapomorphy correction
Row 8 : Rescaled Consistency Index (x1000; RC: Farris)
Row 9 : Average Unit Character Consistency (x1000; AUCC: Sang)
Row 10: Homoplasy Distribution Ratio (x1000; HDR: Sang)
Row 11: Compatible Character State Index (x1000; CCSI: Zandee)

         1     2
    ------------
 1 |     0     0
 2 |     8     8
 3 |     0     0
 4 |     8     8
 5 |   522   510
 6 |   162   141
 7 |  1000  1000
 8 |  1000  1000
 9 |  1000  1000
10 |  1000  1000
11 |   643   643

No-Order Limit for Steps, Extra Steps, RQ, and CI:

   S   ES   RQ   CI
-------------------
  10    2  429  800
```

Table 6.4 Selection criteria for area-cladograms from PLNTAREA.

There is no list of apomorphies for area-cladograms as ‘real characters’ for areas do not obtain. The pattern of putative vicariance events can be deduced from the list of state changes accompanying the area-cladogram (table 6.5).

Beware, however, that these events must coincide with those postulated for other taxonomic groups to be established as general vicariance events. This may come about in the search for a generalized area-cladogram.

```
PlntArea: Area–Cladogram – 2

          /–– 4 Darea
        /––7
      /––8  \–– 5 Earea
   /––9  |
   |  |  \––––– 3 Carea
   |  |
   |  \–––––––– 2 Barea
   |
   \––––––––––– 1 Aarea

PlntArea: Area–Cladogram–2 :
STATE CHANGES
```

Character	Component	Change
1	8	1 –> 0
2	2	0 –> 1
3	3	0 –> 1
4	7	0 –> 1
5	5	0 –> 1
6	7	3 –> 4
	8	2 –> 3
	9	1 –> 2

Components refer to the list of monothetic sets of areas.

Table 6.5 Area-cladogram for PLNTAREA with corresponding state changes.

The following scenario, as derived from the list of state-changes, can be associated with cladogram # 2, and explains the distribution of the taxa in historical terms. The distribution of taxon {Aus} is interpreted as relict in Aarea and Barea. The speciation event that gave rise to {Aus} and {Bus,Cus,Dus,Eus} predated a vicariance event that in itself did not trigger any speciation (split 1 in fig. 6.2b). {Aus} went extinct in {Carea, Darea, Earea}, according to the state-change from 1->0 in character 1 for component # 8, may be at the time when {Bus} speciated from {Cus,Dus,Eus} in response to the break-up of {Barea} vs {Carea, Darea, Earea} (split # 2 in fig. 6.2c). Another distribution follows the fist order explanation of vicariance as the breaking up of {Carea, Darea, Earea} (split # 3 in fig. 6.2d) coincides with the speciation event for {Cus} and {Dus, Eus}. The (sympatric) speciation event that split up the ancestor {Dus, Eus} in {Dus} and {Eus} predated a break-up of {Darea, Earea} (split # 4 in fig. 6.2f), thus resulting in a wide-spread distribution for {Dus}.

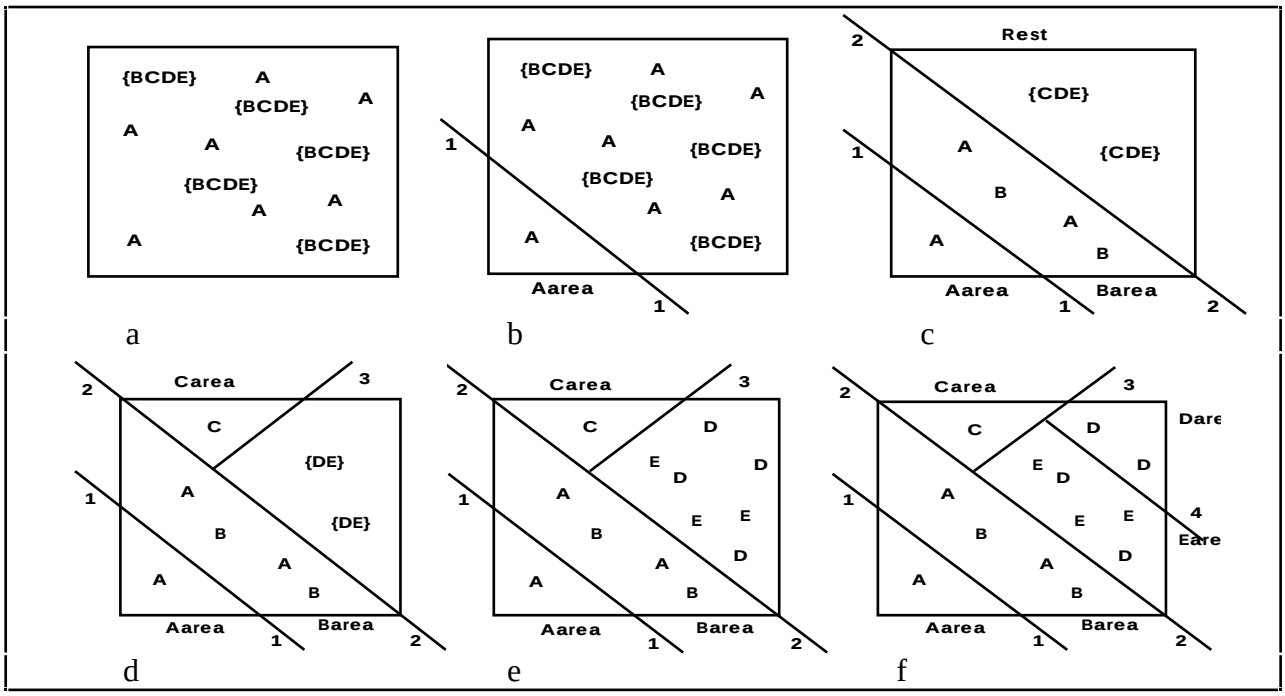


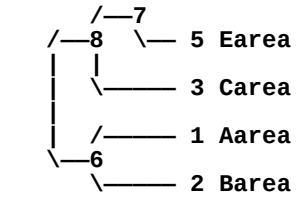
Figure 6.2 Possible vicariance and speciation events for the area-data PlntArea and cladogram 2. In (a) and (e) speciation predates vicariance, thus causing wide-spread distributions for A (=Aus) and D (=Dus).

An alternative explanation in terms of a scenario for the latter wide-spread distribution, involves a vicariance event that breaks up {Darea, Earea} and coincides with the speciation of {Dus, Eus} in {Dus} and {Eus}, followed by dispersal of {Dus} into {Earea}.

It is clear that this scenario (any scenario ?) adds ad hoc elements that are not present in the area-cladogram or its list with state-changes. These ad hoc elements are needed to reconcile the cladogram of the taxa involved with the area-cladogram for the areas, in terms other than vicariance. In this case these elements are the speciation events that do not coincide with the break-up of areas, as well as the break-up of areas that do not trigger speciation. The ad hoc elements as such do not cause extra steps in the area-cladogram, as apparently they are not part of the area-data matrix. One may ask whether they should not be counted one way or another (weighted ?), to be used as an extra criterion to select area-cladograms. As the area-cladogram is not estimated independently from the taxon-cladogram but is derived from it, a comparison of these two diagrams can neither serve as a means of detecting these ad hoc elements nor as a way to decide which set of ad hoc elements is the more likely one, in the way an independent estimate (geological data ?) of the area-cladogram might. I will discuss the latter possibility when dealing with host-parasite co-evolution.

The other area-cladogram (# 1), which is the better one with respect to the RQ, offers an alternative explanation as to the history of the areas involved.

PlntArea: Area—Cladogram — 1



Character	Component	Change
1	6	0 –> 1
2	2	0 –> 1
3	3	0 –> 1
4	7	0 –> 1
5	5	0 –> 1
6	1	2 –> 1
	7	3 –> 4
	8	2 –> 3

PlntArea: Area—Cladogram—1 : STATE CHANGES

Components refer to the list of monothetic sets of areas.

Table 6.6 Alternative most parsimonious area-cladogram for PlntArea.

Almost all explanations to be given here for the present-day distribution of the taxa involved are of the first order, i.e. follow a vicariance model. The first speciation event in PLANT, i.e., {Aus, Bus} vs {Cus, Dus, Eus} (split # 1 in fig. 6.3a) did coincide with a vicariance event. There was a break-up of the original area in separate biota's. One of the next speciation events, i.e., {Bus} vs. {Aus} may have coincided with a break-up of the area with a dispersal of {Aus} in the aftermath, or, as depicted (split # 2 in fig. 6.3c) there was a speciation event first and a break-up of the area later. The next vicariance event (split # 3 in fig. 6.3d) that broke up {Carea} from {Darea, Earea} again coincides with a speciation event; the one that separated {Cus} from {Dus, Eus}.

Only for the wide-spread distributions of {Aus} and {Dus} some allowances must be made. This scenario is, in terms of the area-data matrix concerned, just as parsimonious as the one involved with area-cladogram # 2. It can also be considered just as parsimonious in terms of the scenarios involved, as each scenario has the same type and number of ad hoc elements, needed to explain the widespread distributions of {Aus} and {Dus}.

Again, an alternative scenario involving dispersal after the break-up of the areas instead of sympatric speciation before the break-up, is possible as to the explanation of the wide-spread distribution of {Dus} in {Earea} and {Darea}.

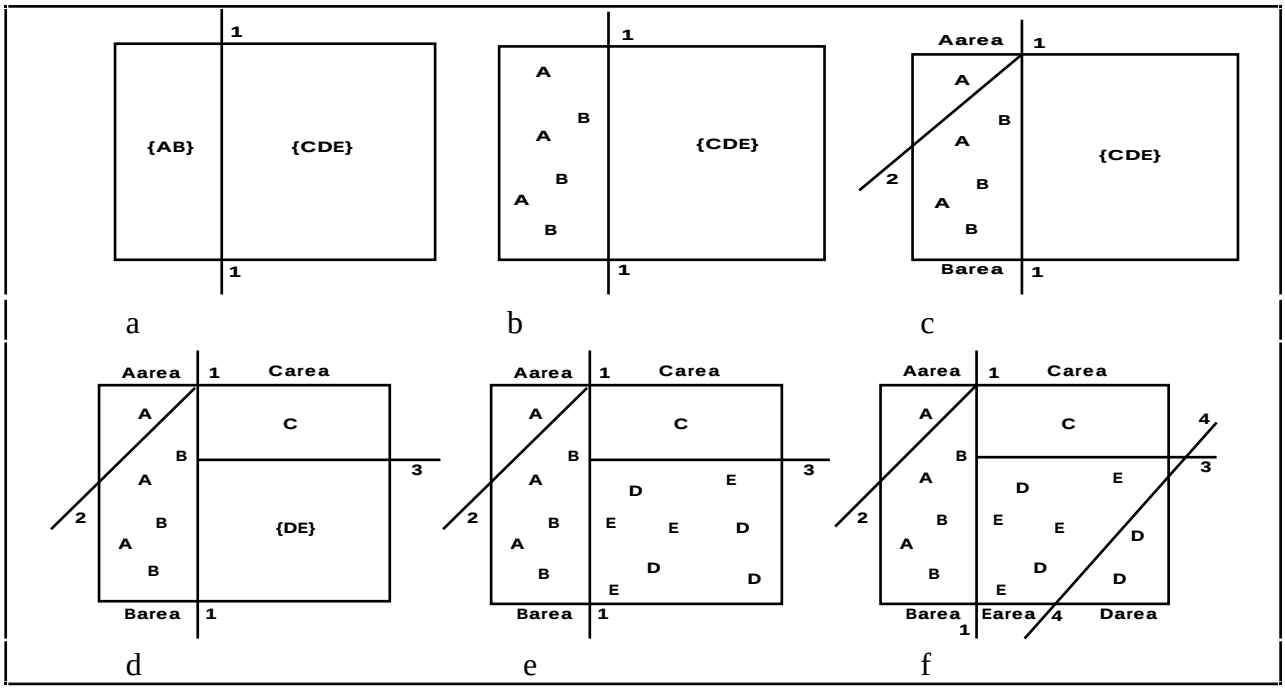


Figure 6.3 Possible vicariance and speciation events for the PlntArea cladogram 1. In (b) and (e) speciation predates vicariance, thus causing wide-spread distributions for A (=Aus) and D (=Dus).

A CRITIQUE OF CCA.

Page (1990, 1993) describes what in his opinion is one of the major drawbacks in *Component Compatibility Analysis* (CCA as defined by Zandee & Roos, 1987) and *Brooks Parsimony Analysis* (BPA,; Wiley 1988 a, b; Brooks, 1990), as already indicated by Zandee and Roos (1987) and Page (1987). The drawback occurs due to area-cladogram optimization. As area-cladograms are optimized in the same manner as normal cladograms, the dispersal of a terminal taxon or a monophyletic group also forces the dispersal of its ancestors, given a strictly binary area-data matrix. According to Page this drawback is caused by the difficulty to combine both vertical (inheritance) and horizontal (dispersal) transmission in a single method. Zandee & Roos (1987, p 312) give the following description of the problem: "Given a contradiction regarding a certain column in the data matrix with respect to a particular area-cladogram, the same contradiction will occur in all other columns that include the areas indicated by the affected column. It follows that contradictions shown by a data matrix with regard to a particular area-cladogram may not be independent."

The following example serves to illustrate the problem and is taken from Page (1993, fig 10, illustrating 'horizontal transmission' in the gene vs taxon or parasite vs host case; fig 3 in Page, 1990, p 126, deals with dispersal in biogeography. Both examples are available in the Xmpls folder on your distribution disk). There are five taxa (or hosts in the case of a problem in co-evolution). Their phylogenetic relationships are described by the following cladogram (a simple hennigian comb)

Aus 100000001
Bus 010000011
Cus 001000111
Dus 000101111
Eus 000011111

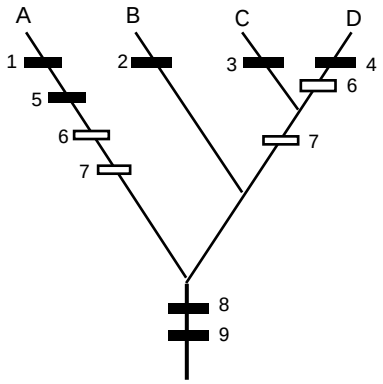
These taxa are distributed over just four areas (or taxa in the case of parasite or gene [co-]evolution), with Aus and Eus both occuring in the same area.

Aarea 10001
Barea 01000
Carea 00100
Darea 00010

The (area-) data matrix resulting from these sets of information looks as follows (table 2 in Page, 1993):

Aarea 100011111
Barea 010000011
Carea 001000111
Darea 000101111

Page (1990, 1993) in illustrating the workings of BPA, maps this matrix, which he still calls a cladogram, onto the cladogram of the areas (or hosts). This mapping, or optimization, results in the picture giver below (Page, 1993, fig 10b). The numbers refer to the columns of the (binary) area-data matrix. The 'drawback' of the method is illustrated by the 'dispersal' of the ancestors of taxon # 5, viz. clada # 6 and 7, into area A. For the ad-hoc element (dispersal) to be accounted for, in an extra step, the presence of either cladon # 6 or # 7 on the branch to A suffices, the other is superfluous. The point made by Page (1990, p 127) is that "... only one or the other of the ancestor-descendant pair need make the journey."

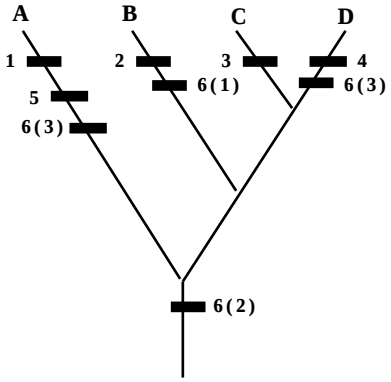


In its recent updates (vs 1.3c etc.), CAFCA avoids the problem as drawn above. CAFCA is able to do that because it no longer treats the columns representing the hierarchical structure of the cladogram (of the taxa) as independent. These columns are treated as a set of additive binary coded characters and substituted by an ordered multi-state character in the multi-state expression of the area-data matrix. This multi-state expression for the example given above looks as follows:

Aarea 100013
Barea 010001
Carea 001002
Darea 000103

It is clear that in Page's table 2 (1993) there are no 4 *different columns* that need to be optimized on the area-cladogram. Actually, there are only *3 different codes* representing the nested sets of monophyletic taxa (column 8 and 9 in the binary area-data matrix are identical). By looking rowwise instead of columnwise, we again see that there are only three different binary codes (those

of Aus and Dus are identical as far as the additive coded part is concerned; see also Brooks, 1990, for indicating species by codes in BPA). This fact is expressed in the multi-state area-data matrix (CAFCA makes the following column partition vector for the binary area-data: 1 1 1 1 1 4). When we map the multi-state area-data on the (area)cladogram given by Page (1993) as the correct one, by treating it as a user-tree, we get the following picture:



This picture shows what really happened; just one dispersal event in A, involving only one, and not both of the ancestors. For this example at least, and most likely also in general, the present version of CAFCA shows no signs of the drawback described by Page as inherent to component compatibility (and BPA). As I showed above, the real drawback is in the neglect of the interdependence of the columns in the area-data matrix that depict the hierachical structure of the taxon-cladogram involved.

Note, however, that CAFCA finds another diagram as the result of its biogeographic analysis (as does PAUP for BPA). As Page's example is really about hosts and parasites, it is only a matter of course that the cladograms shown in his example are independent estimates of the branching events in the evolutionary history of both groups involved, in contrast to the way in which area-cladograms are normally derived. At the end of this chapter I will deal with another of Page's examples concerning co-evolution in a host-parasite relationship to show how these independent cladograms are treated in a group- or component compatibility analysis of the problem, by mapping the host cladogram (in terms of its host-data matrix) onto the parasite cladogram through user-tree evaluation.

Ronquist & Nylin (1990) illustrate the same problem as Page (1990, 1993) does, using data from Brooks (1990, table 11, presented below).

