Cladistics 15, 25–37 (1999)
Article ID clad.1998.0080, available online at http://www.idealibrary.com on IDEAL®

Data Decisiveness, Missing Entries, and the DD Index

Jan De Laet and Erik Smets

Laboratorium voor Systematiek, Instituut voor Plantkunde, K.U. Leuven, Kard. Mercierlaan 92, B-3001 Heverlee, Belgium

Accepted July 18, 1998

The decisiveness of a data set has been defined as the degree to which all possible dichotomous trees for that data set differ in length, and the DD statistic (the data decisiveness index) has been proposed to measure this degree. In this paper, we first discuss an exact nonrecursive formula for the length of indecisive datasets (DD = 0) that consist of informative binary characters in which no missing entries are allowed. Next, the concept of indecisive data sets is extended to data sets in which missing entries may be present. Last, indecisive data sets with missing entries are used as an aid to construct hypothetical data sets that single out some of the factors that influence the DD statistic. On the basis of these examples, it is concluded that the concept of data decisiveness is too elusive to be captured into a single and simple index such as DD.

INTRODUCTION

Goloboff (1991a,b) defined the cladistic decisiveness of a data set as the degree to which all possible resolved trees for the data set differ in length. The decisiveness stands for the information for tree choice that is present in the data: the larger it is, the stronger is the conclusion that the worst cladograms can be safely discarded. Data sets that are fully indecisive are data sets for which every possible dichotomous tree has the same length. For binary characters without missing entries, only data sets that contain every possible informative character state distribution in an equal number are fully indecisive (Goloboff, 1991a; uninformative characters, containing no information for tree choice, may be added in any amount). Goloboff (1991a) defined an indecisive data set for n taxa as a data set for n taxa (the ingroup) to which an all-zero outgroup is added. In this way, an informative character is a character that satisfies both following conditions: (1) at least one terminal taxon of the ingroup has state 0; (2) at least two terminal taxa of the ingroup have state 1.

Two examples of indecisive data sets, one for three and one for four taxa, are shown in Fig. 1. When no missing entries are present, these data sets are essentially the only indecisive data sets that exist for three and four taxa. Besides adding uninformative characters, the only possible variation is to repeat every character for an equal number of times. We will refer to an indecisive matrix that contains all possible informative characters for an ingroup of n taxa precisely once as the minimal indecisive matrix for n taxa or MIM(n). Note that the matrix has $n + 1$ taxa because the all-zero outgroup has to be added. Because all characters are binary the ensemble observed variation of MIM(n), $M(n)$, is equal to its number of characters: $2^n - n - 2$ (Goloboff, 1991a).

Goloboff (1991a) provided equations to calculate

0748-3007/99 $30.00

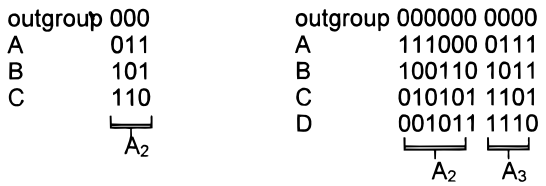

FIG. 1. Minimal indecisive data sets for three and four taxa + all-zero outgroup. An A_i character is a character with i 1-entries. All possible dichotomous trees for the first data set have 5 steps; all possible dichotomous trees for the second data set have 18 steps.

$S(n)$, the length of MIM(n) on a fully resolved tree, and $G(n)$, the length of MIM(n) on an unresolved bush. His equation for $S(n)$ is exact for $n \geq 7$, but it is difficult to calculate because it is recursive and because it contains many nested summation operators. A general ($n \geq 3$) nonrecursive equation (Fig. 2) follows directly from Steel's (1993; see also Steel and Charleston, 1995) exact nonrecursive formula for the expected character length of random binary characters on random trees (see the Appendix for details). An alternative derivation of this equation and its relation to the results of Archie (1989) and Archie and Felsenstein (1993) are discussed in the Appendix. For $G(n)$, Goloboff (1991a:220) provided two exact equations, one for n even and one for n odd, each with one summation operator. Their equivalents without summations were discussed by Steel (1993: 259). The single encompassing equation shown in Fig. 2 is derived in the Appendix.

Goloboff (1991a) restricted his discussion of fully indecisive data sets to data sets without missing entries. Below, we will first discuss how indecisive data sets can be constructed when missing entries may be present. Next we will evaluate the DD index (Goloboff, 1991a), an index that was proposed to measure the decisiveness of data sets. To do so, we will present a

$$S(n) = \frac{1}{9}\left(2^n * (3n+1) - (-1)^n\right) - (n+1)$$

$$G(n) = (n+1) * (2^{n-1} - 1) - \frac{n+1}{2} * \binom{n}{\lfloor (n+1)/2 \rfloor}$$

FIG. 2. Number of steps on a dichotomous tree, $S(n)$, and number of steps on an unresolved bush, $G(n)$, for minimal indecisive data sets with n taxa, $n \geq 3$. $\binom{n}{i}$ stands for $n!/(i!*(n-i)!)$, $[n/i]$ for the integer part of n/i.

number of hypothetical data sets that contain as submatrices indecisive data sets with missing entries. These examples will enable us to single out and illustrate some factors that influence DD.

MISSING ENTRIES AND INDECISIVENESS

Allowing missing entries, the basic observation is that an indecisive data set for $n + 1$ taxa can be produced simply by adding a row of missing entries to an indecisive data set for n taxa. An example, for three taxa, is shown in the left part of Fig. 3. Less trivial cases are obtained by combining several such data sets (Fig. 3, middle), or by combining such sets with indecisive data sets without question marks (Fig. 3, right).

A more elaborate example for five taxa + outgroup is shown in Fig. 4. The data set consists of four indecisive submatrices: MIM(5), MIM(4) + one row of missing entries, and two times MIM(3) + two rows of missing entries.

