?=?

TAXEQ3

Safe Taxonomic Reduction
based on
Taxonomic Equivalence

?=?

Mark Wilkinson
Department of Zoology
The Natural History Museum
London, SW7 5BD
England, U.K.

PREFACE

TAXEQ3 is a simple DOS program that facilitates the implementation of safe taxonomic reduction (Wilkinson, 1995a). It also reports a variety of statistics describing the extent and distribution of missing entries in phylogenetic data. TAXEQ was first released in 1992 and comprised three programs that were developed with the support of SERC grant GR/F 87912 to Dr. Mike Benton of the Department of Geology, University of Bristol. For a subsequent release in 1995 I completely rewrote the code and merged the separate programs into one program that was more reasonable in its demands on input file format and which could handle larger data sets and multistate characters. For TAXEQ3 I have further enhanced the output and added randomisation tests of the distribution of missing entries in phylogenetic data.

Your package should contain four files

1. TAXEQ3.EXE - Executable program file
2. TEST.DAT - An example data set
3. TEST.OUT - An example output file
4. TAXEQ3.PDF – The manual

Restrictions and Citation

TAXEQ3 is distributed for the sole purpose of facilitating and promoting research and is a non-commercial product. The programs and this manual may be freely copied and distributed. The recommended citation is:

> Wilkinson, M. 2001 TAXEQ3: software and documentation. Department of Zoology, The Natural History Museum, London.

Please register as a user and report any bugs to me:

e-mail: *marw@nhm.ac.uk* - Tel: +44 (0)20 7942 5164 - Fax: 7942 5433

INTRODUCTION

When poorly known taxa are included in phylogenetic data sets they are sometimes responsible for high numbers of most parsimonious trees (MPTs). Some phylogeneticists have attempted to reduce the number of MPTs by discarding those taxa with the greatest proportions of missing entries (usually incomplete fossils). While this strategy may succeed in reducing the number of MPTs it also carries the potential risk of ignoring the evidence provided by the fossils. Such evidence, although incomplete, may be significant for reconstructing phylogeny.

Ideally, we should only eliminate taxa if we can be certain that their elimination can have no effect upon the subsequent interpretation of the relationships of the remaining taxa. Adding or deleting a taxon to a parsimony analysis can only effect the parsimonious interpretation of the relationships of the remaining taxa if the taxon has a unique combination of character states that is not replicated in any of the other taxa. It follows that where taxa do not have a unique set of character states duplicates can be eliminated without affecting the parsimonious interpretation of relationships among the remaining taxa. Safe Taxonomic Reduction (STR) is a strategy for reducing numbers of MPTs by eliminating taxa that have non-unique combinations of character states (Wilkinson, 1992a, 1995a). Examples of STR may be found in Wilkinson (1995a), Wilkinson and Benton (1995a,b), and below.

TAXEQ3 allows you to implement STR by exploring relations of taxonomic equivalence between taxa and identifying taxa that can be safely eliminated from subsequent parsimony analyses. It also provides some basic summary statistics describing the extent of missing data and its dispersion across taxa and characters, and randomisation tests of the null hypothesis that missing data is distributed randomly.

TAXONOMIC EQUIVALENCE AND STR

TAXEQ3 reports four kinds of taxonomic equivalence. A pair of taxa are **(A)** *actual* equivalents if they are scored the same for all character states and have no missing entries. Actual equivalence is a *symmetric* relation. *Potential* equivalence occurs when two taxa have the same character states wherever the comparison can be made but might differ because some comparisons are invalidated by missing entries in one or both of the taxa. Potential equivalence can be **(B)** *symmetric* - where neither taxon has any characters scored for it that are coded as missing in the other taxon, or **(C)** *asymmetric all one way* - where just one of the pair has some character states that are scored for it but coded as missing in the other taxon, or **(D)** *asymmetric both ways* where both have characters scored that are coded as missing in the other taxon.

Where there is symmetric equivalence, actual or potential, neither taxon has a unique combination of character states and either can be safely deleted, choice being arbitrary. Where there is asymmetric potential equivalence, only if the symmetry is all one way is there scope for STR. Here the taxon with the least missing entries can be safely deleted.

MISSING AND UNINFORMATIVE DATA

The main routine in TAXEQ3 is an exhaustive pairwise comparison of taxa to determine whether they are taxonomic equivalents and the nature of their equivalence. Some data points in phylogenetic matrices may be analytically equivalent to missing entries, for example character states that are scored for only a single taxon in an unordered multistate character. Such uninformative data can lead to differences in taxa that would otherwise be equivalent and, ideally, such differences should be ignored in assessing taxonomic equivalence. A first pass through the data treats all entries other than missing entries as informative. TAXEQ3 then identifies all uninformative data in the matrix and replaces these with missing entries or with informative character states depending on which is appropriate. If there was any uninformative data TAXEQ3 makes a second pass through the data.

TAXEQ3 also recodes characters to use the lowest possible integer values for character states. Thus if the original data included the two states 2 and 4, they would be recoded as 0 and 1 respectively. The recoding is important for TAXEQ3's interpretation of uninformative data points.

TAXEQ3's interpretation of whether a data point is informative or not depends on the nature of the character of which it is a part:

1. Single state characters are always uninformative.
2. Binary characters are always uninformative if one of the character states is scored for only a single taxon.

Note this does not mean that a polar character with only a single ingroup taxon having the primitive condition is uninformative. TAXEQ3 treats data as intrinsically non-polar but polarity can be represented by including a hypothetical common ancestor coded with the hypothesised primitive character state for polar characters and missing entries for non-polar characters, or by the usual inclusion of one or more outgroups. Hence the informativeness of such characters depends upon the inclusion of a hypothetical ancestor or outgroup which also has this condition.

3. Unordered multistate characters are uninformative if there are not at least two character states that are each scored for two or more taxa. In addition, individual character states are uninformative if they are scored for only a single taxon.

