# A review of numerical methods in bacterial identification

THE LATE W. R. WILLCOX, S. P. LAPAGE, AND B. HOLMES

*National Collection of Type Cultures, Central Public Health Laboratory, Colindale Avenue, London NW9 5HT, UK*

WILLCOX, W. R., The late, LAPAGE, S. P. and HOLMES, B. 1980. A review of numerical methods in bacterial identification. Antonie van Leeuwenhoek **46**: 233–299.

Part A of this review describes the particular computer-assisted identification service operated by the NCTC. In Part B, the use of probability matrices is examined, discussing various methods of calculating likelihoods and the problems that arise when calculating these from probability matrices. Part C describes the alternative numerical methods of constructing identification keys and the supplementary methods of selecting "best sets" of characters to aid identification. Finally, in Part D, the prospects and limitations of numerical methods in bacterial identification are assessed, first with regard to methodology used and then in terms of performance and practical limitations.

CONTENTS

# PART A. A COMPUTER-ASSISTED SERVICE FOR THE IDENTIFICATION OF BACTERIA

## 1. Introduction

The computer-assisted identification service described here is for the identification of gram-negative, mainly rod-shaped bacteria which grow aerobically on nutrient agar and which public health and hospital laboratories have found difficult to identify. The service was experimental from 1966 to 1971 (Lapage et al., 1970; 1973) and has been operated as a service since January 1972.

The methods used in the computer program are described by Willcox et al. (1973) and reviewed in Part B. The identification program was tested by applying it to 1079 reference strains, aberrant and typical strains, whose identity had already been well established (Lapage et al., 1973). Of the 827 strains of fermentative bacteria, 91% were correctly identified, a single strain was incorrectly identified (*Salmonella abortusovis* identified as *S. choleraesuis*), and the remainder were not identified by the program. The results for nonfermentative bacteria were less satisfactory, of 201 strains 82% were correctly identified, the remainder were not identified. Since that trial a separate identification matrix has been developed for nonfermentative bacteria with tests more suitable for these bacteria. The original matrix for fermentative bacteria (Bascomb et al., 1973) has remained in use with minor changes. The taxa and tests currently in the two matrices are given in Tables 1 and 2.

Table 1. Taxa in the identification matrices for fermentative and non-fermentative bacteria

Fermentative bacteria

*Acinetobacter calcoaceticus*[1]
*Actinobacillus equuli*
*Actinobacillus lignieresi*
*Aeromonas formicans*
*Aeromonas hydrophila*
*Aeromonas salmonicida*
*Chromobacterium violaceum*
*Citrobacter freundii*
*Citrobacter koseri*
CDC group EF-4
*Edwardsiella tarda*
*Enterobacter aerogenes*
*Enterobacter cloacae*
*Erwinia herbicola*
*Escherichia adecarboxylata*
*Escherichia coli*
*Hafnia alvei*
*Klebsiella aerogenes* and *K. oxytoca*
*Klebsiella ozaenae*
*Plesiomonas shigelloides*
*Proteus mirabilis*
*Proteus morganii*
*Proteus rettgeri*
*Proteus vulgaris*
*Providencia alcalifaciens*
*Providencia stuartii*
*Salmonella choleraesuis*
*Salmonella ferlac*
*Salmonella gallinarum*
*Salmonella paratyphi* A
*Salmonella pullorum*
*Salmonella* subgenus I
*Salmonella* subgenus II
*Salmonella* subgenus III = *Arizona*
*Salmonella* subgenus IV
*Salmonella typhi*
*Serratia liquefaciens*
*Serratia marcescens*

Table 1. *Continued*

| | |
|---|---|
| *Klebsiella pneumoniae* | *Serratia marinorubra* |
| *Klebsiella rhinoscleromatis* | *Serratia plymuthica* |
| *Kluyvera* spp. | *Shigella sonnei* |
| *Neisseria pharyngis* | *Shigella* spp. other than *sonnei* |
| *Pasteurella haemolytica* A | *Vibrio* spp. other than *parahaemolyticus* |
| *Pasteurella haemolytica* T | *Vibrio parahaemolyticus* |
| *Pasteurella multocida* | *Yersinia enterocolitica* |
| *Pasteurella multocida* (atypical) | *Yersinia pestis* |
| *Pasteurella pneumotropica* | *Yersinia pseudotuberculosis* |
| *Pasteurella ureae* | *Yersinia ruckeri* |
| Non-fermentative bacteria[2] | |
| *Achromobacter* group | *Moraxella phenylpyruvica* |
| *Achromobacter* sp. (biotype 1) | *Moraxella* proteolytic group |
| *Achromobacter* sp. (biotype 2) | *Moraxella saccharolytica* |
| *Achromobacter xylosoxidans* | *Moraxella urethralis* |
| *Acinetobacter calcoaceticus* | *Neisseria meningitidis* |
| *Acinetobacter lwoffii* | *Neisseria* spp. other than *meningitidis* |
| *Agrobacterium tumefaciens* | *Pseudomonas acidovorans* |
| *Agrobacterium rhizogenes* | *Pseudomonas aeruginosa* |
| *Agrobacterium rubi* | *Pseudomonas alcaligenes* |
| *Agrobacterium* yellow group | *Pseudomonas cepacia* |
| *Alcaligenes faecalis* | *Pseudomonas diminuta* |
| *Bordetella bronchiseptica* | *Pseudomonas fluorescens* |
| *Bordetella parapertussis* | *Pseudomonas fragi* |
| *Branhamella* spp. | *Pseudomonas lemoignei* |
| *Brucella* spp. | *Pseudomonas mallei* |
| CDC group IIB | *Pseudomonas maltophilia* |
| CDC group IIF | *Pseudomonas mendocina* |
| CDC group IIJ | *Pseudomonas paucimobilis* |
| CDC group IIK, type 2 | *Pseudomonas pickettii* |
| CDC group IVE | *Pseudomonas pseudoalcaligenes* |
| CDC group VE, type 1 | *Pseudomonas pseudomallei* |
| CDC group VE, type 2 | *Pseudomonas putida* |
| *Chromobacterium lividum* | *Pseudomonas putrefaciens* |
| *Eikenella corrodens* | *Pseudomonas stutzeri* |
| *Flavobacterium breve* | *Pseudomonas taetrolens* |
| *Flavobacterium meningosepticum* | *Pseudomonas testosteroni* |
| *Flavobacterium odoratum* | *Pseudomonas vesicularis* |
| *Kingella* spp. | *Rhizobium meliloti* |
| *Moraxella anatipestifer* | *Xanthomonas hyacinthi* |
| *Moraxella* non-proteolytic group | *Xanthomonas* spp. other than *hyacinthi* |

[1] *Acinetobacter calcoaceticus* is included in both matrices because although it is non-fermentative it resembles some fermentative taxa in certain reactions.

[2] Some of the organisms in the matrix for non-fermentative organisms may be strictly fermenters (i.e. they grow aerobically and anaerobically) but they behave as non-fermenters in the tests we use.

Table 2. Tests in the identification matrices for fermentative and non-fermentative bacteria

Tests in both matrices

Motility at 37°C
Motility at RT[1]
Growth at 37°C
Growth at RT
Pigment production
Growth on MacConkey's agar
Catalase production
Oxidase production
Hugh and Leifson O–F test
Nitrate reduction
Indole production
Methyl red test at 37°C
Methyl red test at RT
Voges-Proskauer test at 37°C
Voges-Proskauer test at RT[1]
Growth on Simmons' citrate
Alkali production on Christensen's citrate
Urease production
KCN tolerance
$H_2S$ production (lead acetate paper method)
$H_2S$ production (triple sugar iron agar method)
Gluconate oxidation
Malonate utilization
β-Galactosidase production (ONPG)
Phenylalanine deamination
Arginine dihydrolase production
Lysine decarboxylase production
Ornithine decarboxylase production
Selenite reduction 0.4 g/100 ml
Gelatinase production (stab method)
Gelatinase production (plate method)
Casein digestion
Deoxyribonuclease production
Acid from glucose PWS[2]
Gas from glucose PWS
Production of acid from
adonitol PWS
arabinose PWS
cellobiose PWS
dulcitol PWS
glycerol PWS
inositol PWS
lactose PWS
maltose PWS
mannitol PWS
raffinose PWS
rhamnose PWS
salicin PWS
sorbitol PWS
starch PWS
sucrose PWS
trehalose PWS
xylose PWS

Tests only in the matrix for non-fermentative bacteria

Arginine desimidase production (Thornley's method)
Tween 20 hydrolysis
Tween 80 hydrolysis
Tyrosine hydrolysis
Pigment production on tyrosine
Nitrite reduction
Growth on β-hydroxybutyrate
Poly-β-hydroxybutyrate inclusion granules
Aesculin hydrolysis
Growth on cetrimide
Fluorescence on King's B medium
Growth at 5°C
Growth at 42°C
3-Ketolactose production
Lecithinase production
Starch hydrolysis
Acid from glucose 10 g/100 ml
Acid from lactose 10 g/100 ml
Production of acid from:
glucose ASS[3]
adonitol ASS
arabinose ASS
cellobiose ASS
dulcitol ASS
ethanol ASS
Production of acid from:
fructose ASS
glycerol ASS
inositol ASS
lactose ASS
maltose ASS
mannitol ASS
raffinose ASS
rhamnose ASS
salicin ASS
sorbitol ASS
sucrose ASS
trehalose ASS
xylose ASS

[1] RT, room temperature (18–22°C) or incubator at 22 or 30°C.
[2] PWS, peptone water sugar.
[3] ASS, ammonium salt sugar.

The service shows that numerical methods of identification and test selection can be used in a flexible way in which the bacteriologist has final control, overriding the automatic scheme if necessary; the examples used to illustrate the service show different ways in which numerical methods can contribute to the identification of bacteria.

## II. Operation of the Service

The sending laboratory will have carried out a number of tests on the strain and these test results are sent together with a culture of the strain. The number of tests done by the sending laboratory varies, typically it is about 20. The test results are entered on an application form which lists the 90 tests in the full computer identification matrices. The source of the strain, clinical details if available and the sending laboratory's tentative identification of the strain are also entered. Space is provided on the form for the results of tests not used in the matrices, e.g. serological data. This additional information helps the bacteriologist to assess the results of the computer identification of the strain.

The operation of the service is most easily described by examples. Figs 1 to 7

```
FOR:                                                    OUR REF:9981/79 RUN W1
DATE: 19/07/79                                                (M518 LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION BASED ON YOUR RESULTS, 37 TESTS DONE:

 IDENTIFIED AS PASTEURELLA MULTOCIDA

YOUR RESULTS USED IN CALCULATION:

 MOTILITY 37     -   1     MACCONKEY       -  25     SIMMONS CITR    -   1
 MOTILITY RT     -   1     CATALASE        +  99     UREASE          -   1
 GROWTH 37       +  99     OXIDASE         +  50     PPA             -   1
 GROWTH RT       +  75     H&L FERM        +  50     GLUCOSE PWS     +  99
 PIGMENT         -   1     NITRATE         +  99     GAS GLUCOSE     -   1

 ADONITOL PWS    -   1     LACTOSE PWS     +   5     SORBITOL PWS    +  90
 ARABINOSE PWS   -  15     MALTOSE PWS     -   1     SUCROSE PWS     +  99
 CELLOBIOSE PWS  -   1     MANNITOL PWS    +  90     TREHALOSE PWS   -  30
 DULCITOL PWS    -  15     RAFFINOSE PWS   -   5     XYLOSE PWS      -  50
 GLYCEROL PWS    -  15     RHAMNOSE PWS    -   1     STARCH PWS      -   1
 INOSITOL PWS    -   1     SALICIN PWS     -   1

 MR 37           -   1     VP 37           -   1     INDOLE          +  99
 MR RT           -   1     VP RT           -   1

DETAILS OF CALCULATION:

        GROUP                              SCORE
 PASTEURELLA MULTOCIDA                    .999847
 PASTEURELLA MULTOCIDA(ATYPICAL)          .000150
```

Fig. 1. A report printed by the identification program. This strain is identified on the sending laboratory's test results ("W" run).

```
FOR:                                                    OUR REF:9982/79 RUN W1
DATE: 18/07/79                                                (M518 LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION BASED ON YOUR RESULTS, 28 TESTS DONE:

 NOT IDENTIFIED, FURTHER TESTS SELECTED

YOUR RESULTS USED IN CALCULATION:

 MOTILITY 37     -  30     H&L FERM        +  99     MALONATE        -   1
 MOTILITY RT     -  55     NITRATE         -  99     ONPG            -  90
 GROWTH 37       +  99     SIMMONS CITR    -   1     PPA             -   1
 GROWTH RT       +  99     UREASE          -   1     ARGININE        -  40
 PIGMENT         -   1     GELATIN PLATE   -   1     LYSINE          +  90
 MACCONKEY       +  99     KCN             -   1     ORNITHINE       +  75
 CATALASE        +  99     H2S TSI         -   1     GLUCOSE PWS     +  99
 OXIDASE         -   1     GLUCONATE       -   1     GAS GLUCOSE     +  90

 ARABINOSE PWS   +  99

 MR 37           +  99     MR RT           +  99     INDOLE          +  95

FURTHER TESTS SELECTED:

   FIRST SET                    SECOND SET
 GLYCEROL PWS     1  95       H2S PAPER        1   5
 MALTOSE PWS      1  99

   SET VALUE =    2/   2        SET VALUE =    1/   2

DETAILS OF CALCULATION:

       GROUP                          SCORE
 ESCHERICHIA COLI                    .998530
 SALMONELLA PULLORUM                 .000894
```

Fig. 2. A report printed by the identification program. This strain cannot be identified on the sender's test results ("W" run) but further tests are selected.

show computer printouts of various strains (the name of the sending laboratory, their reference number for the culture and the patient's name have been deleted from each printout). Before any results can be processed the bacteriologist must decide whether the strain is fermentative or non-fermentative in order to select the appropriate identification matrix. This is usually possible on the results supplied by the sending laboratory but, if not, a Hugh and Leifson Oxidatative–Fermentative (O-F) test is carried out in our laboratory.

The sending laboratory's test results are processed first, printing a report such as in Fig. 1. Each computer analysis of a set of results is called a "run" and the run number for this printout is W1; "W" indicates a run using only the sender's results. In this case the strain is identified by the program as *Pasteurella multocida*. The printout is assessed by a bacteriologist and, unless there was any reason to query the computer identification, a copy of the printout would be returned to the sender. For strains such as this example, which identify on the sender's results, the sending laboratory has carried out a good number of tests and is probably fairly certain of the identity of the strain but requires confirmation because the strain

shows an aberrant reaction. This organism for instance produces acid from lactose which is an unusual reaction for *P. multocida*. The number beside each test result on the printout is the estimate of the probability of a positive result for this test for the most likely taxon. Here the printout shows that 5% of strains of *P. multocida* are expected to be positive in lactose. The result is certainly uncommon though not completely unexpected. This shows how a numerical identification method can use a definition of a taxon which allows for occasional aberrant reactions in particular tests.

Only a minority (about 5%) of the strains received can be identified on the results supplied by the sending laboratory. Fig. 2 shows a case where the strain cannot be identified on these results, but the program has selected some tests as the most useful ones to continue the identification (see Part C). The selected tests are printed in sets; following each test is printed first a theoretical value of the usefulness of the test alone in the set, then the probability figure for that test for that taxon which had achieved the highest identification "score". The total

```
FOR:                                              OUR REF:9982/79 RUN R1
DATE: 19/07/79                                          (MSIB LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION
 BASED ON OUR RESULTS COMBINED WITH YOURS, 30 TESTS DONE:

 IDENTIFIED AS ESCHERICHIA COLI

RESULTS USED IN CALCULATION:

 MOTILITY 37     -  30     H&L FERM        +  99     MALONATE        -   1
 MOTILITY RT     -  55     NITRATE         +  99#    ONPG           (-) 90*
 GROWTH 37       +  99     SIMMONS CITR    -   1     PPA             -   1
 GROWTH RT       +  99     UREASE          -   1     ARGININE        -  40
 PIGMENT         -   1*    GELATIN PLATE   -   1     LYSINE          +  90
 MACCONKEY       +  99     KCN             -   1     ORNITHINE       +  75
 CATALASE        +  99*    H2S TSI         -   1     GLUCOSE PWS     +  99
 OXIDASE         -   1*    GLUCONATE       -   1     GAS GLUCOSE     +  90

 ARABINOSE PWS   +  99     GLYCEROL PWS    +  95<    MALTOSE PWS     +  99<

 MR 37           +  99     MR RT           +  99     INDOLE          +  95

  *=OUR RESULT AGREES WITH YOURS     <=OUR RESULT ONLY
  #=OUR RESULT(CL) DIFFERS FROM YOURS(SL), OURS USED
                    CL  SL
      NITRATE        +   - @

  @=YOUR RESULT UNEXPECTED FOR THIS ORGANISM

DETAILS OF CALCULATION:

        GROUP                          SCORE
 ESCHERICHIA COLI                     .999764
 SHIGELLA SPP.-NOT SONNEI             .000166
```

Fig. 3. A report printed by the identification program. This is the same strain as in Fig. 2. Further tests have been carried out in our laboratory and the strain is identified on our results combined with the sender's results ("R" run).

```
FOR:                                                    OUR REF:9983/79 RUN W1
DATE: 18/07/79                                                (M624 LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION BASED ON YOUR RESULTS, 30 TESTS DONE:

 NOT IDENTIFIED, FURTHER TESTS SELECTED

YOUR RESULTS USED IN CALCULATION:

 MOTILITY 37     +  90     NITRATE         +  99     ONPG            -   1
 MOTILITY RT     +  90     SIMMONS CITR    -  35     PPA             -   1
 GROWTH 37       +  99     UREASE          -  10     ARGININE        -  25
 GROWTH RT       +  99     GELATIN PLATE   -   1     LYSINE          -   1
 YELLOW PIGMENT  +   1     KCN             -   1     ORNITHINE       -   1
 MACCONKEY       +  99     H2S PAPER       +  45     GLUCOSE PWS     -   1
 CATALASE        +  99     GLUCONATE       -   1     GAS GLUCOSE     -   1
 OXIDASE         +  99     MALONATE        -   1     DNASE           -   1
 HUGH & LEIFSON  -  90

 INDOLE          -   1

 KINGS B         -   1     GROWTH 42       -  99     STARCH HYD      -   1
 GROWTH 5        -  10

FURTHER TESTS SELECTED:

   FIRST SET                    SECOND SET
 MALTOSE ASS     25   1       ARABINOSE ASS  20   1
 XYLOSE ASS      25   1       CELLOBIOSE ASS 20   1
 GLYCEROL ASS    10  45       TYROSINE PIG    7  55
 GLUCOSE ASS      6   1
 3-KETOLACTOSE    6   1
 ETHANOL ASS      4  90
 TWEEN 80         3  10
 FRUCTOSE ASS     2  90
 CHRISTEN.CITR    1  25
 ADONITOL ASS     1   1
 LACTOSE ASS      1   1
 TWEEN 20         1  55

   SET VALUE =  85/  90         SET VALUE =  47/  90

DETAILS OF CALCULATION:

        GROUP                                  SCORE
 PSEUDOMONAS PSEUDOALCALIGENES              .388737
 PSEUDOMONAS LEMOIGNEI                      .251968
 AGROBACTERIUM YELLOW GROUP                 .123336
 ACHROMOBACTER XYLOSOXIDANS                 .072439
 AGROBACTERIUM TUMEFACIENS                  .072369
 ACHROMOBACTER SP.(BIOTYPE 2)               .025078
 ACHROMOBACTER GROUP                        .022779
 PSEUDOMONAS ACIDOVORANS                    .014848
 ALCALIGENES SPP.                           .007433
 PSEUDOMONAS VESICULARIS                    .007276
```

Fig. 4. A report printed by the identification program. This strain cannot be identified on the sender's test results ("W" run). Further tests are selected but the value of each of the two sets of tests selected is less than the "key" value.

theoretical value of each set is printed followed by, for comparison the "key value". The key value is the value needed so that, according to the test selection model, a definite identification is probable if these test are carried out. If more than one set has the same value the bacteriologist will choose the most convenient set from the practical point of view. The identification "scores" on the printout show that the strain is very likely to be an *Escherichia coli* though the identification threshold level of 0.999 has not been reached. The probability figures printed beside the test results show that the negative results recorded by the sending laboratory in the nitrate and ONPG tests are not the expected results for *E. coli*; the figures show that 99% and 90% of *E. coli* strains should be positive in these tests respectively. The identification was continued by carrying out the first set of tests shown on the printout and repeating the nitrate and ONPG tests, as the printout suggested that these were suspect. These tests were carried out in our laboratory (also the pigment, catalase and oxidase tests as a standard procedure) and the results obtained were combined with the sender's results on the next computer run for the strain. Fig. 3 shows the printout produced, the run number is R1; "R" indicates a run in which our results are combined with the sender's. The tests which we carried out are indicated by symbols beside the result and, as the printout shows, there was disagreement with the sending laboratory in the nitrate test. Where such conflicting results occur, the NCTC result is used in the calculation. The strain now identifies and again after review by a bacteriologist copies of the W1 and R1 printouts would be returned to the sending laboratory. In this example identification was almost complete on the sender's results and no doubt it was the erroneous result in the nitrate test which caused difficulty. The numerical method was able to guide the identification in the right direction despite an erroneous result in what is a "key character" for conventional identification.

