CABIOS

Vol 2 no 1 1986
Pages 33 – 39

A package of computer programs for handling taxonomic databases

R.J.Pankhurst

Abstract

A package of programs is described which processes taxonomic data, suitable to use when preparing monographs, handbooks, Floras or Faunas, in which species or other taxa are described and identified. There is also an interactive program for specimen identification, and conversion routines which prepare data for numerical taxonomy (clustering and cladistics). The programs are equally suitable for botany or for zoology, or even for non-biological data.

Introduction

A very fundamental aspect of taxonomic research is the data which describe the specimens, species or other taxa from which various conclusions are drawn. By descriptive data is meant morphological information in the widest sense, covering not just the traditional kinds of characters, but also any other kinds of data, such as microscopical or chemical information. Records relating to bibliography, geographical distribution, or citations of specimens are not considered here, since they may be easily handled in other ways by using conventional database software
If the descriptive data are stored in a computer, then they may be called a taxonomic database, and can then readily be revised and communicated as original data, which was scarcely possible in the past, when what was published was only the conclusions based on it. For the purposes of international data exchange, a standard format would be desirable, and the one which appears to be the most comprehensive and flexible has been chosen.

It must be stressed that the package described here is not a database system. Database management systems for taxonomic data in the above sense do not yet exist as such: it is possible that some of the most sophisticated relational systems available on mainframes might be adequate, but this has not been demonstrated See Allkin (1984) for comments on this problem. Efforts are also being made currently to develop a suitable DBMS.

Data format

The format used is called DELTA (DEscription Language for TAxonomy), and was defined by Dallwitz (1984) It has a completely free format, meaning that the user can arrange his data as he wishes, without having to follow rules about putting numbers in the right columns. There is also no restriction on the kinds of characters which can be expressed, and no limit to the description of variability. There is no provision at present for recording probabilistic data, or frequencies of characters, except as comments.

Each character is given a number and a name, e.g.

#23. Wings <presence>/

and, if appropriate, a list of states, e.g.

1. present/
2. absent/

A qualitative character such as this may have two or more states which take the form of strings of words, e.g.

#24. Colour of wings/
1. white/
2. yellow/
3. blue/

A distinction can be made between qualitative characters which are sequential, e.g. size = small, medium or large,and those which are not, e.g. colour, character 24. On the other hand, quantitative characters which involve integers or real numbers, such as e.g.

#16. Number of petals/
#29. Length of mandible/

do not require any states to be defined, although there is the possibility of expressing the units of measurement, e.g.

#29 Length of mandible/mm/

Comments can be placed wherever they are wanted, e.g.

#14. Presence of stem <not counting stolons>/

The descriptions of objects (species, families, diseases or other objects) are preceded by the object number and name, e.g.

#7 Bellis perennis <British form>/

and followed by a list of pairs of character numbers with their states, e.g.

23,1 would mean 'wings present'
23,1/2 would mean 'wings present' or 'wings absent'
16,4–5 would mean 'petals 4 or 5' and
29,2.5–6.3 means 'mandibles 2.5 to 6.3 mm'

In the case of integer characters it is even possible to have discontinuous states, e g.

16,3/5–7 meaning 'petals 3 or 5 to 7'

It is also possible to define characters which consist of text only, which can be useful when preparing descriptions. A character which is unknown can either be left out or marked with a special

Botany Department, British Museum (Natural History), Cromwell Road, London SW7 5BD, UK

state of U Similarly a totally variable character, which shows all the available states, can be marked with a V There is another important kind of character state, when a character is not merely unknown, but impossible e.g. the colour of petals in a plant which has no petals Inapplicable characters such as this are described with a dash, e.g. 24, – . The difference between an unknown and an inapplicable character is often important, as in the construction of identification keys The rules which describe the dependence of characters on each other take a simple form, e g