4. Ordered multistate characters are interpreted as linear character state trees (Wilkinson, 1992b). Terminal character states are uninformative if they are scored for only a single taxon and are converted to the adjacent character state in the character state network.

If your data includes non-linear ordered character state trees then these must be recoded (e.g. with additive binary coding).

TAXEQ3 outputs the original matrix (subject to any recoding of data to ensure the character states correspond to the lowest range of integers). If any uninformative data is identified, TAXEQ3 reports the changes it makes to the original data and outputs the modified matrix.

DISTRIBUTION OF MISSING ENTRIES

TAXEQ3 provides some basic descriptive statistics for the amount and distribution of missing entries in the data. The output includes the amount of missing data in the matrix and in each taxon, reported as absolute numbers and percentages. It also includes average numbers of missing entries per taxon and per character and a summary of the numbers of characters with each observed number of missing entries. TAXEQ3 also reports the number and percentage of pairwise comparisons of the character states of taxa where one or both of the taxa has missing entries. If the matrix includes uninformative data TAXEQ3 provides this output for each pass of through the data.

It is sometimes stated that missing data is non-randomly distributed. In most cases this is both true and obvious. However it may be useful to be able to test this or the judge this in less obvious cases. If you choose the option of testing the distribution of missing entries then TAXEQ3 performs a simple randomisation test using three measures of the concentration of missing entries in the original data (i.e. prior to any recoding of uninformative data). The measures are:

1. **Taxon Concentration.** For each missing entry, TAXEQ3 counts the number of other missing entries in the same taxon. Taxon concentration is the average value. It increases as missing entries are increasingly concentrated among fewer taxa.
2. **Character Concentration.** For each missing entry, TAXEQ3 counts the number of other missing entries in the same character. Character concentration is the average value. It increases as missing entries are increasingly concentrated among fewer characters.
3. **Matrix Concentration.** For each missing entry, TAXEQ3 counts the number of other missing entries in the same taxon or the same character. Matrix concentration is the average value. It increases as missing entries are increasingly concentrated among fewer taxa and/or characters.

The randomisation test compares these values for the original data to the corresponding values for data in which the same total number of missing entries are randomly distributed. The null hypothesis is that the missing entries are distributed randomly by taxon, by characters or by characters or taxa, depending on the measure. The p-value reported is the probability of achieving as high or higher a taxon, character, or matrix concentration respectively as the original data. The randomization test uses 999 trials and the minimum possible p-value is thus 0.001. TAXEQ3 also reports the expected value under the null hypothesis. The tests can be used separately, but they can also be viewed as three alternative tests of a single underlying hypothesis: that the distribution of missing entries is random. Such use raises the issue of multiple testing and user should adopt more conservative significance cut-offs that preserve the overall experiment-wise TYPE I error rate. I hope to publish a fuller description of the measures and the tests eleswhere.

INPUT DATA FILES

The first line should include the number of taxa and the number of characters separated by one or more blank spaces. Any subsequent information included on the first line will be written to the output file.

Each taxon must be represented by a taxon name and the associated character states. Taxon-names can be any length but must include no blank spaces and must be separated from the taxon's character data by one or more blank spaces or by a line break. Character data can occupy any number of lines and blank spaces will be ignored.

TAXEQ3 accepts up to 99 taxa and 600 characters comprising up to six character states. Numerical symbols (0-5) should be used for all characters with missing entries represented by the '?' symbol.

After the data matrix you can include a listing of any multistate characters you want treated as (linear) ordered. Anywhere after the matrix begin a new line with the command '**`ordered`**' (which must be lower case) and list the numbers of the character that are ordered on subsequent lines separated by one or more blanks or line breaks (see example data set below). Without this information TAXEQ3 cannot differentiate between unordered and ordered multistate characters and will interpret them as either all ordered or all unordered.

An example input file, TEST.DAT, from Bennett's (1989) study of Cretaceous pterosaurs is included with the TAXEQ3 package and is also reproduced below.

USING TAXEQ3

To run the program click on its icon under Windows or enter 'TAXEQ3' in DOS. You will be prompted for the input data file name and an output file name. When specifying file names you can use DOS pathnames so long as these do not exceed a length of 40 characters.

If the data includes multistate characters TAXEQ3 will prompt you as follows:

```
Specify treatment of multistate characters

      1. all ordered

      2. all unordered

      3. mixed

Enter number:
```

If the data includes a mixture of ordered and unordered multistate characters then you must specify this at the end of the input file by including an **`ordered`** command followed by a listing character numbers for all ordered multistate characters. TAXEQ3 will read this information if you specify option 3, and will assume that all other multistate characters are unordered.

If the data contains any missing entries, TAXEQ3 will offer the option of testing the randomness of the distribution of the missing entries.

AN EXAMPLE - CRETACEOUS PTEROSAURS

Included in this section is a description of TAXEQ3's application to the systematic data set published by Bennett (1989) for Cretaceous pterosaurs. This serves to illustrate the principles of STR and the input and output formats for TAXEQ3 analyses. Users should be aware, that STR is only one strategy for enhancing phylogenetic inference when data include poorly known and problematic taxa. Other approaches include the use of reduced consensus methods (Wilkinson, 1994, 1995b), and a posteriori pruning of taxa from MPTs (Wilkinson and Benton, 1995b), these methods can be used as alternatives to, or in conjunction with, STR (see below).

Review of the Original Analysis

Bennett's (1989) data matrix comprises 17 ingroup and 2 outgroup taxa (*Rhamphorhynchus* and *Pterodactylus kochi*) and 14 characters (Box 1). Five of these characters (4,5,9,10,11) are three state characters, and the others are binary [note that only two states are listed in Bennett's (1989:676) description of character 11]. In Bennett's original treatment, characters 4, 5 10 and 11 are unordered and character 9 is linear ordered.