Because the addition of one or more rows of missing entries to a data set does not influence S or G, the above equations can be used for the indecisive submatrices, each time substituting the total number of taxa in the submatrix for its number of taxa that do not have missing entries: $n = 5$ for the first submatrix ($S_1 = 51$, $G_1 = 60$), $n = 4$ for the second one ($S_2 = 18$, $G_2 = 20$), and $n = 3$ for the last two ($S_3 = S_4 = 5$, $G_3 = G_4 = 6$). Precisely because the four composing subsets are indecisive, their S and G values can be added to obtain the values of G and S for the complete data set ($S = 79$, $G = 92$).

Knowing that $M = 2^n - n - 2$, and using the above equations for S and G, the upper and lower bounds for the ensemble consistency index CI (CI = M/S; Kluge and Farris, 1969) and the ensemble retention index RI (RI = $(G - S)/(G - M)$; Farris, 1989) for indecisive data sets can be plotted.

The possible ranges for CI are given in Fig. 5. The lower bound (black dots) is achieved when no missing entries are present and decreases as n increases (cf. Fig. 1 in Goloboff, 1991a). The constant upper bound of CI = 0.6 is reached in indecisive data sets that contain only three-taxon statements. Note that the upper bound can be exceeded by adding autapomorphies.

The possible ranges for RI are shown in Fig. 6. The

outgroup	000
A	011
B	101
C	110
D	???

outgroup	000 000
A	011 011
B	101 ???
C	110 101
D	??? 110

outgroup	000000 0000 000
A	111000 0111 110
B	100110 1011 101
C	010101 1101 110
D	001011 1110 ???

FIG. 3. Some examples of indecisive data sets that contain missing entries.

lower bound of RI = 0.2 is reached in indecisive data sets that contain only four-item statements, the upper bound of RI = 0.33 is reached in indecisive data sets that contain only three-item statements. The retention indices of minimal indecisive data sets without missing entries are indicated as black dots (cf. Fig. 3, Goloboff, 1991a); it oscillates asymptotically to 1/3.

MEASURING DECISIVENESS: THE DATA DECISIVENESS INDEX DD

Goloboff (1991a) rejected CI, RI, and RC (the ensemble rescaled consistency index RC = CI*RI; Farris, 1989) as measures for data decisiveness because none of these statistics reaches its minimum in indecisive data sets. Therefore, the homoplasy and the decisiveness of a data set are two different things, and Goloboff (1991a) proposed a new statistic, the DD index or data decisiveness index, to measure the latter. He defined DD as the scaled difference between the length of a most parsimonious tree for a data set (S_{MIN}) and the the mean length of the data set on all possible resolved trees ($\overline{S}$):

$$DD = \frac{\overline{S} - S_{MIN}}{\overline{S} - M}.$$

The scaling is such that data sets without homoplasy ($S_{MIN} = M$) have a DD value equal to 1. Indecisive data sets, having the same length on every resolved tree, have a DD value equal to 0. DD is not directly influenced by the homoplasy of a data set because $\overline{S}$, the mean length of a data set on all possible resolved trees, does not depend on the congruence between the characters in the data set (Le Quesne, 1989), and Goloboff (1991a: 226) concluded that DD directly measures decisiveness: "higher values of DD imply that the possible trees for the matrix differ more in tree length, and therefore [higher DD values] imply more decisiveness" (Goloboff, 1991a: 226). However, below it is shown by example that higher DD values do not necessarily imply more differences in the length of possible trees.

A first pair of hypothetical data sets is presented in Fig. 7. Both data sets consist of an indecisive part and a decisive part. The decisive submatrices are identical in both sets, but the indecisive submatrices are different. Since the decisive parts are identical and consist of characters that are fully congruent among themselves, the same single most parsimonious tree is found in both cases. The indecisive submatrices of data sets 1 and 2 contain only informative characters and are constructed such that they require the same length; the difference lies in the number of characters: the

out	00000000000000000000000000	0000000000	000	000	
A	11110000001111111000001111	1110000111	110	???	
B	10001110001110000111010111	1001101011	???	110	
C	01001001101001101101110111	??????????	101	101	
D	00100101010101011011111101	0101011101	011	???	
E	00010010110010110111111110	0010111110	???	011	
M	25	10	3	3	⇒ 41
G	60	20	6	6	⇒ 92
S	51	18	5	5	⇒ 79

CI = 0.52
RI = 0.25

FIG. 4. A compound indecisive data set for five taxa.

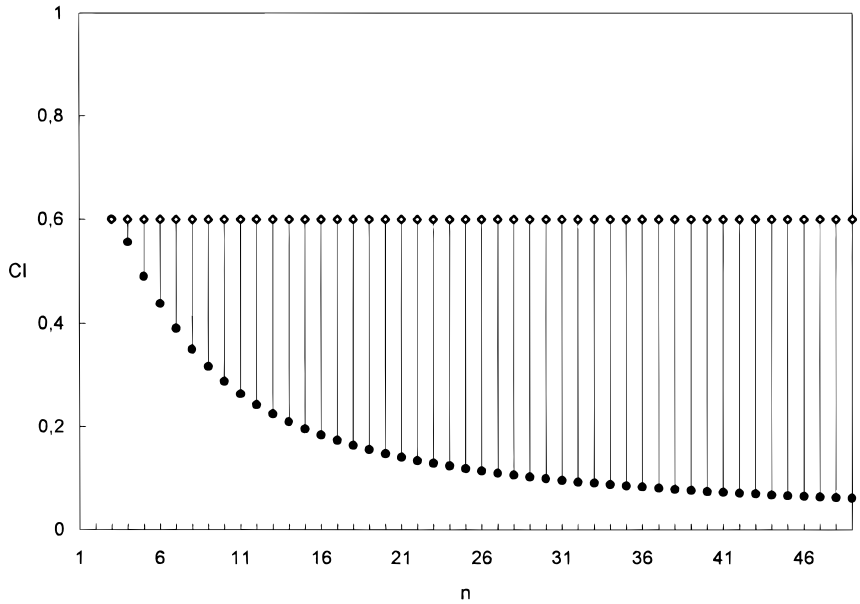

FIG. 5. Possible ranges of ensemble consistency index CI(*n*) for indecisive data sets with *n* taxa; see text for explanation.

indecisive subset of data set 1 has only 25 characters (MIM(5)), while the indecisive subset of data set 2 has 29 characters (2 times MIM(4) + 3 times MIM(3)). As a result, the two data sets have the same minimal length but a different *M* and *G* (see Fig. 7).