In the next example (Fig. 4) the sender's results were again analysed by the program but although 30 tests had been carried out the printout was not very promising. Ten taxa are listed as likely and although two sets of additional tests have been selected the value of the first of these sets is only 85, less than the key value of 90. Also the second and third highest scoring taxa were thought to be very unlikely considering the source of the isolate. Instead of continuing with the tests suggested by the program, a standard set of tests ("a basic set") was used. Two such basic sets of tests have been derived, one for fermentative organisms, and one for nonfermenters. Each set has been chosen to give good discrimination over all the taxa. Fig. 5 shows the analysis of the results of the basic set of 24 tests for nonfermentative organisms. The code letter for the run is "B", indicating a run in which the sender's results are printed but not used in the identification calculation. The strain is now identified as *Pseudomonas putrefaciens* on the results of fewer tests than were originally carried out by the sending laboratory.

Strains are occasionally received with very few test results or none at all. If it is considered that there are too few results, or the results are unsuitable for numerical analysis (because of differences in test methods for example), the

```
FOR:                                                    OUR REF:9983/79 RUN B1
DATE: 19/07/79                                                (M624 LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION BASED ON OUR RESULTS, 24 TESTS DONE:

 IDENTIFIED AS PSEUDOMONAS PUTREFACIENS

RESULTS USED IN CALCULATION:

 MOTILITY 37     +  70*    ORANGE PIGMENT +  99#    SIMMONS CITR    -  20*
 MOTILITY RT     +  99*    CATALASE       +  99*    GELATIN PLATE   +  99#
 GROWTH 37       +  80*    OXIDASE        +  99*    H2S PAPER       +  99*
 GROWTH RT       +  99*    H&L ALKALINE   +  90#

 GLUCOSE ASS     -  45<    MALTOSE ASS    +  45<    XYLOSE ASS      -   1<
 ARABINOSE ASS   +  35<    TREHALOSE ASS  -   1<    ETHANOL ASS     -   1<
 CELLOBIOSE ASS  -  20<

 INDOLE          -   1*

 TWEEN 20        +  99<    TYROSINE PIG   +  55<    PHBA INC        -   1<
 TYROSINE HYD    +  45<    PHBA GROWTH    +  99<

  *=OUR RESULT AGREES WITH YOURS     <=OUR RESULT ONLY
  #=OUR RESULT(CL) DIFFERS FROM YOURS(SL), OURS USED
                    CL  SL
      YELLOW PIGMENT     + @
      ORANGE PIGMENT +
      HUGH & LEIFSON     -
      H&L ALKALINE   +
      GELATIN PLATE  +   - @

YOUR RESULTS NOT USED IN CALCULATION:

 MACCONKEY       +  99     MALONATE       -   1     ORNITHINE       -  95
 NITRATE         +  99     ONPG           -   1     GLUCOSE PWS     -   1
 UREASE          -  10     PPA            -   1     GAS GLUCOSE     -   1
 KCN             -   1     ARGININE       -   1     DNASE           -  99@
 GLUCONATE       -   1     LYSINE         -   1

 KINGS B         -   1     GROWTH 42      -  70     STARCH HYD      -   1
 GROWTH 5        -  25

  @=YOUR RESULT UNEXPECTED FOR THIS ORGANISM

DETAILS OF CALCULATION:

        GROUP                          SCORE
 PSEUDOMONAS PUTREFACIENS             .999982
 PSEUDOMONAS MALTOPHILIA              .000018
```

Fig. 5. A report printed by the identification program. This is the same strain as in Fig. 4. The strain is identified on the results of our "basic set" of tests for non-fermentative organisms, the sender's results are printed but not used in the calculation ("B" run).

sender's results would not be used, but rather a Hugh and Leifson 0-F test is done and then the appropriate basic set of tests. Fig. 6 shows an example of a strain which was identified on the results of the basic set for fermentative organisms. The sender's results were not used at all in the identification and "T" indicates a computer run in which the sender's results are neither printed nor used in the calculation.

```
FOR:                                              OUR REF:9984/79 RUN T1
DATE: 18/07/79                                          (M518 LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION BASED ON OUR RESULTS, 26 TESTS DONE:

 IDENTIFIED AS KLEBSIELLA AEROGENES & K.OXYTOCA

OUR RESULTS USED IN CALCULATION:

 MOTILITY RT     -   1     GELATIN PLATE  -  10     ARGININE       -   1
 GROWTH RT       +  99     KCN            +  99     LYSINE         +  95
 PIGMENT         -   1     GLUCONATE      +  95     ORNITHINE      -   1
 CATALASE        +  99     MALONATE       +  80     GLUCOSE PWS    +  99
 OXIDASE         -   1     ONPG           +  99     GAS GLUCOSE    +  90
 UREASE          +  95     PPA            -   1

 ARABINOSE PWS   +  99     INOSITOL PWS   +  95     SORBITOL PWS   +  95
 DULCITOL PWS    -  40     LACTOSE PWS    +  90     TREHALOSE PWS  +  99

 MR RT           -  30     VP RT          +  70     INDOLE         +  25

DETAILS OF CALCULATION:

        GROUP                          SCORE
 KLEBSIELLA AEROGENES & K.OXYTOCA     .999998
 ENTEROBACTER AEROGENES               .16E-05
```

Fig. 6. A report printed by the identification program. This organism is identified on the results of our "basic set" of tests for fermentative organisms, the senders results were not used in the identification ("T" run).

In the last example (Fig. 7), the strain is identified as *Proteus rettgeri* but an unusual test result is reported. The strain gives a negative result in the test growth on MacConkey agar when at least 99% of strains of *P. rettgeri* should give a positive result. If a strain is identified the program checks for such unusual results which are positive results where the probability figure for the taxon is 1% and negative results where the figure is 99%. In assessing this printout the bacteriologist must decide whether or not the computer identification is acceptable in view of the unusual test result. The example also shows that several cycles of testing and computer analysis may be necessary, the final run shown is R2 so the strain has already had W1 and R1 runs. The example further shows how the numerical method can identify a strain even if it gives very unusual results provided enough evidence is available from the other test results. In this case 51 tests were carried out to identify this aberrant strain.

Finally, a computer printout may indicate that the strain cannot be identified and none of the remaining tests are of any value in continuing the identification or all of the available tests have been carried out. The bacteriologist may be able to identify such a strain, usually to one of the likely taxa suggested by the printout and possibly by using tests not available to the computer program. Some strains (about 10%) however remain which we are unable to identify by any of the means available to us.

```
FOR:                                                OUR REF:9987/79 RUN R2
DATE: 19/07/79                                            (M518 LAB.  -0)

YOUR REF:

COMPUTER IDENTIFICATION
 BASED ON OUR RESULTS COMBINED WITH YOURS, 51 TESTS DONE:

 IDENTIFIED AS PROTEUS RETTGERI
  WITH UNUSUAL RESULTS:
   MACCONKEY      -

RESULTS USED IN CALCULATION:

 MOTILITY 37     -  95*    SIMMONS CITR   -  95*    ONPG           -   5*
 MOTILITY RT     +  99<    CHRISTEN.CITR  +  99<    PPA            +  99*
 GROWTH 37       +  99<    UREASE         +  99*    ARGININE       -   5*
 GROWTH RT       +  99<    GELATIN STAB   -   1<    LYSINE         -   1*
 PIGMENT         -   1<    GELATIN PLATE  -   1*    ORNITHINE      -   1*
 MACCONKEY       -  99<    KCN            +  99<    GLUCOSE PWS    +  99*
 CATALASE        +  99<    H2S PAPER     (+) 50#    GAS GLUCOSE    -  15<
 OXIDASE         -   1*    H2S TSI        -   1<    SELENITE 0.4  (+) 99<
 H&L FERM        +  99<    GLUCONATE      -   1<    DNASE          -   0<
 NITRATE         +  99<    MALONATE       -   1<

 ADONITOL PWS    +  99<    LACTOSE PWS    -   5<    SORBITOL PWS   -   1*
 ARABINOSE PWS   -   5*    MALTOSE PWS    -   1<    SUCROSE PWS    -  50*
 CELLOBIOSE PWS  -   1<    MANNITOL PWS   +  99*    TREHALOSE PWS  -   1<
 DULCITOL PWS    -   1<    RAFFINOSE PWS  -   1<    XYLOSE PWS     -  15<
 GLYCEROL PWS   (-) 95<    RHAMNOSE PWS   +  50*    STARCH PWS     -   1<
 INOSITOL PWS    +  99#    SALICIN PWS    +  50<

 MR 37           +  99<    VP 37          -   1*    INDOLE         +  99*
 MR RT           +  99<    VP RT          -   1<

  *=OUR RESULT AGREES WITH YOURS    <=OUR RESULT ONLY
  #=OUR RESULT(CL) DIFFERS FROM YOURS(SL), OURS USED
                    CL  SL
      H2S PAPER     (+)  -
      INOSITOL PWS   +   - @

  @=YOUR RESULT UNEXPECTED FOR THIS ORGANISM

DETAILS OF CALCULATION:

        GROUP                          SCORE
 PROTEUS RETTGERI                    1.000000
 PROVIDENCIA ALCALIFACIENS            .53E-07
```

Fig. 7. A report printed by the identification program. This strain is identified on our test results combined with the sender's results ("R" run) but an unusual result is reported.

# PART B. NUMERICAL IDENTIFICATION USING PROBABILITY MATRICES

## III. Introduction

Numerical identification methods can be classified according to the different types of data and identification matrices they require (see Part D). The *data matrix* is the information necessary to construct the method and the *identification matrix* is the information actually used by the method. The majority of numerical methods in microbiology use a probability matrix. The entries in the matrix for each taxon are estimates of the probability of observing each character state for that taxon. The advantages of such a matrix are that it is readily compiled from a variety of sources of information and is easily adjusted to allow for new findings. Section IV describes a number of methods using probability matrices, compared with other models and in Section V some detailed problems which arise in applying such methods will be considered.

## IV. Calculation and Use of Likelihoods

*a. Calculation of likelihoods*

All the methods using probability matrices (Dybowski and Franklin, 1968; Lapage et al., 1970; 1973; Friedman et al., 1973; Friedman and Maclowry, 1973; Robertson and MacLowry, 1974; Gyllenberg and Niemelä, 1975*a*, *b*; API, 1977) start by calculating the likelihoods of the taxa on the character states observed for the unknown organism. The likelihood of a taxon on a set of character states is defined as the probability of the states for the taxon. To calculate the probability of a set of character states from a probability matrix it is assumed that the individual states which make up the set are independent in each taxon. The probability of a number of independent states is the product of their individual probabilities so:

$$L_J = P(u/J) = \Pi_i P(x_i/J) \tag{1}$$

where $L_J$ is the likelihood of taxon $J$; $P(u/J)$ is the probability of $u$, the set of character states of the unknown, for taxon $J$; and $P(x_i/J)$ is the probability of $x_i$, the state of the unknown in character $i$, for taxon $J$.

The distinction between likelihood and probability is not easy. Considering equation (1), when values are calculated for different sets of character states for a particular taxon, the values are regarded as the probabilities of the character states for the taxon. When the values for the different taxa for a particular set of character states are compared, they are regarded as the likelihoods of the taxa (Kendall and Stuart, 1963, p. 202). Some authors refer to the likelihoods of the character states for the taxa but the present terminology is preferable (Rao, 1952, p. 150). The phrase "likelihood of taxon $J$" is a convenient shortening of the more

exact "likelihood that the unknown is a member of $J$" and "probability of character states $u$ for taxon $J$" is a shortening of "probability that a member of $J$ will show states $u$".

The probability matrix gives the probabilities of all individual character states for all taxa so the $P(x_i/J)$ are obtained from the matrix and it is easy to calculate using (1) the likelihoods of all the taxa in the matrix for any set of character states. For two-state characters it is usual to express only the probabilities of the positive states in the matrix and the probability of a negative state is calculated as one minus the probability of the positive state. The advantages of this are discussed in Section V.*d* below. The simplicity of the calculation is shown in Table 3(*b*). The example also shows that the method is not affected if there are no results for some of the characters in the matrix, these characters are simply ignored. Although the calculation is very straightforward, in applying the method in practice a number of problems arise and these are considered in Section V.

Equation (1) assumes that the character states are independent in each taxon, i.e. the probability of a particular character state for a particular taxon, $P(x_i/J)$, is always the same, irrespective of the results of the other characters. This will not be true if the characters are correlated within a taxon and this does not seem to have been investigated in any detail (Sneath, 1974). The assumption of independent character states is a serious theoretical objection to the method. It is character correlation *within* taxa which matters, not the character correlation between taxa (itself a consequence of the existence of distinct taxa). Sneath and Sokal (1973, at pp. 103–106) consider other types of character correlation some of which follow from the definition of the characters or from known biochemical relationships between tests. To avoid correlated characters as far as possible, tests for as many different enzymes and biochemical pathways as possible should be chosen for the identification matrix (Lapage et al., 1973) and tests known to be closely related should be excluded. It is not difficult to allow for simple forms of correlation in the calculation (see Section V.*c*). Sneath (1974) considers that the effect of character correlations in bacteria is probably small enough to ignore in identification models.

There are some methods of calculating likelihoods which do not rely on the above assumption. Identification methods which do take account of character correlation must start from a full data matrix, containing for each taxon the character states of a number of sample individuals of the taxon (see Part D). The most direct way of estimating likelihoods is from the frequency of occurrence in the sample data of character state patterns exactly the same as the unknown. This is seldom practicable because of the large number of possible patterns (e.g. about $10^6$ for 20 two-state characters). Even if a particular pattern has been observed before, likelihoods estimated in this way are likely to be unreliable (Gilbert, 1968).

Likelihoods can be obtained from the results of discriminant analysis (Sneath and Sokal, 1973, chapter 8; Darland, 1975) by making other assumptions about

Table 3. Example calculations of methods using likelihoods

(*a*) *Identification matrix*

| | Characters | | | | Prior probability (used only in Bayes' theorem) |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| Taxon A | 0.99 | 0.05 | 0.50 | 0.99 | 0.10 |
| Taxon B | 0.95 | 0.90 | 0.99 | 0.01 | 0.80 |
| Taxon C | 0.01 | 0.01 | 0.75 | 0.05 | 0.10 |

(*b*) *Calculation of likelihoods*

| | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| States of unknown | + | — | + | no result | Likelihood |
| Taxon A | 0.99 × | (1–0.05) × | 0.50 | | = 0.4703 |
| Taxon B | 0.95 × | (1–0.90) × | 0.99 | | = 0.0941 |
| Taxon C | 0.01 × | (1–0.01) × | 0.75 | | = 0.0074 |
| | | | | | Sum = 0.5718 |

(*c*) *Calculation of maximum possible likelihoods*[1, 2]

| | 1 | 2 | 3 | Maximum likelihood | Limit 1[2] | Limit 2[2] |
|---|---|---|---|---|---|---|
| Taxon A | 0.99 | × (1–0.05) × | 0.50 | = 0.4703 | 0.2408 | 0.1613 |
| Taxon B | 0.95 | × 0.90 × | 0.99 | = 0.8465 | 0.4334 | 0.2903 |
| Taxon C | (1–0.01) | × (1–0.01) × | 0.75 | = 0.7351 | 0.3764 | 0.2521 |

Limit 1 = max. likelihood × $(0.8)^3$, Limit 2 = max. likelihood × $(0.7)^3$

(*d*) *Methods of displaying relative likelihoods*

| | Percentage relative likelihood[1] | Identification score[2, 3] |
|---|---|---|
| Taxon A | 0.4703/0.4703 = 100.00% | 0.4703/0.5718 = 0.8225 |
| Taxon B | 0.0941/0.4703 = 20.01% | 0.0941/0.5718 = 0.1646 |
| Taxon C | 0.0074/0.4703 = 1.57% | 0.0074/0.5718 = 0.0129 |

(*e*) *Methods of displaying absolute likelihoods*

| | Estimated frequency of occurrence[4] | Logarithmic probability[2] | Modal likelihood fraction[1] |
|---|---|---|---|
| Taxon A | 1/ 2 | 0.328 | 0.4703/0.4703 = 1.0000 |
| Taxon B | 1/ 11 | 1.026 | 0.0941/0.8465 = 0.1112 |
| Taxon C | 1/135 | 2.131 | 0.0074/0.7351 = 0.0101 |

Table 3. *Continued*

(*f*) *Bayes' theorem using prior probabilities as in* (*a*)

| | Likelihood × prior probability | Posterior probability |
|---|---|---|
| Taxon A | 0.4703 × 0.10 = 0.04703 | 0.04703/0.12305 = 0.3822 |
| Taxon B | 0.0941 × 0.80 = 0.07528 | 0.07528/0.12305 = 0.6118 |
| Taxon C | 0.0074 × 0.10 = 0.00074 | 0.00074/0.12305 = 0.0060 |
| | Sum = 0.12305 | |

(*g*) *Friedman et al. (1973) method, calculation of probabilities excluding taxa*
*Second matrix (for equal prior probabilities)*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Taxon A | (0.95 + 0.01)/2 = 0.48 | 0.455 | 0.87 | 0.03 |
| Taxon B | (0.99 + 0.01)/2 = 0.50 | 0.03 | 0.625 | 0.52 |
| Taxon C | (0.99 + 0.95)/2 = 0.97 | 0.475 | 0.745 | 0.50 |

| | 1 | 2 | 3 | Probability excluding taxon |
|---|---|---|---|---|
| Unknown | + | — | + | |
| Taxon A | 0.48 × (1–0.455) × 0.87 = | | | 0.2276 |
| Taxon B | 0.50 × (1–0.03) × 0.625 = | | | 0.3031 |
| Taxon C | 0.97 × (1–0.475) × 0.745 = | | | 0.3794 |

(*h*) *Friedman et al. (1973) method, without prior probabilities*

| | Relative likelihood score |
|---|---|
| Taxon A | 0.4703/(0.4703 + 0.2276) = 0.6739 |
| Taxon B | 0.0941/(0.0941 + 0.3031) = 0.2369 |
| Taxon C | 0.0074/(0.0074 + 0.3794) = 0.0191 |

(*i*) *Friedman et al. (1973) method, using prior probabilities of 0.3333*

| | Relative likelihood score |
|---|---|
| Taxon A | 0.3333 × 0.4703/(0.3333 × 0.4703 + 0.6667 × 0.2276) = 0.5082 |
| Taxon B | 0.3333 × 0.0941/(0.3333 × 0.0941 + 0.6667 × 0.3031) = 0.1344 |
| Taxon C | 0.3333 × 0.0074/(0.3333 × 0.0074 + 0.6667 × 0.3794) = 0.0097 |

(*j*) *Identification decisions*

| | | |
|---|---|---|
| Lapage et al. (1970) | 0.8225 < 0.999 | Not identified |
| Gyllenberg and Niemelä (1975*a*) | 0.4703 > 0.2408<br>0.8225 < 0.99 | Intermediate |

[1] Dybowski and Franklin (1968).
[2] Gyllenberg and Niemelä (1975*a*).
[3] Lapage, Bascomb, Willcox and Curtis (1970).
[4] API (1977).

the distribution of the character states in the taxa. The usual assumption is of multivariate normal distributions with equal covariance matrices. Other methods which have been developed for medical diagnosis are nearest-neighbour methods (Hills, 1967; Dickey, 1968), which estimate likelihoods by considering sample individuals not necessarily exactly identical but near the unknown ("near" in the sense of similarity or Euclidean distance) and interaction methods (Davies, 1972; Victor, Trampisch and Zentgraf, 1974), which seek to represent the interdependence of the characters in a compact way. In comparison with the simple method which assumes independent character states, the methods allowing for character correlation need a full data matrix which cannot usually be constructed from the literature or incomplete records. They also require much more computation, either in the construction of the identification method, as for discriminant analysis, or in carrying out each identification. Victor et al. (1974) point out that the choice of method depends on the amount of sample data available; theoretically more realistic methods require more parameters to be estimated from the sample data but may give less accurate results if there is insufficient data to give reliable estimates of these parameters. There do not seem to have been any comparative trials of different methods in identification but in medical diagnosis Croft (1972) made an extensive trial of several models applied to the diagnosis of 437 cases of liver disease based on reference data of about 2000 cases. The simple method assuming independent symptoms gave better results than any of the more complex models.