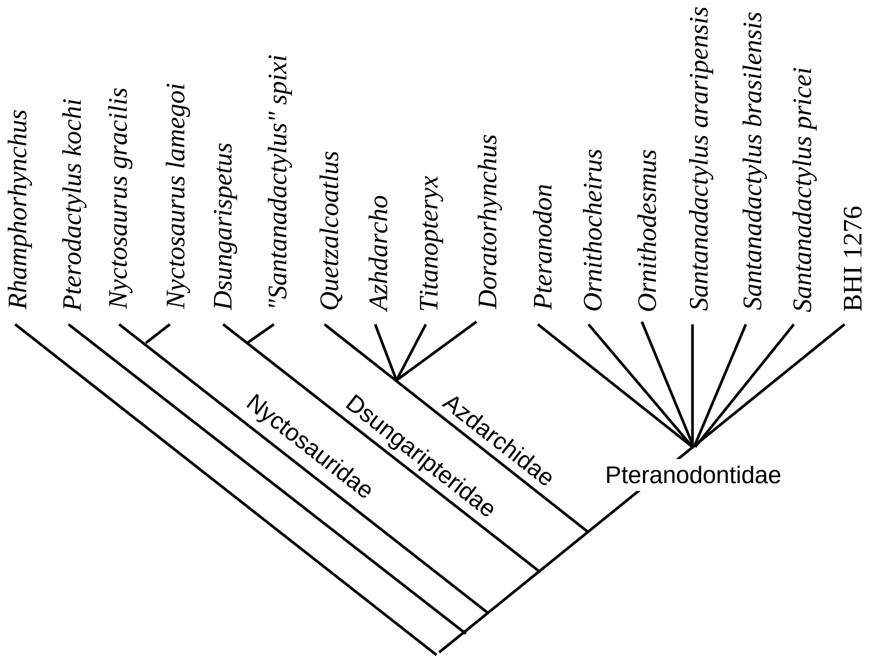

Fig. 1 Bennett's (1989) cladogram of Cretaceous pterosaurs.

Bennett used PAUP 2.4 (Swofford, 1985) to analyse his data and reported finding a single most parsimonious cladogram with length (L) of 19 and consistency index (CI) of 0.947 which is shown in Figure 1. This reported result is somewhat anomalous because PAUP 2.4 outputs only binary (fully dichotomous) trees, and Bennett's tree contains two polytomies. Comparing Bennett's data matrix and cladogram, there are two other unexplained anomalies. Firstly, one of the taxa included in the data matrix, *'Pteranodon' oregonensis* is not included in his cladogram. Following Padian (1984), Bennett considered that this taxon is not a pteranodontid and presumably excluded it from the analysis. Secondly, another taxon, *Santanadactylus brasilensis*, is represented by separate entries for the holotype and paratype in the data matrix, but by only a single terminal in his cladogram. It is not clear from the original treatment whether the single terminal in the cladogram is the holotype, paratype, or a taxon combining the character states of both. The reanalysis below begins with the complete data including *'Pteranodon' oregonensis* and both the holotype and paratype of *Santanadactylus brasilensis*.

Reanalysis of the Complete Data

Using PAUP 3.1.1 (Swofford, 1993), a branch and bound analysis of Bennett's' data yields 120 MPTs (L=20, CI=0.95). A strict component consensus of these 120 MPTs is shown in Figure 2. These results, numbers of trees, tree statistics and consensus tree topology, differ from Bennett's. In particular, the consensus summary is more poorly resolved than his single cladogram.

The differences between these results and Bennett's published results presumably reflect differences in the data that were actually analysed. I attempted to discover these differences by analysing slightly modified data sets that ether included or excluded *'Pteranodon' oregonensis* and included just the holotype or paratype of *Santanadactylus brasilensis* or a taxon comprising data from both the holotype and paratype. The closest match to Bennett's results were found by analyses that excluded both *'Pteranodon' oregonensis* and the paratype of *Santanadactylus brasilensis*. This yielded 10 MPTs (L=19, C=0.947, as in Bennett's analysis). The strict component consensus of these 10 trees is identical to Bennett's tree (Fig. 1).

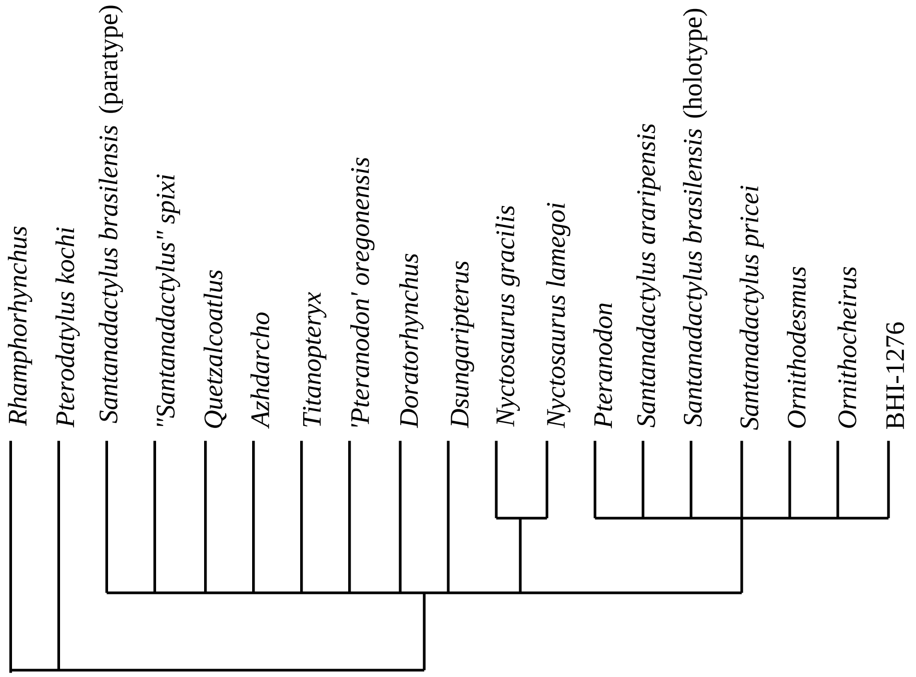

Fig. 2 Strict Component Consensus of 120 MPTs for the full data.