```
                1111111
      1234567890123456
Aarea 1000000011000001
Barea 0100000110100011
Carea 0010001110010111
Darea 0001011110001111
Earea 0000111110000000
```

This example is more complicated than the preceeding one as no single area-cladogram results from the analysis. PAUP finds 3 MPT's with 17 steps for the data in Brooks' table 11, only two of which are depicted by Brooks (1990, fig 19; Ronquist & Nylin fig 11a,c; here fig. 6.4), and 4 MPT's with 19 steps when an all-zero out-area is added to the data matrix.

The same four areagrams (figure 6.4 + figure 6.5) are found and selected as MPT's by CAFCA (without the out-area added to the data) when we use the binary expression of the data . When we use the multi state expression of the

data matrix only the areagrams depicted in figure 6.4a and figure 6.5a are selected as MPT's.

The reconstruction in figure 6.4a produced by parsimony mapping indicates that taxon 16, 15, and 14 become extinct in area E. Ronquist & Nylin (1990) do not consider this a very plausible reconstruction of the history op the species assemblage. They state that if, for instance, taxon 16 became extinct in area E, than taxon 15 would not have originated in that area. Hence, taxon 15 could not have gone extinct in area E unless it colonized it separately after the extinction of taxon 16, something that is not suggested either by the data or by the cladogram. The same reasoning applies to taxon 14 with respect to 15 and 16.

An alternative MP mapping as in figure 6.4b shows extensive parallelisms concentrated in area E, indicating instances of dispersal to area E.

To eliminate the undesirable effects of equal weights parsimony mapping described above, Ronquist & Nylin (1990) propose a weighting scheme where the three different processes (colonization, exclusion, succesive specialization) that change association patterns are given weights relative to the probability of each of these events occuring, in order to allow them to enter the possible explanations of observed deviations of a null model (no change in traits that determine species association). I will not go into the merits or drawbacks of this particular weighting scheme here, but only compare Ronquist & Nylin's results with those obtained by CAFCA.

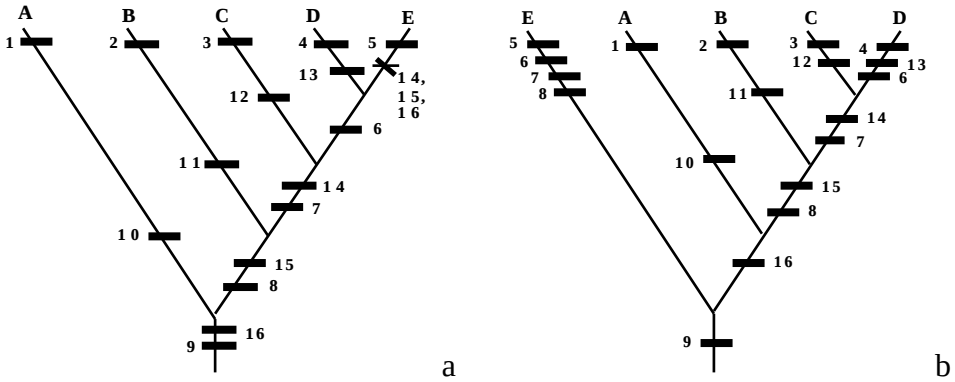


Figure 6.4 Two (out of three) MPT's for the data in table 11 in Brooks (1990), with binary data optimized according to CAFCA (identical to Brooks, 1990, and Ronquist & Nylin, 1990)

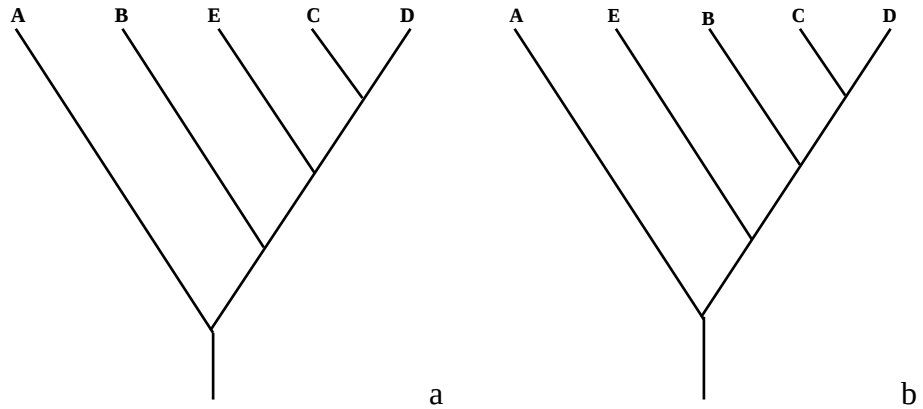


Figure 6.5 The other two MPT's for the data in table 11 (Brooks, 1990), as found by CAFCA and PAUP (if an all-zero aout-area is added), but not depicted by Brooks (1990), nor by Ronquist & Nylin (1990).

The areagrams depicted in figure 6.5 are neither presented by Brooks (1990) nor by Ronquist & Nylin (1990). The latter alternatives for Brooks' so-

lutions regard different optimizations (character mappings), not different topologies, for the areagrams in figure 6.4.

The binary area-data matrix from table 11 in Brooks (1990) can also be represented by a multi state expression by combining the interdependent columns representing the internal nodes of the respective cladograms and replace these columns with a new coding for the areas concerned, as is shown below.

```
                11
       12345678901
Aarea 10000110001
Barea 01000201002
Carea 00100300103
Darea 00010400013
Earea 00001400000
```

This area-data matrix, analysed by the component compatibility algorithm implemented in CAFCA, results in four areagrams two of which are MPT's. The areagram optimization as performed by CAFCA for this multi state matrix is depicted in figure 6.6.

The cladogram in figure 6.6a is the same as the one in figure 6.4a, except that the parsimony mapping of the characters from the multi state expression of the data matrix changes the interpretation as to the history of this species assemblage. The controversial extinctions of both ancestors and their descendants on the same branch to area E in figure 6.4a is now absent and replaced by the single extinction (character 11, state 0) of the last descendant of clade 11, the one that still occurs in areas C and D. This explanation is more akin, although slightly different and more parsimonious, to the one presented by Ronquist & Nylin (1990, their fig. 11b).

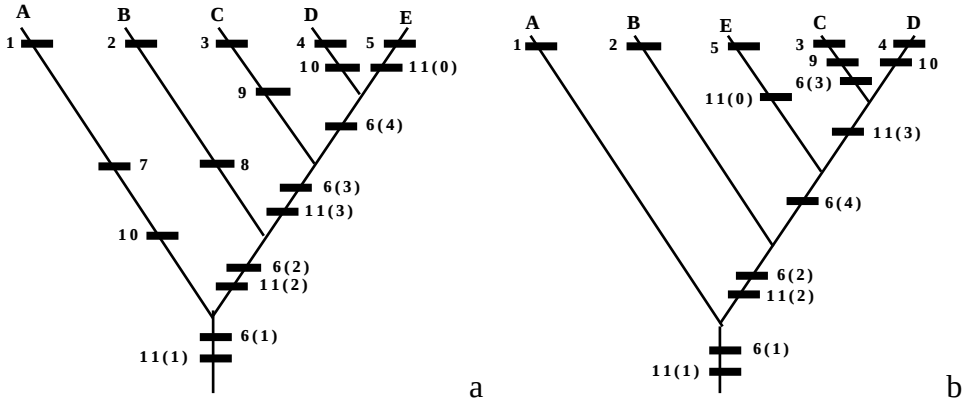


Figure 6.6 Optimization of the two MPT's found by CAFCA from the multi state expression of the data in table 11 (Brooks, 1990).

A COMPARISON WITH BPA.

Another approach to the problems of historical biogeography and the coevolution of parasites and hosts is offered by Wiley (1988 a, b), Brooks (1981, 1990) and Brooks & McLennan (1991, 1993)...Most of the examples offered by Brooks (1990) and in chapter six of Brooks & McLennan (1991) are present in the Xmpls folder on your CAFCA distribution disk. (These files are labeled by the table- or figure-caption number and have the extension **.dst** when they refer to a distribution matrix, the extension **.tre** when they refer to a cladogram, the extension **.bin** when they refer to a binary data matrix, and the extension **.asc** when they refer to a multi state data matrix). In BPA area-data and host-data

matrices are generated in the same way as in CAFCA's CCA, i.e., by computing the boolean inner product of the distribution matrix and the binary representation of a cladogram (Brooks refers to this proces as inclusive ORing). In the most recent protocol for BPA as described by Brooks (1990, p. 24) missing areas or hosts should be indicated by question marks (?), in contrast to CAFCA where primitive absence is prefered as a first order explanation (although CAFCA does produce missing value indicators [? = -1] in the data matrix when they are already present in the distribution matrix to indicate missing taxa). In BPA the area-data or host-data matrix is analysed by a cladogram generating algorithm adhering to the principle of parsimony (e.g, as implemented in PAUP for the Macintosh). Wiley (1988 a, b) opts for an optimization of the cladogram according to the rules for delayed transformation, thus prefering parallelisms (dispersal) over reversals (extinction). Brooks (1990) no longer considers this reasonable because extinction is a real phenomenon. He makes an exception for cases with wide-spread taxa, in which BPA results may support relationships for areas that conflict with the relationships for taxa as depicted in the cladograms used as basic data. Page (1989b), for that matter, considers this a confusion between areas and taxa, in BPA as well as in CCA. According to him a method should allow for the possibility that cladistically unrelated areas can share the same taxon simply because of geographic proximity.

Most results obtained by BPA are consistent with those obtained by CAFCA., as is shown in the tapeworm (*Alcataenia)* and seabird example from Brooks & McLennan (ch 6, table 7.28; B&McLt728.dst and .B&McLt728.tre in the Xmpls folder). This is the distribution matrix for parasites over birds:

```
Laridae          110000000
Fratercula       001000000
Ceorhinca        000100000
Aethia           000010000
Uria_aalge       000001110
Uria_lomvia      000001110
Cepphus_carbo    000000001
Cepphus_colomba  000000001
Cepphus_grylle   000000001
```

And this is the tapeworm cladogram:

