Identification of Bacteria by Computer: Theory and Programming

By W. R. WILLCOX, S. P. LAPAGE, SHOSHANA BASCOMB* AND M. A. CURTIS

Computer Trials Department, National Collection of Type Cultures, Central Public Health Laboratory, Colindale Avenue, London, NW9 5HT

(Received 8 November 1972; revised 19 March 1973)

SUMMARY

The methods incorporated in the computer program used in a trial of computer-aided identification of bacteria are described. The identification method is based on Bayes's theorem and allows for dependent tests and missing data in the probability matrix. It was found useful in developing the method to take account of the occurrence of errors in bacteriological testing. The method suggests a definite identification only if the Bayesian probability of one of the taxa exceeds a threshold level; if not, a separate procedure selects the best tests to continue the identification.

INTRODUCTION

Many applications of statistical methods to medical diagnosis have been reported; Anderson & Boyle (1968) give a review. Most of the methods used, given a set of symptoms or test results, estimate the probability of each of the relevant diseases on the basis of records of previous cases. The disease with the highest probability is diagnosed and the value of the probability indicates the reliability of the diagnosis (Ledley & Lusted, 1959). Some methods (e.g. Nugent, Warner, Dunn & Tyler, 1964) only suggest a definite diagnosis if one of the disease probabilities exceeds a threshold level, and others (Anderson, 1968; Gorry, 1968) allow for the relative seriousness or 'cost' of the various possible misdiagnoses. An extension of these methods is to allow sequential diagnosis (Gorry, 1968); a few results are supplied initially and if a definite diagnosis is not indicated on these data a separate procedure selects the most useful tests to continue the diagnosis.

In microbiology, Möller (1962) and Hill & Silvestri (1962) estimated the reliability of a dichotomous key for ten taxa of the Actinomycetales by calculating the probabilities of all the taxa at each terminal of the key. Dybowski & Franklin (1968) described a complete computerized sequential identification system applied to 38 taxa of the Enterobacteriaceae.

This paper describes the methods used in an extended trial of computer-aided identification of Gram-negative rods (Lapage, Bascomb, Willcox & Curtis, 1973). The theory of the identification method adopted is presented. The method is based on that of Ledley & Lusted (1959), which has been widely used in medical diagnosis. Some problems encountered in applying the identification method and the procedures developed to overcome them are described. The method used for selecting tests when identification was not completed on the initial results is also described. The test selection procedure is similar to methods previously described for selecting tests for diagnostic tables (Gyllenberg, 1963; Rypka, Clapper, Bowen & Babb, 1967).

* Present address: Department of Biochemistry, Imperial College, Imperial Institute Road, London, SW7 2AZ.
THEORY OF THE IDENTIFICATION METHOD

If \( q_1, \ldots, q_n \) are a set of exclusive and exhaustive propositions and \( p \) is a further proposition, then Bayes's theorem (Kendall & Stuart, 1963) states

\[
P(q_i | p) = \frac{P(q_i)P(p | q_i)}{\sum_i P(q_i)P(p | q_i)},
\]

(1)

where \( P(q_i | p) \) is the probability of \( q_i \) given \( p \), \( P(q_i) \) is the probability of \( q_i \), and \( P(p | q_i) \) is the probability of \( p \) given \( q_i \). In terms of the identification of bacteria the propositions are concerned with the alternative taxa, \( t_1 \ldots t_n \), to which identification must be made. Strictly, each proposition should be stated in a form such as 'the organism is a member of taxon \( t_i \)' but it is easier to refer loosely to the probability of a taxon. The proposition \( p \) is the set of test results observed, \( R \), and equation (1) becomes

\[
P(t_i | R) = \frac{P(t_i)P(R | t_i)}{\sum_i P(t_i)P(R | t_i)},
\]

(2)

where \( P(t_i | R) \) is the probability that an organism giving test results \( R \) is a member of taxon \( t_i \), the posterior probability of the taxon; \( P(t_i) \) is the probability than an organism is a member of taxon \( t_i \) before considering any test results, the prior probability of the taxon; and \( P(R | t_i) \) is the probability that a member of taxon \( t_i \) will give test results \( R \).