It follows that both data sets have an identical distribution of tree lengths (Fig. 8), implying that the possible trees for data set 1 do not differ in tree length from the possible trees for data set 2. Nevertheless both data sets have different DD values: DD is equal to 0.10 for

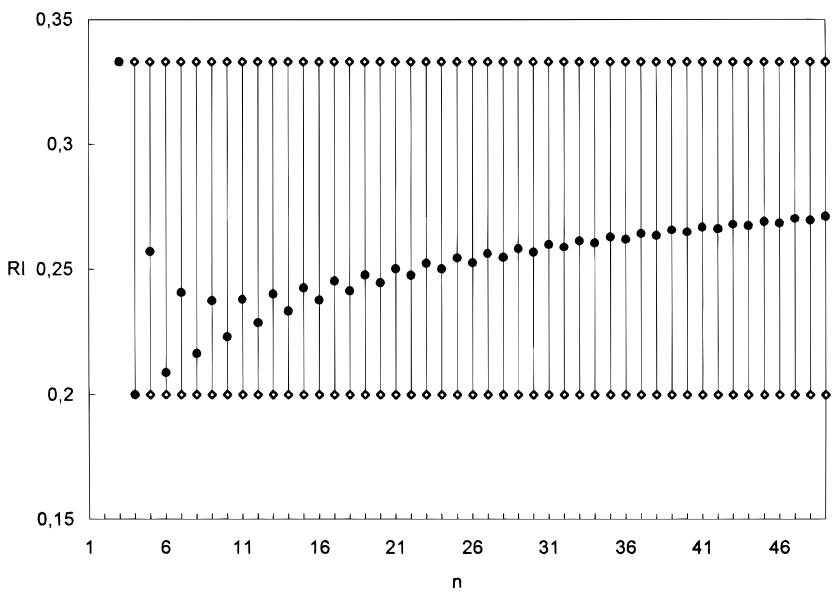

FIG. 6. Possible ranges of ensemble retention index RI(*n*) for indecisive data sets with *n* taxa.

DATA SET 1

out	000000000000000000000000 **0000**	
A	011011110001101111001100 **1111**	
B	101101101010110110101010 **1111**	
C	110110100111011010011001 **1110**	**M = 29**
D	000111011100011101111001 **1100**	**S = 55**
E	000000000011111111111011 **1000**	**G = 68**

DATA SET 2

out	000000000000000000000000000 **0000**	
A	011011110??????????110???110 **1111**	
B	101101101001101111100???110101 **1111**	
C	110110100110110110101011101??? **1110**	**M = 33**
D	??????????1101101001011?????? **1100**	**S = 55**
E	000111011100011101111???011011 **1000**	**G = 66**

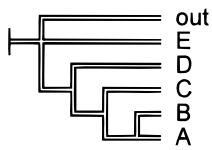

FIG. 7. Two data sets with an indecisive part (left) and a decisive part (right, bold). The decisive part is identical in both data sets and it unambiguously resolves the relationships between taxa *A-E* as on the tree that is shown.

the first but 0.12 for the second (the mean number of steps is 6093/105 in both cases). Put the other way around, the higher DD value of data set 2 does not imply that its possible trees differ more in length than those of data set 1.

Just as data sets 1 and 2, the two data sets shown in Fig. 9 have an identical decisive part and a different indecisive part. The indecisive parts are constructed such that they have the same *M* and *G*, but a different S_{MIN}. As a result, the distributions of possible tree lengths for both data sets have the same shape, but they are shifted with respect to each other (Fig. 10). Even though the distributions are shifted, it can still be argued that the possible trees for data set 3 do not differ more (or less) in tree length than do the possible trees for data set 4. Nevertheless, they have different

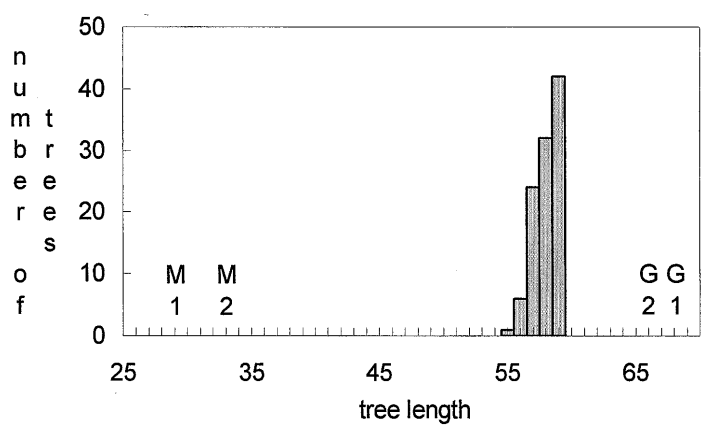

FIG. 8. Data sets 1 and 2 (Fig. 7) have an identical distribution of tree lengths (only fully resolved trees are considered). For each data set, the values of *M* and *G* are indicated.

DATA SET 3

out	00000000000000000000000000000000	**00**	
A	11011011O???11011011O???110110	**11**	
B	101101???11010110 1???110101101	**11**	**M = 32**
C	011???10110101 1???10110101 1???	**10**	**S = 52**
D	???011011011???011011011???011	**00**	**G = 64**

DATA SET 4

out	00000000000000000000000000000000	**00**	
A	01101111000110111100011011 1100	**11**	
B	10110110101011011010101101 1010	**11**	**M = 32**
C	11011010011101101001110110 1001	**10**	**S = 56**
D	00011101110001110111000111 0111	**00**	**G = 64**

FIG. 9. Two data sets with an indecisive part (left) and a decisive part (right, bold). The decisive part is identical in both data sets and it unambiguously resolves the relationships between taxa *A-D* as on the tree that is shown.