23,2·24

means that if character 23 (presence of wings) has state 2 (absent), then character 24 (colour of wings) is impossible These dependency relations are particularly important when diverse organisms are being described e.g. in a Flora, since they can be used to find which characters are comparable, and which are not (Pankhurst, 1983b)

DELTA also includes a number of directives which supply other data, e g

*NUMBER OF CHARACTERS 24

or which give instructions to processing programs, e g

*TRANSLATE INTO NATURAL LANGUAGE

A directive which is important in constructing keys is

*KEY STATES

This relates to quantitative characters only, and shows how these characters are to be used in the construction of a key. Such characters have to be artificially converted into a (pseudo-) qualitative form at some stage in order to have contrasting states. This could be carried out by an algorithm which divides up the range of variation in order to maximise the use of information from the character, but here it is done explicitly, e.g. character 29 might be transformed as

£29 mandibles <length>/
1 0 to 1 mm/
2 1 1 to 2 mm/
3. 2 1 mm or over/

This conversion is specified in the directive as

29,0 – 1/1.1 – 2/2.1

and is used only in the process of key construction. The true values of the range(s) of quantitative variation are preserved both in the data and in the various outputs

Existing programs

There does not yet exist a database system for the DELTA format, but rather a selection of different programs which read the standard data and produce different kinds of output Many of the methods used in the following programs have already been described in a textbook (Pankhurst, 1978a)

Construction of identification keys.

The program constructs and prints keys which, like the traditional kind, are used manually, either in the field or in the laboratory. The method used here is a heuristic one, whereby a tree structure of sets of contrasting questions is built from a choice of the alternative characters available at each level. The selection is made by means of a figure of merit which is described in more detail below, and expresses a compromise between the need to make the key both reliable and short, and the actual distribution of discriminating information. An early version of this program is described by Pankhurst (1971). Both the two common alternative forms of key can be produced (parallel and indented). An option is provided to set the maximum number of characters (NCM) which are to be considered for each lead. In practice, it is usual to search for only one at a time, since the computation time increases greatly with higher numbers. Multiple character leads are still produced, since all the auxiliary characters which correlate with the main character are found immediately afterwards, and added into the lead Weights can be assigned to both characters and objects in order to produce different types of keys from the same data. Incomplete data, where not every taxon is different from every other, can be used to create partial keys Alternatively, the program can list pairs of taxa which are not distinct so that the user can seek additional data. Variable characters are fully supported, and if taxa are very variable, they can be described more than once if need be If it subsequently proves possible to merge different versions of a taxon under the same name, and so to reduce the size of the key, the program will do this.

The figure of merit is arranged so that the 'best' choice of lead will have the lowest score As used currently, it has four components which are added together, in order to make a compromise between the conflicting requirements of a key

(i) The separating power The separation coefficient S of a character is the ratio of the number of pairs of taxa in which this character differs to the total possible number of pairs For N taxa, the total number of pairs is just

$$\frac{N(N-1)}{2} \text{ i.e. } {}^{N}C_{2}$$

Since S increases for 'better' characters, the first component is taken as 1 – S. This has a range between 0 and 1, and of course entirely ignores the ease of observation and the reliability of the character An information statistic could also be used here

(ii) The excess taxa. This is only different from zero if taxa are so variable that a given taxon has to be put into more than one branch. If there are k branches in a lead, and effectively n_1 taxa in the first branch, and so on, with n_k in the last, then the second component is

$$(n_1 + n_2 + \ldots\ldots + n_k) - N$$

This has the effect of discouraging the use of variable characters in favour of the more constant ones.

(iii) Dichotomous branching. This is encouraged, but not enforced. If the user wishes to have a strictly dichotomous key, then all the characters supplied must be binary The third component is

[(2*absolute value of (IC − NCM) + K − 2]²

where IC is the number of characters being considered for the lead, and NCM was defined above. The absolute value of (IC − NCM) will be zero if the desired number of characters in a lead is equal to the actual number of characters, and larger otherwise. Similarly (K − 2) is zero where K = 2, i e. dichotomous branching, and greater otherwise. Hence this component will encourage dichotomous branches with the desired number of characters.