*b. Different methods based on likelihoods*

After calculating the likelihoods of the taxa on the character states of the unknown, a practical identification method needs next to compare these. The results of the likelihood calculation should be displayed in some way, particularly if the method is to be used in a computer-assisted identification procedure (see Part A) in which the results are assessed by a bacteriologist. The method should also incorporate an identification decision element (see Part D). If the identification decision is simply to take the taxon with the highest likelihood as the identity of the unknown then all methods will agree (provided the likelihoods are not modified by prior probabilities, see below). In practice, an identification decision should indicate a definite identification only if the likelihoods meet certain criteria. If they are not met, the likelihoods really indicate that a definite identification is not possible from the character states observed.

The different ways which have been used to display likelihoods and make identification decisions are shown in Table 4 and example calculations by the different methods are given in Table 3 (*c*) to (*j*). Some of the methods are based on Bayes' theorem (Kendall and Stuart, 1963, p. 198):

$$P(J/u) = \frac{P(J)\ P(u/J)}{\sum_J P(J)\ P(u/J)} \qquad (2)$$

where $P(J/u)$ is the probability of taxon $J$ on the character states $u$, known as the posterior probability of the taxon; and $P(J)$ is the probability of $J$ before considering the character states, known as the prior probability of the taxon. The prior probabilities are the frequencies of incidence of the different taxa in the material being identified. Bayes' theorem then allows the posterior probabilities of the taxa to be calculated from their likelihoods and prior probabilities. Theoretically, taking account of prior probabilities should give the highest rate of correct identifications. Requiring that the posterior probability of a taxon exceeds a threshold level before an identification is accepted sets a maximum theoretical error rate for the identification method. The difficulties of estimating the prior probabilites and whether or not their use is desirable in identification are considered in Section IV.*c*. An example calculation using Bayes' theorem is given in Table 3 (*f*) which shows how the listing of the taxa can be changed by taking account of prior probabilities. Friedman and MacLowry (1973) have used unequal prior probabilities in the identification of bacteria, using a modified version of the Bayes' theorem calculation (see below).

Table 4. Identification methods using likelihoods

| Method | Values printed | Identification decisions |
|---|---|---|
| Dybowski and Franklin (1968) | Percentage relative likelihood $= L_J/L_1$ as a percentage<br>Modal likelihood fraction $= L_J/L_J^{max}$ | None |
| Lapage et al. (1970) | Identification score $L_J^* = L_J/\Sigma L_J$ | $L_1^* > 0.999$ Identified<br>$L_1^* \leqslant 0.999$ Not identified |
| Friedman et al. (1973) | Relative likelihood score (does not depend only on likelihoods) | None |
| Gyllenberg and Niemelä (1975*a*) | Normalized probability $L_J^* = L_J/\Sigma L_J$<br>Logarithmic probability $= -\log(L_J)$ | $L_1 \geqslant (l1)^m L_1^{max}$, $L_1^* \geqslant 0.99$ } Identified<br>$L_1 \geqslant (l1)^m L_1^{max}$, $L_1^* < 0.99$ } Intermediate<br>$(l2)^m L_1^{max} \leqslant L_1 < (l1)^m L_1^{max}$ Neighbour<br>$L_1 < (l2)^m L_1^{max}$ Outlier |
| API (1977) | Estimated frequency of occurrence $L_J$ as a fraction | Graded series of decisions based on $L_1$ and $L_1/L_2$ |

$L_J$ likelihood of taxon $J$ ($J = 1$ is taxon with highest likelihood, $J = 2$ is taxon with second highest likelihood), $L_J^{max}$ maximum possible likelihood for taxon $J$ for characters considered, $m$ number of characters considered, $l1$ and $l2$ are parameters setting taxon limits, e.g. $l1 = 0.8$, $l2 = 0.7$ (Gyllenberg and Niemelä, 1975*a*).

If the prior probabilities are set equal for all taxa they cancel out in (2) which becomes

$$P(J/u) = \frac{P(u/J)}{\sum_J P(u/J)} \tag{3}$$

The posterior probabilities of the taxa now refer only to the hypothetical situation of equal prior probabilities and for this reason they have been called "identification scores" (Willcox et al., 1973). In terms of likelihoods (3) is written:

$$L^*_J = \frac{L_J}{\sum_J L_J} \tag{4}$$

where $L^*_J$ is the identification score of $J$. As the $L^*_J$ for all taxa add up to one they have also been referred to as "normalised" values (Lapage et al., 1970; Gyllenberg and Niemelä, 1975*a*, *b*).

Friedman et al. (1973) and Friedman and MacLowry (1973) use the equation:

$$P''_J = \frac{P(J) \prod_i P(x_i/J)}{P(J) \prod_i P(x_i/J) + (1 - P(J)) \prod_i P(x_i/\bar{J})} \tag{5}$$

where $P'_J$ is known as the "relative probability" or "relative likelihood score" of $J$ and $P(x_i/\bar{J})$ is the probability of $x_i$ for an organism not a member of $J$. The $P(x_i/\bar{J})$ are obtained from a second matrix formed from the identification matrix using

$$P(x_i/\bar{J}) = \sum_{K \neq J} P(K)\, P(x_i/K) \,/\, \sum_{K \neq J} P(K) \tag{6}$$

Although this method is based on Bayes' theorem it does not give the same results as the direct use of the theorem as given in (2) nor, for equal prior probabilities, the same results as the identification scores calculation (compare Table 3, *d* and *i*). This would seem to be because equation (5) as well as assuming that the character states are independent in each taxon assumes, in multiplying together the $P(x_i/\bar{J})$, that they are independent excluding each taxon. For equal prior probabilities, the terms in (5) involving prior probabilities do not cancel out unless they are set to 0.5, the value used by Friedman et al. (1973). The use of Bayes' theorem as in (2) requires that the taxa are exclusive and exhaustive and so the prior probabilities must add up to one. Using prior probabilities of $1/q$, where $q$ is the number of taxa, in (5) gives results different from ignoring the prior probabilities (compare Table 5, *h* and *i*). Compared with the other methods, the method of Friedman et al. (1973) requires a second matrix and involves some additional calculations; the method has no immediately obvious advantages though no comparative trials have yet been made.

The different methods, summarized in Table 4, show that both the relative and absolute values of the likelihoods may be displayed and used in the identification decision. Dybowski and Franklin (1968) print the relative likelihood of each taxon as a percentage of the highest likelihood obtained. Lapage et al. (1970) print the identification score based on Bayes' theorem as described above ("normalized probability" in Gyllenberg and Niemelä, 1975*a*, *b*). Friedman et al. (1973) print their relative likelihood score which is difficult to relate to other indices because it is not simply a function of the likelihoods. Percentage relative likelihoods are not so useful as identification scores, for if several taxa are equally likely on a particular set of character states, they will all have percentage relative likelihoods of about 100%, whereas their identification scores will be low. A high identification score is only obtained if the likelihood of one taxon is much greater than any other.

In assessing the absolute likelihoods it is necessary to take account of the differences in variability of the different taxa. The likelihood of a variable taxon will be quite low even for results completely typical of the taxon; for instance a taxon with five matrix entries of 0.5 for the characters considered will always have an absolute likelihood of less than 0.032. To allow for this Dybowski and Franklin (1968) print the likelihood of each taxon as a fraction of the maximum possible likelihood for that taxon. Gyllenberg and Niemelä (1975*a*, *b*) display the absolute likelihoods without adjustment, the negative logarithms of the likelihoods (which can be interpreted as taxonomic distances) and then allow for the different variability of the taxa in making the identification decision. The API (1977) program also prints the absolute likelihoods, without adjustment, as fractions, e.g. 1/10 for a likelihood of 0.1.

Lapage et al. (1970, 1973) base the identification decision on only relative likelihoods, an identification is accepted if the highest identification score exceeds 0.999. An advantage of using these scores based on Bayes' theorem is that although they are only valid as probabilities in the hypothetical situation of equal prior probabilities, regarding them as probabilities does give a guide for the *a priori* setting of the identification parameter. Taking a limit of 0.999 should give a maximum error rate of 1 in 1000 and this limit is valid for any number of characters and so can be used at all times in a sequential identification scheme in which the number of characters determined varies from specimen to specimen. The disadvantage of using the relative likelihood criterion is that a strain may identify with a taxon if it resembles that taxon much more than any other, even though in absolute terms it is quite atypical of the taxon. A theoretical requirement for the use of Bayes' theorem is that the taxa should be exclusive and exhaustive, i.e. any organism considered must belong to one and only one of the taxa. In the identification of bacteria this requirement is not always met; there are some strains which cannot be ascribed to any of the taxa in the matrix. These organisms may belong to known taxa which are not included in the identification matrix, they may belong to as yet unrecognised taxa, or they may reflect the taxonomic situation

suggested by Sneath (1974) and Gyllenberg and Niemelä (1975*a*) in which they are intermediate and aberrant strains between the denser clusters which are recognised as taxa. Lapage et al. (1970, 1973) rely on the judgement of the bacteriologist to reject these false identifications, aided by the reporting of results which are aberrant for the indicated taxon (see Part A). Gyllenberg and Niemelä (1975*a*, *b*) automate this process by basing the identification decision on both relative and absolute likelihoods leading to categories "identified", "intermediate", "neighbour" and "outlier", as shown in Table 4. The API (1977) program also uses both relative and absolute likelihoods in the identification decision but the decision rules are not specified.

A more rigorous formulation of decision rules for probabilistic identification will become possible through statistical decision theory (e.g. Sebestyen, 1962; Gower, 1975). For the present, some of the methods shown in Table 4 have given good results in trial applications (Bascomb et al., 1973; Gyllenberg and Niemelä, 1975*a*) and have been used routinely as identification aids (Part A and API, 1977).

*c. Estimating probabilities for use in identification*

An advantage of probability matrices is that information can be compiled from a variety of sources. Data from the literature can be combined with records of sample strains and the final values entered in the matrix will be based on a subjective assessment of the available information. Although it may be useful to construct an identification matrix in this way for initial trials, Bascomb et al. (1973) found that a matrix based on the results obtained by testing with standardized methods a number of reference strains of each taxon was much more effective. Even if matrix figures are based on the results of sample strains, there is the possibility of adjusting them to take account of *a priori* bacteriological knowledge. For instance, allowance could be made for biovars of a taxon which are known to occur though they have not been encountered in the sample (Lapage, 1974). Estimating the values in the probability matrix thus involves some degree of subjective judgement. Darland (1975) states "extreme care must be taken that the definitions (of the taxa) are based on random samples from the appropriate populations". A more thorough consideration of the relationship of the identification problem to statistical theory is required to clarify the basis for estimating probabilities and the effect of these estimates on the performance of the identification methods. Here we make only some general points on the problems of estimating the probabilities.

Estimating the prior probabilities of the taxa (frequencies of incidence in the material received for identification) is likely to be more difficult than estimating the probabilities in the identification matrix. The prior probabilities will vary from time to time and from place to place due to outbreaks of infection, local conditions and so on. The probabilities of the character states for the taxa are also likely to vary if outbreaks of aberrant strains are involved and geographical variations known to occur (Lapage, 1974) but even so these probabilities will be much more

stable than the prior probabilities. Yankelevitch and Negrete-Martinez (1969) suggest that an identification matrix is mainly a function of evolutionary and genetic effects and can be used in a wide variety of circumstances; the prior probabilities are a function of ecological effects and are only applicable to a particular situation. Ledley and Lusted (1959) present similar conclusions on the use of prior probabilities in medical diagnosis.

If reliable estimates of the prior probabilities can be obtained, their use weights against the rarer taxa (see Table 3). Whether or not this is desirable depends on the type of work which is being carried out (Sneath and Sokal, 1973, at p. 387). In reference laboratories it might be best to base the identification only on the observed characters without considering prior probabilities; computer-assisted identification can recall rare taxa that may be otherwise overlooked (Morse, 1975). In other work it might be more important to identify correctly the highest proportion of organisms overall and the occasional misidentification of a rarity would not be important. In a trial on medical diagnosis, Croft (1972) found, as expected, that using prior probabilities in Bayes' theorem gave the best overall rate of correct diagnosis, while ignoring the prior probabilities improved the diagnosis of the rarer diseases.

If the probabilities in the identification matrix are estimated solely from the results of sample organisms, the most straightforward estimate for each probability is simply the frequency of occurrence of that state in the sample organisms of that taxon. Upper and lower limits to the matrix entries must be set (Section V.*a*) so that probabilities of zero and one are avoided. Willcox et al. (1973) and Sneath (1974) suggest that Laplace's law of succession could be used to estimate the entries giving $P = (m+1)/(n+2)$ where $m$ is the number of positive results observed in a character in a sample of $n$ organism of a taxon and $P$ is the estimated probability of a positive result for the next organisms of the taxon. It is interesting that in practice (Bascomb et al., 1973) the conclusions of Laplace's law are not followed: if only five sample organisms of a taxon were available and all of them gave positive results a matrix entry of 0.99 would be used though Laplace's law indicates an entry of 0.86. There is an *a priori* expectation that taxa will be constant, or nearly so, in their results in a particular character and data presented by Sneath (1974) support this expectation.

The entries in the probability matrix should reflect the behaviour of the organisms received for identification. For example, in the non-fermenter matrix mentioned in Part A, the probability of green pigment production for *Pseudomonas aeruginosa* is entered as 0.10. It is known that the majority of strains of *P. aeruginosa* isolated in medical laboratories produce green pigment but these strains are readily identified by the routine laboratories and are not sent to the reference laboratory. Few of the strains of this taxon received for computer identification produce green pigment. This shows the importance of basing the matrix entries on samples of the appropriate population of organisms as pointed out by Darland (1975). However, Lapage et al. (1973) and Lapage (1974) give

reasons against adjusting the matrix figures *automatically* for each strain identified by a numerical method.

*d. Comparison with taxon-radius methods*

Taxon radius identification methods (Gyllenberg, 1965; Sneath and Sokal, 1973, chapter 8; Sneath, 1974; Gyllenberg and Niemelä, 1975*a*, *b*) are based on a geometrical model. Each taxon is defined by a central point, its centroid, and one or two radii and identification is made by finding the distances of the unknown from the taxon centroids and comparing these distances with the radii of the taxa. The model can be applied in the original attribute space or in a transformed space. Sneath (1969) and Gyllenberg and Niemelä, (1975*a*, *b*) point out the close analogy between the taxon-radius method in original attribute space and the calculation of likelihoods assuming independent character states. (For the method in transformed space, the corresponding analogy is with discriminant analysis; Sneath and Sokal, 1973, chapter 8.) For two-state characters the identification matrix required by the taxon-radius method, the centroid matrix, is identical with the probability matrix.

The Euclidean distance between the unknown and the centroid of taxon $J$,$\Delta_J$ is calculated by:

$$\Delta^2{}_J = \sum_i (x_i - \bar{x}_{iJ})^2 \tag{7}$$

where $x_i$ is the state of the unknown in characters $i$ and $\bar{x}_{ij}$ is the coordinate of the centroid of $J$ for $i$, i.e. the average value of the character for the taxon. For two-state characters, giving the states values 1 and 0 equation (7) can be written:

$$\Delta^2{}_J = \sum_{i^*} (1 - P_{iJ})^2 + \sum_{i^{**}} (P_{iJ})^2 \tag{8}$$

and compared with

$$-log\,(L_J) = \sum_{i^*} -log\,(P_{iJ}) + \sum_{i^{**}} -log\,(1 - P_{iJ}) \tag{9}$$

where $P_{ij}$ is the proportion of 1 states in character $i$ for taxon $J$, $i^*$ represents summation over characters for which the unknown had a 1 state, and $i^{**}$ summation over characters with a 0 state, and $L_J$ is the likelihood of $J$ calculated as in (1). The analogy then is between the weight against the taxon represented by the squared Euclidean distance and the negative logarithm of the likelihood. The contribution to this weight of a single positive state for different values of $P_{ij}$ is shown in Fig. 8. Also plotted in Fig. 8 is the weight resulting from the model suggested by Sneath (1974) in which the contribution to the squared distance for each character is scaled by the variance of the character in the taxon. The functions are plotted in Fig. 8 so that $P_{ij} = 1$ gives a zero weight and $P_{ij} = 0.5$ gives a weight of one. Compared in this way the main difference between the models is seen as the differential weight resulting from a mismatch in a nearly constant character. For

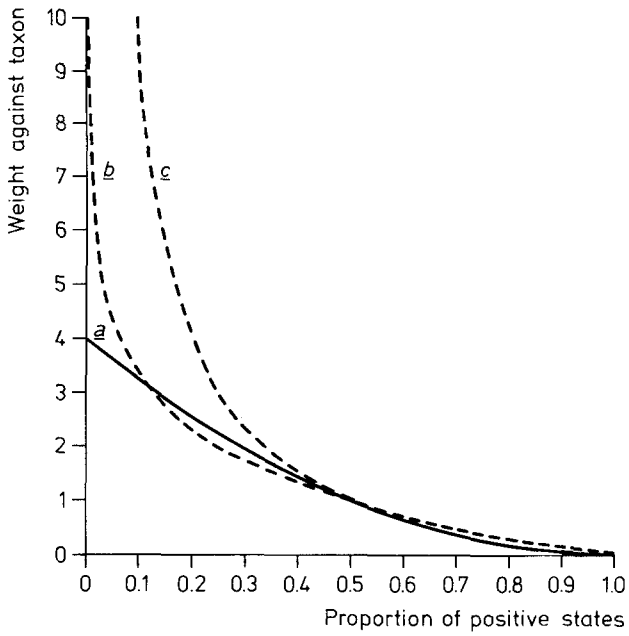



Fig. 8. The weight against a taxon contributed by a single positive character state plotted against the proportion of positive states in the character for the taxon ($P_{iJ}$). (*a*) Euclidean distance model $[(1-P_{iJ})^2]$. (*b*) Likelihood model $[-\log (P_{iJ})]$. (*c*) Scaled Euclidean distance (Sneath, 1974) $[(1-P_{iJ})/P_{iJ}]$. Functions are plotted so that weight is zero at $P_{iJ} = 1.0$ and weight is one at $P_{iJ} = 0.5$. Curves *b* and *c* are indistinguishable on this graph below $P_{iJ} = 0.5$.

the Euclidean distance model, this reaches a maximum of four whereas for the other models it increases indefinitely as $P_{ij}$ approaches zero. For this reason it is necessary to set limits on the entries in a probability matrix to prevent the likelihood identification method becoming monothetic (Section V.*a*). The lower limit used by most workers is $P_{ij} = 0.01$ giving a weight of about 6.6 for the likelihood model and 99 for Sneath's model. This comparison suggests that identification methods using likelihoods or Sneath's model are likely to be more powerful than Euclidean distance methods, in the sense of reaching an identification with fewer characters, but, unless the matrix limits are carefully chosen, they will be more susceptible to errors in testing and the definitions of the taxa. Gyllenberg and Niemelä (1975*b*) report close agreement between likelihood and Euclidean distance methods in the identification of 223 isolates of bacteria.

