Computer Methods for Describing Groups from Binary Phenetic Data: Modification of Numerical Taxonomy Programs to Increase Flexibility

CYNTHIA A. WALCZAK AND MICAH I. KRICHEVSKY
Microbial Systematics Section, National Institute of Dental Research, National Institutes of Health, Bethesda, Maryland 20205

The logic of modifications to numerical taxonomy programs is described. The programs modified are to calculate and store interstrain similarities, to arrange the data by cluster analysis, and to calculate inter- and intragroup statistics. The modifications included increasing the efficiency of the similarity calculations, storing intermediate matrices on disks for later use, and permitting the calculation of group statistics on any arbitrary groups of strains as well as taxonomic groups without recalculating interstrain similarities. Changing communication paths of information allowed using the programs flexibly for ecological as well as taxonomic analyses, reduced program execution costs, increased program versatility, and reduced errors.

The process by which numerical descriptions of groups of strains are derived from matrices of phenotypic data is similar for ecological and taxonomic purposes. The major difference is the method by which groups are defined and presented for calculation to the computer. The overall procedures may be divided into five main tasks: (i) preliminary summary and editing of the data before analysis; (ii) calculation and storage of interstrain similarities; (iii) arrangement of the data according to the grouping logic; (iv) calculation of inter- and intragroup statistics; and (v) calculation of feature frequencies and other group descriptors based on frequencies. This communication will be limited to a discussion of the programs used in tasks ii and iii (if cluster analysis is performed to define groups) and task iv.

In both ecological and taxonomic analyses, the final analytic programs (tasks iv and v) are told which strains belong in which groups. For taxonomy, the original order of the strains is rearranged by various cluster analysis procedures, such as single, unweighted average or complete linkage, so that highly phenotypically similar strains are near each other in the list of strains. This ordering process is automatic once the desired clustering algorithm is selected. When taxonomy is not involved, the strain list is ordered by some process external to the numerical procedure itself. That is, the phenetic data themselves drive the ordering process in cluster analysis, whereas in the other cases some factor(s) extrinsic to the phenetic characteristics leads to the desired grouping. Examples of such extrinsic factors are deoxyribonucleic acid homology groups, serotypes, geographic or anatomic site of isolation, associated disease process, and temporal variations.

Modification of computer programs designed for numerical taxonomy to utilize any arbitrarily ordered list of strains allowed the additional use of the programs for ecological analyses. Advantages of other modifications include increased program efficiency with consequent cost reduction, general improvement in versatility of programs, and decreased user error. The logic of the program modifications is described.

MATERIALS AND METHODS

Calculation and storage of strain similarities. The program we implemented and modified to suit our needs was TAXAN5, developed by P. H. A. Sneath and M. J. Sackin under the auspices of the Medical Research Council of Great Britain. The techniques used are those described by Sneath and Sokal (2). The program was written in ALGOL for a small computer. The modified program uses the SAIL language (a dialect of ALGOL) on the Digital Equipment Corporation PDP-10 time-sharing system at the National Institutes of Health. The modified programs are interactive; i.e., all control parameters are entered in response to computer prompts. Because of the nature of the PDP-10 system, these programs also are executable as batch programs without modification. A highly condensed presentation of the logic flow of the original program is summarized in Fig. 1.

The majority of data sets we analyze are large and composed of strictly binary information. Matrices of 300 to 500 strains described by 150 to 300 features are common. Since the original program was very expensive to run on such large data sets, we decided to analyze the algorithm used in the program.

Because the strain information is binary and we edit
A; otherwise, proceed to the next feature for these strains.

coefficient. Store scores, by finding A, the advantage of these conventions in our algorithm to optimize the similarity values for all pairs of strains, using information on the matches and mismatches of scores for features of the two strains; similarity coefficients are used to calculate the similarity coefficients when binary information is considered.

The calculation of the similarity coefficients was optimized by comparative timing of a variety of methods. This resulted in a substantial savings of both computer time and cost (Table 1). An outline of the resulting arithmetic comparison routine for the simple matching and Jaccard similarity coefficients follows.