TAXEQ3 Analysis

Bennett's data is illustrated in a format that can be read by TAXEQ3 in Box 1. The ordered command below the character data tells TAXEQ3 that there is a list of ordered characters, the following '9' is the single ordered character.

BOX 1. Example data included in the TEST.DAT file.

```
19 14 Bennett's Cretaceous Pterosaur Data
Rhamphorhynchus              0 0 0 0 0 0 0 0 0 0 0 0 0 0
Pterodactylus_kochi          0 0 0 0 0 0 0 0 0 0 0 0 0 0
Pteranodon                   1 1 1 1 0 1 1 1 2 1 2 1 0 0
Santanodactylus_araripensis  ? ? ? ? ? ? ? ? ? 1 2 1 0 ?
S._brasilensis_(holotype)    ? ? ? ? ? ? ? 1 ? 1 2 ? ? ?
S._brasilensis_(paratype)    ? 1 1 0 1 ? ? ? ? ? ? ? ? ?
S._pricei                    ? ? ? ? ? ? ? ? ? 1 ? 1 0 ?
"S."_spixii                  ? ? ? ? ? ? ? ? ? ? ? ? 1 ?
Ornithodesmus                ? 1 ? ? ? 1 ? ? ? 1 ? ? ? ?
Ornithocheirus               1 1 1 1 0 1 1 1 2 1 2 1 0 0
BHI_1276                     ? ? ? ? ? ? ? ? ? 1 2 1 ? ?
Quetzalcoatlus               1 1 1 2 2 1 1 1 2 0 ? 0 0 0
Azhdarcho                    1 1 1 2 2 1 ? 1 ? ? ? ? ? ?
Titanopteryx                 ? 1 1 2 2 ? ? ? ? ? ? ? ? ?
"Pteranodon_orogonensis"     ? ? ? ? ? ? ? ? ? 0 1 0 ? ?
Doratorhynchus               ? 1 1 2 2 ? ? ? ? ? ? ? ? ?
Dsungaripterus               1 1 0 1 0 1 1 1 1 0 ? 0 1 0
Nyctosaurus_gracilis         1 1 1 1 0 1 0 1 0 2 1 0 0 1
N._lamegoi                   ? ? ? ? ? ? ? ? ? 2 ? ? ? ?
ordered
9
```

The first part of the TAXEQ3 output is just basic information about the file including the name of the input file, the numbers of taxa and characters, any other information included on the first line of the input file, and a list of the taxa included in the input file with assigned numbers (Box 2). TAXEQ3 uses these numbers to refer to taxa in the output file. The next block (Box 3) lists the original character data and how each character is treated (0 and U for ordered and unordered respectively. The distinction is meaningless for binary characters and is U by default).

BOX 2

```
TAXEQ3 : Safe Taxonomic Reduction for
test.dat 13/4/2001 19:15
19 Taxa and 14 characters
 Bennett's Cretaceous Pterosaur Data

Terminal Taxa
          1 = Rhamphorhynchus
          2 = Pterodactylus_kochi
          3 = Pteranodon
          4 = Santanodactylus_araripensis
          5 = S._brasilensis_(holotype)
          6 = S._brasilensis_(paratype)
          7 = S._pricei
          8 = "S."_spixii
          9 = Ornithodesmus
         10 = Ornithocheirus
         11 = BHI_1276
         12 = Quetzalcoatlus
         13 = Azhdarcho
         14 = Titanopteryx
         15 = "Pteranodon_oregonensis"
         16 = Doratorhynchus
         17 = Dsungaripterus
         18 = Nyctosaurus_gracilis
         19 = N._lamegoi
```

BOX 3

```
Original Matrix

1.   00000000000000
2.   00000000000000
3.   11110111212100
4.   ?????????1210?
5.   ???????1?12???
6.   ?1101?????????
7.   ?????????1?10?
8.   ????????????1?
9.   ?1???1???1????
10.  11110111212100
11.  ?????????121??
12.  1112211120?000
13.  111221?1??????
14.  ?1122?????????
15.  ?????????010??
16.  ?1122?????????
17.  1101011110?010
18.  11110101021001
19.  ??????????2????
-------------------
Type:UUUUUUUUOUUUUU
```

The next blocks of output describe the dispersion of missing entries in the data (Box 4).

BOX 4

```
Dispersion of Missing entries across Characters

None   0 characters
 5     1 character
 7     1 character
 8     4 characters
 9     1 character
10     3 characters
11     1 character
12     3 characters

130 Missing entries (48.872%)
Averages
     6.842 per taxon
     9.286 per character
Uninformative between-taxon pairwise comparisons = 1773 (74.060%)
```

The dispersion of missing entries is described by the number of characters having a particular number of missing entries. Uninformative between-taxon pairwise comparisons is the number of comparisons for which one or both taxa have a missing entry.
Information on taxonomic equivalence and the scope for STR is included in the next block (Box 5).