Sneath (1969) suggests that likelihood identification methods may be susceptible to vigor and pattern effects (Sneath and Sokal, 1973) and gives a way of correcting for the effects in the likelihood model. Gyllenberg and Niemelä (1975*a*, *b*) give an identification method using correlation coefficients which should allow for these effects. As yet there is no evidence that vigor and pattern effects have caused difficulty in numerical identification but they may do so as numerical

identification is applied to other groups of bacteria.

The taxon-radius and likelihood models were compared above in terms of two-state qualitative characters. Multistate qualitative characters are readily handled by the likelihood model (Section V.*d*) but not by the taxon-radius model whereas for continuous quantitative characters the reverse is true (Gyllenberg and Niemelä, 1975*a*, *b*). Continuous quantitative characters can be used in the likelihood model if a probability distribution can be postulated for each character. The parameters of the distribution for each taxon are held in the identification matrix and the probability density for the value of the unknown used in the likelihood calculation. In his trial Croft (1972) found that this procedure did not give as good results as converting each quantitative character to a two-state character by setting a dividing level.

## V. Problems in Calculating Likelihoods from a Probability Matrix

### *a. Limits on matrix entries*

Suppose none of the organisms of a particular taxon which have been examined showed a particular character state. If the matrix entry was given a value of zero then an organism of that taxon having a character state never before observed could never be identified as a member of that taxon, whatever states it showed in the other characters (in equation (1) if any $P(x_i/J) = 0$ then $L_J = 0$). This monothetic behaviour is undesirable since any previous sampling cannot have discovered all possible character states for a given taxon and, furthermore, the result in question might have been due to an error in testing. This problem can be avoided by setting a lower limit to matrix entries, Dybowski and Franklin (1968) use 0.05, Lapage et al. (1970), Friedman et al. (1973) and Gyllenberg and Niemelä (1975*a*) use 0.01. Upper limits to matrix entries are similarly set, e.g. 0.99, to give a corresponding minimum probability for negative results, as one minus the probability of a positive result. The values of these limits have been chosen arbitrarily, but they can be justified by considering errors in testing (Section V.*b*). Willcox et al. (1973) suggest that matrix entries could be estimated using Laplace's law of succession (Section IV.*c*) as probabilities estimated in this way are never zero but this has not been used.

### *b. Unknown matrix entries*

There may be no information available on the behaviour of some of the taxa in some of the characters so the matrix entries are unknown. This may occur if a test is included in the matrix because it is useful in discriminating between some of the taxa but has not been tested on strains of the other taxa. The matrix should be preferably completed by testing sample strains, but methods which allow for unknown values in the matrix enable such tests to be used while further data is being collected.

Friedman et al. (1973) could not obtain sufficient data to establish 48 of the 1,292 elements in their matrix of two-state characters and they set the unknown elements to 0.5. This is reasonable for such a small proportion of unknown elements but could give misleading results if there were many unknowns or if they were concentrated in a few taxa or tests. Willcox et al. (1973) use the following procedure in their identification score calculation.

Unknown matrix elements are given a special value (zero) so that the program can recognise that they are unknown. If the probability of some character states for a taxon, $P(x_i/J)$ of equation (1), are unknown they are assumed to be 0.99 (if this is the upper limit to matrix entries, see Section V.*a*). If one of the resulting identification scores exceeds the threshold level a tentative identification to that taxon is made. If there were no unknown matrix entries in the calculation for the tentatively identified taxon then the identification is accepted because the values assumed for the unknown entries of the other taxa were those most favourable to these taxa and thus the least favourable to the identifying taxon. If, on the other hand, unknown matrix entries for the tentatively identified taxon had been encountered, the calculation must be repeated without assuming values favourable to this taxon, a process termed "rescoring". The identification is accepted only if the score of the same taxon exceeds the threshold level after rescoring. Two strategies can be used for rescoring. In the "lenient strategy" the character states for which the matrix entries for the tentatively identified taxon are unknown are simply ignored for all taxa. In the "stringent strategy" values are assumed for these entries which are the least favourable to the tentatively identified taxon. In either strategy, unknown entries for taxa other than the tentatively identified taxon are treated as before. The lenient strategy seems most suitable for most unknown matrix entries but the stringent strategy is required when dealing with linked characters (Section V.*c*).

*c. Linked characters*

The method assumes that the character states are independent in each taxon. For some bacteriological tests it is known that some results are governed by a simple logical relationship valid for all taxa. For instance if a strain does not grow at 37 °C the results for motility at 37 °C and methyl red at 37 °C must be negative. The procedure given here for taking account of linked characters is described in more detail in Willcox et al. (1973).

Considering the likelihood calculation, equation (1), and supposing that the first three characters are not independent of each other but are independent of the other characters, then

$$P(u/J) = P(x_1, x_2, x_3/J)\ P(x_4/J)\ P(x_5/J)... \qquad (10)$$

so the usual method of multiplying the probabilities of the individual character states can be retained as long as the dependent states are taken separately and their joint probability found. Now,

$$P(x_1, x_2, x_3/J) = P(x_1/J)\ P(x_2/x_1, J)\ P(x_3/x_1, x_2, J) \tag{11}$$

where, for example, $P(x_3/x_1,x_2,J)$ is the probability of state $x_3$ for taxon $J$ and having observed states $x_1$, and $x_2$. These joint probabilities must be calculated by a special procedure in the program. For two-state characters and writing $+_i$ for a positive result in character $i$ and $-_i$ for a negative result, then for the term $P(x_2/x_1,J)$, to allow for all possibilities, $P(+_2/+_1,J)$ and $P(+_2/-_1,J)$ must be calculated; $P(-_2/+_1,J)$ can be found as $1 - P(+_2/+_1,J)$ and $P(-_2/-_1,J)$ as $1 - P(+_2/-_1,J)$. For the linkage $-_1$ implies $-_2$, a negative result in the first test implies a negative result in the second,

$$P(+_2/-_1, J) = 0 \tag{12}$$

follows immediately from the linkage, to obtain the other value

$$\begin{aligned} P(+_2/J) &= P(+_1, +_2/J) + P(-_1, +_2/J) \\ &= P(+_1/J)\ P(+_2/+_1, J) + P(-_1/J)\ P(+_2/-_1, J) \end{aligned}$$

so

$$P(+_2/+_1, J) = P(+_2/J) / P(+_1/J) \tag{13}$$

The required probabilities are obtained as zero or by a simple calculation from the matrix entries for the two tests. The results $-_1, +_2$, given a zero probability by (12) are impossible according to the character linkage, and should be rejected by an editing procedure before reaching the stage of calculating the likelihoods.

For three characters with the same linkage, $-_1$ implies $-_2$ and $-_3$, it is easy to show that

$$P(+_3/-_1, -_2, J) = P(+_3/-_1, J) = 0 \tag{14}$$

and

$$\begin{aligned} P(+_3/+_1, +_2, J) &= P(+_3/+_1, -_2, J) \\ &= P(+_3/+_1, J) = P(+_3/J) / P(+_1/J) \end{aligned} \tag{15}$$

so the same procedure using (12) and (13) will deal with two, three or any number of characters linked in the way $-_1$, implies $-_2$ and $-_3$ etc.

An exception to this straightforward situation occurs if the first character of a linked series of 3 or more characters has not been observed. If there is no result for the first character, $P(+_2/J)$ is obtained from the matrix as for an unlinked test, $P(+_3/+_2,J)$ can be obtained from (15) as usual, since the first test must have been positive, but a new equation is needed to obtain $P(+_3/-_2,J)$. This can be easily derived but is more complex and becomes increasingly so when four and more linked tests are considered.

The problem can be avoided by saying that if two or more of a series of linked characters have been determined but the first of the series has not, then the result of the first character can be assumed to be positive. It then becomes unnecessary to

include a number of additional equations in the program for situations which occur, in practice, very infrequently. The program also remains completely general and can handle any number of characters linked in this way. This form of linkage is the only one considered by Willcox et al. (1973) but other forms could be analysed in the same way.

Consider the calculation of the probability of the results $+_1, -_2$ for two linked tests:

$P(+_1, -_2/J) = P(+_1/J)\,P(-_2/+_1, J) = P(+_1/J)[1 - P(+_2/J) / P(+_1/J)]$.

If $P(+_1/J) = P(+_2/J)$ then $P(+_1, -_2/J) = 0$ which is unacceptable for the reasons given in Section V.*a*. In practice (Willcox et al., 1973) this situation usually occurs when $P(+_1/J) = P(+_2/J) = 0.99$ or 0.01. Reasonable results can be obtained by recognising that, because of the limits on matrix entries an entry of 0.99 represents a probability somewhere between 1 and 0.99 and an entry of 0.01 a probability between 0 and 0.01. Then using the procedure for unknown matrix entries (Section V.*a*), the most favourable values are first assumed; for $P(+_1/J)$ and $P(+_2/J)$ entered as 0.99 assume $P(+_1/J) = 1$, $P(+_2/J) = 0.99$ giving $P(-_2/+_1, J) = 0.01$; for $P(+_1/J)$ and $P(+_2/J)$ entered as 0.01 assume $P(+_1/J) = 0.01$, $P(+_2/J) = 0$ giving $P(-_2/+_1, J) = 1$. Because values have been assumed for matrix entries, rescoring can be necessary but only when one or more results are unexpected for the tentatively identified taxon (e.g. $+_1$ where $P(+_1/J) = 0.01$). The lenient strategy of rescoring is not applicable here because it would mean ignoring results known to be unexpected. The stringent strategy is not immediately applicaple either as the least favourable assumptions give $P(+_1/J) = P(+_2/J)$ and

Table 5. Example calculations for linked characters (negative result in character 1 implies negative results in 2 and 3)

| | Characters | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Matrix entries | 0.99 | 0.99 | 0.99 |
| Results of unknown | + | − | − |
| Likelihood calculation | 0.99 × | ?[1] × | ?[1] |
| Assumed probabilities | 0.99 × | 0.01 × | 0.01 |
| Rescore probabilities | 0.99 × | 0.01 × | 0.01 |
| | 1 | 2 | 3 |
| Matrix entries | 0.01 | 0.01 | 0.01 |
| Results of unknown | + | + | + |
| Likelihood calculation | 0.01 × | ?[1] × | ?[1] |
| Assumed probabilities | 0.01 × | 1 × | 1 |
| Rescore probabilities | 0.01 × | 0.01 × | 0.01 |

[1] ?: usual procedure gives unacceptable value.

$P(+_1, -_2/J) = 0$. A simple procedure for rescoring can be derived by using the stringent strategy and taking account of test errors (Willcox et al., 1973). The calculation for the tentatively identified taxon is repeated, setting the probability of any results of linked characters which are unexpected to 0.01 irrespective of the value given by the linked characters equations.

The example calculations in Table 5 show that this procedure gives reasonable results. In the second example, the linked characters were growth at 37 °C, motility at 37 °C and methyl red at 37 °C; the matrix figures show that strains of this taxon able to grow at 37 °C were rarely encountered. If a strain was found to grow at 37 °C, past experience does not indicate whether or not it would be positive in the other tests at this temperature. First it is assumed that all such strains would be positive, giving probabilities of one; then, for rescoring, it is assumed that few if any of these strains would be positive, giving probabilities of 0.01.

*d. Multistate characters*

In applications of this method most of the characters have been two-state. For such characters the probability of a negative result for a given taxon is not stored in the matrix but is calculated as one minus the probability of a positive result. In general terms, for an $n$-state character, either $n-1$ probabilities can be stored and the probability of the $n$th calculated as required; or $n$ probabilities can be stored and obtained directly. For two-state characters the first approach halves the amount of computer store required, for multistate ($n > 2$) characters the saving in store is proportionately less and more calculation is necessary to obtain the $n$th probability. For two-state characters the first approach is usually preferable, for multistate characters the second. The same calculating procedure can be used for both types of character if each multistate character is entered in the matrix as $n$ component characters. Any result of a multistate character is shown as a positive result in the appropriate component character, the other components being scored "not done" and hence ignored in the calculation.

*e. Taking account of test errors*

Results of bacteriological tests are not completely reproducible and therefore any test result has a certain probability of being incorrect (Sneath and Johnson, 1972; Sneath and Collins, 1974). Sneath (1974) considers the effects of these errors on identification methods. The likelihood method can be adapted to allow for test errors and use the results obtained to justify the limits set on matrix entries (Section V.*a*). The reasoning followed is that character states which are unexpected for a particular taxon may be due to errors in testing and hence the minimum probability for such states is the probability of an error.

If $p_i$ is the probability of an error in character $i$, $P(+_i/J)$ the "true" probability of a positive result in $i$ for taxon $J$ and $P^a(+_i/J)$ the equivalent apparent probability allowing for test errors, then

$$P^a(+_i/J) = P(+_i/J)\ (1 - p_i) + (1 - P(+_i/J))p_i \tag{16}$$

for $p_i = 0.01$

$$P^a(+_i/J) = 0.98\ P(+_i/J) + 0.01$$

and similarly for a negative result

$$P^a(-_i/J) = 0.99 - 0.98\ P(+_i/J) \tag{17}$$

If equations (16) and (17) were used in the identification method and $P(+_i/J)$ obtained from the matrix there would be no need to set limits on matrix entries as $P^a(+_i/J) = 0.01$ for $P(+_i/J) = 0$ and $P^a(-_i/J) = 0.01$ for $P(+_i/J) = 1$. The matrix entries are estimates of the true probabilities, the results on which they are based are assumed to be free of errors. In practice there are difficulties in using these equations directly but the effect of them is very similar to that of the original method of setting limits on matrix entries. Table 6 shows that the matrix entries and apparent probabilities differ only slightly over most of the range of probabilities and the greatest difference (a factor of about two at a probability of 0.01) can be considered negligible as estimates of such low probabilities are likely to be only approximate.

Table 6 also shows that the limits on matrix entries mean that for entries of 0.09 and 0.01 the true probability is only known to lie within a range of values. To allow for this the approach used for unknown matrix entries (Section V.*a*) is followed. The most favourable values are first assumed for all taxa, e.g. for a positive result and a matrix entry of 0.99, assume $P(+_i/J) = 1$ giving $P^a(+_i/J) = 0.99$, while for an entry of 0.01, assume $P(+_i/J) = 0.01$ giving $P^a(+_i/J) = 0.0198$. For rescoring the lenient strategy is not appropriate, in this case since the probabilities are not completely unknown but are known to lie within limits, instead the least favourable values are assumed for the tentatively identified taxon $P^a(+_i/J) = 0.9802$ for a matrix entry of 0.99 and $P^a(+_i/J) = 0.01$ for an entry of 0.01. The difference for an

Table 6. The values entered in the identification matrix (applying upper and lower limits of 0.99 and 0.01) and the apparent probabilities (allowing for an 0.01 probability of an error in testing) for various probabilities of a positive result in a given character for a given taxon

| True probability | Matrix entry | Apparent probability |
|---|---|---|
| 1 | 0.99 | 0.99 |
| 0.99 | | 0.9802 |
| 0.5 | 0.5 | 0.5 |
| 0.1 | 0.1 | 0.108 |
| 0.05 | 0.05 | 0.059 |
| 0.01 | 0.01 | 0.0198 |
| 0 | | 0.01 |

entry of 0.99 is clearly negligible and the difference for an entry of 0.01, a factor of about two, can also be ignored as there will usually be few such unexpected results for the tentatively identified taxon. Rescoring is therefore unnecessary.

This approach can be applied in a similar way to linked characters. For such characters it is found that rescoring is sometimes required and the results obtained (Willcox et al., 1973) give the simple procedure for rescoring these characters which was described in Section V. *c*.

The difficulty in applying this approach directly is in estimating the error rates and these were assumed above to be 0.01 for all tests. Data compiled by Sneath (1974) show a typical average error rate of 0.02 with significant differences between tests. If results obtained in different laboratories are used in numerical identification (see Part A), differences in media and methods between laboratories are a further source of variation and lead to an error rate of about 0.06 (Willcox et al., 1973).

Finally, in carrying out an identification some tests may be repeated and the probability of an error in a confirmed result is obviously much reduced. A full application of this approach then should use different error rates for different tests, different rates for results obtained in different laboratories and allow for confirmed results.

# PART C. NUMERICAL METHODS FOR CONSTRUCTING IDENTIFICATION KEYS AND SELECTING SETS OF CHARACTERS

## VI. Introduction

Numerical methods can be used to select characters for use in identification whether or not the identification itself is carried out by a numerical method. Numerical character selection can be applied to choose characters to form an identification key or sets of characters for diagnostic tables. Alternatively, numerical selection can be combined with numerical identification to form a flexible, sequential identification scheme (see Part A).

The methods for numerical character selection differ according to the different types of data matrices to which they can be applied (Part D). Most of the methods use non-probabilistic matrices where entries for each taxon are the expected character states. Methods for constructing keys from such data matrices are described in Section VII and methods for selecting sets of characters in Section VIII. Some methods which use probability data matrices and full data matrices are reviewed in Sections IX and X, respectively.

## VII. Constructing Keys from Non-Probabilistic Data Matrices

### *a. General aims and constraints*

Computer programs for constructing identification keys from non-probabilistic data matrices have been described by Pankhurst (1970), Gower and Barnett (1971), Morse (1971), Watson and Milne (1972), Dallwitz (1974), Hall (1975) and Payne (1975). They differ in the numerical functions used for evaluating characters (reviewed by Hill, 1974) and in practical refinements (reviewed by Pankhurst, 1974).

Some of the features of key constructing programs are illustrated by the two example keys derived from the data matrix of Fig. 9 (*a*) and represented schematically in Fig. 9 (*b*) and (*c*). The usual aim in constructing a key is to minimize the average length of the key, i.e. the number of characters necessary to identify a specimen averaged over all the taxa. Key (*b*) has an average length of 16/6 and key (*c*) a length of 17/6, so on this criterion (*b*) is preferable. In taking these averages each taxon was given equal weight but if the prior probabilities of the taxa are known to be different, this should be taken into account in assessing the keys. For instance, if taxon 3 was known to be particularly abundant, key (*c*) would be preferred.