(iv) Equal distribution of taxa in each branch. As a rule of thumb, it may be said that the most efficient key (having the shortest average path to identify a taxon) will often have as nearly as possible an equal number of taxa in each question of a lead. That this is not strictly true can easily be shown by constructing counter examples, but it is true often enough to be useful The average number of taxon in each question is N/k for k branches, whereas the actual number is n_i Hence each ratio of n_i divided by N/k i.e. kn_i/N should be ~1. The fourth component used is the sum of all i from 1 to k of

$$1 - \frac{kn_i}{N}$$

with the sign removed For weighted taxa, n_i and N reflect w_i, where w_i is the weight of taxon i, by counting the ith taxon w_i times, instead of once.

Weights for characters are positive integers, but are not included in the figure of merit. Instead, at each stage in the key where a new lead is being selected, the characters are considered in order of their weight Provided that the character(s) of the highest weight do in fact contain states which distinguish the taxa, and satisfy the dependency rules, these will be considered to the exclusion of all the rest Naturally, no degree of weighting will force the use of characters which do not contain sufficient data, or which are logically impossible in the context Hence, if all characters are given different weights, the user may completely dictate the sequence of the key, and if all the weights are equal, the computer will decide by itself Usual practice is to allot about 3 to 5 different weights so as to provide overall control whilst leaving the computer to decide the details.

Interactive identification

This program is a question-and-answer procedure which performs a step-by-step elimination of taxa according to the characters of the specimen which is described (Pankhurst, 1978a, chapter 3). Any selection of characters can be specified in any sequence, which is a great practical advantage for incomplete, damaged or fragmentary specimens. Another very important feature is that it is possible to interrogate the computer to find the 'best' characters to be used next (in the identification of an unknown), or to ask for the most diagnostic characters (when trying to prove whether the specimen is a particular taxon or not). It is also possible to vary the accuracy of the identification, so that a certain degree of error can be tolerated. Commands which just retrieve information from the data are also provided It would not be difficult to invent a plethora of commands for this program in order to retrieve and manipulate information in a great variety of ways. However, experience has shown which commands are the most useful and it is these which have been retained. The program usually operates in two parts, such that part 1 checks and compresses the data and part 2 is the interactive portion. This is because part 2 often operates in microcomputers where memory is more limited.

Characters can be entered one at a time, or several together, if the user knows in advance which he wishes to feed in. However, it is often more effective to get the computer to offer assistance in choosing characters The BEST command calculates the separation number (see above) for each of the available characters, and sorts these into descending order The user can then prompt the machine for characters one at a time, until a character is found which can conveniently be used Notice that the 'goodness' of the character calculated here is a measure only of its value for distinguishing pairs of taxa, and does not include considerations such as ease of observation, availability or reliability. If it is suspected that the unknown specimen belongs to a particular taxon, then the DIAGNOSE command may be used. This operates like BEST, but uses a different figure of merit which takes into account (i) the occurrence of unusual character states in the specified taxon, and (ii) the occurrence of varying but non-exclusive character states e g if one taxon has states A and B for a character, and another has only A, then state B, if it occurs in the specimen, will make a useful distinction.

The command TAXA will show at any time a list of taxa which have not been eliminated. Initially, it is assumed that the specimen must agree exactly with the taxon with which it is identified, and the list contains only those taxa which do not differ from the specimen However, the LIMIT command can be used to change this, by stating how many character differences are to be allowed If the limit is 2, for example, then the taxa in the list will be those which have zero differences i e agree exactly, or 1 or 2 differences. If there are no taxa which agree exactly (which is a frequent and sometimes embarrassing occurrence!) then it may be possible to find one which has significantly fewer disagreements than the rest, which might be the right answer At any stage, characters can be deleted or replaced in any order, so that corrections can be made immediately without having to retrace, as in a key.