DD values: DD equals 0.0740 for data set 3 and 0.0625 for data set 4 (the mean number of steps is 53.6 and 57.6, respectively).

In the two above comparisons, the different DD values are caused by the different homoplasies of the data sets being compared. More precisely, the sensitivity of DD to the amount of homoplasy in data sets follows from the presence of *M* in the scaling factor (DD's denominator) that scales the decisiveness to 1 for data sets without homoplasy. This scaling factor, the difference between the mean step number over all trees and *M*, is by definition the mean homoplasy of the data set over all trees. In this way, DD can be rewritten as the complement of the ratio of minimal and mean homoplasy (see also Archie, 1994):

$$\mathrm{DD} = 1 - \frac{H_{\mathrm{MIN}}}{\overline{H}}.$$

To remove DD's sensitivity to *H*, DD could be redefined such that it refers only to factors that describe the distribution of tree lengths, which would turn DD

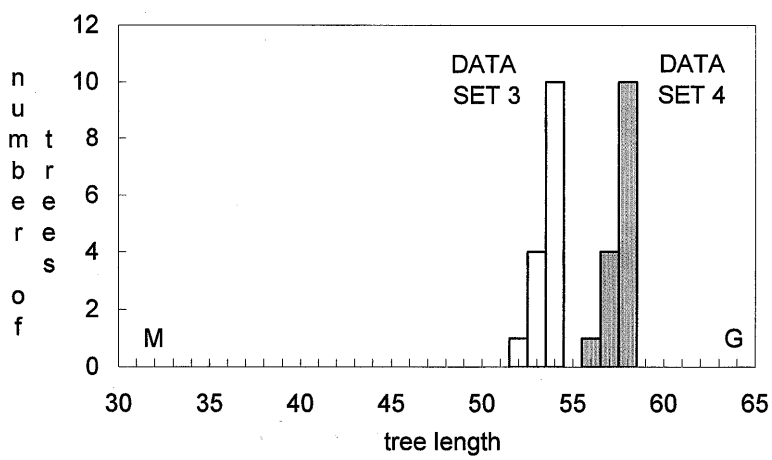

FIG. 10. The distributions of tree lengths (fully resolved trees only) for data sets 3 and 4 (Fig. 9) have an identical shape, but they are shifted with respect to each other. *M* and *G* are the same for both data sets.

into a descriptor of some aspect of the shape of that distribution. However, if this were done one would have to conclude that data set 4 (Fig. 9; DD = 0.0625) and data set 5 (Fig. 11; no homoplasy, so DD = 1), consisting of the decisive part of data set 4, would have the same power to discriminate between trees, because the shapes of their tree length distributions are identical. Assuming that all characters in a data set are solid hypotheses of primary homology (de Pinna, 1991), this seems difficult to defend (even though both data sets yield identical Bremer supports; Bremer, 1988; Farris, 1996): in data set 5 all available evidence is congruent with the same single branching pattern, while data set 4 is full of contradicting evidence. Therefore, even though data set 4 contains little evidence (characters) compared to data set 5, its overall data quality seems to be much higher, which is reflected in the different DD values.

So, if it is accepted that the overall level of homoplasy of a data set influences its decisiveness (e.g., data set 5 is more decisive than data set 4), the questions arise how this influence must be taken into account, and if DD does it in a sensible way. Consider the data sets of Fig. 12: both have a minimal homoplasy of 2 and a mean homoplasy of 2.66, and as a result they have the same DD value (0.25). However, because of the distribution of possible homoplasies, it might be reasonably argued that data set 6 is more decisive than data set 7: it allows at least one possible tree to be discarded (the second one) rather safely because the amount of homoplasy in this tree is twice as much as the amount of homoplasy in the two other trees. Data set 7 has only a single most parsimonious tree (the third one), but the two other trees for this data set are only one step worse, which is a smaller difference than in data set 6.

DATA SET 5

out	00	
A	11	
B	11	**M = 2**
C	10	**S = 2**
D	00	**G = 4**

FIG. 11. Data set 5 consists of the decisive part of data set 4 (Fig. 9) and therefore data sets 4 and 5 have a distribution of possible tree lengths with an identical shape.

DISCUSSION

Goloboff (1991a) proposed the DD index to measure how much the possible trees for a data set differ in length. However, as shown above, data sets with identical distributions of tree lengths (e.g., data sets 1 and 2) or data sets with identically shaped but shifted distributions of tree lengths (e.g., data sets 3 and 4) do not necessarily have the same DD value. The distributions of tree lengths of data sets with identical DD values (e.g., data sets 6 and 7), on the other hand, may be markedly different. Both types of examples show that DD does not adequately measure the degree to which possible trees differ in length.

From a technical point of view, the presented examples are easily explained. The DD value for a given data set is completely determined by (1) the distribution of tree lengths for the data set (from which S_{MIN} and $\overline{S}$ can unequivocally be calculated); and (2) $S - M$, the homoplasy of the data set (which in combination with the tree length distribution yields the scaling factor). The pairs of hypothetical data sets that are contrasted in Figs. 7–11 are constructed such that each time the shape of the tree length distribution is constant (yielding identical $\overline{S} - S_{MIN}$) while the homoplasy varies. Therefore, even if in each pairwise comparison the differences in length of the possible trees are identical, a different DD value is obtained. Given that each time the lowest DD value is obtained for the data set with highest homoplasy, it can be argued that the examples presented in Figs. 7–11 illustrate a desirable property rather than a defect: for the same difference in possible tree lengths, a data set with more homoplasy has lower overall data quality than a data set with less homoplasy.