BOX 5

```
==============================================================
Taxonomic Equivalence and Scope for Safe Taxonomic Reduction
==============================================================
Index  ? % (n)  :      Equivalents
--------------------------------------------------------------
 1   0.00% (0)  : 2A*
 2   0.00% (0)  : 1A*
 3   0.00% (0)  : 4C* 5C* 7C* 9C* 10A* 11C*
 4  71.43% (10) : 3E 5D 6D 7C 9D 10E 11C 13D 14D 16D
 5  78.57% (11) : 3E 4D 6D 7D 8D 9D 10E 11D 13D 14D 16D
 6  71.43% (10) : 4D 5D 7D 8D 9D 11D 15D 19D
 7  78.57% (11) : 3E 4E 5D 6D 9D 10E 11D 13D 14D 16D
 8  92.86% (13) : 5D 6D 9D 11D 13D 14D 15D 16D 17E 19D
 9  78.57% (11) : 3E 4D 5D 6D 7D 8D 10E 11D 13D 14D 16D
10   0.00% (0)  : 3A* 4C* 5C* 7C* 9C* 11C*
11  78.57% (11) : 3E 4E 5D 6D 7D 8D 9D 10E 13D 14D 16D
12   7.14% (1)  : 13C* 14C* 15D 16C*
13  50.00% (7)  : 4D 5D 7D 8D 9D 11D 12E 14C 15D 16C 19D
14  71.43% (10) : 4D 5D 7D 8D 9D 11D 12E 13E 15D 16B 19D
15  78.57% (11) : 6D 8D 12D 13D 14D 16D 17D
16  71.43% (10) : 4D 5D 7D 8D 9D 11D 12E 13E 14B 15D 19D
17   7.14% (1)  : 8C* 15D
18   0.00% (0)  : 19C*
19  92.86% (13) : 6D 8D 13D 14D 16D 18E
==============================================================
```

The listing of taxonomic equivalence includes each taxon (the *Index*), the percentage and absolute number of missing entries it has, each taxon with which it is equivalent and a letter indicating the nature of the taxonomic equivalence. Codes A, B, C all indicate that the taxon can be safely deleted provided the index taxon is retained. TAXEQ3 output includes a key to these codes which is given in Box 6.

BOX 6

```
KEY to Taxonomic Equivalence
A – Actual equivalents (symmetric)
    can be safely deleted if Index is retained
B – Potential equivalents (symmetric)
    can be safely deleted if Index is retained
C – Potential equivalents (asymmetric all one way)
    can be safely deleted if Index is retained
D – Potential equivalents (asymmetric both ways)
    cannot be safely deleted
E – Potential equivalents (asymmetric all one way)
    Index can be safely deleted if equivalent is retained
* - Taxon must originate from same node as Index taxon in
    any MPT (assuming no arbitrary resolutions)
! – Taxon has no informative data
```

TAXEQ3 reveals that there is much scope for STR. For example, taxa 1 and 2 (the two outgroups) are identical: one is sufficient to root the tree. In addition there are many cases of asymmetric potential equivalence with the asymmetry all one way (indicated by a C). A total of 12 (2, 4, 5, 7, 8, 9, 10, 11, 13, 14,16, 19) taxa have their character states duplicated in other taxa and add nothing to the analysis. These taxa can be eliminated under the constraint of STR.

Taxa marked with an * are ones that, in addition to showing equivalence to the respective index taxon sufficient to allow their safe deletion, are not also potentially equivalent to any taxa that have characters coded differently from the index taxon. In other words, their combination of character states is subsumed in the index taxon and no other taxon also subsumes that combination of character states in combination with some other character states (that are not missing entries) different from those of the index taxon. In any parsimony analysis, such taxa would be expected to originate from the same node as the index taxon and can be separated from that node only as a result of arbitrary resolutions (see Wilkinson, 1995c).

Thus, independent of any parsimony analysis, TAXEQ3 reveals that there are several groups of taxa that should be associated with each other in any MPT, and that each group can be represented by a single taxon in order to expedite the analysis, as set out below in Table 1.

TABLE 1. Summary of TAXEQ3 analysis showing groups of associated taxa that must arise from the same node in any MPT. For each group only a single taxon is retained in subsequent analyses.

Group	Included	Excluded
1.	*Rhamphorhynchus*	*Pterodactylus kochi*
2.	*Pteranodon*	*Santanadactylus araripensis*, *S. brasilensis* (holotype), *S. pricei*, *Ornithodesmus*, *Ornithocheirus*, and BHI 1276
3.	*Santanadactylus brasilensis* (paratype)	
4.	*Quetzalcoatlus*	*Azhdarcho, Titanopteryx, and Doratorhynchus*
5.	*'Pteranodon' oregonensis*	
6.	*Dsungaripterus*	*"S." spixi*
7.	*Nyctosaurus gracilis*	*N. lamegoi*

TAXEQ3 then modifies the input data so that uninformative data points will be ignored in the investigation of taxonomic equivalence and in determining the scope for STR. In this case TAXEQ3 detects uninformative data in Bennett's matrix and outputs a list of changes that it makes (Box 7) and the modified matrix (Box 8).

BOX 7

```
Recoded T6 C5 -> ?
Character 14 is parsimony uninformative
```

Uninformative data points, other than those represented by '?' in the original matrix are recoded as '?' in the modified data matrix.

For example, the last character (character 14) is a binary character in the original data but has only a single taxon (18) scored with state 1 and is therefore uninformative.

Character 5 is a three state character in the original matrix. It has a single taxon (has a single taxon (6) coded with state 1. Given that this character is treated as unordered the data point is uninformative and is replaced by a '?' in the modified data matrix.

The changes to the matrix are highlighted in Box 7. An uninformative character becomes all '?'s. T and C stand for taxon and character.

BOX 8

```
Modified Character Matrix

1.   0000000000000?
2.   0000000000000?
3.   1111011121210?
4.   ?????????1210?
5.   ???????1?12???
6.   ?110??????????
7.   ?????????1?10?
8.   ????????????1?
9.   ?1???1???1????
10.  1111011121210?
11.  ?????????121??
12.  1112211120?00?
13.  111221?1??????
14.  ?1122?????????
15.  ?????????010??
16.  ?1122?????????
17.  1101011110?01?
18.  1111010102100?
19.  ?????????2????
------------------
Type:UUUUUUUUOUUUUU
```

There are slightly more uninformative data points in the data than indicated solely by '?'s, primarily because character 14 is uninformative. With this example, taking account of all uninformative data points rather than just the '?'s changes the relations of taxonomic equivalence slightly (see the file TEST.OUT which includes the results of TAXEQ3's second pass through the data) but has no effect upon the scope for STR.