Commands are provided so that the character differences between any pair of taxa, or between a taxon and the specimen can be listed. This answers questions such as 'how do X and

Y differ?', or 'why is it not taxon X?'. It is possible to obtain a list of the available characters at any time Lastly, there is a command EXPAND which tabulates the states of a character over the current set of taxa, a command DESCRIBE which produces a description of the specimen as given so far, and a command RETURN which clears the machine memory in preparation for the next specimen

Description printing

The data from the DELTA format is converted to printed descriptions in conventional language with sentences, paragraphs, punctuation and summaries of characters of taxonomic groups Full allowance is made for character variation, character dependencies, the elimination of repetitive phrases and the expression of numerical characters. An earlier version of this program is described by Pankhurst (1978b)

This program accepts additional data which defines which characters are to be used and in what order, and how they are to be arranged in sentences and paragraphs The way in which the characters and states are defined in words in DELTA is critical to the style of the resulting text, and a little thought needs to be given to this. If a sequence of characters is assigned to a sentence, then it will be assumed that the characters can all be given the same name e.g if characters 23 and 24 (above) are put into the same sentence, then a sentence such as 'wings present, wings yellow' would be simplified to 'wings present, yellow' If the name string 'wings' had not been the same in both characters, a warning message would have appeared. The distinction between sequential and non-sequential multi-state characters is important in descriptions, since variation in sequential characters can be abbreviated e.g. 'wings small, or medium, or large' will be shortened to 'wings small to large', whereas non-sequential characters have to be spelt out in full, with 'or' preceding the alternative states. If successive states have a phrase in common, like 'leaves shorter than stem, or as long as stem', this will be shortened to 'leaves shorter than, or as long as stem'. Dependent characters may give rise to a special situation if the controlling character varies, e g 'wings absent or present, wings yellow' meaning yellow if present, can be printed by the program as 'wings absent, or present and yellow', provided that they are put into the same sentence in the right order As options, comments from the DELTA data can also be printed and the character numbers can be added to the output

It is possible to obtain character summaries when, for example, the taxa are members of a genus which is divided into subgenera. As part of the additional data, the user can divide the taxa into groups with titles, and then each group will have a character summary printed. This will list all characters which are constant within the group, and also those which occur most frequently with the adverb 'usually' added. Here 'usually' means in X% or more of cases,where X may be set to, say, 80% In this way, accurate descriptions of higher level taxa may be obtained. Quantitative characters are summarised so as to give a range the lowest and highest values in the group.

Identification by matching

This program takes a description of a specimen to identify and compares it with a matrix of taxon descriptions (Pankhurst, 1975) It is particularly useful for dealing with a large number of very similar taxa, such as a set of crop cultivars, and can be of use in naming hybrids. It may also show a degree of resistance to errors in the specimen data, according to the quality of the database.

For each taxon, a generalised similarity coefficient is calculated (see below) and a list is printed of those taxa which have the highest similarity. From this, the user should be able to decide on the identity of the unknown. To help further, the program reports on agreement with important characters, similarity with intermediate taxa (e.g. the most similar subgenus), and the scores for the taxon which the user expects the unknown to belong to.

The descriptions of specimens which are read by the program are exactly like the descriptions of taxa in DELTA. The output from this program is a list of the first N highest scoring taxa, where N can be chosen by the user, arranged in descending order of similarity between the specimen and the taxa. Also printed is a count of the number of characters common to the taxon and the specimen, on which the similarity is based This should not be too low, or else the similarity may be inaccurate. As an option, the agreement of 'special characters' is printed. The user may feel, for example, that the identification of a pink-flowered taxon with a white-flowered specimen is unacceptable If 'flower colour' is designated as a special character, then a sign is printed to show whether this character is correct or not, giving a subjective assessment to help the identification Another option allows the intermediate taxon to be identified e g. if the taxa are members of a genus, the intermediate taxon could be the subgenus Additional input data defines the subgenera, and the program assigns the specimen to the subgenus which has the highest average similarity, and puts an asterisk beside every taxon in the output list which belongs to that subgenus It is also possible to assign 'special taxa' e g. if the user suspects the correct identification is taxon 6, then 6 is made a special taxon, and its score will always be printed, whether or not it appears in the main list This helps to answer questions like 'I thought it was number 6, but what happened to it?' The 'correct' identifiction is decided on the grounds of the highest score, the best set of special characters, and the right intermediate taxon (asterisked), and is not necessarily the highest scorer. If the specimen does not belong to one of the taxa in the database as it is supposed to, then this may be detected by low similarity scores. Also, if the specimen is a hybrid, then at least one of its parents should score quite highly in the list.