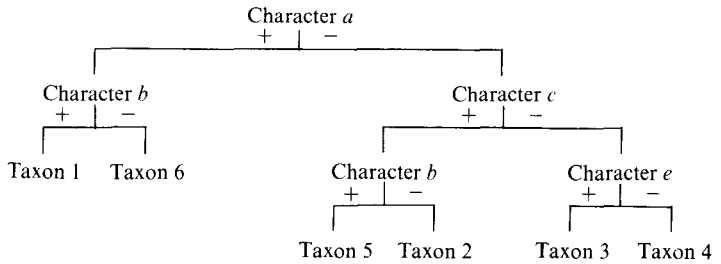
Some programs allow prior probabilities to be specified and use them in the numerical evaluation of the characters. If the prior probabilities are unknown, a sensible objective would be to minimize the longest path through the key. Key (*b*) is the best by this criterion and key (*c*) would be particularly unsuitable if it turned out that taxon 6 was very abundant in the material examined. The different

(*a*)

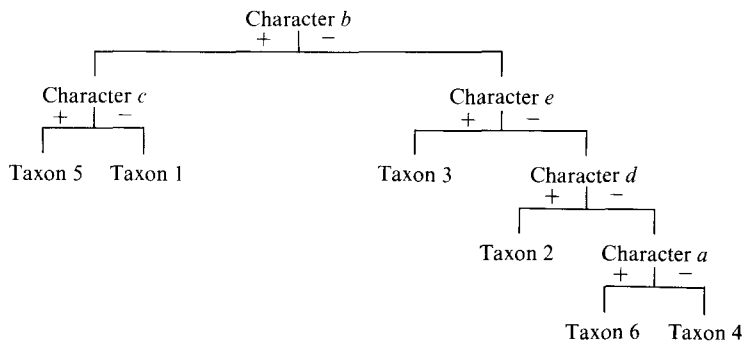
| | Characters | | | | |
|---|---|---|---|---|---|
| | *a* | *b* | *c* | *d* | *e* |
| 1 | + | + | − | − | d |
| 2 | − | − | + | + | − |
| Taxa 3 | − | − | − | − | + |
| 4 | − | − | − | − | − |
| 5 | − | + | + | − | − |
| 6 | + | − | d | − | − |

d = members of the taxon differ in their results

(*b*)




(*c*)




(*d*)

| | Characters | | | |
|---|---|---|---|---|
| | *a* | *b* | *c* | *e* |
| 1 | + | + | − | d |
| 2 | − | − | + | − |
| Taxa 3 | − | − | − | + |
| 4 | − | − | − | − |
| 5 | − | + | + | − |
| 6 | + | − | d | − |

(*e*)

| | Characters | | | |
|---|---|---|---|---|
| | *a* | *b* | *d* | *e* |
| 1 | + | + | − | d |
| 2 | − | − | + | − |
| Taxa 3 | − | − | − | + |
| 4 | − | − | − | − |
| 5 | − | + | − | − |
| 6 | + | − | − | − |

Fig. 9. Two keys and two diagnostic tables constructed from the same hypothetical data. (*a*) Data matrix, (*b*) and (*c*) keys, (*d*) and (*e*) diagnostic tables.

characters can also be given different weights in assessing the keys. Thus if character *a* was inconvenient (e.g. time consuming, difficult to observe or unreliable) key (*c*) would be preferred as it only uses this character in distinguishing two taxa while key (*b*) uses it for all taxa.

Characters can be allotted numerical "costs" to account for these practical considerations and the program aims to find a key with the minimum average cost per identification. The costs can be used directly in some character evaluating functions; alternatively the costs can be used to divide the characters into blocks, the program evaluates the characters in the least costly block first and only considers the characters in the next block if the key cannot be completed without them. In microbiology, the characters are not usually determined sequentially because of the time required to determine them. Instead, a set of tests is carried out and a key which contains the fewest characters overall is advantageous. Key (*c*) contains all the characters but character *c* could be replaced by character *a* to reduce the number of characters. Some programs seek to minimize the overall number of characters by giving preference to characters already used in other branches.

*b. General procedure*

The general procedure for constructing a key is shown in the example of Fig. 10. The first character is chosen according to a numerical evaluation of the characters applied over all taxa; this character divides the taxa into two subsets. The next character for each subset is found by a similar evaluation of the characters applied to the taxa present in that subset and this is continued until the key is complete. Different functions used to evaluate the characters are discussed in Section VIII.*c*. (the example uses that of Rypka et al., 1967). The data matrix used in Fig. 9 (*a*) contains some "d" entries for variable where different strains in a taxon give different responses in a character. The formula used in the example is not applicable when variable entries occur so only the characters with no such entries for the taxa remaining at a particular stage are evaluated. This is only satisfactory if variable entries occur infrequently in the matrix, otherwise it will be impossible even to start constructing the key. Some programs use evaluating functions which allow for variable entries.

If a character with variable entries is used in a key, some taxa will be present in both subsets of the key and hence will appear at more than one end-point of the key. Matrix entries may be specified as "unknown" or as "inapplicable". Unknown entries can be treated as variable entries but characters with some inapplicable entries should not be selected as their presence in the key might confuse the user.

The example data matrix has only two-state characters and the resulting key is dichotomous. Some programs use functions which evaluate multistate characters creating polychotomous keys with more than two subsets at some stages. Polychotomous divisions often give shorter keys yet dichotomous keys are usually

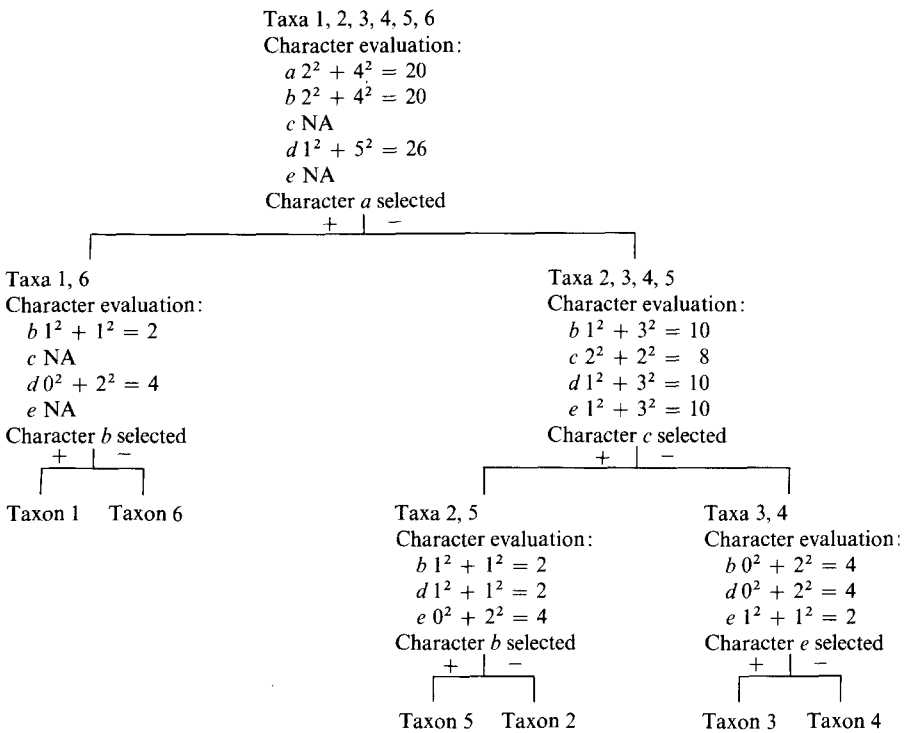



Fig. 10. Example showing the construction of a key by a numerical method. The data matrix is as in Fig. 9 (*a*) and the key constructed is as in Fig. 9 (*b*). The character evaluating formula is $(n_+)^2 + (n_-)^2$, where $n_+$ is the number of positive entries for the character and $n_-$ is the number of negative entries; the character with the lowest value is selected. The formula is not applicable (NA) when variable (d) entries occur.

preferred on practical grounds (Pankhurst, 1970; Dallwitz, 1974), probably because dichotomous keys are easier to use in printed from and in some cases two-state characters are easier to observe and more reliable than multistate. For these reasons, some functions which evaluate multistate characters have a deliberate bias to favour two-state characters. Combinations of characters can also be evaluated by considering all possible combinations of a set number of characters and evaluating each choice as a single multistate character. Some programs can also use quantitative characters by selecting the best point in the range of values to make a dichotomous division.

At some stages in the example of Fig. 10 two characters are equally good choices according to the numerical evaluation; in these instances the first character was arbitrarily taken. By using a different rule to decide between equal values the same method can produce a number of different keys; the key of Fig. 9 (*c*) is produced by reversing the order in which the characters are considered. This shows that

numerical methods which construct keys sequentially, selecting the best character at each stage, do not necessarily produce the best possible key. Key (*b*) has a shorter average length than key (*c*), yet with the reversed character order the automatic method produces only key (*c*).

To ensure the best key is found necessitates investigation of all possible arrangements of characters, but this is computationally impracticable. Selecting characters sequentially one at a time is the simplest strategy for searching for the best characters. The search could be increased in "breadth" by considering in turn all the characters which are equally good choices at any stage. Exactly equal values may not occur very frequently in problems of practical interest but characters with values within a specified interval from the best value could be considered. The search could also be increased in "depth" by evaluating each character not only on the division it creates but also on the further divisions one or more stages further down the key.

Programs so far developed use the simple strategy but a number of alternative keys can often be produced from the same data by changing, for example, the character costs which may indicate the best key for a particular purpose.

Some key-constructing programs allow particular characters to be used in particular positions in the key so that the key will reflect important taxonomic relationships; it is also another way of generating alternative keys. Other features allow keys to be constructed for special purposes by specifying that certain taxa need not be distinguished or certain characters must not be used. Keys can then be produced for particular ecological situations or seasons of the year and so on.

Some programs search for confirmatory characters for each division (in Fig. 9 (*b*) character *d* confirms character *b* in distinguishing taxa 2 and 5) and some print with each taxon name the appropriate states of the characters not used in keying out the taxon. These refinements aim to overcome the monothetic nature of keys where a mistake in observing a character early in the key can lead to a gross misidentification.

Finally, some programs print the key in a form suitable for immediate use or even publication with control over the style of printing.

*c. Character evaluating functions*

A character to be evaluated at a particular stage in the key divides the taxa into subsets corresponding to the alternative states of the character. The average cost of completing an identification from that stage if that character is used is:

$$C = c + \sum_b p_b \, l_b \, c_b \tag{18}$$

where $c$ is the cost of the character, $p_b$ is the probability of state $b$ of the character at that stage, $l_b$ is the average number of additional characters required to complete an identification following state $b$ of the character, and $c_b$ is the average cost of these characters. To minimize the average cost of an identification, the best

character is the one giving the lowest $C$. If there are no "d" entries in the data matrix, $p_b$ is the sum of the frequencies of incidence of the state $b$ taxa present divided by the sum of the frequencies of all the taxa. As $l_b$ and $c_b$ cannot be found without actually completing the key, they must be estimated and Dallwitz (1974) sets $c_b$ to $c_{min}$ (which is the smallest cost for any of the characters being considered) and estimates $l_b$ by $L(q_b)$, where $q_b$ is the number of taxa with state $b$ and $L(q)$ is a function which gives the expected average number of characters to identify a specimen in a key for $q$ taxa. Then (18) becomes:

$$C = c + c_{min} \sum_b a_b L(q_b) / \sum_b a_b \tag{19}$$

where $a_b$ are the frequencies of the taxa (or "abundances", Dallwitz, 1974) a that stage which give state $b$. To obtain $L(q)$ consider a key made up of two-state characters which successively divide the taxa into equal subsets; this is the most efficient arrangement of two-state characters and $m$ characters will distinguish $2^m$ taxa. Even when the number of taxa is not an exact power of 2 the average number of characters per taxon in a key for $q$ taxa is very close to $log_2(q)$ and using this function in (19) gives:

$$C1 = c + c_{min} \sum_b a_b \log_2(q_b) / \sum_b a_b \tag{20}$$

which is the function used by Dallwitz (1974). Dallwitz points out that for equal character costs and taxon frequencies (20) reduces, apart from constants, to

$$C1 = \sum_b q_b \log_2(q_b) \tag{21}$$

which is equivalent to the function proposed by Maccacaro (1958).

The logarithmic function was the only one used by Dallwitz (1974) but other functions for $L(q)$ might be considered. The least efficient use of two-state characters in a key is to successively divide off a single taxon at a time and for such a key for $q$ taxa the average number of characters per taxon is $(q-1)(q+2) \,/\, 2q$ which is closely approximated by $q/2$. Using this function in (19) gives:

$$C2 = c + \tfrac{1}{2} c_{min} \sum_b a_b q_b / \sum_b a_b \tag{22}$$

which for equal character costs and taxon frequencies reduces, apart from constants, to

$$C2' = \sum_b (q_b)^2 \tag{23}$$

a function suggested by Sneath (1974) and equivalent to the function used by Rypka et al. (1967). The general equation (19) thus produces alternative evaluating functions according to the choice of $L(q)$ and this choice depends on the number of characters available compared with the number of taxa. The more characters available the more likely that keys approaching the ideal assumed by

the logarithmic function will be produced.

If only two-state characters with no "d" entries are evaluated and character costs and taxon frequencies are equal, all the functions which have been proposed agree in selecting the character which has the most nearly equal numbers of positive and negative states for the taxa considered. With these conditions function (19) is not necessarily a minimum at the equal division point for all $L(q)$ which are reasonable (i.e. monotonic increasing functions of $q$).

If there are "d" entries, the $p_b$ of (18) cannot be found because the probabilities of some states for some taxa are not known. To use (18) values could be assumed for the probabilities of the different states for taxa with "d" entries, e.g. for two-state characters assuming probabilities of $\frac{1}{2}$ for positive and negative states gives:

$$C = c + c_{min}\,[(a_+ + \tfrac{1}{2}a_d)\,L(q_+ + q_d) + (a_- + \tfrac{1}{2}a_d)\,L(q_- + q_d)] \\ /[(a_+ + \tfrac{1}{2}a_d) + (a_- + \tfrac{1}{2}a_d)] \qquad (24)$$

where $q_+$ is the number of taxa with positive states, $a_+$ the sum of the frequencies of these taxa, and similarly $q_-$ and $a_-$ for taxa with negative states, and $q_d$ and $a_d$ for taxa with "d" entries. The taxa with "d" entries occur in both subsets so the numbers of taxa left to discriminate are $q_+ + q_d$ and $q_- + q_d$ respectively. For equal character costs and taxon frequencies and using $q/2$ for $L(q)$ (24) reduces to

$$C3' = (q_+ + \tfrac{1}{2}q_d)\,(q_+ + q_d) + (q_- + \tfrac{1}{2}q_d)\,(q_- + q_d) \qquad (25)$$

which is equivalent to the function used by Morse (1971).

The frequencies of the taxa appear in (24) but if a character with "d" entries has been used early in the key, some taxa will not be present at a later stage with their full frequencies because some of the members of these taxa will be following a different branch or subset of the key. This effect can be allowed for by again assuming probabilities for the "d" entries and reducing the frequencies of the taxa with these entries at later stages of the key.

Dallwitz (1974) did not use assumed probabilities but instead used (20) directly taking $q_b$ as the number of taxa in the subset indicated by state $b$ and $a_b$ as the sum of the frequencies of these taxa. Retaining the notation of (24) this gives for two-state characters:

$$C = c + c_{min}\,[(a_+ + a_d)\,L(q_+ + q_d) + (a_- + a_d)\,L(q_- + q_d)] \\ /[(a_+ + a_d) + (a_- + a_d)] \qquad (26)$$

which for equal character costs and taxon frequencies and using $log_2(q)$ for $L(q)$ is nearly, though not exactly, equivalent to the information criterion of Gower and Barnett (1971). Using (20) in this way is equivalent to entering each possible variant of a taxon with "d" entries and giving each variant a frequency equal to that of the entire taxon but not requiring the key to separate each variant. This procedure enables those programs whose character evaluation functions cannot

accept "d" entries to analyse matrices with such entries. This procedure will give good results unless some of the taxa have a particularly high proportion of "d" entries.

Evaluation of multistate characters with "d" responses requires the exact specification of such responses. For example, if the states are say, red, green and blue then a taxon with red, green, but no blue members will appear in the subsets indicated by red and green states but not in the blue subset (see Section VIII).

Finally, the problem of constructing a key to minimize the maximum cost of an identification is a sensible objective if the prior probabilities of the taxa are completely unknown. The appropriate character evaluation function is now:

$$C^* = c + c_{min}\, max\, [L(q_b)] \tag{27}$$

where *max* means the maximum value over all $b$. As the $p_b$ do not appear in this function it can be used with "d" entries without qualification provided $q_b$ is the number of taxa in the subset indicated by state $b$ i.e. including taxa with "d" entries. For equal character costs (27) reduces to $max[L(q_b)]$ which is equivalent to $max[q_b]$ as $L(q)$ is always a monotonic increasing function so for this case a single function is obtained independent of the choice of $L(q)$.

As well as the character evaluating function a program for constructing keys must contain an inference element which decides which taxa are present at a particular stage of the key. The method of inference followed in identifying a specimen with a key is set by the method used in constructing the key (Part D). The inference in the example (Fig. 10) is a simple monothetic method in which a taxon is eliminated as soon as a character state is observed which is not as expected for that taxon. This seems to be the only inference method used in programs so far though other methods could be used and Hill and Silvestri (1962) and Möller (1962) describe the construction of a key with a check on the final inference by probability calculation (see Section IX).

Gower and Payne (1975) develop a number of criteria to be met by functions for selecting characters for keys (and a new generalised function), but their approach is not entirely compatible with the generalised form of Dallwitz (1974; see also Brown, 1977).

## VIII. Selecting Sets of Characters from Non-Probabilistic Data Matrices

To choose a set of characters to be observed simultaneously for use in a diagnostic table, or other identification methods which require a set of characters, combinations of characters have to be evaluated. When combinations of characters are used in keys (Section VII) only small numbers of characters can be involved as examination of all possible combinations of larger numbers of characters is impracticable. With large sets, characters are selected one by one.

*Selection of characters according to separation values*

Initial values ($n_+ \times n_-$) *a* 8, *b* 8, *c* 6, *d* 5, *e* 4
Character *a* selected
Completes separation of: 1, 2 (reduce value of *b*, *c*, *d*); 1, 3 (ditto *b*); 1, 4 (ditto *b*); 1, 5 (ditto *c*); 2, 6 (ditto *d*); 3, 6 (ditto *e*); 4, 6; 5, 6 (ditto *b*)
New separation values *a* 0, *b* 8 −4=4, *c* 6 −2=4, *d* 5 −2=3, *e* 4 −1=3
Character *b* selected
Completes separation of: 1, 6; 2, 5 (reduce value of *d*); 3, 5 (ditto *c*, *e*); 4, 5 (ditto *c*)
New separation values *a* 0, *b* 0, *c* 4 −2=2, *d* 3 −1=2, *e* 3 −1=2
Character *c* selected
Completes separation of: 2, 3 (reduce value of *d*, *e*); 2, 4 (ditto *d*)
New separation values *a* 0, *b* 0, *c* 0, *d* 2 −2=0, *e* 2 −1=1
Character *e* selected
Completes separation of: 3, 4
New separation values *a* 0, *b* 0, *c* 0, *d* 0, *e* 0
Set of characters complete

*Separation matrix*

Initial

| Taxa | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | | 1 | 1 | 1 | 1 |
| 3 | | | 1 | 1 | 1 |
| 4 | | | | 1 | 1 |
| 5 | | | | | 1 |

After character *a* selected

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

After character *b* selected

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

After character *c* selected

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

After character *e* selected

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

Fig. 11. Example showing the selection of a set of characters by a numerical method. The data matrix is as in Fig. 9 (*a*) and the set of characters selected is as used in the diagnostic table of Fig. 9 (*d*). Characters are selected according to their separation values i.e. the number of pairs of taxa they separate. After each character is selected the values of the other characters are reduced according to the pairs of taxa now completely separated. The separation matrix records the number of characters still required to separate each pair of taxa. Initially this is set to the specified minimum number of characters, one in this example.

The method described here is based on that of Gyllenberg (1963) as used by Willcox et al. (1973). Taxa are separated by the characters if they show different character states but not if they have the same state or if a taxon has a "d" entry. The method consists of the following steps (see example in Fig. 11).

(1) Set the elements of a taxon by taxon "separation matrix" to the minimum number of characters required to separate each pair of taxa.
(2) For each character count the number of positive and negative entries in the data matrix and calculate the value of the character as the product of these numbers. (This is the "separation figure" of Gyllenberg, 1963.)
(3) Find the character with the highest value; if more than one character has this value take the first encountered. If the highest value is zero character selection is complete, otherwise the character with the highest value, character $x$ say, is added to the set of selected characters. The value of $x$ is set to zero so it will not be selected again.
(4) For all pairs of taxa not already completely separated (i.e. with non-zero entries in the separation matrix) determine whether $x$ separates the pair. If so subtract one from the separation matrix entry for that pair and if the entry is now zero inspect each character not already selected and if the character also separates the pair reduce its value by one.
(5) Return to step (3).