Box 9 shows the results of the randomisation tests of the concentration of missing entries in the original (unmodified) data. Missing entries are non randomly distributed, but this is due to concentration in particular taxa. The distribution appears random with respect to characters.

BOX 9

```
*** Distribution of Missing Entries in Original Data ***

Matrix Concentration = 18.4462 (p =  0.001, expected = 15.0906)

Taxon Concentration = 9.7231 (p =  0.001, expected = 6.3283)

Character Concentration = 8.7231 (p =  0.700, expected = 8.7623)
```

Reanalysis after STR

Analysis of the reduced data set after elimination of those taxa that can be safely deleted yields 9 MPTs (L=20, CI=0.95). As expected, STR does not alter tree statistics (Wilkinson, 1995a). The strict component consensus of these 9 MPTs is completely unresolved, however, the considerable reduction from 120 to 9 MPTs facilitates inspection of the MPTs (Fig. 3).

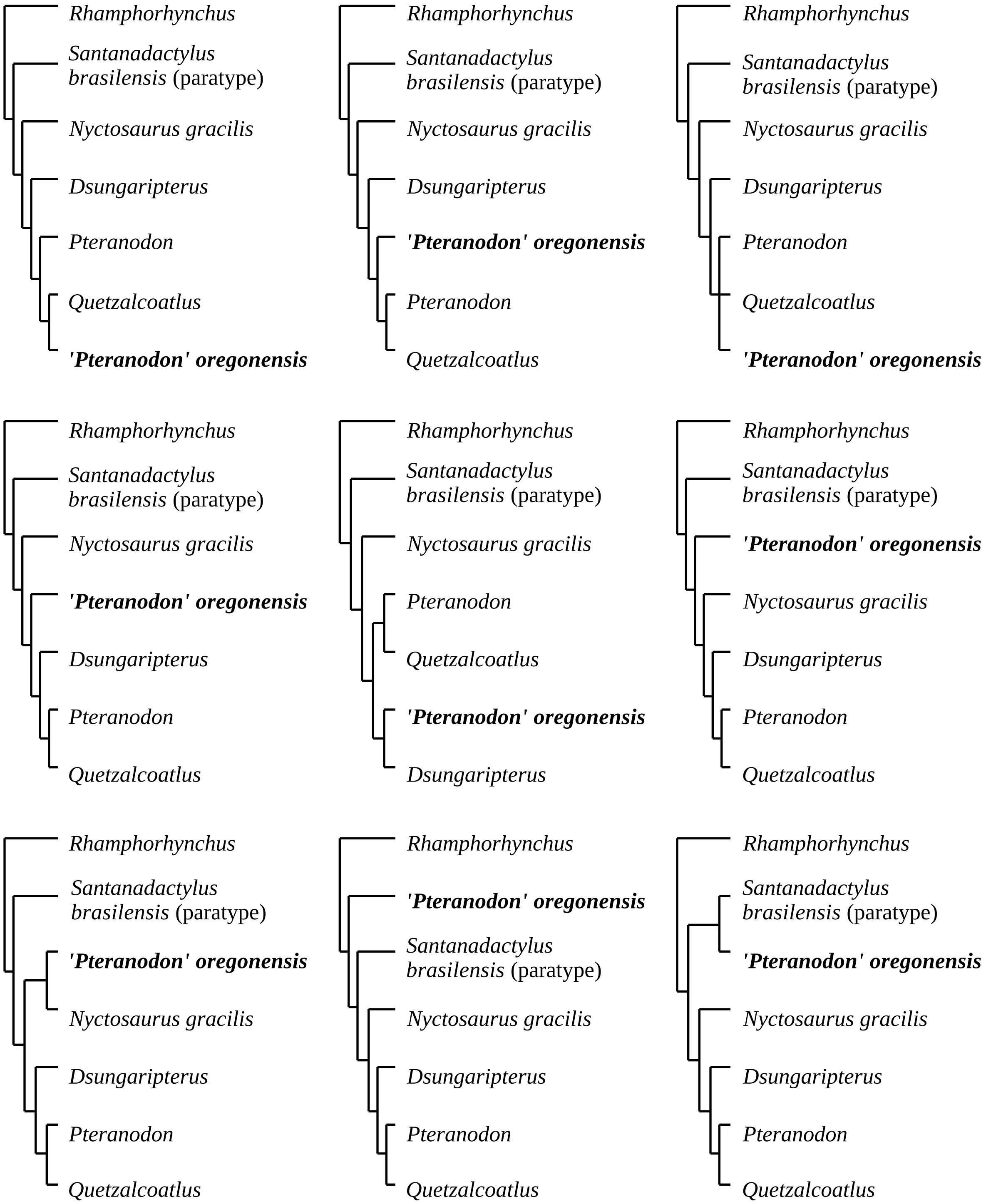

Fig. 3. The nine MPTs found after the first round of STR. Note these trees differ only in the position of*'Pteranodon' oregonensis.*

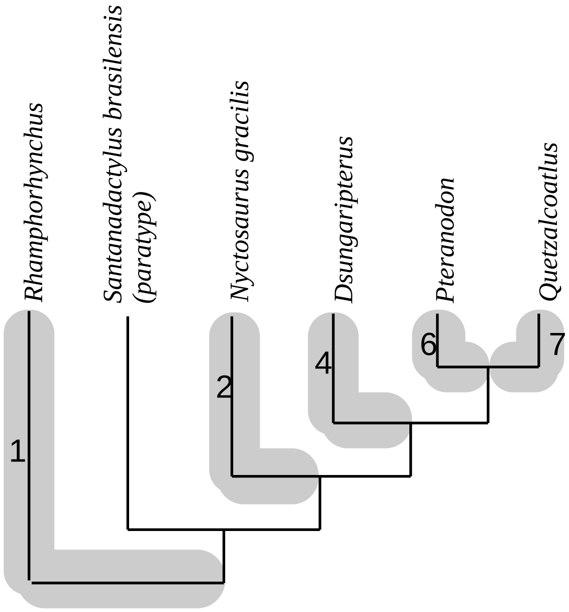

Fig. 4. Strict reduced cladistic consensus of the nine MPTs of Fig. 3, forrmed by pruning *'Pteranodon' oregonensis*. The positions of associated convex groups are indicated by the grey areas with numbers corresponding to those of Table 1.