If a special taxon is specified, then the program gives a report on the discrepancies between the specimen and the taxon, in the form of a list of characters and the states which disagree This is intended to assist in the task of updating the database by means of adding new data from freshly identified new material, as when making a taxonomic revision If the specimen has data for a character which is missing from the database, then this is reported in case it is desired to update the data. If the specimen disagrees in certain characters, these are reported in case this represents a new observation of different states, which ought also to be added to the data. On the other hand, the new range of variation may be only abnormal or aberrant, and would obliterate useful existing distinctions between taxa, were it added to the database For this reason, a list is also printed of the taxa whose distinctiveness would be reduced if the extra data were accepted.

The characters used in the similarity calculations are weighted in two ways. Firstly, the qualitative characters may have two states (binary) or more (multi-state). The latter can be approximated as two or more binary characters, and so are really worth more The weighting used is the logarithm of the number of states to the base of 2, so that, for example, a 4-state character will receive a weight of 2 The other kind of weighting relates to the occurrence of unusual characters in the specimen, so that up to a limit, the rarer the state, the higher the weighting. This gives greater importance to rare states when they occur Quantitative characters are converted into qualitative characters by means of the KEY STATES directive, and are then treated in the same way

Polyclaves

This program has been used to prepare punched-card keys (Pankhurst and Aitchison, 1975). Each card represents a character state and contains columns and rows of holes to correspond with each taxon which can agree with that state A set of cards which corresponds to the characters of the specimen is put together and held up to the light, so that the hole(s) which show through give the taxon or taxa which agree This type of key can be used in the field and where no computers are available. The equipment which is needed to make the cards is becoming obsolete, which is unfortunate, since this type of key is increasing in popularity

Diagnostic descriptions

A diagnostic description for a taxon is a set of characters which is possessed by that taxon and no other. There will usually be more than one such description for each taxon, differing in the choice and number of characters Some diagnostic descriptions may contain relatively few characters and ones which are easy to observe, and such descriptions are very useful for confirming (or denying) the identification of the taxon. Such character sets are often italicised in taxonomic descriptions, but sets involving more than a few characters have always been difficult to find, even by computer. There is now an algorithm (Pankhurst, 1983a) which can find all the diagnostic character sets for a taxon, given a range of set sizes, and will guarantee a minimum number of character differences, if the data is good enough.

The method by which the diagnostic sets are found is not easily described in few words, but is based on a method first described by Kautz (1968), in an entirely different context. The first step is to examine all the possible pairs of taxa which include the specified taxon, and for each pair, to list the characters in which these taxa differ. This is not difficult to do, even with the wide range of different character types which DELTA allows For a database of N taxa, a set of N − 1 sets of characters is obtained. Evidently all the diagnostic sets will be found if we examine all the possible ways in which it is possible to make a new set by taking one character from each of the sets and putting it into a new set, since this will guarantee that every pair of taxa is distinct These new sets will not necessarily all have N − 1 characters in them, since some of the characters will repeat, and only need to be counted once. The smallest of these new sets will be the diagnostic set(s) we are interested in, and there may be more than one. If the first set contains n_1 characters, the second n_2, etc., then the total number of sets will be n_1 times n_2 times etc, which will be a large number. If we wish to insist that the diagnostic sets contain at least two characters which distinguish each pair of taxa, then we have to take two characters from each of the original sets, and the number of possibilities will be larger still