```
((((((8,9),6,7),5),4),2,3),1);
```

CAFCA generates a host-data matrix (table 6.7) from the parasite distribution and their cladogram. This data matrix is the same as presented in Brooks & McLennan (1991, p. 270, table 7.28).

Data Matrix (binary) : BMcL728

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Laridae	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1
Fratercula	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1
Ceorhinca	0	0	0	1	0	0	0	0	0	0	0	0	1	1	1
Aethia	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1
Uria_aalge	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1
Uria_lomvia	0	0	0	0	0	1	1	1	0	1	1	1	1	1	1
Cepphus_carbo	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Cepphus_colomba	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Cepphus_grylle	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1

Column Partitioning Vector :
1 1 1 1 1 1 1 1 1 6

Data Matrix (multi–state) : BMcL728

	1	2	3	4	5	6	7	8	9	10
Laridae	1	1	0	0	0	0	0	0	0	1
Fratercula	0	0	1	0	0	0	0	0	0	1
Ceorhinca	0	0	0	1	0	0	0	0	0	2
Aethia	0	0	0	0	1	0	0	0	0	3
Uria_aalge	0	0	0	0	0	1	1	1	0	4
Uria_lomvia	0	0	0	0	0	1	1	1	0	4
Cepphus_carbo	0	0	0	0	0	0	0	0	1	4
Cepphus_colomba	0	0	0	0	0	0	0	0	1	4
Cepphus_grylle	0	0	0	0	0	0	0	0	1	4

Table 6.7 Host-data matrix for seabirds, as derived from the distribution of their tapeworm parasites and the tapeworm cladogram.

The host-cladogram as found by CAFCA through CCA (table 6.8) is also the same as found by BPA.

As regards the the interpretation of the cladogram and its state changes I refer to the book by Brooks & McLennan (1991). There is a small difference between CAFCA's solution presented above and that of BPA. It is found in the state changes on the host-cladogram. CAFCA treats the hierarchical part of the parasite phylogeny as **one** ordered multistate character. This results in a difference in number of steps as found by CAFCA vs those found by BPA. BPA treats the same hierarchy as 6 additive coded binary characters and finds 15 steps in the cladogram, without homoplasies. CAFCA finds 12 steps, also without homoplasies.

```
BMcL728B: Host—Cladogram — 1

          /—— 7 Cepphus_carbo
          |
        /—11— 8 Cepphus_colomba
        |  |
      /—12 \—— 9 Cepphus_grylle
    /—13 |
/—14 |   |  /—— 5 Uria_aalge
|   |   |  \—10
|   |   |     \—— 6 Uria_lomvia
|   |   |
|   |   \————————— 4 Aethia
|   |
|   \———————————— 3 Ceorhinca
|
|—————————————————— 1 Laridae
|
\—————————————————— 2 Fratercula
```

BMcL728B: Host—Cladogram—1 : STATE CHANGES

Character	Component	Change
1	1	0 —> 1
2	1	0 —> 1
3	2	0 —> 1
4	3	0 —> 1
5	4	0 —> 1
6	10	0 —> 1
7	10	0 —> 1
8	10	0 —> 1
9	11	0 —> 1
10	12	3 —> 4
	13	2 —> 3
	14	1 -> 2

Table 6.8 Host cladogram as derived from tapeworm distribution and phylogeny.

In Brooks & McLennan (1991) mapping the parasite phylogeny, in terms of the host-data matrix, onto the independent cladogram of the hosts (when known) is sometimes used to generate an *a posteriori* interpretation in terms of processes (e.g., host switching) for the patterns found. In general, this proces of mapping data onto a tree is known as user-tree evaluation or parsimony mapping, and was dealt with in chapter 5. Its use in studies of co-evolution is treated at the end of this chapter.

Sometimes other differences between BPA and CCA may occur. These differences are related to the different cladogram-finding algorithms that are used, i.e., Wagner parsimony in BPA and component-compatibility in CCA, and the constraints that obtain when searching and optimizing the cladogram. Let's treat a hypothetical example first and then see if I can make a case.

Suppose we have a biogeographical problem for one or more groups of taxa in which at least one (ancestral) species is absent in two areas, say, species A occuring in N-America, S-America, Europe, and continental Asia, and species B only occuring in N-America and Europe as it has become extinct in S-

America and continental Asia. If an tree-finding algorithm unites the latter two areas as sister areas due to common absence of species B, geology shows proof of the incorrectness as S-America is of Gondwana origin and continental Asia of Laurasian origin. Even if the areas would have had a common history, the extinction of taxa in these areas would still be no proof for that common history. The question remains which algorithm under which protocol may result in this kind of apparently false hypotheses of common history. The point to be made is that component-compatibility never will and that under particular protocols other parsimony algorithms might, as components are equivalent to monothetic sets of taxa and therefore always based on the presence of taxa.

Van Welzen (1989) observes this apparent anomaly in BPA in his analysis of the genus *Guioa*. However, his conclusions may be premature due to his use of midpoint rooting for the area-cladograms.

NEED ANOTHER EXAMPLE HERE...

I will use the well know data on the fish *Heterandria* and *Xiphophorus* to demonstrate another difference in the results obtained by CCA and BPA due to the differences in cladogram-finding algorithm. As this actually concerns a generalized analysis as two different species groups (genera) are involved, I will deal with this example later on in the paragraphs on generalised area-cladograms (p. 105).

EXAMPLE OF AN ANALYSIS UNDER ASSUMPTION 1.

To run a biogeographic analysis using assumption 1 for the derivation of the area-data matrix, you must follow the same steps as outlined in the first example of this chapter, except for step **11** where you must click **1** in the **Assumption** dialog. What follows now is a discussion of the results.

With regard to the preparation of the area-data matrix, assumption 1 implies that the historical relationships among areas occupied by a widespread taxon might be misconstrued and that consequently this might also be the case for the taxon itself.

Maybe a widespread taxon does not represent a single taxon but two taxa, or even more, and maybe each of these two or more taxa does not even find its closest relative among its break-up companions but in the original sistergroup. A data matrix then should allow for these possibilities as to phylogenetic relations differing from those expressed in the original cladogram for taxa.

This is exactly what happens in the (hidden columns # 10-17 of the) area-data matrix (table 6.6). These columns are hidden because the data represented by them are not ‘real’ but reflect assumptions. As such, these hidden columns do not enter the computations for cladogram optimisation and selection criteria. They only surve the purpose to generate the components, i.e., building blocks for area-cladograms, that reflect the assumptions.

There are two widespread taxa, # 1 and 4. Taxon 1 occurs in the areas 1 and 2. If due to the wide distribution of taxon 1 the historical relations among the areas 1 and 2 will be misconstrued if not remedied, the area-data matrix must allow each of these areas to get separated from the other (= break up their sister area relation) but nevertheless retain their relations relative to the other areas (3, 4, and 5). That is, areas which contain widespread taxa, like 1 and 2 do, either should maintain their sister area relationship or they should branch off sequentially in the area-cladogram (see also figure 6.1).

The hidden columns of the area-data matrix (table 6.10) now assure that both these things may happen as columns 10 and 11 show area 1 and 2 separated, and columns 12 and 13 joins area 1 and 2 each on its turn with the distribution of the sistergroup of taxon 1 (i.e., the group {2 3 4 5} [see table 3.6]), that is, the joint areas 2, 3, 4, and 5.

```
Area-Data Matrix (binary) : PlntArea

     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17
  +--------------------------   ----------------------
1 |  1  0  0  0  0  0  0  0  1  0  1  0  1  0  0  0  0
2 |  1  1  0  0  0  0  0  1  1  1  0  1  1  0  0  0  0
3 |  0  0  1  0  0  0  1  1  1  0  0  1  1  0  0  0  0
4 |  0  0  0  1  0  1  1  1  1  0  0  1  1  0  1  0  1
5 |  0  0  0  1  1  1  1  1  1  0  0  1  1  1  0  1  1

Column Partitioning Vector :
1 1 1 1 1 4

Data Matrix (multi—state) : PlntArea

          1  2  3  4  5  6
        ———————————————————
Aarea |   1  0  0  0  0  1
Barea |   1  1  0  0  0  2
Carea |   0  0  1  0  0  3
Darea |   0  0  0  1  0  4
Earea |   0  0  0  1  1  4
```

Table 6.10 Area-data matrix with hidden columns for a biogeographic analysis on PLNTAREA using assumption 1.

The same happens in the (hidden) columns 14-17 for areas 4 and 5 (for widespread taxon 4) although here the effect is obscured by the fact that taxon 5, the sistergroup of taxon 4, also occurs in area 5.

All in all the widespread taxa do not add to the contradictions already present in the area-data matrix used in the assumption zero analysis, as only 2 new distributional types are created in allowing for assumption 1, i.e., solo's for area 1 and 5 (columns 11 and 14). Apart from the intricacies of the additional (compared with table 6.1) hidden columns of the data matrix all other output is identical to that obtained by assumption zero.

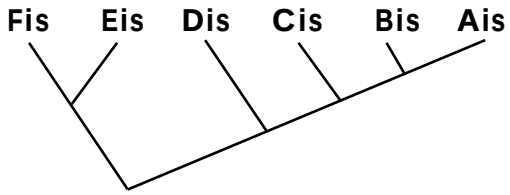
MISSING AREAS AND REDUNDANCY

Before we can turn to a generalised analysis we need the phylogeny as well as the distribution of a second (independent) monophyletic group. This also gives us the opportunity to introduce two anomalies, other than wide-spread species, in the analysis of historical biogeography, viz. missing areas (or lack of occurrence) and redundancy (information doubling as regards the relationship of areas). In the following example I will see how CAFCA handles these situations.

To complicate matters beyond the problem of widespread taxa I made up an example, ANIMAL, for 6 taxa for which the distribution matrix expresses lack of occurrence (ANIMAL taxa do not occur in area 3, PLANT taxa do) and redundancy of information (group {AB} and {EF} in the cladogram imply an identical relation as to areas 4 and 5). The distribution matrix (ANIMAL.DST in the Xmpls folder on your distribution disk) looks as follows:

```
    A B C D E  -area
Ais 0 0 0 1 0
Bis 0 0 0 0 1
Cis 0 1 0 0 0
Dis 1 0 0 0 0
Eis 0 0 0 1 0
Fis 0 0 0 0 1
```

The cladogram for ANIMAL looks like this:



Its binary image (ANIMAL.TRE in the Xmpls folder on your distribution disk) is as follows:

```
Ais  1 0 0 0 0 0 1 1 1 0 1
Bis  0 1 0 0 0 0 1 1 1 0 1
Cis  0 0 1 0 0 0 0 1 1 0 1
Dis  0 0 0 1 0 0 0 0 1 0 1
Eis  0 0 0 0 1 0 0 0 0 1 1
Fis  0 0 0 0 0 1 0 0 0 1 1
```

In order to provide the data necessary for a generalised analysis, eventually, you should run a biogeographic analysis for ANIMAL, following the same steps as in the first example (PlntArea), although with the following changes.

In step **2**, use **AnimArea**, for example, as the name for the area-data matrix.

In step **6**, the name of the distribution matrix to use in this analysis is ANIMAL.DST.

In step **7**, copy the cladogram matrix from an ASCII file, in stead of from the OutputFile system. You will find the ASCII file in the examples folder on your distribution disk. The name to select in step **8** (File selector box) is ANIMAL.TRE.

The area-data matrix for a biogeographic analysis on ANIMAL using assumption zero is given in table 6.11. The resulting area-cladogram is presented in table 6.12. The area-data matrix from this analysis (ANIMAL) as well as the matrix from the analysis on PLANT (table 6.1) is used as input for a generalised analysis.

Area–Data Matrix (binary) : AnimArea

	1	2	3	4	5	6	7	8	9	10	11
Aarea	0	0	0	1	0	0	0	0	1	0	1
Barea	0	0	1	0	0	0	0	1	1	0	1
Carea	0	0	0	0	0	0	0	0	0	0	0
Darea	1	0	0	0	1	0	1	1	1	1	1
Earea	0	1	0	0	0	1	1	1	1	1	1

Column Partitioning Vector :
1 1 1 1 1 1 5

Data Matrix (multi–state) : AnimArea

	1	2	3	4	5	6	7
Aarea	0	0	0	1	0	0	1
Barea	0	0	1	0	0	0	2
Carea	0	0	0	0	0	0	0
Darea	1	0	0	0	1	0	3
Earea	0	1	0	0	0	1	3

Table 6.11 Area-data matrix for taxa and areas in ANIMAL.

But first let's take a quick look at the output from the biogeographic analysis for areas and taxa in ANIMAREA. We have not yet seen how missing areas are treated, in this case area 3 (table 6.11).