The example presented in Fig. 12 is constructed differently: the two data sets have identical M, S_{MIN}, and $\overline{S}$, and hence identical DD values. Nevertheless, they differ clearly in how much their possible trees differ in tree length and homoplasy, and therefore they should not be considered equally decisive. Goloboff (1991a: 227) rightfully pointed out that decisiveness of a data set and confidence in its most parsimonious trees are not the same thing, and it might seem that these properties are confounded in this example. However, this is not the case: the possible trees for data set 7 clearly differ less in tree length and in homoplasy

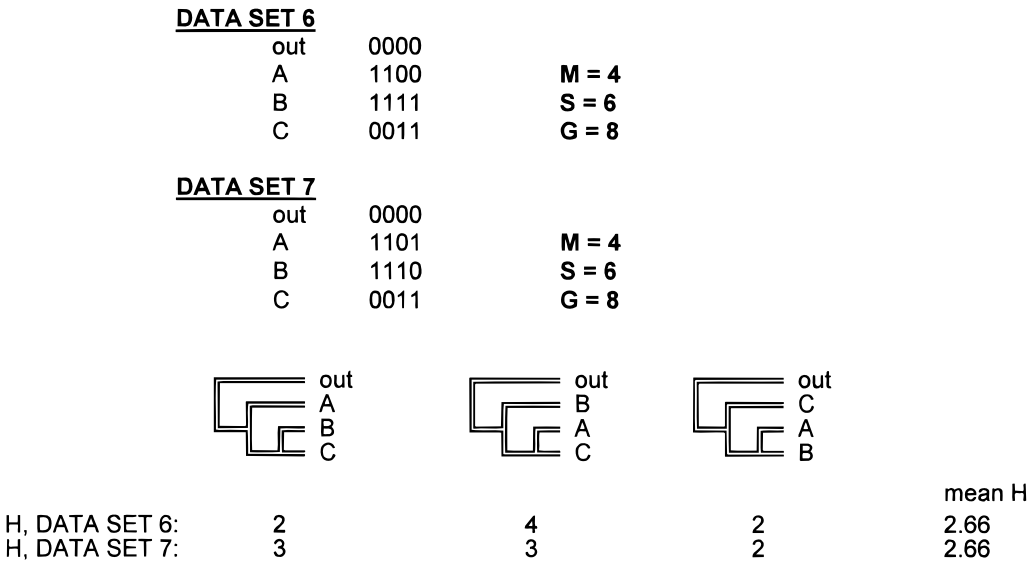

FIG. 12. Data sets 6 and 7 have the same minimal and mean homoplasy but different distributions of tree lengths (listed as homoplasies for the three possible trees).

than the possible trees for data set 6, and therefore data set 7 offers less information for tree choice than data set 6 and it should be considered less decisive.

One could try to modify DD to reflect differences in decisiveness as illustrated in Fig. 12, but rather than to propose and discuss such modifications, we want to point out that this would be a fruitless exercise. As the discussed examples indicate, the concept of data decisiveness seems to be too complex and elusive to be captured in a single and simple index such as DD. Moreover, even if it were possible to devise an index that captures the essence of decisiveness, such a measure would not be of great help. Indeed, what systematists are really interested in is not how safely the worst cladograms for a data set can be discarded, but how strongly the groups that appear in most parsimonious cladograms are supported by the data. As Goloboff (1991a: 227) was well aware, this question is not answered by simply considering data decisiveness. Other approaches (e.g., Felsenstein, 1985; Bremer, 1988, 1994; Kallersjo *et al.*, 1992; Farris *et al.*, 1994, 1996) are better suited to address this question.

ACKNOWLEDGMENTS

We thank Drs. James S. Farris, Pablo Goloboff, and Mike Steel for their constructive comments on (parts of) earlier versions of the manuscript. This research was supported by *Vlaamse Leergangen Leuven* (Travel Grant 11.7.94-94/69 to J.D.L.). J.D.L. is a postdoctoral fellow of the F.W.O., the Fund for Scientific Research—Flanders (Belgium).

APPENDIX: *S* AND *G* FOR INDECISIVE DATA SETS

The number of characters in MIM(n), the minimal indecisive matrix for n taxa + all-zero outgroup ($n \geq 3$; see Introduction), depends only on n and can be obtained as follows (Goloboff, 1991a): let A_i denote a binary character with i 1-entries for a given suite of taxa. Since there are $\binom{n}{i}$ different A_i characters, the total number of characters is

$$\sum_{i=2}^{n-1} \binom{n}{i},$$

which equals $2^n - n - 2$ ($\binom{n}{i}$ stands for $n!/(i!*(n-i)!)$, the number of different ways in which the i 1-entries can be assigned to the n taxa of the ingroup). In the following, square brackets are used to indicate the integer part of a ratio; e.g., $[i/2]$, with i an integer, denotes the integer part of $i/2$.

$S(n)$

According to Steel and Charleston (1995: 371; see also Steel, 1993; Goloboff, 1991b: 396–397), the mean character length of an indecisive data set sensu Goloboff (1991a) is equal to the expected length of a random binary character on a random fully bifurcating tree, for which they provided an exact nonrecursive equation. However, a correction for the absence of uninformative characters in indecisive data sets is necessary, as will be discussed below. First we present an alternative derivation.

The total number of steps or total length, $S(n)$, for MIM(n) is the same on any possible resolved cladogram. In the following derivation, we will assume a completely pectinate cladogram in which the first taxon of the matrix is the sister group of all the following taxa and so on. The logic of the argument is as follows: the number of 1-entries in MIM(n) (denoted as S_{MAX}) provides an upper limit for S. This maximum length is achieved when every occurrence of state one is counted as a single step. However, character state distributions may contain patterns of 1- and 0-entries that require less steps than 1-entries, which leads to a reduction of the required number of steps (compared to the number of 1-entries in the pattern). As will be shown, the patterns that lead to step reduction can be classified into three types, and by summing the occurrences of patterns of these types over the indecisive data matrix, the total number of step reduction can be calculated. If this number is subtracted from S_{MAX}, $S(n)$ results.