A program written to simply enumerate all the possible sets like this rapidly runs out of memory or time, or both, even for small numbers of taxa and characters The actual algorithm uses a number of short cuts and simplifications in order to reduce an impossibly large problem to just a large problem. In particular, it can exploit the fact that only the smallest diagnostic sets, plus a limited range of larger sizes, are all that are really needed. A test sample with 193 taxa and 52 characters gave diagnostic sets for one taxon in ~3 min of computation in ~300K bytes of storage.

Conversion programs

There are currently two programs available which convert DELTA data to other formats One of these produces a lower triangular matrix of similarity coefficients for use with clustering algorithms, in the form required by the CLUSTAN package (Wishart, 1978) The generalised method for calculating the similarity combines the contributions of different kinds of characters between two taxa If the characters are qualitative, then the agreement is 0 if there are no states in common, or 1 if the states are the same. If the two taxa are variable but not identical, then the agreement is the ratio of the number of common states to the total number of different states Multi-state char-

acters are weighted as described above for the matching program. Quantitative characters are treated in a similar fashion, such that the proportion of overlap between the ranges of the character is taken as the agreement — e.g if taxon 1 has an integer character 5 – 10, and taxon 2 has 8 – 15, then the common overlap is 10 – 8 = 2, and the total range is 15 – 5 = 10, so the agreement is 0.2. Real variables are treated similarly, and there is allowance for special cases where in one or both of the taxa the character has a single value, i.e it is without a range, e g if one taxon is measured as 5 cm (although in fact there is very likely to be a range of variation) and another as 4 – 7 cm, then the agreement is set to 1. The documentation shows how to alter the algorithm in order to produce different coefficients to suit other clustering algorithms.

The second program converts DELTA data into the (somewhat simpler) form required by the PAUP cladistics program (Swofford, 1984) There is not currently any interface to statistical packages, but this would be quite simple to provide

Applications

In principle the package is applicable to the complete range of biological taxonomy and beyond No survey has been conducted among the users, so the summary below includes only those applications known to the author and excludes examples where earlier program versions were or are used Unless otherwise stated, it should be assumed that most or all of the programs have been applied to the data in each case.

Within the institution of origin, and in the Department of Botany, the programs are being used by the author for databases of various groups of British flowering plants, such as the genera *Rubus*, *Taraxacum*, *Euphrasia* and a vegetative key to grasses. The polyclave to grasses (Pankhurst and Allinson, 1985) and a new key for the monocotyledonous families (Rao and Pankhurst, 1986) have been published; the polyclave to world plant families of Hansen and Rahn (1969) will be republished An experimental database is being prepared for the British flora (Pankhurst, 1983b). Another botanical example is the database for the brown seaweeds (Phaeophyta, Pankhurst and Tittley, 1978). Other examples in this museum are a database for genera of protozoa, and another for the types of a family of beetles (Elateridae, Coleoptera).

Botanical examples outside the museum include a database of cultivars of the garden pea, for cultivars of oats (Canada), a key to wood anatomy (South Africa), a local Flora project (Iowa, USA) and keys to genera of the Compositae (Nigeria). Perhaps the most important of these is the Flora of Veracruz project (Gómez-Pompa *et al.*, 1984). Several very impressive publications have been produced with one of the alternative packages, such as that for Australian grasses (Watson and Dallwitz, 1980).

Non-botanical projects outside the museum include a database for biting midges (Ceratopogonidae), a palaeological application in petroleum prospecting, and a teaching application (interactive identification) in a university department of archaeology. It is likely that the actual known number of applications for this package is very much less than its actual and potential usefulness.