```
AnimArea: Area—Cladogram — 1

              /—— 4 Darea
            /—6
          /—7  \—— 5 Earea
        /—8  |
        |  |  \———— 2 Barea
        |  |
        |  \—————— 1 Aarea
        |
        \———————— 3 Carea
```

```
AnimAreaB Area—Cladogram—1
COMPATIBILITIES
————————————————————————————
 Component |Character| State
————————————————————————————
     1     |    4    |   1
     2     |    3    |   1
     4     |    1    |   1
           |    5    |   1
     5     |    2    |   1
           |    6    |   1
     6     |    7    |   3
           |    7    |   3
     7     |    7    |   2
     8     |    7    |   1
           |    7    |   1
————————————————————————————
```

```
AnimArea: Area—Cladogram—1 :
STATE CHANGES
————————————————————————————
Character|Component|Change
————————————————————————————
    1    |    4    | 0 —> 1
    2    |    5    | 0 —> 1
    3    |    2    | 0 —> 1
    4    |    1    | 0 —> 1
    5    |    4    | 0 —> 1
    6    |    5    | 0 —> 1
    7    |    6    | 2 —> 3
         |    7    | 1 —> 2
         |    8    | 0 —> 1
————————————————————————————
Components refer to the list of
     monothetic sets of areas.
```

Table 6.12 Area-cladogram and corresponding compatibilities and state changes for AnimArea.

As none of the monophyletic groups in ANIMAL occurs in area 3 and the data matrix does not give us any clue whether these taxa have ever been there or not, this absence is interpreted as primitive absence. That is, none of the monophyletic groups occurs in area 3 because their common ancestor did not occur there. As a consequence area 3 is placed at the outgroup node in the area-cladogram (table 6.12).

Another possible anomaly in the area-data matrix on ANIMAL concerns redundancy (groups {AisBis} and {EisFis} in the cladogram for ANIMAL imply an identical relation as to areas 4 and 5). Note that this phenomenon is not treated any different than redundancy of information is treated in the case of normal characters that indicate an identical relationship as to the grouping of taxa. This kind of redundancy merely serves to strengthen the indicated relationship, as long as the characters concerned can be considered independent, of course. The redundancy is indicated in the list of compatibilities in table 6.12. Character 7 state 3 and state 1 are listed twice for both component 6 and 8, respectively. As for an area-data matrix, the case for independence is questionable indeed. However, the fact that {AisBis} and {EisFis} are the products of separate lineage's may support the case for (partial) independence.

That redundancy is, in this case, a possible anomaly also follows from the not at all straightforward explanation to be generated for the present distribution of taxa Eis and Fis. Most distributions, except the missing area Carea and that for group {EisFis}, follow from vicariance events that coincide with speciation events as a first order explanation. The first speciation event, {EisFis} vs. {AisBisCisDis}, apparently did not coincide with a break up of areas (fig 6.10a). {EisFis} more or less lingered in that part of the total area where only later Darea + Earea came into existence as separate areas (or biota's). The next speciation event, {Dis} vs. {AisBisCis} did coincide with a vicariance event, viz. area Aarea versus the rest (fig 6.10 b). Also the next vicariance coincided with speciation: {Cis} vs. {AisBis} and {Barea} vs. {Darea+Earea} (fig. 6.10 c). Only when Darea separated from Earea the response in {EisFis} to {Eis} vs. {Fis} followed, jointly with the speciation of {Ais} vs. {Bis} (fig 6.10 d).

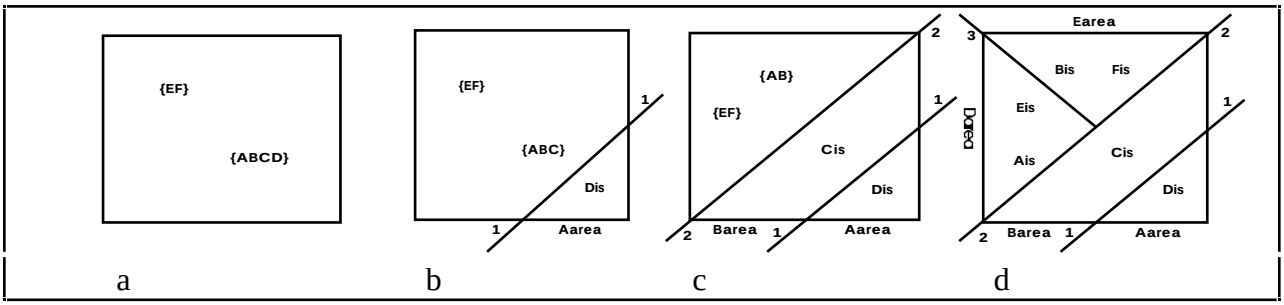


Figure 6.10 Possible vicariance and speciation events for the area-data AnimArea and cladogram 1. In (a) speciation predates vicariance, in the end causing redundant information as to the relation of Earea vs Darea by {Ais,Bis} and {Eis,Fis}; A= Ais, B=Bis, C=Cis, D=Dis, E=Eis, F=Fis.

Another, more complicated and thus less likely explanation for the current distribution of {Eis} and {Fis} might imply an unknown area, let's say Garea, that indeed was involved in the speciation {EisFis} vs. {AisBisCisDis} and the vicariance event {Garea} vs. {A-. B- C-, D-, E-area}, but the ancestor {EisFis} later on dispersed to {Darea + Earea} and got extinct in Garea.

You may have noted that in contrast to other methods, like in Brooks Parsimony Analysis (BPA, Brooks 1981, 1990) for instance, CAFCA does not *by default* use missing value indications in the area-data matrix for taxa that are absent in one or more areas. If mssing value codes are to be used at all for missing taxa, in CAFCA I prefer to indicate terminal taxa with a zero and only the internal nodes of the cladogram with a question mark. If, however, you indicate *all* entries for missing taxa with a question mark, CAFCA will show them in the area-data matrix but nevertheless ignore these entries for terminal taxa and substitute them with zero's in deriving components for area-cladograms. Only for the internal nodes of the cladogram will the question marks be effective.

GENERALISED AREA-CLADOGRAM.

INTRODUCTION.

As stated before, a generalised area-cladogram expresses the historical relations among areas as deduced from the phylogeny of several (unrelated) groups of taxa occurring in these areas. Several groups of taxa imply several cladograms, thus complicating the derivation of a joint area-data matrix allowing the extraction of a generalised area-cladogram.

However, as each of the groups of taxa has already been analysed separately in order to derive a normal area-cladogram for each group, a generalised area-data matrix can easily be composed by joining the separate area-data matrices of each group. This also lightens the task of allowing for assumptions 1 and 2, eventually, by means of additional hidden columns in the area-data matrix as this will already be accomplished in the area-data matrices for the separate groups. I will now exemplify the procedure for a two-group situation.

AN EXAMPLE.

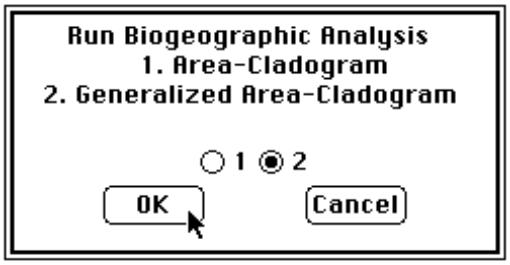
As an example, I will use the results obtained above on the data matrices PLANT and ANIMAL to run a generalised analysis.

TUTORIAL

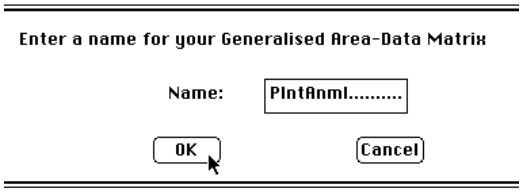
To run a generalised analysis follow the next sequence of steps.

1. Select **Biogeographic Analysis** in the **Run** menu.

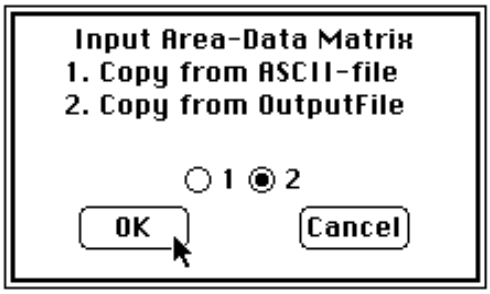
2. Click **2** (**Generalized area-cladogram**) in the next dialog.



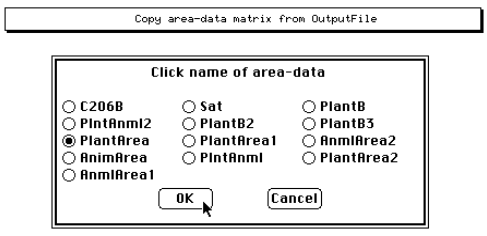
3. Type a name for the resulting area-data matrix, **PlntAnml** for example, to be used by CAFCA in this analysis



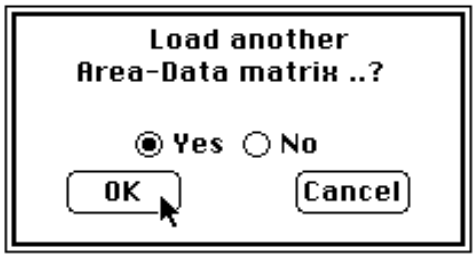
4. Click **2** (**Copy from OutputFile**) in the next dialog prompting for the source of the first area-data matrix.



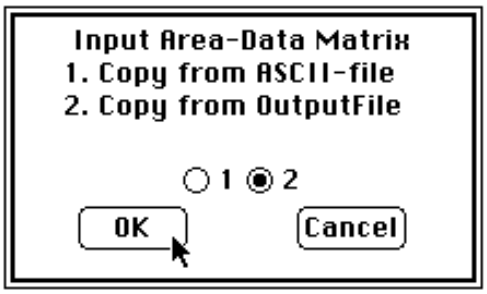
5. Click the button of appropriate name, i.e., the name (e.g. PlntArea) you used in the first example of a biogeographic analysis of PLANT, in the next dialog.



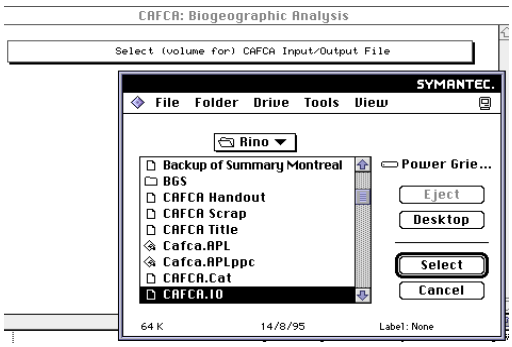
6. Because we need at least two area-data matrices to run a generalised analysis click **OK** for the default **Yes** in the next dialog.



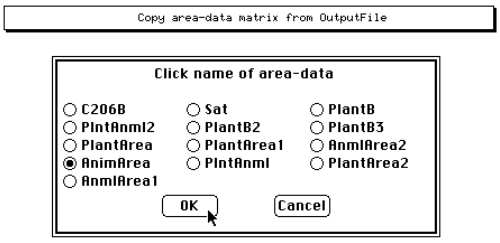
7. Click **2** (**Copy from OutputFile**) in the next dialog prompting for the source of the second area-data matrix.



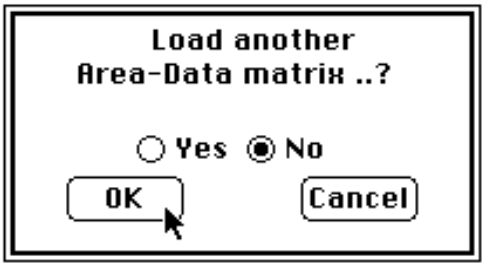
8. You must choose one of the possibly many Outputfiles you prepared. You can do so in the following dialog.



9. The next dialog will show you all the files you saved in the particular Outputfile you just chose. Click the name you used in the biogeographic analysis on ANIMAL. Note that you could have used a different name than the one used in this particular example.



10 As in this run we confine the generalised analysis to only two area-data matrices, click **No** in the next dialog.



11. You went through all of the next steps before, either in the primary analysis or in the preceding biogeographic analysis, so they should be familiar. You do not need to save the results of the analysis, as long as you do print them for sake of the next discussion.

DISCUSSION OF RESULTS

If we now look at the output for the generalised analysis the first thing we notice is the size of the binary data matrix (table 6.14). In fact the first nine columns are those of PLNTAREA, and the last 11 columns are from ANIMAREA.

By joining these matrices as used earlier for a standard analysis all columns, i.e., their patterns displayed, can now interact freely. That is, all possible co-variation among components from both PLNTAREA and ANIMAREA can now be observed and analysed. As a consequence the number of components from PLNTANML (table 6.15) is larger than either the number from PLNTAREA or from ANIMAREA and the increased co-variation causes more generalised area-cladograms to be possible.

Area—Data Matrix (binary) : PlntAnml

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Aarea	1	0	0	0	0	0	0	0	1	0	0	0	1	0
Barea	1	1	0	0	0	0	0	1	1	0	0	1	0	0
Carea	0	0	1	0	0	0	1	1	1	0	0	0	0	0
Darea	0	0	0	1	0	1	1	1	1	1	0	0	0	1
Earea	0	0	0	1	1	1	1	1	1	0	1	0	0	0

	15	16	17	18	19	20
Aarea	0	0	0	1	0	1
Barea	0	0	1	1	0	1
Carea	0	0	0	0	0	0
Darea	0	1	1	1	1	1
Earea	1	1	1	1	1	1

Column Partitioning Vector :
1 1 1 1 1 4 1 1 1 1 1 1 5

Data Matrix (multi—state) : PlntAnml

	1	2	3	4	5	6	7	8	9	10	11	12	13
Aarea	1	0	0	0	0	1	0	0	0	1	0	0	1
Barea	1	1	0	0	0	2	0	0	1	0	0	0	2
Carea	0	0	1	0	0	3	0	0	0	0	0	0	0
Darea	0	0	0	1	0	4	1	0	0	0	1	0	3
Earea	0	0	0	1	1	4	0	1	0	0	0	1	3

Table 6.14 Area-data matrix for a generalized biogeographic analysis for PLANT and ANIMAL.

Partial Monothetic Sets of areas in PlntAnml	
1	1
2	2
3	3
4	4
5	5
6	1 2
7	4 5
8	3 4 5
9	2 4 5
10	2 3 4 5
11	1 2 4 5
12	1 2 3 4 5

Partial Monothetic Sets of Monophyletic Groups (= Components) in PlntAnml	
1	13
2	2 12
3	3
4	10 14
5	5 11 15
6	1
7	4 6 16 19
8	7
9	17
10	8
11	18 20
12	9

Table 6.15 Components for a generalized area-cladogram for both PLANT and ANIMAL.

Monophyletic groups on root for PlntAnml

Rownumbers refer to index numbers of cladograms
Columnnumbers refer to columns of multistate data matrix.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	0	0	0	0	2	0	0	0	0	0	0	1
5	1	0	0	0	0	1	0	0	0	0	0	0	1

Table 6.16 Indication of ancestral monophyletic groups present in ancestral area.

States on the root (table 6.16) can be interpreted as ancestral monophyletic groups present in an ancestral area.

Selection criteria for cladograms of: PlntAnmlÆB
Column numbers refer to numbers of cladograms

Row 1 : Total number of homoplasous events
Row 2 : Total number of single origins (Support)
Row 3 : Corrected Extra Length (x1000; CEL: Turner + Zandee)
Row 4 : Total number of state changes (S: Steps)
Row 5 : Redundancy Quotient (x1000; RQ: Zandee + Geesink)
Row 6 : Rescaled Redundancy Quotient (x1000; RQc)
Row 7 : Consistency Index (x1000; CI), with autapomorphy correction
Row 8 : Rescaled Consistency Index (x1000; RC: Farris)
Row 9 : Average Unit Character Consistency (x1000; AUCC: Sang)
Row 10: Homoplasy Distribution Ratio (x1000; HDR: Sang)
Row 11: Compatible Character State Index (x1000; CCSI: Zandee)

	1	2	3	4	5
1	0	1	2	1	0
2	17	14	13	16	17
3	0	1000	2000	1000	0
4	17	18	19	18	17
5	436	437	426	433	450
6	47	47	29	41	69
7	1000	944	895	944	1000
8	1000	472	0	472	1000
9	1000	981	942	962	1000
10	1000	654	452	308	1000
11	538	564	513	538	538

No—Order Limit for Steps, Extra Steps, RQ, and CI:

S	ES	RQ	CI
19	2	409	895

Table 6.17 Selection criteria for generalized area-cladograms for PLANT and ANIMAL.

Two out of the five area-cladograms found are minimum step diagrams (table 6.17). PAUP finds a third (# 2) if an all-zero outgroup is added to the multi-state area-data matrix (unordered multi-states !). The RQ could guide your choice among these, or the likelihood of the different scenarios associated with each area-cladogram.

To interpret the state changes in the context of a possible scenario one has to backtrack to the original cladograms, to have a picture of the branching events in the evolutionary history of the taxa concerned. In figure 6.11 these cladograms are presented once again, now with an indication of the columns from the binary area-data matrix on the nodes as well as an indication of the columns and character state from the multi-state area-data matrix alongside the nodes.

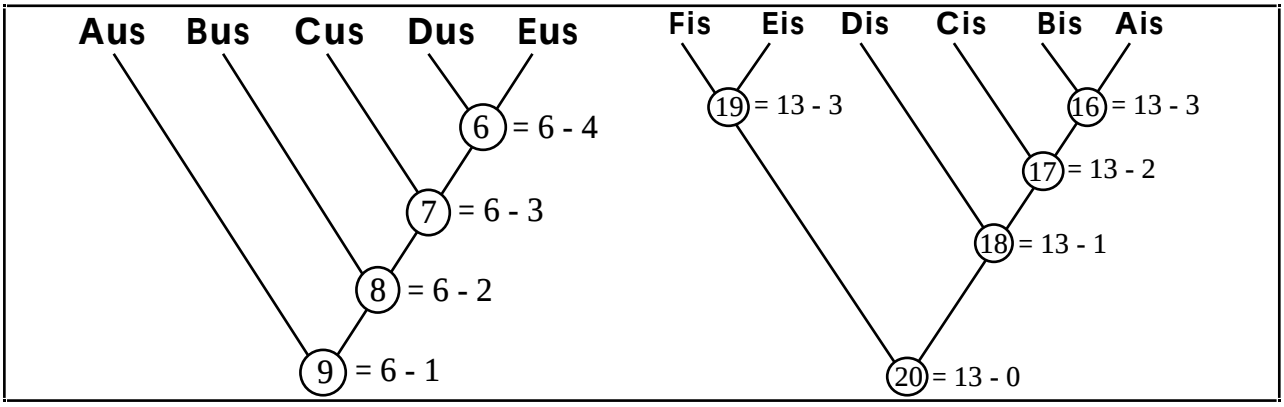


Figure 6.11 Cladograms for PLANT and ANIMAL with an indication of the columns from the generalised area-data matrix. in table 6.14

Area-cladograms # 1 and 5 in table 6.17 are identical to the solutions found in the separate analysis of the area-data matrix PlntArea. Area-cladogram # 2 represents the solution found in the separate analysis of AnimArea. The question is which area-cladogram offers the best overall explanation in historical terms for the present day distribution of all taxa involved.

As I have already discussed in the treatment of PlntArea, area-cladogram # 1 does not offer an overall first order explanation for the data at hand. Adding AnimArea does not exactly contribute to the transparency of the explanation as it prompts the same scenario for Cis and Dis in relation to their occurence in Aarea and Barea, as for Aus and Bus in the case of PlntArea.

The same is true for area-cladogram # 2 as Aarea and Barea are again sister areas. That leaves area-cladogram # 5 as the best choice as the PlntArea part of the generalised area-data matrix mostly obeys a first order explanation in terms of vicariance events, with two anomalies that do not cause extra steps, while for AnimArea only one anomaly obtains, i.e., the one for Ais + Bis and Eis + Fis in Darea + Earea (see also previous discussions of PlntArea and AnimArea).