Since every A_i character has by definition i 1-entries, the calculation of S_{MAX} is straightforward (the summation operators that appear in the following equations can be eliminated by using finite sum equations as can be found in, e.g., Prudnikov *et al.* (1988):

$$S_{\mathrm{MAX}}(n) = \sum_{i=2}^{n-1} \binom{n}{i} * i = n * (2^{n-1} - 2).$$

A character state distribution will be described as a concatenation of the symbols 0, 1, or x (x stands for either 0 or 1). A subscript i to a symbol or a group of symbols indicates that the symbol or group of symbols is repeated i times. The order of the symbols refers to the order of the taxa in the data matrix. As an example, $x1_3(01)_201_2$ stands for the state distributions 11110101011 or 01110101011 for 11 taxa $A - L$ (assuming that the taxa appear in alphabetical order in the data set).

The first type of step reduction concerns character state distributions of the form

$$x_{n-i-1}01_i \quad \text{with } 2 \le i \le n-1.$$

The 1_i groups of this type appear at the distal end of a pectinate cladogram and hence they require only a single step each (see Fig. 1A for some general examples; exhaustive lists for MIM(6) and MIM(7) are given in Fig. 2A), resulting in a step reduction of $(i - 1)$. The total step reduction of such patterns in the full matrix can be calculated by enumerating all possible i values, and within each i value all possible assignments to the x positions. This yields

$$\mathrm{SR}_1(n) = \sum_{i=2}^{n-1} (i-1) * 2^{n-i-1} = 2^{n-1} - n.$$

The second type of step reduction concerns character state distributions of the forms

$$x_{n-i-j-2}01_i0x_j \qquad \text{with } 2 \le i \le n-2 \text{ and } 0 \le j \le n-i-2,$$

or

$$1_i0x_{n-i-1} \qquad \text{with } 2 \le i \le n-1$$

In this type of pattern (see Figs. 1A and 2A for examples), the 1_i group is a pectinate group of taxa having state 1 that is at both sides delimited by a taxon having state 0 (the proximal zero-taxon in the second subpattern is the outgroup). Each such group with i members requires only two steps, giving a step reduction of $(i - 2)$. For the calculation of the total step reduction in these patterns it has to be taken into account that the 01_i0 group involved in the first subpattern can appear in $(n - i - 1)$ different positions within the string of the state distribution. This yields the total

$$\mathrm{SR}_2(n) = \sum_{i=2}^{n-2} (i-2) * (n-i-1) * 2^{n-i-2} + \sum_{i=2}^{n-1} (i-2) * 2^{n-i-1} = (n-4) * 2^{n-3} + 1.$$

In the third type of step reduction, two groups of the first or the second pattern are separated by a $0(10)_i$

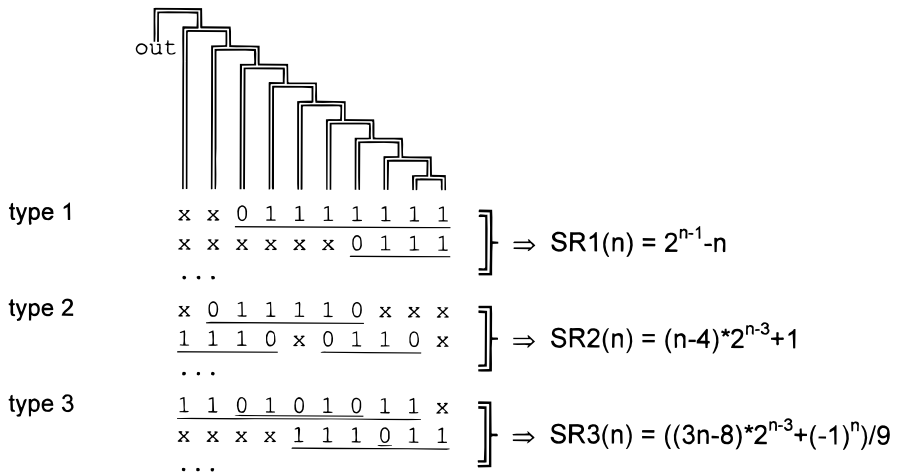

FIG. 1A. Some examples of the three types of patterns in character state distributions that lead to step reduction. The outgroup (out) has state zero. Note that a single character can simultaneously contain patterns of all three types (cf. the second character shown for type 3), and that a single character can contain more than one type 2 or type 3 pattern (cf. the second character shown for type 2).

group ($i \geq 0$). Every such case gives a single supplementary step of reduction in addition to the reduction that is present in the two other patterns that are involved (see Figs. 1A and 2A for examples). Character state distributions that satisfy these conditions are of the following form:

$$x_{n-2^*i-j-5}110(10)_i11x_j \quad \text{with } 0 \leq 2^*i \leq n-5 \text{ and } 0 \leq j \leq n-2^*i-5.$$

As a limiting case, the monophyletic group involved can consist of only a single taxon:

$$x_{n-2^*i-4}110(10)_i1 \quad \text{with } 0 \leq 2^*i \leq n-4.$$

Taking into account that the $110(10)_i11$ group involved in the first subpattern can appear in $(n-2i-4)$ different positions within the string of the state distribution, the following total is obtained:

$$SR_3(n) = \sum_{i=0}^{[(n-5)/2]} (n-2i-4)^*2^{n-2i-5} + \sum_{i=0}^{[(n-4)/2]} 2^{n-2i-4}$$

$$= \frac{1}{9}((3n-8)^*2^{n-3} + (-1)^n).$$

In all other patterns of 0- and 1-entries, every 1-entry is at least at one side (proximally or distally) separated by at least two 0-entries from the next 1-entry. As a result none of these remaining patterns will yield step reduction, and $S(n)$ is obtained as $S_{MAX}(n) - SR_1(n) - SR_2(n) - SR_3(n)$:

$$S(n) = \frac{1}{9}(2^n*(3n+1) - (-1)^n) - (n+1).$$

The length of an indecisive matrix with n taxa in the ingroup can as well be expressed as a function of $t = n + 1$, the total number of taxa in the matrix, i.e., with the all-zero outgroup included:

$$S(\text{MIM}(t-1)) = \frac{1}{9}(2^{t-1}*(3t-2) + (-1)^t) - t.$$

Random Data and Indecisive Data Sets

The same result follows in a straightforward way from Steel (1993), who derived an exact nonrecursive formula for the distribution of the length of binary characters (no missing entries allowed) on fully resolved trees, together with exact nonrecursive formulas for the mean and the variance of this distribution. As discussed by Steel and Charleston (1995: 371), this mean value μ, $(3t - 2 - (-2)^{1-t})/9$, is the expected length of a random binary character on a random fully resolved tree (with random characters defined as characters in which $P(0)$, the probability that a taxon has