Implementation

All the above programs are written in FORTRAN IV and FORTRAN 77 using only standard language features They are all provided with detailed documentation, which includes details of the algorithms and a complete worked example, which uses the same trial data set for each program. The programs all operate on mainframes, mini- and microcomputers Not all microcomputers are suitable, however, and most of the programs can only be implemented easily on 16-bit micros with not less than 256K of RAM The interactive program will however run in Z80 CP/M micros for databases of moderate size, provided that the database is pre-processed on a larger machine

Acknowledgements

Mike Dallwitz, the originator of DELTA, has written his own package of programs which he calls CONFOR (Dallwitz, 1984) These are similar, but different in nature and scope from those described above The routines which are used to read DELTA in the package described here were adapted from CONFOR, with the author's permission

References

Allkin,R (1984) Handling taxonomic descriptions by computer In Allkin,R and Bisby,F A (eds), *Databases in Systematics* Systematics Association Special Vol 26, Academic Press, London and Orlando, pp 263–278

Dallwitz,M J (1984) User's guide to the DELTA system, Report No 13, CSIRO Division of Entomology, PO Box 1700, Canberra City, ACT 2601, Australia Copies from author on request

Gómez-Pompa,A , Moreno,N P , Gama,L , Sosa,V and Allkin,R (1984) Flora of Veracruz Progress and Prospects In Allkin,R and Bisby,F A (eds), *Databases in Systematics* Systematics Association Special Vol 26, Academic Press, London and Orlando, pp 165–174

Hansen,B and Rahn,K (1969) Determination of angiosperm families by means of a punched-card system *Dansk Botanisk Arkiv*, 26(1)

Kautz,W H (1968) Fault testing and diagnosis in combinatorial digital circuits *IEEE Trans Comput* , **C17**, 352–366

Pankhurst,R J (1971) Botanical keys generated by computer *Watsonia*, **8**, 357–368

Pankhurst,R J (1975) Identification by matching In Pankhurst,R J (ed), *Biological Identification with Computers* Systematics Association Special Vol 7, Academic Press, London and New York, pp 79–91

Pankhurst,R J (1978a) *Biological Identification the Principles and Practice of Identification Methods in Biology* Edward Arnold, 104 pp

Pankhurst,R J (1978b) The printing of taxonomic descriptions by computer *Taxon*, **27**, 35–38

Pankhurst,R J (1983a) An improved algorithm for finding diagnostic taxonomic descriptions *Math Biosci* , **65**, 209–218

Pankhurst,R J (1983b) The construction of a floristic database, *Taxon* **32**, 193–202

Pankhurst,R J and Aitchison,R R (1975) A computer program to construct polyclaves In Pankhurst,R J (ed), *Biological Identification with Computers* Systematics Association Special Vol 7, Academic Press, London and New York, pp 73–78

Pankhurst,R J and Allinson,J M (1985) British Grasses – a punched card key to grasses in the vegetative state *Field Studies Council Occasional Publication No 10*, and *British Museum (Natural History)*, 76 pp and 124 cards

Pankhurst,R J and Tittley,I (1978) The application of computers to the identification of phaeophyta In Irvine,D E G and Price,J H (eds), *Modern Approaches to the Taxonomy of Red and Brown Algae*, Systematics Association Special Vol 10, Academic Press, London and New York, pp 325–337

Rao,C K and Pankhurst,R.J (1986) A polyclave to the Monocotyledonous Families of the World A computer generated Identification key *British Museum (Natural History)*, 59 pp and 235 cards

Swofford,D L (1984) Phylogenetic analysis using parsimony Version 2 2 Illinois Natural History Survey, 607 East Peabody Drive, Champaign, IL 61820, USA

Watson,L. and Dallwitz,M J (1980) *Australian Grass Genera Anatomy, Morphology and Keys* Australian National University, Canberra, 209 pp

Wishart,D (1978) CLUSTAN User Manual, 3rd edition, Program Library Unit, Edinburgh University, 18, Buccleugh Place, Edinburgh EH8 9LN, UK

Received on 20 September 1985, accepted on 11 November 1985