```
PlntAnml: Area—Cladogram — 1

          /—— 4 Darea
       /——7
    /——8  \—— 5 Earea
    |  |
    |  \————— 3 Carea
    |
    |  /————— 1 Aarea
    \——6
       \————— 2 Barea
```

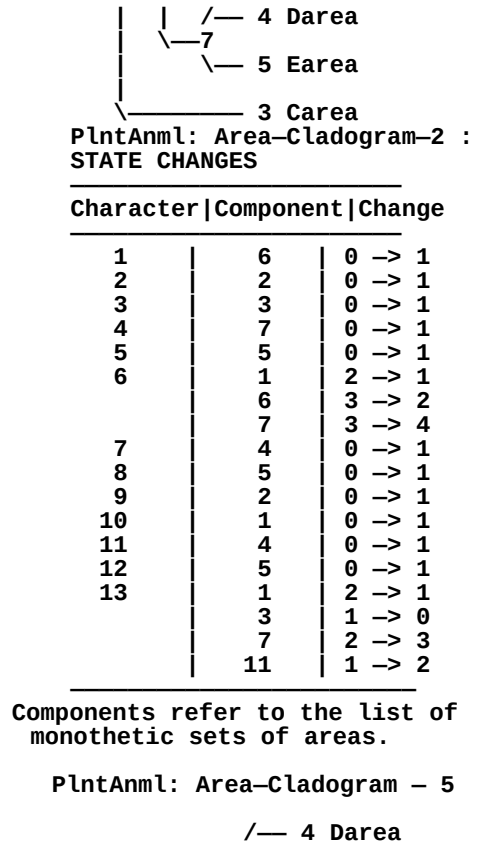
PlntAnml: Area—Cladogram—1 : STATE CHANGES

Character	Component	Change
1	6	0 —> 1
2	2	0 —> 1
3	3	0 —> 1
4	7	0 —> 1
5	5	0 —> 1
6	1	2 —> 1
	7	3 —> 4
	8	2 —> 3
7	4	0 —> 1
8	5	0 —> 1
9	2	0 —> 1
10	1	0 —> 1
11	4	0 —> 1
12	5	0 —> 1
13	2	1 —> 2
	3	1 —> 0
	7	1 —> 3

Components refer to the list of monothetic sets of areas.

```
PlntAnml: Area—Cladogram — 2

          /—— 1 Aarea
       /——6
       |  \—— 2 Barea
    /——11
```



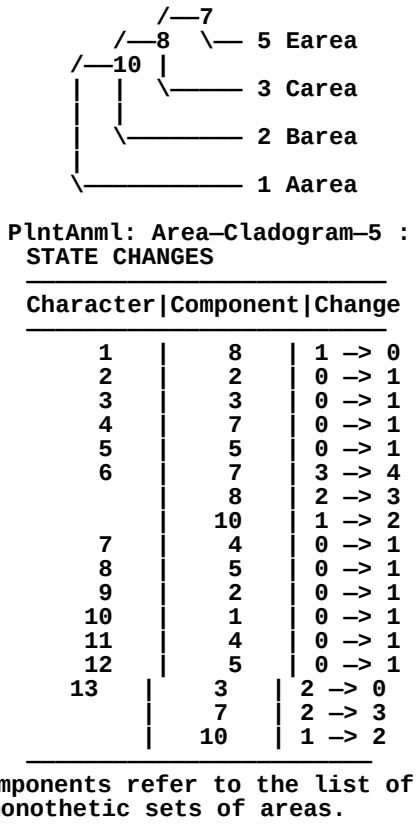


Table 6.17 All most parsimonious generalized area-cladograms for PLNTANML, with corresponding state changes .

MULTI-STATE AGAINST BINARY AREA-DATA

When we run a generalised analysis on only the binary representation of the area-data matrix PlntAnml, i.e., with a partitioning vector that is all-one, we get the same results in terms of area-cladograms but their number of steps (state-changes) and therefore the interpretations differ.

```
Selection criteria for cladograms of: PlntAnml
Column numbers refer to numbers of cladograms
-----------------------------------------------
Row 1 : Total number of homoplasous events
Row 2 : Total number of single origins (Support)
Row 3 : Corrected Extra Length (x1000; CEL: Turner + Zandee)
Row 4 : Total number of state changes (S: Steps)
Row 5 : Redundancy Quotient (x1000; RQ: Zandee + Geesink)
Row 6 : Rescaled Redundancy Quotient (x1000; RQc)
Row 7 : Consistency Index (x1000; CI), with autapomorphy correction
Row 8 : Rescaled Consistency Index (x1000; RC: Farris)
Row 9 : Average Unit Character Consistency (x1000; AUCC: Sang)
Row 10: Homoplasy Distribution Ratio (x1000; HDR: Sang)
Row 11: Compatible Character State Index (x1000; CCSI: Zandee)

         1     2     3     4     5
    -------------------------------
 1 |     1     1     2     2     1
 2 |    19    19    18    18    19
 3 |  1700  1700  2750  2750  1700
 4 |    20    20    21    21    20
 5 |   468   481   461   468   480
 6 |    99   120    88    99   119
 7 |   909   909   833   833   909
 8 |   814   814   646   646   814
 9 |   975   975   950   950   975
10 |   500   500   475   475   500
```

```
11 |   538   564   513   538   538

No-Order Limit for Steps, Extra Steps, RQ, and CI:

   S   ES   RQ   CI
------------------
  26    7  409  588
```

Table 6.18 Selection criteria for the cladograms from the binary part of the area-data matrix PlntAnml.

We see in the table with selection criteria values for cladograms that none of the cladograms has a CI equal to one, as was the case with multi-state data. It seems as if the binary data do not fit the cladograms as well as their multi-state expressions. What exactly is going on ?

```
PlntAnml: Cladogram – 1

          /–– 4 Darea
       /––7
    /––8  \–– 5 Earea
    |  |
    |  \––––– 3 Carea
    |
    |  /––––– 1 Aarea
    \––6
       \––––– 2 Barea
```

```
PlntAnml: Cladogram–1: STATE CHANGES
──────────────────────────
Character|Component|Change
──────────────────────────
    1    |     6   |  0 –> 1
    2    |     2   |  0 –> 1
    3    |     3   |  0 –> 1
    4    |     7   |  0 –> 1
```

```
   5   |     5   |  0 –> 1
   6   |     7   |  0 –> 1
   7   |     8   |  0 –> 1
   8   |     1   |  1 –> 0
   9   |         |
  10   |     4   |  0 –> 1
  11   |     5   |  0 –> 1
  12   |     2   |  0 –> 1
  13   |     1   |  0 –> 1
  14   |     4   |  0 –> 1
  15   |     5   |  0 –> 1
  16   |     7   |  0 –> 1
  17   |     2   |  0 –> 1
       |     7   |  0 –> 1
  18   |     3   |  1 –> 0
  19   |     7   |  0 –> 1
  20   |     3   |  1 –> 0
──────────────────────────
Component refer to the list of
monothetic sets of areas.
```

Table 6.19 Cladogram # 1 and its state-changes; generalised analysis on binary area-data PlntAnml.

Column 17 has two state changes. It is a binary character and thus a CI = 0.5 results. In the multi-state expression of the area-data matrix it represents one of the states in column 13. The sequence of states in this multi-state expression perfectly fits the cladogram if the character is unordered (which, of course, it is **not** when it represents a hierarchy in a part of the original cladogram for ANIMAL).

In a new (binary) scenario the distribution of the cladon {AisBisCis} over the areas {Barea, Darea, Earea} can now be explained by assuming a dispersal for Cis to Barea, as indicated by the extra step in binary character 17, in contrast to the underlying non-response to a vicariance event in our former multi-state scenario, which as such does not represent an extra step.