Inspection reveals that the MPTs differ only in the placement of *'Pteranodon' oregonensis* and are otherwise identical. At this stage, the results can be summarised by pruning *'Pteranodon' oregonensis* to produce a fully resolved (reduced cladistic) consensus summary of relationships among the remaining taxa that can be further elaborated by considering the associations of these taxa with those excluded through STR (Fig. 4), particularly those that must arise from the same node as the retained taxon in any MPT. Ignoring any arbitrary resolutions, each of these groups must be *convex* (Meacham and Duncan, 1987) relative to the fully resolved reduced cladistic consensus because they either arise from the last common ancestor of their index taxon and its sister group in the consensus or from a node closer to the index in which case the convex group is also a clade. Whether these groups are also clades requires additional information and depends on the parsimonious optimisation of their character states identifying derived characters that separate each group from the stem of the tree. Considering the strict component consensus of the 120 MPTs for the full data (Fig. 2) it is apparent that two of these groups (2 and 7) are also clades in each of the MPTs.

One way of exploring whether the other groups are also clades is to *'Pteranodon' oregonensis* from the 120 MPTs for the full data. This yields just 10 distinct pruned MPTs. The pruned trees differ only in the relationships within the Pteranodontidae, and one of them, which is also their strict component consensus (Fig. 5), indicates that all of these convex groups are also clades. Possible positions of *'Pteranodon' oregonensis* are also indicated in Figure 5. These results are very similar to Bennett's original tree (Fig. 1) with the exception that the paratype of *Santanadactylus brasilensis* is recovered as the sister group of the other ingroup taxa.

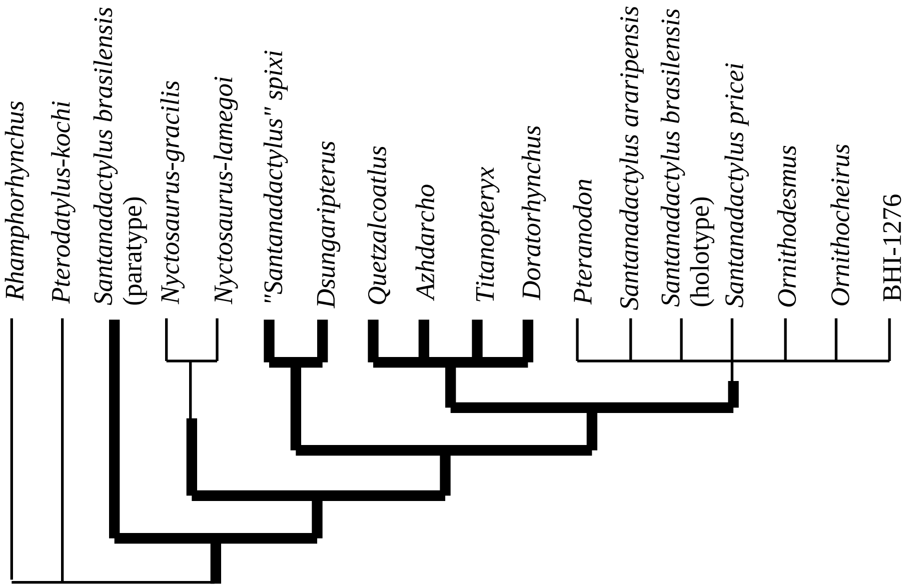

Fig. 5 Tree summarizing relationships supported by the parsimonious interpretation of Bennett's (1989) data. This tree was among the 10 distinct topologies produced by pruning *'Pteranodon' oregonensis* from the 120 MPTs for the full data, and is also the strict component consensus for the 10 pruned trees and the primary strict reduced cladistic consensus of the full set of 120 MPTs. Heavy lines indicate the possible positions of *'Pteranodon' oregonensis.*

Reapplication of TAXEQ3

The reduced data set of seven taxa can also be reanalysed with TAXEQ3. This produces interesting results, and an incomplete summary of the TAXEQ3 output is shown in Box 10. TAXEQ3 identifies that, with respect to the reduced data set, *'Pteranodon' oregonensis* has no informative characters and can be safely deleted. This complete lack of evidence explains why this taxon has such variable positions in the previous analysis. Similarly, *Quetzalcoatlus* groups with *Pteranodon* and can also be excluded. Note that this reflects the grouping of these taxa in the previous analysis.