This procedure is designed for two-state characters but can be adapted to allow for multistate characters. Each multistate character is represented by a number of two-state "component characters", one for each state. The appropriate entries, +, − or "d", are determined as shown in the example (Fig. 12); once a + entry is made for a particular taxon the following entries are all "d". The resulting entries are such that not more than one of the component characters separates a given pair of taxa. The entries also represent the behaviour of a multistate character for taxa

| Taxon | States of multistate character | Component characters | | | |
|---|---|---|---|---|---|
| | | *a* | *b* | *c* | *d* |
| 1 | *a* | + | d | d | d |
| 2 | *b* | − | + | d | d |
| 3 | *c* | − | − | + | d |
| 4 | *d* | − | − | − | + |
| 5 | *a* or *b* | d | d | − | − |
| 6 | *b*, *c* or *d* | − | d | d | d |
| 7 | *a*, *b*, *c* or *d* | d | d | d | d |

Fig. 12. Example showing the representation of a four-state character in a non-probabilistic matrix by four two-state component characters. Once a positive state is entered for a particular taxon the following entries are all "d".

which vary in their response, e.g. a taxon which can show state *a* or *b* is separated from one which shows state *c* only (see Section VII). At step (3) above the values of the component characters are added together to give the value of the multistate character itself and if the multistate character is selected step (4) is repeated with each component character in turn as *x*.

Rypka et al. (1967) used a similar method for selecting sets of characters but found the character values at each stage by an algebraic formula rather than the procedure described above. The formula of Rypka et al. (1967) is a generalization of Gyllenberg's (1963) separation figure but is only applicable when there are no "d" entries and when the minimum number of characters to separate each taxon pair is one. Niemelä, Hopkins and Quadling (1968) use a different formula which has the same restrictions. Advantages of the present method are that it is not affected by "d" entries and it allows the identification performance of the resulting diagnostic table (Part D) to be increased by increasing the minimum number of characters required to separate each pair of taxa.

This sequential character selection method does not necessarily find the set with the fewest possible characters, as shown in the example of Fig. 11 by changing the data matrix entry for taxon 2, character *b* to "d"; the method then fails to find the minimum set of *a*, *b*, *d* and *e*. A method which always finds the set with the fewest characters is described by Willcox and Lapage (1972).

A program implementing this character selection procedure can incorporate a number of practical refinements. It can indicate if any pairs of taxa cannot be completely separated by the available characters or it can allow the user to specify that certain pairs of taxa need not be separated. The user can specify that some characters should not be used unless it is impossible to complete the set without them or, conversely that some characters must appear in the set (the program would then start with these characters). Certain characters may always be required in the set as the only ones to separate particular pairs of taxa and the program can discover these necessary characters first and then complete the set in the usual way. The program can produce alternative sets of characters either by the user changing some of the above options or automatically by taking in turn the characters which are equally good choices at a given stage (or within a specified interval from the best value).

## IX. Methods Using Probability Matrices

The first approach to selecting characters from a probability matrix is to convert the matrix to a non-probabilistic form and use the methods described in Sections VII and VIII. Hill and Silvestri (1962) with Möller (1962) produced a key in this way and Willcox et al. (1973) used it to select sets of characters in a flexible, sequential identification scheme (Part A). For two-state characters probabilities of positive states $\geqslant 85\%$ were converted to +, probabilities $\leqslant 15\%$ to −, and

remaining probabilities to "d", these limits having been chosen subjectively. Upper and lower probability limits can be set for each data matrix according to the frequency distribution of the probabilities (Möller, 1962). The nearer the limits are to 100% and zero the more reliable will be the resulting identification method, though more characters will be required, and if the limits are set too wide potentially useful characters may be ignored. In practice the distribution of the probabilities of positive results is strongly U-shaped in most data matrices (Sneath, 1974) so the values of the limits will probably not be critical. Though the probability matrix was reduced for character selection, probability calculations were used by Möller (1962) to check the reliability of the key produced, and by Willcox et al. (1973) as the inference element of their sequential identification scheme. To use this method of character selection in sequential identification, the identification scores (Part B) of the taxa are first calculated on the initial character states. If the highest score does not exceed the identification level (0.999) a set of likely taxa is formed by taking taxa in order of their scores until the sum of the scores is greater than the identification level. The characters available for selection are determined by excluding those characters already observed and characters whose outcome can be predicted from results already known (e.g. motility at 37 °C would be excluded if growth at 37 °C was known to be negative). A set of characters is selected from the available characters to discriminate between the likely taxa by forming a subsidiary, non-probabilistic, data matrix for these characters and taxa and using the method described in Section VIII. Willcox et al. (1973) find that each pair of taxa must be separated by at least two characters to give a reasonable chance of identification by the probability calculation when the results of the selected characters are combined with the initial results.

Sneath (1962) suggests that the ability of a two-state character to discriminate between two taxa can be measured by the algebraic difference ($G$) in the probabilities of positive states for the taxa. The method of Section VIII based on separation of pairs of taxa could employ $G$; a character would be considered to separate a taxon pair if the absolute value of $G$ exceeded a threshold level. The consequences of doing this rather than using thresholds to convert probabilities to +, − and "d" have not been investigated.

These same procedures are applicable to multistate characters if these are entered in the matrix as a number of component two-state characters (Part B and Section VIII).

A second approach to selecting characters from a probability matrix is to calculate the probabilities of the taxa on particular sets of character states assuming the states are independent within taxa and using Bayes' theorem (Part B). The current situation at any stage of an identification can be represented by the probabilities of the taxa on the character states known up to that stage. Observing an additional character will change the probabilities of the taxa. If the value of a given situation can be measured, the expected utility of the additional character can be obtained as the average value of the new situations. For a two-state

character

$$U = P(+/S_0)\ V(S_+) + P(-/S_0)\ V(S_-) \tag{28}$$

where $U$ is the expected utility of the character, $P(+/S_0)$ is the probability of a positive state of the character given the current situation $S_0$, $V(S_+)$ is the value of the situation $S_+$ resulting when a positive state of the character is added to the character states already known, and similarly $P(-/S_0)$ and $V(S_-)$ for a negative state. $P(+/S_0)$ can be found as $\sum_J P_0(J)\ P(+/J)$ where $P_0(J)$ is the probability of taxon $J$ on the states already known and $P(+/J)$ is the probability of a positive state of the character for taxon $J$, i.e. the appropriate matrix entry. Good (1970) suggests a number of possible measures for $V(S)$, for example the negative entropy

$$V(S) = \sum_J P(J)\ log\ P(J) \tag{29}$$

where $P(J)$ is the probability of taxon $J$ in situation $S$.

The above procedure can be extended to multistate characters and, in theory, to sets of characters. To evaluate a set of characters each possible pattern of character states must be considered and the probability of the pattern and the value of the resulting situation found. For $m$ characters there will be $2^m$ patterns and the amount of computation can become too great for routine use (Willcox et al., 1973).

## X. Methods Using Full Data Matrices

Very little experience has been reported with methods of character selection for full data matrices, containing the character states of sample individuals of each taxon.

For two-state characters, Gyllenberg (1963) calculates the value of chi-square for each character in each taxon, comparing the observed numbers of individuals giving each state with an equal distribution of states. Only character states with a frequency significantly different from $\frac{1}{2}$ are entered in the non-probabilistic matrix used in character selection (Section VIII). Ross (1975) also uses a chi-square calculation, for multistate as well as two-state characters, comparing the observed frequency of each state of a character in a taxon with the frequency over all the taxa. For quantitative characters, Ross (1975) calculates the $t$-statistic in a similar way but he points out that it is not statistically valid to use chi-square and $t$ values in tests of significance if the taxa have been formed by a numerical classification based on the same data matrix. However the values are useful guides to the characteristic features of each taxon.

Some discriminant analysis techniques can be used to select a subset of charac-

ters. The analysis is carried out in steps, selecting at each step the character which adds most to the discrimination between the taxa measured as the ratio of between-taxon variance to within-taxon variance. Spicer, Jones and Jones (1973) applied this method to medical diagnosis. Hills (1967) gives a step-by-step character selection procedure used with the nearest-neighbour identification method (Part B).

If the taxa are based on a numerical classification, the classification method itself may indicate discriminatory characters (Sneath and Sokal, 1973, p. 385). Hill and Silvestri (1962) give a method for assessing the contribution of each character to the formation of taxa in a numerical classification (the "taxonomic significance" of the character). Hill and Silvestri (1962) applied this method and independently constructed an identification key from the same bacterial data and so compared the taxonomic significance of the characters with their use in identification. The same group of characters (i.e. physiological–biochemical rather than morphological) were more important in both instances, but there was no close correspondence for individual characters.

# PART D. THE APPLICATION OF NUMERICAL METHODS TO IDENTIFICATION: PROSPECTS AND LIMITATIONS

## XI. Introduction

In comparison with classification, much less work has been done on developing numerical methods for identification. Numerical methods have been applied to different identification problems in different fields of biology but little has been done to compare methods theoretically and still less to compare results given by different methods applied to the same problem, reviewed by Sneath and Sokal, 1973 (chapter 8); Hill, 1974; Pankhurst, 1974; Morse, 1975.

Some characteristic features of different schemes for identifying specimens are considered in Section XII. The problems of evaluating and predicting the performance of an identification scheme are discussed in Section XIII together with the related problem of assessing the performance required by the users of a scheme. Some of the practical considerations which seem to limit the application of numerical methods to identification are discussed in Section XIV. Section XV describes the use of numerical codes in particular identification schemes.

## XII. Identification Methods

### *a. Types of identification method*

Identification methods can be divided into types according to how the identification is carried out (Table 7). The first distinction is between intuitive (or subjective) and automatic methods. Intuitive methods rely on the judgement of the scientist. Sometimes a specimen is immediately recognized by an expert without any conscious mental process. The expert may compare the specimen with named specimens, illustrations or printed descriptions. Determination by an expert is still regarded as the most reliable of all identification methods. In bacteriology, diagnostic tables giving the test results expected for various taxa are often printed without any definite rules for matching the results of the unknown strain against the table. The use of such tables is to some extent subjective since

Table 7. Types of identification method

| | | |
|---|---|---|
| *Intuitive* | – | immediate recognition, comparison, expert determination |
| *Automatic* | | |
| Printed schemes | – | keys, diagnostic tables (with matching or elimination rules), numerical indices |
| Mechanical | – | various devices to aid matching or elimination, punched card schemes |
| Computer/numerical | – | require use of computer for each identification |

often the results of the unknown do not match exactly just one of the taxa so the "best taxon" must be assessed by the bacteriologist. A similar situation can arise in keys with more than one character at each stage (Sneath and Sokal 1973, at p. 391). Intuitive methods are not necessarily completely reproducible, as different scientists may reach different conclusions on the same evidence or even the same scientist at different times. However, expert determination is usually taken as the final opinion and is the standard against which other methods are assessed.

Automatic methods will always give the same identification when considering the same evidence. (An alternative name for such methods is algorithmic; a defined procedure or algorithm is followed.) Automatic methods can be divided into three types according to how they are implemented. Printed schemes include keys, diagnostic tables (if a definite procedure is adopted for matching the results of the unknown with the entries in the table) and numerical codes which are becoming quite widely used in bacteriology in conjunction with commercially-produced testing kits (see Section XV).

A second category of automatic identification methods comprises the use of mechanical devices. Cowan (1974) gives examples of several devices designed either to assist in matching the character states of the unknown organism against the entries in a diagnostic table or to reach an identification by successively eliminating taxa if their expected character states disagree with those observed. None of these devices have been widely adopted. Another mechanical approach is the use of punched cards. These can be edge-punched card systems which use one card for each taxon; the cards are sorted, usually with the aid of a needle, to extract the taxon cards whose expected character states are compatible with the states observed. Feature card ("peek-a-boo") systems similarly use one card for each character state with holes punched corresponding to the taxa which can show this state. As the characteristics of the unknown are observed, the appropriate cards are layered on top of each other and the positions of any holes which remain unobscured indicate the taxa which remain as possible identifications, Morse (1971) and Cowan (1974) refer to several such systems and they are usually implementations of the multiple-entry key method of identification, i.e. the identification is carried out sequentially as in a key but the sequence of characters used is determined by the user.

Another category of automatic methods comprises those which use a computer. These are the methods normally implied by the term "numerical identification". Of course, simple numerical methods can be applied without the use of a computer and some printed schemes involve simple calculations, but numerical identification methods usually means those which are impracticable to use without a computer.

Although automatic methods in following a defined procedure will always produce the same result when considering the same evidence, in actual use the scientist will often exercise a degree of judgement. In the computer-assisted identification service (Part A) the automatic procedure is controlled and the

results are assessed by the bacteriologists responsible for the identifications. It is desirable that an automatic identification method should encourage this expert contribution.

Computers have been used to construct identification keys and to choose sets of characters for diagnostic tables (Part C), and to generate tables for numerical code identification schemes (Robertson and MacLowry, 1975). Programs have been described which produce sets of punched cards forming multiple-entry keys (Pankhurst and Aitchison, 1975*a*). In these applications a computer is used to construct the identification method but is not required for each specimen identified. Computers can also be used to construct schemes which are themselves computerised. Gyllenberg and Niemelä (1975*a*, *b*) described the use of a computer to form and evaluate the groups to be used in a numerical identification method and Darland (1975) described the use of computer-calculated discriminant functions in numerical identification.

*b. Strategies of identification*

The diagram in Fig. 13 represents the general process of identification divided into a series of steps. Different identification methods can be characterized in terms of this diagram according to the overall strategy of how they carry out the individual steps of the process.

To identify a specimen some of its properties must be observed and the first step is to decide which characters should be determined at the outset. The next step is to observe the states of these characters shown by the specimen. The character states are next analysed to give some expression of the current situation (a step called "inference" in the terminology of Gorry, 1968). The current situation might be expressed in simple terms, such as taxa 1, 2 and 3 are possible identifications, other taxa have been eliminated, or in a more complex numerical representation. An identification decision is then made, is it possible to identify the specimen on the characters observed so far? (Gyllenberg and Niemelä, 1975*a*, include in their identification decision the choice between categories such as "intermediate", "neighbour" and "outlier" for specimens which cannot be unequivocably identified.) If not, the identification may be continued by observing further characters and the next step is then to select these. The final step is to review the results of the character selection to decide whether to continue. The character selection may indicate that none of the remaining characters will advance the identification or are so difficult to observe that it is not worthwhile continuing. If this is so, the identification is ended with such conclusions as can be drawn from the current situation. Alternatively the decision would be to observe the states of the selected characters and start another cycle of the identification process.

Sneath (1969) distinguishes two basic strategies for identification: sequential and simultaneous. In sequential methods, exemplified by keys, single characters are considered in a sequence. In simultaneous methods, such as diagnostic tables, a number of characters are observed and their results evaluated simultaneously. In

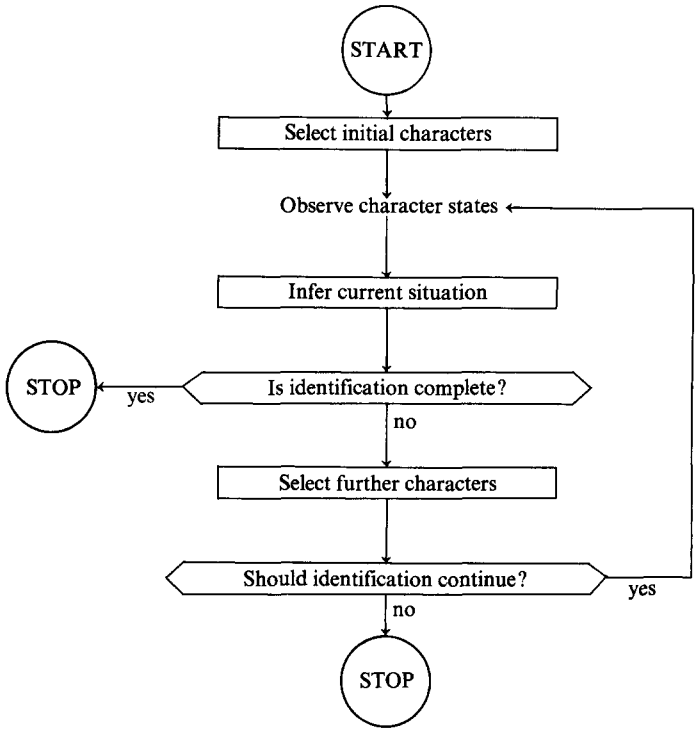



Fig. 13. A representation of the process of identification. Defined procedures for carrying out the boxed elements form an identification scheme; together with a method of observing the character states they form an identification system.

the representation of Fig. 13 each cycle of the process can be called a stage and sequential then means an identification which is carried out in a number of cycles of the loop. The term simultaneous refers to the simultaneous observation and analysis of a number of characters at a particular stage. The two strategies are often used together of course, for a key may contain more than one character at each stage and diagnostic tables may be arranged in several stages. Only in a key using just one character at each stage is the strategy purely sequential and only in a single-stage method using a fixed set of tests is the strategy purely simultaneous.

Identification methods can be compared in terms of the number of stages used and the number of characters used at each stage. The more stages the fewer characters needed altogether, because at each stage the best characters can be selected to continue the identification. On the other hand, if the observation of the characters involves a time delay, the number of stages must be restricted if the total time to identify a specimen is not to become too long. Although the sequential and simultaneous strategies are often used together, most identification schemes tend to one extreme or the other. Keys use many stages with one or two characters at

each stage whilst diagnostic tables use few stages with many tests at each stage. This is obviously conditioned by the different requirements for botanical and zoological work, where keys are widely used, and for microbiology where diagnostic tables have been more successful. In microbiology the time delay involved in carrying out most tests limits the number of stages. In botany and zoology, immediately observable morphological characters are more usual so there is no objection to a large number of stages.

Computer identification methods show the same division between methods using many stages for botanical applications (e.g. Morse, 1971; Pankhurst and Aitchison, 1975*b*) and those using one or a few stages for microbiology (e.g. Gyllenberg and Niemelä, 1975*a*; Part A). Some numerical methods of inference are not suitable for sequential operation which may require a numerical test selection method. Finally, computer methods using many stages require "on-line" access to the computer but for methods using a few stages with a time delay between stages computer access is required only periodically.

*c. Elements of identification methods*

Identification methods can also be compared according to how they carry out the separate steps shown in Fig. 13. A defined way of carrying out each step can be called an element of an identification scheme. Three elements can be considered; character selection, inference, and decision.

Most printed schemes such as keys and diagnostic tables have a fixed *character selection element*. The characters to be used at each stage are printed in the key or table and must be determined to continue the identification. This fixed character selection can be a disadvantage if one of the characters concerned is unobservable because the specimen is damaged or in a particular state of development. Methods have been used which are similar in operation to keys but which allow the user a completely free choice of the sequence of characters. These are multiple-entry keys and they have been produced in printed form and also as punched card systems. Computer methods which operate sequentially also allow a free choice of characters and some assist the user in this choice by recommending characters chosen numerically (see Part C). Other computer methods require a fixed set of tests to be applied to each specimen which, especially in bacteriology, is advantageous if the tests are well chosen. Testing procedures in the laboratory can be adapted to these tests and the use of test "kits" or automated testing may be possible. If the method is completely rigid in this requirement, however, identification may be delayed if one test result is missing on a particular specimen because of a technical failure. Thus at least three types of character selection can be distinguished: *fixed* character selection (a given sequence of characters or set of tests), *free* character selection (free choice of characters permitted or minor deviations from a given set of characters tolerated), and *computer-assisted* character selection (results of a numerical evaluation of the characters are available to guide the selection). Morse (1971, 1975) uses the term "polyclave" for any method which does not have fixed

character selection but other authors apply polyclave only to multiple-entry keys.