```
PlntAnml: Area-Cladogram – 2

          /–– 1 Aarea
       /––6
       |  \–– 2 Barea
    /––11
    |  |  /–– 4 Darea
    |  \––7
    |     \–– 5 Earea
    |
    \––––––– 3 Carea
```

```
PlntAnml: Area-Cladogram–2:
STATE CHANGES
──────────────────────────
Character|Component|Change
──────────────────────────
    1    |     6   |  0 –> 1
    2    |     2   |  0 –> 1
    3    |     3   |  0 –> 1
    4    |     7   |  0 –> 1
```

```
   5   |     5   |  0 –> 1
   6   |     7   |  0 –> 1
   7   |     6   |  1 –> 0
   8   |     1   |  1 –> 0
   9   |         |
  10   |     4   |  0 –> 1
  11   |     5   |  0 –> 1
  12   |     2   |  0 –> 1
  13   |     1   |  0 –> 1
  14   |     4   |  0 –> 1
  15   |     5   |  0 –> 1
  16   |     7   |  0 –> 1
  17   |     1   |  1 –> 0
       |    11   |  0 –> 1
  18   |    11   |  0 –> 1
  19   |     7   |  0 –> 1
  20   |    11   |  0 –> 1
──────────────────────────
Components refer to the list of
monothetic sets of areas.
```