Similarity coefficients are used to calculate the similarity values for all pairs of strains, using information on the matches and mismatches of scores for features of the two strains being compared.

Let: \( A \) = the number of positive matches for all features of the two strains; \( B + C \) = the number of mismatches; and \( D \) = the number of negative matches. Then the Jaccard coefficient can be expressed as \( A / (A + (B + C)) \), and the simple matching coefficient can be expressed as \( (A + D) / (A + (B + C) + D) \).

Positive scores are represented by "1"; negative scores, by "0"; and missing data, by "3." We took advantage of these conventions in our algorithm to find \( A, (B + C), \) and \( D \) when comparing two strains. For each feature of strains 1 and 2, do the following:

(i) Add the score for the feature of strain 1 to that of strain 2.

(ii) If the sum is 0, then increment \( D \); if the sum is 1, then increment \( (B + C) \); if the sum is 2, increment \( A \); otherwise, proceed to the next feature for these strains.

After all features are compared, calculate similarity of the pair of strains by simple matching or Jaccard coefficient. Store all similarity values as a triangular table. Analogous logic is applicable to the other similarity coefficients when binary information is considered.

This arithmetic comparison algorithm saves time over a logical-comparison algorithm because the number of comparisons needed to determine \( A, (B + C), \) and \( D \) are reduced. Because the data can be positive, negative, or missing, there are potentially nine cases to check for logically per comparison: 11, 10, 01, 00, 31, 30, 33, 13, and 03. But by "adding" the corresponding features of the two strains to be compared, this is reduced to four cases per comparison. This results in a substantial savings of computer time, since for \( n \) strains and \( t \) features the number of comparisons needed to create the similarity triangle is \( (t)(n^2 - n) / 2 \). For 50 strains and 50 features, this is 61,250 comparisons.

As part of the clustering algorithm in the original program, some of the actual similarity values were replaced by group linkage levels. That is, the similarity triangle stored in the computer contains the calculated similarity levels between the groups rather than the actual interstrain similarities. Although this modification process has no effect on the resulting dendrogram (and is part of the process of constructing it), it does change the values in the stored similarity triangle; the resulting values represent linkage levels instead of actual interstrain similarity values. We wished to have the original similarity triangle available for later use with the strain list in the original input order after program termination (see below). Sneath and Sackin stated in their original program documentation: "If the original similarity matrix is to be preserved it must be dumped before clustering." This only required writing a copy of the original similarity triangle on the disk before the sorting and modification process of clustering. The modified similarity triangle is scanned to establish the "taxonomic" order of the strains as well as cluster linkage levels; the original values are then recopied into the computer, rearranged in the taxonomic order, and printed. Since both copying operations are sequential in nature, they are efficient and cheap.

Modifications to the program for communication of information to other programs are the storage on disk

Table 1. Time required to calculate similarities, using the simple matching coefficient

<table>
<thead>
<tr>
<th>No. of strains</th>
<th>No. of features</th>
<th>Arithmetic comparison</th>
<th>Logical comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>10</td>
<td>0.8</td>
<td>2.4</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>3.3</td>
<td>9.1</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>13.3</td>
<td>36.4</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>3.3</td>
<td>10.6</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>12.9</td>
<td>42.0</td>
</tr>
<tr>
<td>200</td>
<td>50</td>
<td>52.4</td>
<td></td>
</tr>
</tbody>
</table>

"The two best methods were chosen for detailed evaluation. The evaluation was made by writing programs which only computed the similarities and interrogated the internal clock of the computer.

Fig. 1. TAXANS flow chart.
of (i) the strain designations in the original input order and (ii) the strain order derived from the clustering algorithm. A final modification, which greatly increases efficiency, is to allow the program to bypass calculation of the similarity triangle and use one previously stored on disk. The purpose of this change is to allow use of more than one clustering method without having to recompute the similarity triangle. The altered program logic (now called TAXON) is summarized in Fig. 2.