MIM(6)

	A_2	A_3	A_4	A_5	
O	000000000000000	00000000000000000000	0000000000000000	000000	
A	111110000000000	11111111110000000000	0000011111111111	111110	
B	100001111000000	11110000001111110000	0111100001111111	111101	
C	010001000111000	10001110001110001110	1011101110001111	111011	
D	001000100100110	01001001101001101101	1101110110110011	110111	
E	000100010010101	00100101010101011011	1110111101101010	101111	
F	000010001001011	00010010110010110111	1111011110110100	011111	
					TOTAL
sr1	000000000000001	00000000010000010012	321002100100000	001234	26
sr2	000000000000000	10000000001000001000	000120001000112	321000	17
sr3	000000000000000	00000000000000000100	001100010011000	011100	9

MIM(7)

	A_2	A_3
O	00000000000000000000	00000000000000000000000000000000000000
A	11111100000000000000	11111111111111100000000000000000000000
B	10000011111000000000	11111000000000011111111110000000000
C	01000010000111000000	10000111100000011110000001111110000
D	00100001000100011000	01000100011100010001110001110001110
E	00010000100010010101	00100010010011001001001101001101101
F	00001000010001001010	00010001001010100100101010101011011
G	00000100001000100101	00001000100101100010010110010110111
sr_1	00000000000000000001	00000000000000100000000010000010012
sr_2	00000000000000000000	10000000000000010000000001000001000
sr_3	00000000000000000000	00000000000000000000000000000000100

	A_4	A_5	A_6	
O	00000000000000000000000000000000	00000000000000000000	0000000	
A	00000000000000011111111111111111111	00000011111111111111111	0111111	
B	00000111111111110000000000111111111	01111100000111111111	1011111	
C	01111000011111100001111110000001111	10111101111000001111111	1101111	
D	10111011100011101110001110001110001	11011171011101110001111	1110111	
E	11011101101100110110110010110010010	11101111011101101100111	1111011	
F	11101111011010101101101010101010100	11110111101110110101010	1111101	
G	11110111011010011101101001101001000	11111011110111011010100	1111110	
				TOTAL
sr1	32100210010000021001000001000000000	43210032100210010000	5432100	57
sr2	00012000100011200010000010000001112	00012300012000111223	0001234	49
sr3	00110001001100000100000000000010000	00111000110012101100	0011110	23

FIG. 2A. All patterns of step reduction in MIM(6) (M = 56, S = 128, S_{MAX} = 180) and MIM(7) (M = 119, S = 305, S_{MAX} = 434), each time assuming a pectinate cladogram with the taxa ordered as in the data matrices. Italic: the 1_i group of type 1 patterns; underscored: the 1_i group of type 2 patterns; bold: the connecting $0(10)_i$ group of type 3 patterns.

state 0, and $P(1)$, the probability that a taxon has state 1, are both equal to 1/2). As mentioned above, Steel and Charleston's (1995) observation that this mean is also equal to the mean character length of an indecisive data set sensu Goloboff (1991a) needs qualification. The close relationship between random data ($P(0) = P(1) = 1/2$) and indecisive data sets arises because the universe of all possible random characters constitutes an indecisive data set (Goloboff, 1991b), which we will denote as $U_1(t)$ (with t the total number of terminals). In this way the expected character length of a random character on a random fully bifurcating tree is equal to the mean character length of $U_1(t)$. $U_1(t)$, containing all 2^t different characters for t taxa, is composed of the following subsets:

- all possible informative characters for t taxa:
  - the $2^{(t-1)} - t - 1$ characters of MIM($t - 1$)
  - the $2^{(t-1)} - t - 1$ characters of the complement of MIM($t - 1$) (i.e., all character states, including those of the all-zero outgroup are switched)
- all possible uninformative characters for t taxa:
  - t characters in which only one taxon has state 0
  - t characters in which only one taxon has state 1
  - one character with all taxa having state 0
  - one character with all taxa having state 1

S(MIM($t - 1$)) is then obtained as $1/2(\mu^* 2^t - 2t)$.

Goloboff (1991b) discussed the relation between random data and indecisiveness on the basis of Archie's (1989) approximate equation for the expected character length on random trees. Archie's approximation, $(3t - 2)/9$, differs only by $(-2)^{1-t}/9$ from Steel's (1993) exact result. This difference is largest for small numbers of taxa, but even then the deviation is relatively small: e.g., for $t = 3$ the exact and the approximate values are 0.7500 and 0.7778, respectively (rounded to four decimals), a difference of $-1/36$; for $t = 10$, 3.1113 (exact) and 3.1111 (approximate), a difference of only 1/4608. Goloboff (1991b: 397) reported a much greater discrepancy between Archie's result and his own calculations, which were based on his (Goloboff, 1991a) recursive formula for the length of indecisive matrices that contain only informative characters, which is exact for $n \geq 7$. Therefore Goloboff (1991b: 397) at the time concluded that the large discrepancy is due to the fact that Archie's equation is only approximate. However, the discrepancy partly follows from a different convention of counting the taxa (considering only ingroup versus considering ingroup + all-zero outgroup), and partly from an inaccurate modification to take into account uninformative characters (Goloboff, pers. comm.).