The *inference element* of identification is the evaluation of the observed characters to obtain some statement of the current position in the identification. The terms sequential and simultaneous can also be used to distinguish different types of inference. In sequential inference, some taxa are eliminated completely at some stages in the identification. The term simultaneous inference describes the simultaneous comparison of a number of character states with the definitions of a number of taxa. Again, the two types of inference are often used together. An identification key uses sequential inference, for in moving to a subset of the key those taxa not in the subset are eliminated. Diagnostic tables involve simultaneous inference, as the results of a number of tests on the specimen are matched against the entries in the table. If such tables are arranged in stages then sequential inference is used as well. Of computer identification methods, some use sequential inference (Morse, 1975, calls these elimination methods) whereas others consider all the taxa throughout the identification. In most methods, especially keys and multi-stage tables, the sequence of inference follows the sequence of the overall identification process. Some computer methods, however, operate sequentially without using sequential inference (e.g. Part A, where the results of the first characters observed may favour one taxon but as further results are considered another taxon may finally be given as the identity).

Another distinction in methods of inference is between monothetic and polythetic inference (Sneath and Sokal, 1973, Chapter 8; Morse, 1975). In monothetic inference a single character state can be sufficient to exclude an organism from a particular taxon whatever the outcome of other characters. In polythetic methods, if the unknown resembles a particular taxon over many characters one or two discrepant character states may be accepted. This does not mean that the characters are not given different weights but no single character is given an overwhelming weight. Many identification keys are monothetic, a single aberrant character or error will lead to mis-identification. Diagnostic tables can be polythetic, as aberrant specimens can often be identified by finding the best match. In many tables however some pairs of taxa only differ in a single character and for these taxa the table is monothetic. In computer identification, inference is usually polythetic but in probability methods special precautions may be needed to avoid monothetic inference (Part B). Different types of numerical inference are described in Section XII.*d*).

The inference used in identifying a specimen with a key is set by the inference method used in constructing the key. The key itself merely determines the outcome of this inference method for a particular specimen. Sequential monothetic inference has almost invariably been used in constructing keys but other methods could be used (Part C).

The *decision element* in keys and diagnostic tables is made automatically on reaching a terminal or final match. The scientist may not necessarily accept the automatic decision, if he thinks it is suspect. Alternatively, before reaching a termi-

Table 8. Types of data and identification matrices

| | | |
|---|---|---|
| Full matrix | – | character states for individuals of each taxon |
| Centroid matrix | – | co-ordinates of centroid of each taxon |
| Probability matrix | – | probabilities of character states for each taxon |
| Non-probabilistic matrix | – | character states for each taxon |

nal or final match an identification to genus level, say, may be sufficiently precise for the particular specimen. The characters necessary to complete the identification may be inconvenient and, again, the user may decide not to continue. Computer methods usually indicate when sufficient evidence has accumulated to identify the specimen. If these methods use numerical character selection they will also indicate when none of the additional characters are likely to further the identification. Again, the scientist may choose to override the automatic decisions (e.g. Part A).

How the automatic identification decision is made depends on the type of inference used. If it is by elimination, an identification is indicated when only one taxon remains. Numerical methods of inference require more complex decision rules and different methods using the same inference element may differ in their decision elements (Part B).

*d. Types of data and identification matrices*

The term identification matrix has been applied to both the information available for constructing a method and the information actually used by the method (Sneath and Sokal, 1973, chapter 8; Gyllenberg and Niemelä, 1975*a*). It is clearer to distinguish between these two kinds of matrix: the information available for constructing a method is the *data matrix* and the information used by the method is the *identification matrix*. The data matrix must be assembled before constructing a scheme and, for computer methods, the identification matrix is stored in the computer and used by the identification program.

Several types of data and identification matrices can be distinguished as shown in Table 8. A full matrix contains for each taxon the character states of a number of individual members of the taxon. A centroid matrix gives the co-ordinates of the centroid of each taxon in either the original, or transformed, space. Unless otherwise specified, a centroid matrix refers to original space and each co-ordinate is then the average value for members of the taxon in that character. A probability matrix contains for each taxon and character state an estimate of the probability that a member of the taxon will show that character state. For two-state characters the probability of only one of the states need be entered. A nonprobabilistic matrix gives for each taxon and character the expected character state for the taxon which are those shown by all or most of the members of a taxon, other characters are recorded as "variable" or "d".

Table 9. Types of data and identification matrices required by different identification methods

| Method | Minimum data matrix | Identification matrix |
|---|---|---|
| Numerical classification (Oliver and Colwell, 1974) | Full | Full |
| Similarity with individuals (Ross, 1975) | Full | Full |
| Exact match with individuals – numerical indices (Section XV) | Full[1] | Full (printed) |
| Discriminant analysis (Sneath and Sokal, 1973; Darland, 1975) | Full | Discriminant function coefficients |
| Taxon-radius in transformed space (Gyllenberg, 1965) | Full | Centroid in transformed space |
| Taxon-radius (Sneath, 1974; Gyllenberg and Niemelä, 1975*a*, *b*) | Centroid | Centroid |
| Likelihoods assuming independent characters (Part B) | Probability | Probability |
| Similarity with taxa (Campbell, 1975; Pankhurst, 1975*a*) | Non-probabilistic | Non-probabilistic |
| Taxon elimination (Morse, 1975) | Non-probabilistic | Non-probabilistic |
| Keys (Part C) | Non-probabilistic (printed scheme) | Not applicable |
| Multiple-entry keys – punched cards (Pankhurst and Aitchison, 1975*a*) | Non-probabilistic | Not applicable (punched card scheme) |
| Diagnostic tables (Part C) | Non-probabilistic | Non-probabilistic (printed) |

[1] Can be produced by simulation from other types of data matrix.

Only in a full matrix are the character states of sample individuals available, other matrices contain summaries of the properties of the taxa. For two-state characters, the centroid and probability matrices are identical (Gyllenberg and Niemelä, 1975*a*). However, continuous quantitative characters are readily represented in a centroid matrix but not in a probability matrix, whereas for multistate qualitative characters the reverse is true (Part B).

Identification methods can be characterised by the type of data matrix required to construct the method and by the type of identification matrix used by the method (Table 9). The range of methods available for a particular problem depends on the type of identification matrix available. All the methods are available with a full identification matrix because it can always be reduced to one of the other types. However, if it is not possible to obtain a full data matrix, one of the other forms of data matrix must then be used. As these require only a summary of the properties of each taxon it can be assembled from the literature or incomplete records. If the information available is sufficient the properties can be expressed numerically in a centroid or probability matrix, otherwise only qualitatively in a non-probabilistic matrix. Practical considerations also restrict the choices. A full identification matrix will usually be larger than other types of matrix thus demanding more computer store and processing time. For methods

such as discriminant analysis or the taxon-radius method in transformed space (both constructed from a full data matrix though they do not use a full identification matrix), it may be difficult to adjust the method to allow for new findings. Whenever a property needs to be changed, or a new taxon added, the identification matrix must be computed again from the data matrix. Finally, in data matrices the correlation of characters within taxa is lost so any methods based on data matrices cannot take account of these correlations.

The different methods in Table 9 will only be outlined briefly here. The first method is to carry out a numerical classification of reference and unknown specimens and identify the unknowns according to the reference specimens with which they are clustered; the many methods of numerical classification are then available for identification. This approach is obviously more suited to research studies (Oliver and Colwell, 1974) but for routine use the amount of computation required makes it impracticable. Somewhat similar is to calculate the similarities between each unknown and all the reference specimens but not forming a full similarity matrix nor carrying out cluster analysis. Ross (1975) describes a computer program which incorporates this identification method.

Another method using a full identification matrix is to seek an exact match between the character states of the unknown and those of one of the reference specimens. Provided a fixed set of characters, not exceeding 20 or so, is applied to all specimens this can be done without a computer by using a numerical code identification scheme (see also Section XV). The index of code numbers is a full identification matrix as each number represents the particular pattern of a reference specimen. It is not always clear from the description of such schemes whether the entries in the index refer to actually observed patterns or to patterns theoretically derived from some other form of data matrix. Patterns can be easily generated from non-probabilistic matrices by allowing for all possible combinations of states for the "variable" entries and it can be done for probability matrices too (Young, 1975). For this particular identification method, a full identification matrix of simulated specimens can be generated from another form of data matrix. Rypka (1975) describes the use of the exact matching identification method in a computer program and considers the selection of characters and collection of data for this method.

In discriminant analysis methods computation is carried out once on the full data matrix to obtain for each taxon the coefficients of the discriminant functions. These coefficients form the identification matrix. Sneath and Sokal (1973) refer to a number of applications of this powerful method but point out that application to qualitative characters can cause computational difficulties. It has been used in microbiology in identifying bacteria on their antibiotic susceptibility, expressed quantitatively by measuring the zones of inhibition (Darland, 1975) or by light-scattering measurements (Sielaff, Johnson and Matsen, 1976).

The taxon-radius methods use a geometrical model where each taxon is defined by a central point (centroid) and a radius. If the distance between the unknown and

the centroid of a taxon is less than the radius of the taxon the unknown is assigned to that taxon. Gyllenberg (1965) describes this method in a transformed space obtained by principal component analysis of the full data matrix. Gyllenberg and Niemelä (1975*a*, *b*) apply the model in original space; they set two radii for each taxon, the outer radius defines unknowns as "neighbours" of the taxon, and permits the recognition of "intermediates" belonging to two or more taxa. Sneath (1974) examines some properties of these methods and in particular the effect of reducing the number of available characters and the effect of test errors. When the taxon-radius model is applied in the original space only a centroid data matrix is required. The method then assumes that there is no correlation between the characters within taxa, as it treats the taxa as hyperspheres, and it is comparable with the calculation of likelihoods assuming independent characters (Part B). Sneath and Sokal (1973, Chapter 8) point out the analogy between the taxon-radius model in transformed space and discriminant analysis; discriminant analysis effectively produces a transformed space in which the taxa are as nearly hyperspherical as possible.

Widely used in the identification of bacteria is to calculate the likelihoods of the taxa, assuming the characters are independent within taxa. The likelihood of a taxon on a set of character states is defined as the probability of the character states for the taxon and is easily calculated from a probability matrix. Methods using this approach differ in the way they treat the likelihoods in making an identification decision (see Part B).

The entries in a non-probabilistic matrix are the character states of the taxa and the similarity of the unknown with each taxon can be measured by similarity coefficients. Any "variable" entries in the matrix are treated as "no comparison". This has been used by Campbell (1975) and Pankhurst (1975*a*) in bacteriology and botany respectively. Another method applied to non-probabilistic matrices by Morse (1971) is to eliminate a taxon if the states of the unknown differ in more than a set number of characters. This approach is suitable for sequential identification.

Most work on computer construction of keys and diagnostic tables, (see Part C) has involved non-probabilistic data matrices. Pankhurst and Aitchison (1975*a*) describe a computer method for producing punched card multiple-entry keys from such matrices. Keys, including most multiple-entry keys, are usually based on a monothetic taxon elimination method of inference. Diagnostic tables often have no definite rules for their use; if rules are required either the similarity with taxa or taxon elimination approach can be used.

## XIII. Identification Performance

### *a. Evaluating the performance of identification methods*

Identification is essentially a practical exercise, the aims of which can be simply stated. "The objects of any identification scheme are ease and certainty of

identification. All other considerations are secondary" (Sneath and Sokal, 1973, p. 383). Identification schemes should be evaluated in those terms, especially if different methods are to be compared. Commercial identification systems for bacteria are now available, each system comprising a testing "kit" and its own identification scheme. The growing literature on kit evaluation and comparison (e.g. Nord, Lindberg and Dahlbäck, 1974) has brought out some of the general problems of evaluating identification methods.

The obvious way to evaluate performance is to carry out a trial with a number of organisms which should be different from the reference organisms used to construct the scheme, but separately identified by an independent method, and should be representative of the organisms for which the scheme has been designed. It is often difficult to meet all these requirements. Because of the general problem of assembling sufficient data to construct a scheme, it may be difficult to set aside a proportion of the data for use in trials. Establishing the 'true' identity of the trial organisms may also be difficult. In bacteriology it is sometimes possible to confirm an identification based on biochemical tests by, say serology, but often the only independent identification is an assessment by an expert of much the same information as used by the identification method under trial. Reference strains can be used as their identity will be well-established but they may not be representative of isolates encountered in clinical laboratories either in the proportions of strains of the various taxa or in their reactions.

With a trial completed, an established "true" identity of the organisms and the identification obtained by the method under trial, the results must be assessed. Some organisms may be identified to the incorrect taxon, obviously a failure, but some misidentifications may be more serious than others, in terms of the *degree* of misidentification. Sneath (1974) suggests two possible bases for measuring this; firstly the phenetic discrepancy involved in the misidentification (the taxonomic significance of the error) and, secondly, the cost to the user. In bacteriology the different costs of misidentifications can be easily appreciated but it may be difficult to assign numerical values to them. The identification of a strain of *Salmonella* to the wrong subgenus might not be considered important if the identification is checked by serology, but misidentification of a *Salmonella typhi* strain as *Citrobacter freundii* would be very serious.

The situation is more complicated if the method returns a "not identified" decision. It is less serious to fail to identify an organism than to identify it incorrectly but a measurement of the cost difference would complete the evaluation. Methods which allow several grades of identification are yet more complex. Finally there may be organisms in a trial which should be recorded as unidentifiable, namely those of taxa not included in the scheme.

The difficulties in evaluating the performance of a particular method obviously affect the comparison of different methods. One method might identify a higher proportion of organisms but make more incorrect identifications; different methods may identify to different taxonomic levels so while one method identifies

a high proportion of organisms but only to genus level, another identifies a lower proportion but to species level. Performance can also be altered by changing the values of the parameters of the method. To compare different numerical methods it may be necessary to distinguish the different elements of the methods shown in Fig. 13. One method may give a better performance because of a superior method of inference, or a better set of characters, or a more realistic identification matrix. To compare different methods of numerical inference, as such, the other elements would have to be removed.

The results of a trial thus present a complex picture (Lapage, 1974). Where different methods are available (e.g. commercial kit systems for Enterobacteriaceae), the development of a single "index of performance" following the lines suggested by Sneath (1974), would be very useful.

*b. Predicting the performance of an identification method*

How performance can be predicted, and adjusted, can be illustrated by the construction of a diagnostic table, using numerical character selection as reviewed in Part C. A parameter of the selection procedure is the minimum number of characters required to discriminate between each pair of the taxa. If this number is set to one, then some pairs of taxa will differ in only one character: an aberrant result would lead to mis-identification. Setting the number to two prevents an organism aberrant in one character being misidentified, but the organism could match two taxa equally well. If at least three character differences are required, an organism aberrant in only one character can always be identified by its best match. The performance of the table, expressed in this rather inexact way, can thus be increased by changing that one parameter.

In the method of Lapage et al. (1973), the identification decision depends on a single parameter, the "identification level". A trial on 1028 strains of bacteria showed the expected effect that as the identification level was increased fewer incorrect identifications were made but more strains were unidentified (Fig. 14). In routine use, an identification level of 0.999 was adopted to obtain a low rate of misidentifications. In other circumstances a lower level may be preferable to identify a higher proportion of strains at the expense of more mistakes. The trial of this method also showed that the proportion of strains identified varied from taxon to taxon. Much of this variation could be accounted for by considering how well the probability matrix discriminated between the taxa (Bascomb et al. 1973; Willcox and Lapage, 1975). It is possible to predict for any matrix which particular taxa are not sufficiently distinguished from the other taxa to ensure a high identification rate.

Another aspect of predicting performance is the estimation of the effect of errors. Sneath (1974) distinguishes two types of errors, (a) errors in the reference descriptions of taxa and (b) errors in the character states of the unknown organisms. Sneath considers the effects of different types and rates of error and in particular gives a numerical evaluation of the effect of type (b) errors on taxon-

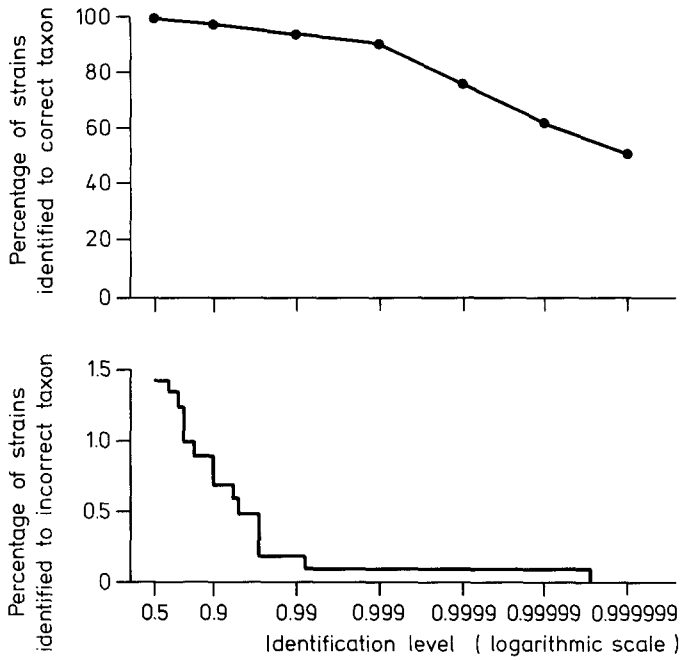



Fig. 14. Effect of changing the value of the identification decision parameter (identification level) on the computer identification of 1028 reference strains (Lapage et al., 1973).

radius identification methods.

These examples show how performance can be predicted approximately and how the expected performance of some methods can be changed according to user-requirements. A more exact approach may be possible through statistical discrimination theory (Gower, 1975). Given numerical estimates of the prior probabilities of the taxa and the costs of misidentifications, the theory provides rules of identification which minimize the expected cost of mis-identifications.

*c. Assessing the performance required of an identification method*

Evaluation of performance and the design of identification methods to have a particular performance both depend on a statement of the costs of misidentifications to the user. Evaluation and prediction of identification performance can only be useful if the performance is related to the needs of the user.

Costs can be visualised in a general way but difficult to quantify. If a method is to be used in a well-defined situation, statements about the performance required can be quite specific. For example, in phage-typing *Pseudomonas aeruginosa* a biochemical screening procedure may be needed to exclude other species of *Pseudomonas*. It could be specified that very few strains of other species are identified as *P. aeruginosa* (say a maximum 1 in 1000). Strains of taxa other than *P.*

*aeruginosa* need not be identified to species, but only excluded, and a proportion of *P. aeruginosa* strains falsely rejected (say 1 in 50) could be tolerated. Given such specifications and prior probabilities an identification method could be designed to have the required performance yet use as few biochemical tests as possible. In contrast, the laboratory providing a reference identification service may have little knowledge of the ultimate use of the identifications and it will be difficult to specify the performance required.

To use the costs of misidentification, estimates of the prior probabilities are required. The prior probabilities can be incorporated into present methods of numerical identification and effectively weight against the rarer taxa (see Part B). Whether this is desirable depends on the type of work being carried out. In some situations it might not be desirable because the rarer taxa are precisely those it is hoped to detect. In other work it may be more important to identify correctly the highest overall proportion of organisms and the occasional misidentification of a rare taxon is unimportant. In terms of different costs of misidentification, in the first case a high cost is placed on failing to identify a member of a rare taxon, but in the second all misidentifications are given about the same cost and the method aims at identifying the most frequently occurring organisms. The majority of applications of numerical identification have been in reference situations, where prior probabilities and the costs of misidentification are difficult to estimate. If numerical methods are applied to more routine situations prior probabilities and costs should be taken into account.

The problems of measuring performance and assessing the performance required are unresolved and limitations on the use of numerical methods make it difficult to compare different methods.

## XIV. Practical Limitations

Experience with numerical identification methods (e.g. Lapage et al., 1973; Pankhurst, 1975*a*) shows that an accurate and complete data matrix is required before the methods become useful. A data matrix can be compiled from a number of sources (e.g. the literature, laboratory records). Alternatively, specimen organisms can be obtained and examined afresh and the results used to form the matrix. Pankhurst (1975*b*) describes the difficulties of both these approaches, in a botanical context. In using the literature, descriptive terms were not used consistently so it was difficult to define consistent character states. There were discrepancies in the character states ascribed to a particular taxon even within the same publication and about 30% of the data matrix remained uncompleted. By examining herbarium specimens afresh the data matrix could be completed and some reduction of the inconsistencies achieved but there was a difficulty in obtaining sufficient suitable specimens. A major practical limitation in botany seems to be the difficulty of constructing a data matrix. This problem indicates a

need for further theoretical work in determining just how much data is required and how many specimens of each taxon need be examined to obtain a reliable data matrix.