```
PlntAnml: Area-Cladogram – 5

                /—— 4 Darea
             /—7
          /—8  \—— 5 Earea
       /—10 |
       |  |  \————— 3 Carea
       |  |
       |  \———————— 2 Barea
       |
       \————————————— 1 Aarea
```

PlntAnml: Area-Cladogram–5 : STATE CHANGES

Character	Component	Change
1	8	1 –> 0
2	2	0 –> 1
3	3	0 –> 1
4	7	0 –> 1
5	5	0 –> 1
6	7	0 –> 1
7	8	0 –> 1
8	10	0 –> 1
9		
10	4	0 –> 1
11	5	0 –> 1
12	2	0 –> 1
13	1	0 –> 1
14	4	0 –> 1
15	5	0 –> 1
16	7	0 –> 1
17	3	1 –> 0
	10	0 –> 1
18	3	1 –> 0
19	7	0 –> 1
20	3	1 –> 0

Components refer to the list of monothetic sets of areas.

Table 6.19 Cladogram # 2 and its state-changes; generalised analysis on binary area-data PlntAnml.

However, as we have seen previously in our discussion of the example presented by Page (1990, 1993), illustrating a 'drawback' of BPA and CCA, we should prefer a multi-state expression of the area-data matrix as a means to overcome the problem of interdependence among columns indicating the hierarchical nature of the relationships among the taxa involved. Neglect of this interdependence may indeed cause a multiplication of ad hoc elements (dispersal, extinction) in the explanation of the distribution of taxa over areas (or genes over taxa, or parasites over hosts).

One may ask whether the multi-state coding and optimising as employed in CAFCA is the best available. Alternative coding schemes have been described by O'Grady & Deets (1987) and O'Grady, Deets, & Bentz (1989). These schemes, known as nonredundant linear coding or mixed ordinal-additive binary coding (Pimentel & Riggins, 1987), were developed to describe cladogram topologies, such that the coding is most efficient and at the same time avoids unjustified weighting of certain branches of a topology.

An important question remains as to the use of this coding scheme in biogeographical analyses or coevolutionary studies. In order to obtain a data matrix, the phylogenetic information on the taxa concerned as contained in the cladogram must be linked to the distributions (over areas or hosts) of the same taxa. This linking proceeds by a process called inclusive ORing (O'Grady & Deets, 1987), which is computationally identical to obtaining the Boolean inner product of the cladogram and the distribution matrix (Zandee & Roos, 1987). For this process to take place it is essential that the data concerned (cladogram and distributions) are represented in a binary (0/1) way. It is not yet clear how a cladogram coded in a nonredundant linear scheme, i.e., by employing multistate expressions, can be used in this computation. Funk and Brooks (1990, table 8) in their treatment of the example on *Heterandria* and *Xiphophorus*, show how the result of inclusive ORing an additive binary coded cladogram and a binary distribution matrix can be recoded by hand into a nonredundant linear scheme. As we will see, the result of this recoding process differs slightly from the result obtained by the automated procedure currently used in CAFCA.

A COMPARISON WITH BPA.

I promised another example on the differences in results obtained by CCA vs BPA. This example is based on Rosen's (1978) well known data on the fish

genera *Heterandria* and *Xiphophorus*, also used in Zandee & Roos (1987), Page (1988a, 1990a,b 1993), and Funk & Brooks (1990), among others.

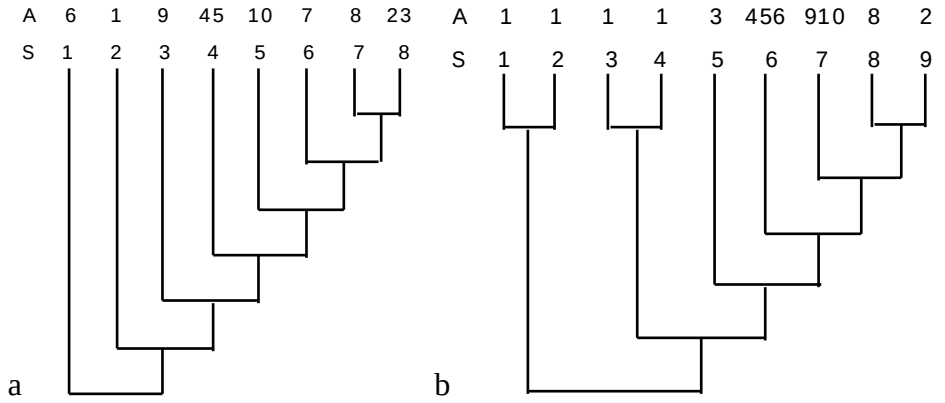


Figure 6.12: Cladogram for 8 species (S) of *Heterandria*, and cladogram for 9 species (S) of *Xiphophorus*, with an indication of their distribution over areas (A).

The cladograms and distributions of the species of *Heterandria* and *Xiphophorus* are presented in figure 6.12 a and b, respectively.

From these cladograms and distributions, area-data matrices for both genera can be derived according to the procedures outlined earlier in this chapter. These two area-data matrices can be joined columnwise to make one area-data matrix suitable for a generalized analysis (table 6.22).

Data Matrix (binary) : HetXiphB (Columns represent character states)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Area_1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Area_2	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
Area_3	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0
Area_4	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0
Area_5	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	0
Area_6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Area_7	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	-1
Area_8	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	0
Area_9	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	0
Area_10	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	0

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Area_1	1	1	1	0	0	0	0	0	1	1	0	0	0	0	1	1
Area_2	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1	1
Area_3	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1
Area_4	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1
Area_5	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1
Area_6	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1
Area_7	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Area_8	0	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1
Area_9	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1
Area_10	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1

Column Partitioning Vector :
1 1 1 1 1 1 1 1 7 1 1 1 1 1 1 1 1 1 8

Data Matrix (multi–state) : HetXiph (Columns represent characters)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Area_1	0	1	0	0	0	0	0	0	2	1	1	1	1	0	0	0
Area_2	0	0	0	0	0	0	0	1	7	0	0	0	0	0	0	0
Area_3	0	0	0	0	0	0	0	1	7	0	0	0	0	1	0	0
Area_4	0	0	0	1	0	0	0	0	4	0	0	0	0	0	1	0
Area_5	0	0	0	1	0	0	0	0	4	0	0	0	0	0	1	0
Area_6	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
Area_7	0	0	0	0	0	1	0	0	6	-1	-1	-1	-1	-1	-1	-1
Area_8	0	0	0	0	0	0	1	0	7	0	0	0	0	0	0	0
Area_9	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	1

Area_10	0	0	0	0	1	0	0	0	5	0	0	0	0	0	0	1

	17	18	19
Area_1	0	0	1
Area_2	0	1	5
Area_3	0	0	2
Area_4	0	0	3
Area_5	0	0	3
Area_6	0	0	3
Area_7	-1	-1	-6
Area_8	1	0	5
Area_9	0	0	4
Area_10	0	0	4

Table 6.22: Area-data matrix for *Heterandria* and *Xiphophorus.* Neagative integers indicate missing values.

The binary area-data matrix from table 6.22 is used in CCA as well as in BPA as the same protocol for the derivation of area-data matrices (assumption 0) applies in both methods. Note, however, that CAFCA would not *by default* use missing value indications for taxa that are absent in an area, as is the case in table 6.22 for area 7 and the species of *Xiphophorus*. I do so in this case in order to make a full comparison between BPA and CCA possible. Also note that if missing value codes are to be used at all for missing taxa, in CAFCA I prefer to indicate terminal taxa with a zero and only the internal nodes of the cladogram with a question mark.

In BPA (PAUP), 10 most parsimonious area-cladograms counting 35 steps are found (one area-cladogram contains a trichotomy). In CCA (CAFCA) 25 area-cladograms are found, 4 of which are most parsimonious (35 steps) when we use an all-one partitioning vector in the evaluation of the cladograms. It appears that BPA using PAUP is better equiped to find more (= all) most parsimonious area-cladograms. If we partition the columns of the binary area-data as indicated in table 6.22 CCA finds a selection of 10 most parsimonious cladograms with 28 steps. Entering this multi-state expression in PAUP we find 72 area-cladograms with 27 steps. These findings by BPA (PAUP) and CCA (CAFCA) are different, mainly due to the different algorithms used in cladogram finding, and partly due to the different partitioning of the columns in the area-data matrix during optimization. I will now describe and discuss these differences.

Comparing the cladograms found by BPA (PAUP) and CCA (CAFCA) on the basis of the binary data it appears that out of the 10 area-cladograms found by PAUP only 2 are validated by CAFCA in the sense that they are built on components that are all defined by the presence of at least one unique species or monophyletic taxon (i.e., these components are partially monothetic sets).

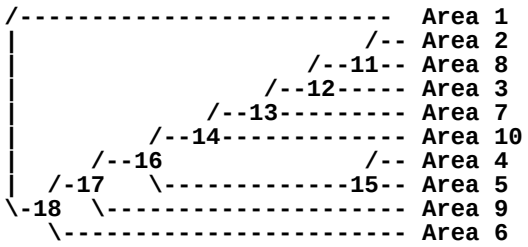


Figure 6.13: Area-cladogram found by PAUP using the binary data in table 6.22, also found by CAFCA due to the presence of components that are all partially monothetic

Figure 6.13 presents one of the two area-cladograms found by PAUP as well as CAFCA, as all its components are partially monothetic, i.e., completely correspond with at least one species or taxon in the binary area-data matrix.

The other 8 area-cladograms contain components that are not based on the presence of at least one unique monophyletic taxon. Figure 6.14 presents an example for the latter case.

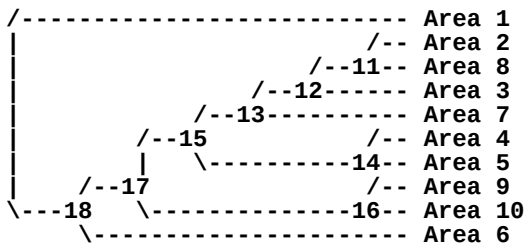


Figure 6.14: Area-cladogram found by PAUP using the binary data in table 6.22, but not found by CAFCA due to the presence of non-monothetic components.

Node 13 in this area-cladogram is component {2 3 7 8}, defined by column 10 in table 6.22, which indicates the monophyletic group {6,7,8} in *Heterandria* (fig 6.12). However, for node 15, the component {2 3 4 5 7 8}, no corresponding monophyletic taxon can be found. To make matters more complicated, there is not one character in table 6.22 for which a state-change occurs on this node. The branch leading to node 15 is empty and it should be collapsed (as in fig 65 in Funk & Brooks, 1990).

In yet another area-cladogram from the set of 8, 'strange' character optimization takes place in order to keep branches from being empty (fig 6.15).

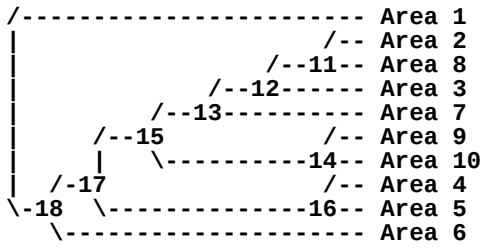


Figure 6.15 Area-cladogram found by PAUP using the binary data in table 6.22, but not found by CAFCA due to the presence of non-monothetic components.

Node 15 corresponds to component {2,3,7,8,9,10}. It has no unique defining species or monophyletic taxon. On the other hand, it is characterized by a change 0->1 in character 11, a change 1->0 in character 21, and a change 0->1 in character 28. Character 11 corresponds to the clade {5,6,7,8} in *Heterandria*, thus going extinct in area {9} (and in 7 as well, if not primitively absent). Character 21 corresponds to species {6} in *Xiphophorus*, thus going extinct in area {2,3,7,8,9,10}. Character 28 corresponds to clade {7,8,9} in *Xiphophorus*, thus originating in area {2,3,8,9,10} and later going extinct in area 3.

To summarize so far, I may state that the area-cladograms found by CCA are more constrained as to the definition of their constituent components in terms of occurence of species of monophyletic taxa. In contrast, BPA's definition of a component appears to be more relaxed and as a result more area-cladograms are found.

When we analyse the binary data in table 6.22 *without* missing value indications for area 7 in *Xiphophorus*, i.e., with CCA's default 0 for primitive absence, CCA finds three area-cladograms, only one of which is most parsimonious with 29 steps (fig 6.16) when we optimize the multi-state expression of the binary data. This area-cladogram, but with area 1 and 7 interchanged, is also the one with the highest RQ among the 10 most parsimonious results obtained by CAFCA when missing values *are* included. It is not present in the set of area-cladograms found by BPA.

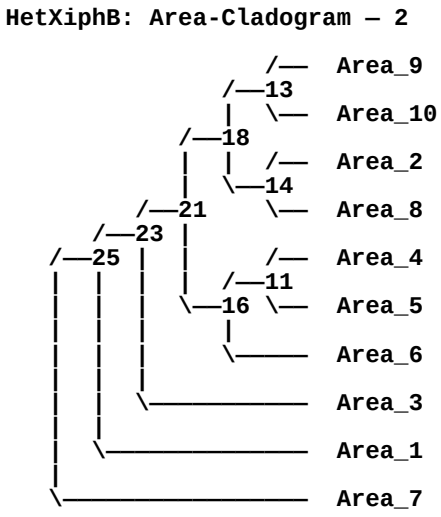


Figure 6.16: Best cladogram obtained by CCA from table 6.22, optimizing the multi-state data using zero's for area 7 in *Xiphophorus*.

Assuming primitive absence for *Xiphophorus* in area 7 this area-cladogram implies all the sistergroup relations for the species of *Xiphophorus* as depicted in figure 6.12, without any contradiction. For *Heterandria* it implies one ad-hoc statement (contradiction), i.e., the dispersal of species # 8 into Area_3, and also three unique events, i.e., the occurences in areas 1, 9, and 7 for the species 2, 3, and 6, respectively, that can not be explained by vicariance and allopatric speciation (Zandee & Roos, 1987, p. 324).

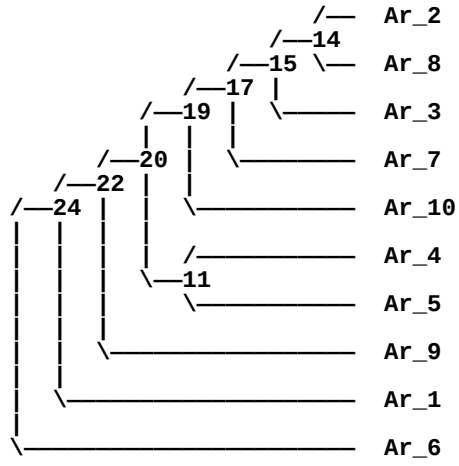


Figure 6.17: Best cladogram obtained by CCA from table 6.22, optimizing the binary data using 0's for area 7 in *Xiphophorus*.

When we run the same analysis in CCA (no missing values) but optimise the binary data instead of their multi-state expression, the area-cladogram in figure 6.17 is just one step shorter (39 steps vs 40) then the one in figure 6.16 (see also fig 8b vs 8c in Zandee & Roos, 1987). Interchanging the position of Area_6 and Area_1 costs no extra steps and results in one of the two area-cladograms on which BPA and CCA agree (fig 6.13) on the basis of the area-data in table 6.22, including missing values.

However, this area-cladogram needs more ad-hoc explanations for contradictions then the one in figure 6.16 (although the differences are minor). As for *Heterandria* there is extinction of species #8 in area 8. Assuming that there is global primitive absence for *Xiphophorus* in area 7 we observe dispersal for species # 6 in area 6, for species # 7 in area 9, and for species # 5 in area 3.

As we have seen before in other examples, it appears that the optimization of the multi-state expression of the area-data matrix as opposed to its binary expression leads to a choice of area-cladograms with less ad-hoc statements.

In conclusion, CCA's best result (fig 6.16) is not included in the set of area-cladograms found by BPA (following its own protocol), whether we use missing value indications for missing areas or not but always optimise the multi-state expression of the binary data. Even when we change the current BPA protocol (Brooks 1990) by using the multi-state expression of the data matrix (without missing values), the resulting 38 area-cladograms (with 28 steps) all contain the clade {2,3}, and only 16 area-cladograms are fully resolved. Including missing values for area 7 in *Xiphophorus* results in 72 area-cladograms (with 27 steps). They all contain the clade {2,3}, but only 30 area-cladograms are fully resolved.

Ance	**00000001000000001**
One	**01000002121200002**
Two	**00000008000000007**
Three	**00000008000010003**
Four	**00010004000001004**
Five	**00010004000001004**
Six	**10000001000001004**
Seven	**00001106999999999**
Eight	**00000017000000016**
Nine	**00100003000000105**
Ten	**00001005000000105**

Table 6.23 Nonlinear redundant coding for the *Heterandria* and *Xiphophorus* area-data matrix, as presented by Funk & Brooks (1990). Area 7 is coded as missing for *Xiphophorus*.

An alternative coding for the *Heterandria* and *Xiphophorus* area-data is offered by Funk & Brooks (1990, table 8), based on O'Grady & Deets' (1987) nonlinear redundant coding scheme. It is presented here in table 6.23. In contrast to the other area-data (table 6.22) an overall ancestral area (outarea ?) is included.

Analysing this table by BPA (PAUP) results in 4725 area-cladograms with 28 steps when the multi-state characters are considered unordered. I did not check whether the 10 area-cladograms found by CCA on the basis of the multi-state date from table 6.22 are present in this set. Ordering the characters results in the 10 area-cladograms (with 37 steps) presented in Funk & Brooks (1990, figs 56-65).

One may ask whether CCA's best result is included in the set of results obtained by still other alternative approaches, and what is to be considered the best result anyway.

A COMPARISON WITH *COMPONENT* 2.0.

The application of Page's (1990a, b, 1993) new methodology for problems in biogeography and co-evolution brings yet another set of results, different from the one presented above. First to be considered is the area-cladogram as found by Page (1993, fig 6.18a) as the reconciliation between the two cladograms of the genera concerned and a possible area-cladogram. This solution agrees with the area-cladogram in figure 6.17, except for the interchanged positions of areas 3 and 8 and of areas 1 and 6. Relative to the multi-state expression of the area-data matrix in table 6.22 it counts 32 steps (vs 33 for fig 6.17), and is therefore more parsimonious. The area-cladogram(s) preferred by Page (1993, fig 6.18b) as the final solution to the problem is (are) one step longer (33 steps). It also closely resembles the one in figure 6.17 (counting 33 steps), ex-

cept for the position of area 3. Relative to the binary area-data matrix these area-cladograms in fig 6.18a and fig 6.18b take 40 and 42 (43) steps, respectively (remember that the area-cladogram in figure 6.17 counts 39 steps when measured against the binary matrix; fig 6.16 counts 40 steps).

Reconciled Tree Analysis (RTA) considers the area-cladogram in figure 6.18d optimal for *Xiphophorus* when taken separately (Page 1993, his fig 19 b). In our opinion the almost identical area-cladogram in fig. 6.16 (except for the basal trichotomy in 6.18d) is optimal for both genera taken together as it introduces just one ad hoc statement and three unique events for *Heterandria*, and none for *Xiphophorus*.

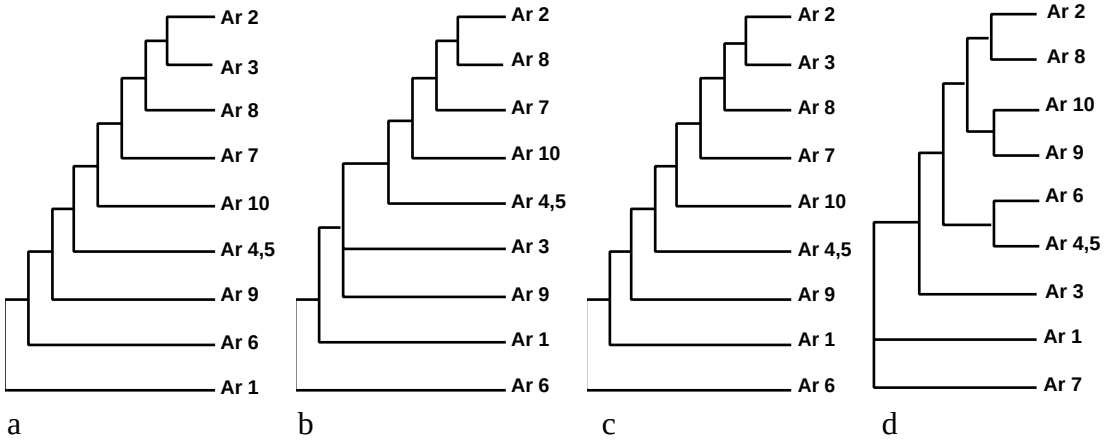


Figure 6.18 Page, 1993, figures 17 a (a), 17 b (b), 19 a (c), and 19 b (d) redrawn. (**a**) Optimal area-cladogram for *Heterandria* and *Xiphophorus* when taken together; (**b**) idem but area 3, 6, and 9 sored for endemic occurences only (**c**) Optimal area-cladogram for *Heterandria* when taken separately; (**d**) idem for *Xiphophorus* .

For comparisons sake I must emphasize that for Reconciled Tree Analysis as implemented in COMPONENT 2.0 to reach its optimal solution (Page 1993, fig 6.18b), the areas about whose relationships the two fish genera disagree (i.e., 3, 6, and 9) get a special treatment. These areas are deleted from the range of each wide-spread species (i.e., area 3 in *Heterandria*; areas 6 and 9 in *Xiphophorus*), and only enter the analysis for the genus that has endemic species for these areas.

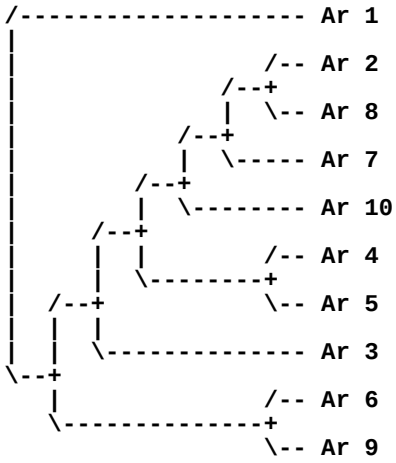


Figure 6.19: Area-cladogram obtained by BPA from table 6.22, deleting area 3, 6, and 9 from the range of wide-spread taxa and using ?'s for area 7 in *Xiphophorus*.

Applying this protocol to BPA we find the area-cladogram depicted in fig 6.19. Except for the component {6,9} and the relative position of area 1 this area-cladogram is almost identical to one of the possibilities implied by Page's

(1993) fig 17b (here 6.18b). It is 34 steps long ! (29 steps for the multi-state expression of the data)

Applying the same protocol to CCA we find the area-cladogram in figure 6.20.

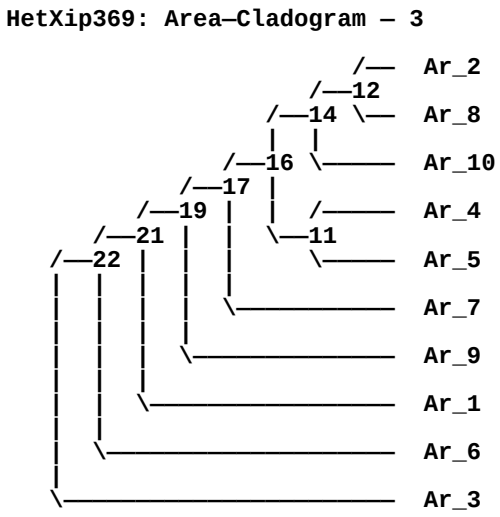


Figure 6.20: Area-cladogram obtained by CCA from table 6.22, deleting area 3, 6, and 9 from the range of wide-spread taxa and using 0's for area 7 in *Xiphophorus*.

This area-cladogram takes 29 steps when measured against the multi-state expression of the area-data matrix, amended for the deletion of the areas 3, 6, and 9 from the range of the wide-spread taxa. Page's optimal solution(s) (fig 6.18b) also take 29 steps.

The results obtained by different methods are different but equally parsimonious. I have summarised this comparison in table 6.23 and figure 6.21.

		RTA		BPA		CCA	
		a	b	- out	+ out	1	2
	Binary	**35**	38	**35**	39	**35**	38
	Multi state	30	31	33	37	**28**	31
				28-31	29-31		
Only endemic	Binary	43	36	**34**	37	40	39
3, 6, 9	Multi state	**29**	**29**	**29**	**29**	**29**	**29**

Table 6.23 Lengths of area-cladograms found by RTA (Reconciled Tree Analysis), BPA and CCA for the *Heterandria* - *Xiphoporus* problem. RTA a and b refer to figure 6.18 a and b, respectively. BPA -out and +out refer to the absence and presence of an outarea in the datamatrix. CCA 1 and 2 refer to the two area-cladograms in figure 6.16 and 6.17, respectively.

So the question remains: What's best in terms of parsimony and area-data, and which method should be used to obtain it ? All solutions found by RTA, BPA, and CCA, except one in the set of 10 MPA's found by BPA, relate the areas 1, 2, 4+5, 8, and 10 in the same way (figure 6.21). As Page (19xx) notes, this also what Rosen (1978) found as the general solution of the problem, as did Platnick (19xx) in his re-analysis of Rosen's data. In this respect there is no difference between CCA, BPA, and RTA. They all agree on the non-problematic part of the data. The point is, what do they do with the messy part of the data, i.e. the data concerning areas 3, 6, and 9 where widespread species occur, and what do they do with incomplete data, i.e. the data concerning area 7 where no species of *Xiphophorus* occur.

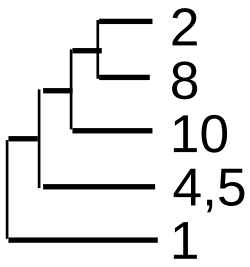


Figure 6.21 The common structure in the area-cladograms found by RTA, BPA (9 out of its 10 solutions), and CCA for the *Heterandria* - *Xiphophorus* problem.

The ad hoc removal of possible discordance or sources of discordance beforehand, as is done in Page's approach (RTA: Reconciled Tree Analysis), appears to be a futile excercise as it does not lead to better results in terms of parsimony. Rather paradoxically, the RTA protocol under its own assumptions renders solutions which are at best as parsimonious (in terms of the multi-state expression of the data) as those obtained by BPA and CCA, and at worst 2 to 9 steps longer (in terms of the binary expression of the data) than the solution obtained by BPA. Based on unaltered data (assumption 0), RTA's solutions are at best as parsimonious (in terms of the binary expression of the data) as those obtained by either BPA or CCA, or at worst 2-3 steps longer (in terms of the multi-state expression of the data) compared to BPA and CCA solutions. BPA consistently presents the most parsimonious solutions using either the binary or the multi state expression of the data, whether the data are adapted to accomodate certain assumptions or not.

CO-EVOLUTION

INTRODUCTION

In previous chapters and paragraphs I claimed that the component compatibility method (CCA) can also be used as a tool to solve problems in coevolution (of parasites and hosts). Before I present an example, I should point out some possible differences in the approach of problems in historical biogeography vs those in parasite-hosts relationships. In the latter case cladograms can be estimated separately for hosts and parasites from morphological or molecular data, ect...In general, this is not the case for areas or biotas. Only when areas of endemism have a very clear circumscription in a geological context (islands, island arcs, parts of islands known to be composite) we might try to reconstruct the historical hierarchical pattern on the basis of geological data. In most cases however, the historical relationships of areas of endemism or biotas must be estimated indirectly on the basis of the general patterns extracted from the phylogenies of taxonomically independent groups of organisms and the distribution of these taxa in the areas concerned.

How, then, do we proceed in the case of problems of coevolution of parasites and hosts? How do we handle the two independent cladograms to generate a solution of the problem that originates when there is no one-to-one relationship between these two cladograms as regards the number of taxa as well as the branching pattern?

EXAMPLE

An example is offered by Page (1993), as taken from a study by Hafner and Nadler (1988). It concerns pocket gophers and (one of) their parasite groups, chewing lice.

Host–Data Matrix (binary) : GopLiceB

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
T_talpoides	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0
T_bottae	0	1	0	1	0	0	0	0	0	0	1	1	1	0	0
G_bursarius	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
O_hispidus	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
O_cavator	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
O_underwoodi	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
O_cherriei	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1
O_heterodus	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0

	16	17	18	19
T_talpoides	0	0	1	1
T_bottae	0	0	1	1
G_bursarius	0	0	1	1
O_hispidus	0	1	1	1
O_cavator	1	1	1	1
O_underwoodi	1	1	1	1
O_cherriei	1	1	1	1
O_heterodus	1	1	1	1

Column Partitioning Vector :
1 1 1 1 1 1 1 1 1 9

Data Matrix (multi–state) : GopLice

	1	2	3	4	5	6	7	8	9	10	11
T_talpoides	1	0	1	0	0	0	0	0	0	0	1
T_bottae	0	1	0	1	0	0	0	0	0	0	3
G_bursarius	0	0	0	0	1	0	0	0	0	0	2
O_hispidus	0	0	0	0	0	1	0	0	0	0	4
O_cavator	0	0	0	0	0	0	1	0	0	0	6
O_underwoodi	0	0	0	0	0	0	0	1	0	0	6
O_cherriei	0	0	0	0	0	0	0	1	1	0	7
O_heterodus	0	0	0	0	0	0	0	0	0	1	5

Table 6.24 Binary and multi-state expression of the data matrix generated from the cladogram for the chewing lice and their distribution over the species of gophers.

Starting point for the analysis are the cladograms for both groups concerned, and the distribution of the parasites over their hosts (these data are available in the Xmpls folder on your distribution disk: Gopher.tre, Lice.tre, and Lice.dst, respectively). We can run a biogeographic analysis, by entering the distributional data for the Lice as well as their cladogram to generate an 'area-data' matrix for Lice over Gophers ('taxa over areas', as it were).

The first 10 columns in the host-data matrix from table 6.24 represent the 10 species of chewing lice:

1. T_wardi
2. T_minor
3. T_thomomyus
4. T_actuosi
5. T_ewingi
6. G_chapini
7. G_panamensis
8. G_setzeri
9. G_cherriei
10. G_costaricensis

They are distributed over the gopher species as follows:

```
                    1
               1234567890
               ----------
T_talpoides    1010000000
T_bottae       0101000000
G_bursarius    0000100000
O_hispidus     0000010000
O_cavator      0000001000
O_underwoodi   0000000100
O_cherriei     0000000110
O_heterodus    0000000001
```

The last 9 columns in the host-data matrix represent the 9 inner-nodes of the lice-cladogram. The analysis of this data matrix results in the following cladogram and state changes.

```
GopLiceB: Area—Cladogram — 1
          /—— 6 O_underwoodi
        /—9
      /—14 \—— 7 O_cherriei
   /—15 |
/—16 |  \——————— 5 O_cavator
|  |  |
|  |  \————————— 8 O_heterodus
|  |
|  \———————————— 4 O_hispidus
|
|     /————————— 1 T_talpoides
|  /—10
\—13 \————————— 2 T_bottae
   |
   \———————————— 3 G_bursarius
```

GopLiceB: Cladogram—1 : STATE CHANGES

Character	Component	Change
1	1	0 –> 1
2	2	0 –> 1
3	1	0 –> 1
4	2	0 –> 1
5	3	0 –> 1
6	4	0 –> 1
7	5	0 –> 1
8	9	0 –> 1
9	7	0 –> 1
10	8	0 –> 1
11	1	2 –> 1
	2	2 –> 3
	7	6 –> 7
	14	5 –> 6
	15	4 –> 5
	16	2 –> 4

Components refer to the list of monothetic sets of areas.

Table 6.25 Cladogram resulting fom the data in table 6.24.

In a biogeographic context this would not represent a general solution to the problem at hand. We would need more parasitic taxa, other than this particular genus of lice, from this gophers to arrive at a generalized cladogram for the hosts (‘areas’). That would be one track to follow to generate a (generalized) solution for this coevolutionary problem. However, in this case a cladogram for the gophers is already available as obtained from other independent sources (either morphological or molecular data). A comparison of this independent cladogram with that just obtained from the analysis above may also indicate the moments of cospeciation (the codivergent nodes, as coined by Page, 1993). This comparison results from using the independent gopher cladogram as a user-tree for the data matrix obtained from the lice cladogram and lice distribution. A recipe for the complete procedure is now presented, using the data available in the Xmpls folder on the CAFCA distribution disk.

RECIPE

1. Select **Biogeographic Analysis** from the **Run** menu.
2. Type a name for the host-data matrix to be used by CAFCA.
3. Take the default option 1 (**Area Cladogram**)
4. Take the default option 1 (**Generate from distribution and cladogram matrices**).
5. Take the default option 1 (**Copy from ASCII file**).

6. In the file selector box, select **Lice.dst** as the file containing the distribution matrix .
7. Take the default option 1 (**Copy from ASCII file**).
8. In the file selector box, select **Lice.tre** as the file containing the cladogram.
9. Click **No** in the dialog asking if you want to see the cladogram.
10. Take the default option in the next dialog (**Assumption 0**).
11. Take the default option in the next dialog (No clipping).
12. Take all defaults in the **Set CAFCA Parameters** dialog box and click **OK**.
13. Let the analysis run untill completed.
14. Select **User-tree evaluation** from the **Run** menu.
15. Take the default option 1 (**User-tree from ASCII file**).
16. Select **Gopher.tre** as the file containing the user-tree in the following file selector box.
17. Click **No** and **OK** for not viewing the user-tree.
18. Click **No** and **OK** for **Ancestral state indicated by zero.**
19. Clik **Yes** and **OK** for **Collapse empty branches in computation of RQ.**
20. Let the analysis run untill completed.
21. Select **All results** from the **Print** menu.

Nota Bene:

As indicated in this recipe it is **recommended** that you run this user-tree evaluation *immediately* after the analysis of the host-area data. You preferably should **not** run the latter analysis first, save its results in the OutputFile system, and then start the user-tree evaluation by copying back the host-data matrix from the OutputFile system. The reason is that the *binary* image of the data matrix, if absent in the OutputFile, is computed from the *multi-state* image instead of the other way around (as is the case in the procedure outlined in the recipe). The pitfall that may be present is due to the fact that a multi-state coding of a binary data matrix always implies a **linear** ordering if a multi-state character is indicated as ordered. In the case of the analysis of host-parasite relationships the host-data matrix reflects the phylogenetic structure of the parasite group. The cladogram of the parasites may very well have another topology than a strict hennigian comb, but only the latter topology can be represented by the **linear** sequence of states from a multi-state character. In all other cases this linearity breaks down. Therefore the states of the character in the multi-state expression of the host-data matrix only represent a **code**, which is treated as if ordered, with the single purpose to make the optimization of the binary data on the independent host cladogram possible. This code should in its turn **not** be translated back into a binary representation because the result will, in general, be **different** from the original binary representation of the host-data matrix due to the effect of linear-ordering.

DISCUSSION OF RESULTS

Running a user-tree evaluation with the original independent cladogram for the gopher species using the data matrix presented in table 6.24, results in the list of state changes given in table 6.26 below.

```
GopLiceBtree: Cladogram – 1

                 /—— 7 O_cherriei
              /—10
           /—11 \—— 8 O_heterodus
        /—12 |
     /—13 |  \——— 6 O_underwoodi
  /—14 |  |
  |  |  |  \————————— 5 O_cavator
  |  |  |
  |  |  \———————————— 4 O_hispidus
  |  |
  |  \——————————————— 3 G_bursarius
  |
  |  /——————————————— 1 T_talpoides
  \—9
     \——————————————— 2 T_bottae
```

GopLiceBtree: Cladogram–1 : STATE CHANGES

Character	Cladon	Change
1	1	0 –> 1
2	2	0 –> 1
3	1	0 –> 1
4	2	0 –> 1
5	3	0 –> 1
6	4	0 –> 1
7	5	0 –> 1
8	8	1 –> 0
	11	0 –> 1
9	7	0 –> 1
10	8	0 –> 1
11	2	1 –> 3
	7	6 –> 7
	8	6 –> 5
	12	4 –> 6
	13	2 –> 4
	14	1 –> 2

Cladon nrs refer to the list of monothetic sets of terminal taxa.

Table 6.26 Gopher cladogram as user-tree and its state changes for the host-data matrix in table 6.25

A comparison of the topologies of the cladogram in table 6.25 as found by the analysis of the host-data matrix (table 6.24) with the independent Gopher cladogram as a user-tree (table 6.26) shows that they differ in the placement of one of the Gopher species, *Geomys bursarius*. The cladogram in table 6.25 fits the data matrix perfectly.

The user-tree is one step longer. Lice species # 8 (*Geomydoecus setzeri)*, which has a widespread distribution over *O. underwoodi* and *O. cherriei*, originates on the ancestral node # 11 and goes into extinction on node # 8 (*O. heterodus*). It follows that this lice species # 8 (and its sister species, for that mather) must be at least as old as or older than the ancestral species of gopher on node # 11.

As we can see from the Lice cladogram given in figure 6.22, species # 8 has a sister species, # 7 *G_panamensis*, occuring on the gopher *O_cavator*. Apparently, when their ancestor (node # 14) speciated this co-occured with the speciation event on node # 12 in the gopher cladogram (table 6.26). Assuming a constant evolutionary rate on each branch of the cladograms concerned, the relative age for lice versus gophers as stated above is in agreement with the small Manhattan distances between nodes # 12, 11, 10, and their terminals in the gopher cladogram, as given in Page's figure 11 (Page, 1993), compared with the larger Manhattan distances between the nodes # 13, 14, and their terminals in the lice cladogram.

```
Lice: Cladogram – 1

                  /——  9 G_cherriei
               /—13
               |  \—— 10 G_costaricensis
            /—16
         /—17 |  /——  7 G_panamensis
         |  |  \—14
         |  |     \——  8 G_setzeri
      /—18 |
      |  |  \————————  6 G_chapini
      |  |
      |  |     /———— 4 T_actuosi
      |  |  /—12
      |  \—15 \———— 5 T_ewingi
      |     |
      |     \————————  3 T_thomomyus
      |
      |  /————————————  1 T_wardi
      \—11
         \————————————  2 T_minor
```

Figure 6.22: Cladogram for chewing lice on gophers.

As a corollary we may tentatively suggest that the six steps in the multi-state character reflecting the hierarchy in the gopher user-tree all correspond with a co-speciation event, i.e., they point to codivergent nodes. In its turn, this agrees with the result obtained by Page for the number of codivergent nodes in the cladograms for gophers and lice when he relaxes the constraint of strict association by descent. However, in our approach the result is obtained *without* deleting the taxa for which we assume or have evidence that they have dispersed. CAFCA implicitly allows for dispersal (horizontal transmission) and extinction events, and as a result always seeks to maximize the number of hypotheses of cospeciation.

Brooks & McLennan (1991) also use the technique of mapping the parasite phylogeny onto the host tree to generate an *a posteriori* interpretation in terms of processes (e.g., host switching) for the patterns found. From the many examples in their book we will use the data on amphilinid flatworms and their hosts (B&McLt727.xxx in CAFCA's example folder).

The data matrix used by CAFCA in this analysis has a multi-state character representing the parasite phylogeny, in contrast to BPA where this part of the data is represented in an additive binary coded fashion.

	1	2	3	4	5	6	7	8	9
Acipensiformes_1	1	0	0	0	0	0	0	0	1
Acipensiformes_2	0	1	0	0	0	0	0	0	1
Siluriformes	0	0	0	0	0	0	1	0	3
Osteoglossiformes_1	0	0	0	0	0	0	0	1	4
Osteoglossiformes_2	0	0	0	1	0	0	0	0	5
Osteoglossiformes_3	0	0	0	0	1	0	0	0	5
Perciformes	0	0	1	0	0	0	0	0	2
Chelonia	0	0	0	0	0	1	0	0	2

Table 6.27 Multi-state data matrix for hosts of amphilinid flatworms, derived from the phylogeny of the worms and their distribution over hosts.

From the mapping (table 6.28) of the characters in the data matrix (table 6.27) onto the independent host phylogeny (fig 6.23) we can derive a posteriori interpretations in terms of processes.

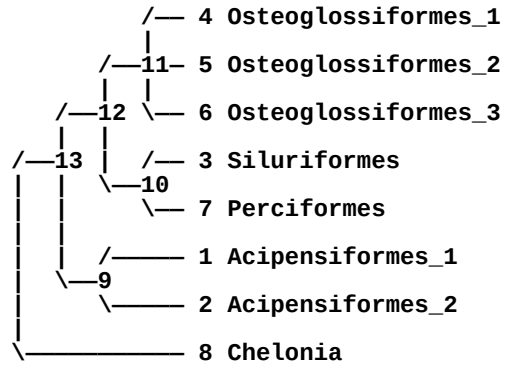


Figure 6.23 Independent host cladogram for amphilinid flatworms (after Brooks and McLennan 1991, fig 7.56).

AmphilinidsÆBtree: Area—Cladogram—1 : STATE CHANGES

Character	Component	State Change
1	1	0 —> 1
2	2	0 —> 1
3	7	0 —> 1
4	5	0 —> 1
5	6	0 —> 1
6	8	0 —> 1

7	3	0 –> 1
8	4	0 –> 1
9	4	5 –> 4
	7	3 –> 2
	9	2 –> 1
	11	3 –> 5
	12	2 –> 3

Table 6.28 Character state changes for the data matrix in table 6.27 on the cladogram for host of amphilinid flatworms (figure 6.23).

Character state 2 for character 9 originated two times independently. This implies a host-switch for flatworm # 3, *Gigantolina elongata*, to host # 7, the Perciformes.

Although CAFCA differs from BPA in the use of a multi state expression of the binary data obtained after inclusive ORing, the results in terms of interpretations are identical in this example.

THIS PAGE INTENTIONALLY LEFT BLANK