Archie (1989; see also Archie and Felsenstein, 1993) discussed also a second universe of random characters, obtained by assigning the states of characters at random with the probability that a particular taxon has state 0, state 1, or a polymorphism [01] equal to 1/3 (the polymorphism can equally well be interpreted as a missing entry or an inapplicable character, both commonly represented by a question mark). We will denote this universe as $U_2(t)$. In this model of random data, the exact equation for the expected number of steps per character on a random dichotomous tree is $(2t - 2)/9$ (Archie 1989: 256). This equation is easily verified with the above results (Archie and Felsenstein, 1993: 62, used the inverse of the following argument to derive an exact but recursive formula for the mean character length of $U_1(t)$). $U_2(t)$ contains all 3^t different characters with states 0, 1, and ?, and it can be generated in the following way: first consider $U_1(t)$, which contains all $U_2(t)$ characters without question marks. Next consider $U_1(t - 1)$ and add a single row of question marks in all $\binom{t}{1}$ possible positions. The result is a matrix with $\binom{t}{1}$ times as many characters as $U_1(t - 1)$, and this matrix contains all $U_2(t)$ characters with precisely one question mark. This can be generalized for characters with any number of question marks, and since adding rows of question marks does not change the number of steps of a U_1 matrix, the total number of steps of $U_2(t)$ is obtained as the sum of the number of steps of all composing $U_1(i)$ matrices, $0 \geq i \geq t$:

$$S(U_2(t)) = \sum_{i=0}^{t} \binom{t}{i} * S(U_1(t - i)) = \frac{2}{9}(t - 1) * 3^t.$$

Division by 3^t, the total number of characters, confirms Archie's (1989) result.

$G(n)$

For every character A_i with $i \leq [n/2]$, the maximal number of steps equals i. For every character A_i with

$i > [n/2]$, the maximal number of steps equals $n - i + 1$ (recall that there are $n + 1$ taxa: the all-zero outgroup must be taken into account also). Summation over all possible informative characters gives

$$G(n) = \sum_{i=2}^{[n/2]} \binom{n}{i} * i + \sum_{i=[n/2]+1}^{n-1} \binom{n}{i} * (n - i + 1).$$

Since $\binom{n}{i} = \binom{n}{n-i}$ this can be expressed as

$$\begin{aligned} G(n) &= \sum_{i=2}^{[n/2]} \binom{n}{i} * i + \sum_{i=n-[n/2]-1}^{n-(n-1)} \binom{n}{i} * (n - (n - i) + 1) \\ &= \sum_{i=2}^{[n/2]} \binom{n}{i} * i + \sum_{i=1}^{n-[n/2]-1} \binom{n}{i} * (i + 1). \end{aligned}$$

This equation is equivalent to the pair

$$\begin{cases} G(n_{\text{even}}) = \sum_{i=2}^{n/2} \binom{n+1}{i} * i \\ G(n_{\text{odd}}) = \sum_{i=2}^{(n-1)/2} \binom{n}{i} * (2i + 1) + 2n, \end{cases}$$

which, after elimination of the summations, equals

$$\begin{cases} G(n_{\text{even}}) = (n + 1) * (2^{n-1} - 1) - \dfrac{n+1}{2} * \binom{n}{n/2} \\ G(n_{\text{odd}}) = (n + 1) * (2^{n-1} - 1) - n * \binom{n-1}{(n-1)/2}. \end{cases}$$

Two similar formulas, for the length of the random universe $U_1(t)$ on an unresolved bush, are provided by Steel (1993: 259). Algebraic manipulation ultimately yields

$$G(n) = (n + 1) * (2^{n-1} - 1) - \frac{n+1}{2} * \binom{n}{[(n+1)/2]}.$$

Or in terms of t, the number of taxa with the all-zero outgroup included

$$G(t) = t * (2^{t-2} - 1) - \frac{t}{2} * \binom{t-1}{[t/2]}.$$

REFERENCES

Archie, J. W. (1989). Homoplasy excess ratios: New indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. *Syst. Zool.* **38,** 253–269.

Archie, J. W. (1994). Measures of homoplasy. *In* "Homoplasy. The recurrence of similarity in evolution." (M. J. Sanderson and L. Hufford, Eds.), pp. 153–188. Academic Press, San Diego.

Archie, J. W., and Felsenstein, J. (1993). The number of evolutionary steps on random and minimum length trees for random evolutionary data. *Theor. Pop. Biol.* **43,** 52–79.

Bremer, K. (1988). The limits of amino-acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42,** 795–803.

Bremer, K. (1994). Branch support and tree stability. *Cladistics* **10,** 295–304.

de Pinna, M. C. C. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7,** 367–394.

Farris, J. S. (1989). The retention index and the rescaled consistency index. *Cladistics* **5,** 417–419.

Farris, J. S. (1996). Names and origins. *Cladistics* **12,** 263–264.

Farris, J. S., Källersjö, M., Kluge, A. G., and Bult, C. (1994). Permutations. *Cladistics* **10,** 65–76.

Farris, J. S., Albert, V. A., Källersjö, M., Lipscomb, D., and Kluge, A. G. (1996). Parsimony jackknifing outperforms neighbor-joining. *Cladistics* **12,** 99–124.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39,** 783–791.

Goloboff, P. A. (1991a). Homoplasy and the choice among cladograms. *Cladistics* **7,** 215–232.

Goloboff, P. A. (1991b). Random data, homoplasy and information. *Cladistics* **7,** 395–406.

Källersjö, M., Farris, J. S., Kluge, A. G., and Bult, C. (1992). Skewness and permutation. *Cladistics* **8,** 275–287.

Kluge, A. G., and Farris, J. S. (1969). Quantitative phyletics and the evolution of the Anurans. *Syst. Zool.* **18,** 1–32.

Le Quesne, W. J. (1989). Frequency distributions of lengths of possible networks from a data matrix. *Cladistics* **5,** 395–407.

Prudnikov, A. P., Brychkov, Y. A., and Marichev, O. I. (1988). "Integrals and series." Gordon and Breach Science Publishers, New York. (Translated from the Russian by N. M. Queen).

Steel, M. A. (1993). Distributions on bicoloured binary trees arising from the principle of parsimony. *Discr. Appl. Math.* **41,** 245–261.

Steel, M. A., and Charleston, M. (1995). Five surprising properties of parsimoniously colored trees. *Bull. Math. Biol.* **57,** 367–375.