Arranging data according to group logic. In the usual numerical taxonomy procedures, the order in which the strains are presented to the program is not a significant factor in the results. This is because the clustering algorithm reorders the strains according to their phenetic similarities. All subsequent calculations make use of this reordered strain list.

If the original strain list and associated data are ordered by some logical grouping other than phenetic, and if the similarity triangle is stored in this original order, then all needed conditions are met to calculate group statistics by both phenetic and alternate logic systems. This dual analysis capability is made possible by altering the programs to store or use either strain order (and groups formed within those orders) for the calculations. Various strategies for sorting the strains into the desired input order are available.

Calculation of inter- and intragroup statistics. A second program, IGROUPS, was obtained from P. H. A. Sneath and M. J. Sackin. This program calculates feature frequencies, mean intra- and intergroup similarities, and their standard deviations for groups which are indicated by the program user. It shares many internal procedures with TAXAN5. Because of limitations in computer size, it had to recalculate the same similarity triangle as TAXAN5. The logic flow of IGROUPS is summarized in Fig. 3.

The first modification was to separate the calcula-

![Flowchart of TAXON](image-url)
READ IN DATA

SELECT SIMILARITY COEFFICIENT

CALCULATE SIMILARITY TRIANGLE

READ IN (ONE BY ONE) WHICH STRAINS
ARE IN WHICH GROUPS

CALCULATE INTRA- AND
INTERGROUP STATISTICS

CALCULATE FEATURE FREQUENCIES

Fig. 3. IGROUPS flow chart.

The next modification was to eliminate the recalculation of the similarity triangle and new strain order. GPSTAT reads the similarity triangle stored on the disk by the previous program, TAXON. Since TAXON also writes the original strain list and the strain list in taxonomic order on the disk, GPSTAT may be told to use either list for further calculations. Thus, either phenotypic similarity or other logic may be used to define the groups. The modifications are summarized in Fig. 4.

It should be noted that this last modification (use of two strain order lists) is not needed in the original IGROUPS program. Each strain in a group is entered in the IGROUPS program. If a strain is in more than one group, it must be entered as many times as it appears. With very large data sets, the potential for an error is great.

We chose to have the user enter only the first and last strain of each group. The advantages of this method for large data sets are fewer errors and faster entries. The disadvantage of requiring two lists is minimal, since the lists are generated automatically within the program TAXON.

RESULTS AND DISCUSSION

Various advantages have resulted from changing the program logic flow. The costs of analyzing large data matrices have been markedly reduced (3- to 10-fold for TAXON).

A consequence of storing the similarity matrix on a disk is the ability to restart the program TAXON without having to recalculate the similarity values. This is important in those rare, but potentially expensive, cases of computer failure during the latter stages of program execution. A more important use of an existing similarity triangle is to apply multiple clustering algorithms for comparison purposes. A further use is the availability of the similarity values to other programs.

By dissociating the feature frequency calculations from the others, we eliminated the restriction that the same data matrix be used for all calculations. That is, the features included in feature frequency calculations can be more inclusive than the set used for calculation of phenotypic similarity. The features can even include nonphenetic information such as source of isolation, homology group, and other characteristics quite inappropriate to phenotypic analysis.

Suitable modifications to numerical taxonomy programs permit the calculation of the intra- and intergroup statistics of any arbitrary groups of strains. This flexibility becomes important in comparing group statistics of phenotypic (i.e., taxonomic) groups with those of ecologically derived groups. The ecological principle that diversity decreases as ecological stress increases (1) can be tested by comparing the intragroup similarities from increasingly stressed groups (such as a progressing abscess, silage, or an enrichment culture).

Collectively, these modified programs give the microbiologist the capacity to analyze strain data in ways well beyond the original intent of such programs. The programs and directions for their use in a PDP-10 are available from us.
REPRINT REQUESTS

Address reprint requests to: Dr. Micah I. Krichevsky, Microbial Systematics Section, National Institute of Dental Research, National Institutes of Health, Bethesda, MD 20205.

LITERATURE CITED