In microbiology there is less difficulty in accumulating data by the second approach. Reasonably large numbers of strains can be obtained for most species from culture collections, reference laboratories or field isolates and can be fully tested in one laboratory, with standard methods, to obtain the results for the data matrix. Lapage et al. (1973) found that a data matrix constructed in this way was more effective than construction from the literature. The practical limitations on numerical methods in microbiology lie in carrying out the identifications. Routine laboratories cannot carry out a sufficient number of tests to make numerical methods worthwhile except for a few isolates. Furthermore, differences in media composition and methods between laboratories means that results obtained in routine laboratories are not always suitable for centralised analysis. Commercial testing kits (e.g. Nord et al., 1974) and the development of automated testing methods (e.g. Hedén, 1975) should allow more extensive testing of strains with methods standardised between laboratories. More use of numerical methods would then be possible. The methods of Darland (1975) and Sielaff et al. (1976) for the presumptive identification of bacteria on their antibiotic susceptibility illustrate another approach to increasing the contribution of numerical methods to routine medical microbiology.

A second practical aspect is access to a computer, different numerical methods have different computer requirements. The computer may be used once only to construct, say, a key (often the case in botany and zoology), but other methods require 'on-line' computer use for each identification. The third type of method, met with in microbiology, requires only periodic use to analyse the results of a set of tests read simultaneously. This sort of computer access is becoming quite feasible for many laboratories using simple terminals linked to a central computer through the public telephone network. Another development is the increasing power of 'micro-computers' which could mean that these may become capable of numerical identification work.

The results of numerical identification can be made widely available without access to a computer through numerical coding schemes (Section XV). If a fixed set of not more than 20 or so characters is observed on all organisms, the results can be converted to a numerical code which is looked up in an index. The entries in the index give the results of a numerical identification method applied to the pattern of character states represented by each code number. The results of numerical identification for several hundred of the most commonly expected character patterns are then available in printed form (see, for example API, 1977).

Turning to the prospective role of numerical identification methods, irrespective of their theoretical and practical limitations, Morse (1975) concludes that most biologists have little need of numerical identification in day-to-day work with organisms which are their special interest. The main contribution in botany and

zoology is likely to be the production of keys and similar identification aids. Pankhurst (1975*b*) has constructed about 50 keys by computer, but considers (Pankhurst, 1974) that computer constructed keys are not yet as effective as the best keys produced by taxonomists. However, computer methods have made a valuable contribution in rapidly producing reliable keys either for groups for which no key was previously available, or for specialised keys for particular ecological situations, seasons of the year, and so on. Morse (1975) also suggests that numerical methods might be useful for identifying types of specimen which are often regarded as unidentifiable, e.g. incomplete, sterile or immature specimens.

In routine diagnostic microbiology probably most isolates will continue to be identified by highly specialised procedures using very small sets of biochemical tests, selective media, serology and similar techniques. The reliability of these procedures, often relying on presumptive identification based on the source and clinical information, has not been systematically investigated. Some isolates will occur which cannot be identified by these procedures and to identify these reliably without a disproportionate amount of labour is likely to require the use of testing kits or automated testing and computer analysis. The potential clinical value of identifying such strains has been demonstrated (e.g. Holmes, Lapage and Malnick, 1975).

## XV. Numerical Code Identification Schemes

In numerical code identification schemes a fixed set of characters is observed on each unknown organism and the results are converted to a numerical code. The code number is looked up in an index which gives the appropriate identification or additional characters to complete the identification. If a particular code is not in the index the scheme cannot identify that organism. These schemes all relate the code numbers to conventional taxonomic names in contrast to proposals for replacing taxonomic nomenclature with numerical coding systems, reviewed by Sneath and Sokal (1973, p. 412).

Fey (1959) used seven, two-state, biochemical tests to identify gram-negative bacteria. Each test is allotted a score, 5, 10, 20 and so on and the code number for a pattern of test results is obtained by adding the scores of the positive tests, as shown in Table 10. Baer and Washington (1972) developed Fey's scheme using the same coding method while Dito et al. (1972) used a similar coding method in a 10-test scheme for identifying Enterobacteriaceae with test scores of 1, 2, 4, 8 etc. The

Table 10. The numerical coding scheme of Fey (1959)

| Test results | + | − | + | + | + | − | + |
|---|---|---|---|---|---|---|---|
| Test scores | 320 | 160 | 80 | 40 | 20 | 10 | 5 |
| Numerical code | 320 | | +80 | +40 | +20 | | +5 = 465 |

Table 11. Codes used by Cowan (1965) for blocks of three tests

| Test results | Number | Test results | Number |
|---|---|---|---|
| − − − | 0 | + + − | 4 |
| + − − | 1 | + − + | 5 |
| − + − | 2 | − + + | 6 |
| − − + | 3 | + + + | 7 |

Table 12. The API numerical coding scheme (API, 1977)

| Test results | + − + | + + − | − − − | − − + | + + + | − + − | + − − | |
|---|---|---|---|---|---|---|---|---|
| Test scores | 1 2 4 | 1 2 4 | 1 2 4 | 1 2 4 | 1 2 4 | 1 2 4 | 1 2 4 | |
| Numerical code | 5 | 3 | 0 | 4 | 7 | 2 | 1 | i.e. 5304721 |

identification scheme provided with the Enterotube kit of 11 tests (ENCISE, 1973) uses this same coding method.

Cowan (1965) divided tests into blocks of three and each block of results is represented by one digit in the final code number. Cowan allotted digits 0 to 7 to the eight possible combinations of results, as in Table 11. Farmer (1970) used a similar method in a scheme for representing bacteriocin and bacteriophage reaction patterns but a different way of numbering the eight patterns.

The coding system used in conjunction with the API 20E test kit (API, 1977) also divides the tests into blocks of three but the digit representing each block of results is obtained by adding the scores of the positive results in the block, as shown in Table 12.

Most of these coding schemes have been described without any theoretical basis. A pattern of results of two-state tests becomes a binary number if one (for plus) and zero (for minus) are written. Dito et al. (1972) give this as the basis of their method which converts this binary number to a decimal number. The test scores they use are the decimal values of successive binary digits, see Table 13. As the number of binary digits increases, the conversion from binary to decimal becomes

Table 13. Converting test results to decimal and octal numbers

| Test results | + | − | + | + | + | − | + | |
|---|---|---|---|---|---|---|---|---|
| Binary number | 1 | 0 | 1 | 1 | 1 | 0 | 1 | |
| Decimal value of binary digits | 64 | 32 | 16 | 8 | 4 | 2 | 1 | |
| Decimal number | 64 | | +16 | + 8 | + 4 | | + 1 = 93 | |
| | | | | | | | | |
| Binary number | 1 | 0 | 1 | 1 | 1 | 0 | 1 | |
| Octal value of binary digits | 1 | 4 | 2 | 1 | 4 | 2 | 1 | |
| Octal number | 1 | 2 + 1 = 3 | | | 4+1 = 5 | | | i.e.135 |

rather laborious and a common technique in computer programming is to convert to an octal (base 8) number. Successive blocks of three binary digits convert to one octal digit, using the values shown in Table 13. A reasonable theoretical basis for numerical coding schemes is then the conversion of unwieldy binary numbers to more convenient decimal or octal numbers. Provided the number of tests does not exceed 10 or so conversion to decimal is not difficult but conversion to octal is always easier. The API method (API, 1977) is similar to the octal number method, except the test scores in each block are reversed which should not affect the ease of using the scheme. Some aspects of the production of the indices for numerical code identification schemes are discussed in Sections XII and XIV above.

*Received* 11 *March* 1980

## References

API 1977. Analytical Profile Index: Enterobacteriaceae. — Analytab Products, New York.

Baer, H. and Washington, L. 1972. Numerical diagnostic key for the identification of Enterobacteriaceae. — Appl. Microbiol. **23**: 108–112.

Bascomb, S., Lapage, S. P., Curtis, M. A. and Willcox, W. R. 1973. Identification of bacteria by computer: identification of reference strains. — J. Gen. Microbiol. **77**: 291–315.

Brown, P. J. 1977. Functions for selecting tests in diagnostic key construction. — Biometrika **64**: 589–596.

Campbell, I. 1975. Numerical analysis and computerized identification of the yeast genera *Candida* and *Torulopsis*. — J. Gen. Microbiol. **90**: 125–132.

Cowan, S. T. 1965. Development of coding schemes for microbial taxonomy. — Adv. Appl. Microbiol. **7**: 139–167.

Cowan, S. T. 1974. Cowan and Steel's manual for the identification of medical bacteria. 2nd Ed. — Cambridge Univ. Press, London.

Croft, D. J. 1972. Is computerized diagnosis possible? — Comput. Biomed. Res. **5**: 351–367.

Dallwitz, M. J. 1974. A flexible computer program for generating identification keys. — Syst. Zool. **23**: 50–57.

Darland, G. 1975. Discriminant analysis of antibiotic susceptibility as a means of bacterial identification. — J. Clin. Microbiol. **2**: 391–396.

Davies, P. 1972. Symptom diagnosis using Bahadur's distribution. — Int. J. Bio-Med. Comput. **3**: 307–312.

Dickey, J. M. 1968. Estimation of disease probabilities conditioned on symptom variables. — Math. Biosci. **3**: 249–265.

Dito, W. R., Bulmash, J., Campbell, J. and Roberts, E. 1972. A numerical coding and identification system for the Enterobacteriaceae. — American Society of Clinical Pathologists Commission on Continuing Education, Chicago.

Dybowski, W. and Franklin, D. A. 1968. Conditional probability and the identification of bacteria: a pilot study. — J. Gen. Microbiol. **54**: 215–229.

Encise1973. Enterotube numerical coding and identification scheme for Enterobacteriaceae. — Hoffman-LaRoche, Basle.

Farmer, J. J. 1970. Mnemonic for reporting bacteriocin and bacteriophage types. — Lancet **1970, 2**: 96.

Fey, H. 1959. Differenzierungsschema für gramnegative aerobe Stäbchen. — Schweiz. Z. Allg. Pathol. Bacteriol. **22**: 641–652.

FRIEDMAN, R. B. and MACLOWRY, J. 1973. Computer identification of bacteria on the basis of their antibiotic susceptibility patterns. — Appl. Microbiol. **26**: 314–317.

FRIEDMAN, R. B., BRUCE, D., MACLOWRY, J. and BRENNER, V. 1973. Computer-assisted identification of bacteria. — Am. J. Clin. pathol. **60**: 395–403.

GILBERT, E. S. 1968. On discrimination using qualitative variables. — J. Am. Stat. Assoc. **63**: 1399–1412.

GOOD, I. J. 1970. Some statistical methods in machine intelligence research. — Math. Biosci. **6**: 185–208.

GORRY, G. A. 1968. Strategies for computer-aided diagnosis. — Mathematical Biosciences **2**: 293–318.

GOWER, J. C. 1975. Relating classification to identification. p. 251–263. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

GOWER, J. C. and BARNETT, J. A. 1971. Selecting tests in diagnostic keys with unknown responses. — Nature **232**: 491–493.

GOWER, J. C. and PAYNE, R. W. 1975. A comparison of different criteria for selecting binary tests in diagnostic keys. — Biometrika **62**: 665–672.

GYLLENBERG, H. G. 1963. A general method for deriving determination schemes for random collections of microbial isolates. — Ann. Acad. Sci. Fenn. Ser. A, IV. **69**: 1–23.

GYLLENBERG, H. G. 1965. A model for computer identification of micro-organisms. – J. Gen. Microbiol. **39**: 401–405.

GYLLENBERG, H. G. and NIEMELÄ, T. K. 1975*a*. Basic principles in computer-assisted identification of microorganisms. p. 201–223. *In* C.-G. Hedén and T. Illeni (eds), New approaches to the identification of microorganisms. — John Wiley, New York.

GYLLENBERG, H. G. and NIEMELÄ, T. K. 1975*b*. New approaches to automatic identification of micro-organisms, p. 121–136. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

HALL, A. V. 1975. A system for automatic key-forming. p. 55–63. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

HEDÉN, C.-G. 1975. The modular approach to the automation of the microbiological routines. p. 13–37. *In* C.-G. Hedén and T. Illeni (eds), New approaches to the identification of micro-organisms. — John Wiley, New York.

HILL, L. R. 1974. Theoretical aspects of numerical identification. — Int. J. Syst. Bacteriol. **24**: 494–499.

HILL, L. R. and SILVESTRI, L. G. 1962. Quantitative methods in the systematics of Actinomycetales. III. The taxonomic significance of physiological–biochemical characters and the construction of a diagnostic key. — G. Microbiol. **10**: 1–28.

HILLS, M. 1967. Discrimination and allocation with discrete data. — Applied Statistics **16**: 237–250.

HOLMES, B., LAPAGE, S. P. and MALNICK, H. 1975. Strains of *Pseudomonas putrefaciens* isolated from clinical material. — J. Clin. Pathol. **28**: 149–155.

KENDALL, M. G. and STUART, A. 1963. The advanced theory of statistics. Vol. 1. — Griffin, London.

LAPAGE, S. P 1974. Practical aspects of probabilistic identification of bacteria. — Int. J. Syst. Bacteriol. **24**: 500–507.

LAPAGE, S. P., BASCOMB, S., WILLCOX, W. R. and CURTIS, M. A. 1970. Computer identification of bacteria. p. 1–22. *In* A. Baillie and R. J. Gilbert (eds), Automation, mechanization and data handling in microbiology. Soc. Appl. Bacteriol., Techn. Ser. No. 4. — Academic Press, London.

LAPAGE, S. P., BASCOMB, S., WILLCOX, W. R. and CURTIS, M. A. 1973. Identification of bacteria by computer: general aspects and perspectives. — J. Gen. Microbiol. **77**: 273–290.

LEDLEY, R. S. and LUSTED, L. B. 1959. Reasoning foundations of medical diagnosis. — Science, New York **130**: 9–21.

MACCACARO, G. A. 1958. La misura della informazione contenuta nei criteri di classificazione. — Ann. Microbiol. Enzimol. **8**: 231–239.

MÖLLER, F. 1962. Quantitative methods in the systematics of the Actinomycetales. IV. The theory and application of a probabilistic identification key. — G. Microbiol. **10**: 29–47.

MORSE, L. E. 1971. Specimen identification and key construction with time-sharing computers. — Taxon **20**: 269–282.

MORSE, L. E. 1975. Recent advances in the theory and practice of biological specimen identification. p. 11–52. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

NIEMELÄ, S. I., HOPKINS, J. W. and QUADLING, C. 1968. Selecting an economical binary test battery for a set of microbial cultures. — Can. J. Microbiol. **14**: 271–279.

NORD, C.-E., LINDBERG, A. A. and DAHLBÄCK, A. 1974. Evaluation of five test-kits – API, AuxoTab, Enterotube, PathoTec and R/B for the identification of Enterobacteriaceae. — Med. Microbiol. Immunol. **159**: 211–220.

OLIVER, J. D. and COLWELL, R. R. 1974. Computer program designed to follow fluctuations in microbial populations and its application in a study of Chesapeake Bay microflora. — Appl. Microbiol. **28**: 185–192.

PANKHURST, R. J. 1970. A computer program for generating diagnostic keys. — Comput. J. **13**: 145–151.

PANKHURST, R. J. 1974. Automated identification in systematics. — Taxon **23**: 45–51.

PANKHURST, R. J. 1975*a*. Identification by matching. p. 79–91. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

PANKHURST, R. J. 1975*b*. Identification methods and the quality of taxonomic descriptions. p. 237–247. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

PANKHURST, R. J. and AITCHISON, R. R. 1975*a*. A computer program to construct polyclaves. p. 73–78. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

PANKHURST, R. J. and AITCHISON, R. R. 1975*b*. An on-line identification program. p. 181–194. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

PAYNE, R. W. 1975. Genkey: a program for constructing diagnostic keys. p. 65–72. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

RAO, C. R. 1952. Advanced statistical methods in biometric research. — Wiley, New York.

ROBERTSON, E. A. and MACLOWRY, J. D. 1974. Mathematical analysis of the API enteric 20 profile register using a computer diagnostic model. — Appl. Microbiol. **28**: 691–695.

ROBERTSON, E. A. and MACLOWRY, J. D. 1975. Construction of an interpretive pattern directory for the API 10S kit and analysis of its diagnostic accuracy. — J. Clin. Microbiol. **1**: 515–520.

ROSS, G. J. S. 1975. Rapid techniques for automatic identification. p. 93–102. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

RYPKA, E. W. 1975. Pattern recognition and microbial identification. p. 153–180. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

RYPKA, E. W., CLAPPER, W. E., BOWEN, I. G. and BABB, R. 1967. A model for the identification of bacteria. — J. Gen. Microbiol. **46**: 407–424.

SEBESTYEN, G. S. 1962. Decision-making processes in pattern recognition. — Macmillan, New York.

SIELAFF, B. H., JOHNSON, E. A. and MATSEN, J. M. 1976. Computer-assisted bacterial identification utilizing antimicrobial susceptibility profiles generated by Autobac 1. — J. Clin. Microbiol. **3**: 105–109.

SNEATH, P. H. A. 1962. The construction of taxonomic groups. p. 289–332. *In* G. C. Ainsworth and P. H. A. Sneath (eds), Microbial classification. Twelfth Symposium of the Society for General Microbiology. — Cambridge Univ. Press, Cambridge.

SNEATH, P. H. A. 1969. Computers in bacteriology. — J. Clin. Pathol. **22**, suppl. 3: 87–92.

SNEATH, P. H. A. 1974. Test reproducibility in relation to identification. — Int. J. Syst. Bacteriol. **24**: 508–523.

SNEATH, P. H. A. and COLLINS, V. G. 1974. A study in test reproducibility between laboratories: Report of a *Pseudomonas* working party. — Antonie van Leeuwenhoek **40**: 481–527.

SNEATH, P. H. A. and JOHNSON, R. 1972. The influence on numerical taxonomic similarities of errors in microbiological tests. — J. Gen. Microbiol. **72**: 377–392.

SNEATH, P. H. A. and SOKAL, R. R. 1973. Numerical taxonomy. — W. H. Freeman, San Francisco.

SPICER, C. C., JONES, J. H. and JONES, J. E. L. 1973. Discriminant and Bayes analysis in the differential diagnosis of Crohns disease and proctocolitis. — Methods Inf. Med. **12**: 118–122.

VICTOR, N., TRAMPISCH, H. J. and ZENTGRAF, R. 1974. Diagnostic rules for qualitative variables with interactions. — Methods Inf. Med. **13**: 184–186.

WATSON, L. and MILNE, P. 1972. A flexible system for automatic generation of special-purpose dichotomous keys, and its application to Australian grass genera. — Aust. J. Bot. **20**: 331–352.

WILLCOX, W. R. and LAPAGE, S. P. 1972. Automatic construction of diagnostic tables. — Comput. J. **15**: 263–267.

WILLCOX, W. R. and LAPAGE, S. P. 1975. Methods used in a program for computer-aided identification of bacteria. p. 103–119. *In* R. J. Pankhurst (ed.), Biological identification with computers. Syst. Assoc. Spec. Vol. No. 7. — Academic Press, London.

WILLCOX, W. R., LAPAGE, S. P., BASCOMB, S. and CURTIS, M. A. 1973. Identification of bacteria by computer: theory and programming. — J. Gen. Microbiol. **77**: 317–330.

YANKELEVITCH, G. and NEGRETE-MARTINEZ, J. 1969. El uso del contenido de informacion de las caracteristicas en la identificacion taxonomica automatizada. — Bol. Estud. Med. Biol. (Mexico City). **26**: 73–79.

YOUNG, W. D. 1975. Computer simulation of bacterial cultures. — Int. J. Syst. Bacteriol. **25**: 143–149.