BOX 10

```
TAXEQ3 : Safe Taxonomic Reduction for test2.dat 13/4/2001 19:25
Terminal Taxa                                Original Matrix          Modified Matrix
          1 = Rhamphorhynchus                  1. 00000000000000        1. ??00??0?0?????
          2 = Pteranodon                       2. 11110111212100        2. ??11??1?2?????
          3 = S._brasilensis_(paratype)        3. ?1101?????????        3. ??10??????????
          4 = Quetzalcoatlus                   4. 1112211120?000        4. ??1???1?2?????
          5 = "Pteranodon_oregonensis"         5. ?????????010??        5. ??????????????
          6 = Dsungaripterus                   6. 1101011110?010        6. ??01??1?1?????
          7 = Nyctosaurus_gracilis             7. 11110101021001        7. ??11??0?0?????
                                             -------------------      -------------------
                                             Code:UUUUUUUUOUUUUU      Code:UUUUUUUUOUUUUU
==============================================================
Taxonomic Equivalence and Scope for Safe Taxonomic Reduction
==============================================================
Index  ? % (n)  :      Equivalents
--------------------------------------------------------------
 1  71.43 (10)  : 5C!
 2  71.43 (10)  : 4C* 5C!
 3  85.71 (12)  : 4D 5C!
 4  78.57 (11)  : 2E 3D 5C!
 5 100.00 (14)  : 1E 2E 3E 4E 6E 7E
 6  71.43 (10)  : 5C!
 7  71.43 (10)  : 5C!
==============================================================
```

Discussion and Summary

This restudy of Bennett's (1989) data illustrates the use of TAXEQ3. It also produces results that are not entirely consistent with the original analysis. The first application of TAXEQ3 identifies taxa that can be eliminated because their character states are represented in other taxa. This reduces the number of MPTs from 120 to 9, facilitating the individual inspection of the MPTs and the production of a clear consensus representation of relationships among the remaining taxa that are unambiguously supported by the parsimonious interpretation of the complete data. Taxa that were excluded under STR, but which are identified as grouping with other included taxa can be incorporated into the consensus. Those groups that were identified as necessarily arising from the same node must be convex if arbitrary resolutions are suppressed. To determine if these convex groups are also clades requires additional work. In this example, pruning *'Pteranodon' oregonensis* from the original 120 MPTs clarifies the relationships among the remaining taxa. The analysis allows the poor resolution of the strict component consensus tree for the 120 MPTs supported by the complete data to be attributed to (1) the lack of data sufficient to resolve relationships within the Pteranodontidae and Azdarchidae, and (2) the obfuscatory effect of the problematic *'Pteranodon' oregonensis*.

Note that although *'Pteranodon' oregonensis* is the main source of problems with the complete data TAXEQ3 only identified it as a candidate for STR in the second iteration, after other taxa have been removed. Although *'Pteranodon' oregonensis* can be safely eliminated from the reduced data this does not guarantee that its removal from the complete data will also be safe. Also, most of the taxa that are eliminated in the first round of STR can be reattached to the consensus that excludes *'Pteranodon' oregonensis*. Users should bear in mind that not all taxa that satisfy the constraints of STR need be problematic and that TAXEQ3 analysis may be most profitably used as a guide to candidates for experimental a priori deletion of taxa from the analysis or a posterior pruning of taxa from trees.

Differences between my results and Bennett's original tree concern the relationships of *'Pteranodon' oregonensis* and the paratype of *Santanadactylus brasilensis*, both of which appear to have been excluded from Bennett's (1989) analysis. Interestingly, the paratype of *Santanadactylus brasilensis* does not appear to be closely related to the holotype, but is inferred to be the sister taxon of all other pteranodontids and cannot be included in any of the four families recognised and diagnosed by Bennett on the basis of his analysis. As Bennett noted, the holotype and paratype specimens came from the same locality but from different concretions. Given this, the apparent diversity of pterosaurs from the Santana formation, and the results of the phylogenetic analysis it seems probable that the holotype and paratype are not the same species. However, the paucity of data and incompleteness of the fossils means that any systematic conclusions should be tempered with extreme caution. The view that '*Pteranodon' oregonensis* is not a pteranodontid is neither supported nor challenged by the analysis because it is equally parsimonious to place this taxon within or outside the pteranodontids (Fig. 5). The analysis indicates that '*Pteranodon' oregonensis* is not a member of Pteranodontidae or Nyctosauridae. It also suggests that this taxon is more closely related to the pteranodontids than are either of the outgroups, but it should be borne in mind that this is primarily a result of how the tree is rooted.

Acknowledgements

I am grateful to Mike Benton and SERC for their support in the development of the inmitial ideas behind and implementation of TAXEQ. Thanks also to Ginny Shires for help in analysing the pterosaur data, and the University of Glasgow New Initiatives Fund for providing funds for her research assistantship.

REFERENCES

Bennett, S. C. 1989. A pteranodontid pterosaur from the Early Cretaceous of Peru, with comments on the relationships of Cretaceous pterosaurs. *Journal of Vertebrate Paleontology* **63**:669-677.

Meacham, C. A. and Duncan, T. 1987. The necessity of convex groups in biological classification. *Systematic Botany* **12**:78-90.

Padian, K. 1984. A large pterydactyloid pterosaur from the Two Medicine Formation (Campanian) of Montana. *Journal of Vertebrate Paleontology* **4**:516-214.

Swofford, D. L. 1985. *PAUP: phylogenetic analysis using parsimony*, version 2.4, Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Swofford, D. L. 1993. *PAUP: phylogenetic analysis using parsimony*, version 3.1.1, Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Wilkinson, M. 1992a. Consensus, compatibility and missing data in phylogenetic inference. *PhD thesis, Dept. of Geology, University of Bristol*.

Wilkinson, M. 1992b. Ordered versus unordered characters. *Cladistics* **8**:375-385.

Wilkinson, M. 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* **43**:343-368.

Wilkinson, M. 1995a. Coping with missing entries in phylogenetic inference using parsimony. *Systematic Biology*. **44**:501-514.

Wilkinson, M. 1995b. More on reduced consensus methods. *Systematic Biology.* **44**:435-439.

Wilkinson, M. 1995c. Arbitrary resolutions, missing entries and the problem of zero-length branches in parsimony analysis. *Systematic Biology* **44**:108-111.

Wilkinson, M. and Benton, M. J. 1995a. Missing data and rhynchosaur phylogeny. *Historical Biology* **10**:137-150.

Wilkinson, M. and Benton, M. J. 1995b. Sphenodontid phylogeny and the problems of multiple trees. *Philosophical Transactions of the Royal Society B* **351**:1-16.