Bayes's theorem as given above is easily derived from the rules of the mathematical theory of probability, but the application of the theorem to practical problems has given rise to much controversy among statisticians. Kendall & Buckland (1957) describe 'probability' as 'a basic concept which may be taken either as undefinable, expressing in some way a "degree of belief", or as the limiting frequency in an infinite random series'. They continue: 'both approaches have their difficulties and the most convenient axiomatization of probability theory is a matter of personal taste. Fortunately both lead to much the same calculus of probabilities'. Thus, although the mathematical form of Bayes's theorem is not controversial, the meaning of 'probability' and the sort of practical situations to which the term can be applied is in dispute.

The difficulty is most clear-cut in the problem of deciding between several scientific hypotheses, the \( q_1, \ldots, q_n \) of equation (1), relevant to an experimental result, \( p \) of equation (1) (Edwards, 1969). Many such hypotheses are expressed in terms of the frequency of events in the long run, giving the \( P(p | q_i) \) of equation (1). The prior probabilities of the hypotheses \( P(q_i) \), on the other hand, can only be considered as degrees of belief; they are subjective probabilities. Edwards (1969) illustrates the spectrum of opinion among statisticians as to the use of subjective probabilities and how their values can be determined. If subjective probabilities are not accepted, Bayes's theorem cannot be used in this problem and a separate 'principle of maximum likelihood' (Kendall & Stuart, 1963) is postulated to make inferences on the basis of the non-controversial \( P(p | q_i) \). The likelihood of a hypothesis given the experimental results is defined as the probability of the results given the hypothesis, \( P(p | q_i) \); then the hypothesis with the maximum likelihood is selected. (Many authors refer to the likelihood of the results given the hypothesis; for the present problem the terminology adopted (Good, 1959) seems clearer.) If the subjective prior probabilities are accepted, but are set equal for all the hypotheses, as is usually done if there is no reason to do otherwise, they cancel out in equation (1), which becomes

\[
P(q_i | p) = \frac{P(p | q_i)}{\sum_i P(p | q_i)} \propto L(q_i | p),
\]

(3)
where \( L(q_i | p) \) is the likelihood of the hypothesis. Thus when the prior probabilities are equal, the posterior probabilities of the hypotheses are proportional to their likelihoods and Bayes's theorem and the principle of maximum likelihood will give the same result as far as accepting one or other of the hypotheses is concerned. However, because Bayes's theorem yields probabilities which can be manipulated according to the rules of the theory of probability, more complicated inferences can be made concerning the hypotheses which are not possible using likelihoods.

The use of Bayes's theorem in deciding between scientific hypotheses was examined here in an attempt to illustrate the controversial aspects of the use of the theorem and its relationship to the principle of maximum likelihood. In medical diagnosis and bacteriological identification the choice between Bayes's theorem and maximum likelihood is not so clear-cut. The identification problem can be stated in a non-controversial way: the unknown organism has been selected at random from a population in which the taxa \( t_1 \ldots t_n \) are present in proportions \( P(t_1) \ldots P(t_n) \) and the test results are a random selection from the possible results for each taxon, the proportions being \( P(R | t_1) \ldots P(R | t_n) \). In this statement of the problem the prior probabilities of the taxa and the conditional probabilities of the test results are seen to be of the same type, as compared with the scientific hypothesis problem in which they were of different types. The difficulty remains of obtaining the values of these probabilities; they must be estimated by sampling the populations involved. It has been recognized (Ledley & Lusted, 1959) that the estimation of the prior probabilities of the taxa (or diseases for medical diagnosis) is likely to be more difficult than the estimation of the probabilities of the test results. The former are likely to vary with time, place, type of laboratory, etc., due to outbreaks, local conditions, interest of participants, etc. The probabilities of different test results in a given taxon could also vary as there could be outbreaks of aberrant strains, but it can be argued that they will be more stable than the prior probabilities. If the difficulties in estimating the prior probabilities are considered much greater than those of estimating the probabilities of the test results their use may be rejected and the principle of maximum likelihood used (Dickey, 1968; Dybowski & Franklin, 1968). Boyle et al. (1966) found that using prior probabilities decreased their success rate because the probabilities were estimated on a sample drawn from a population different from that on which diagnosis was attempted.

In order to use either Bayes's theorem or the principle of maximum likelihood the conditional probabilities of the test results \( R \) given the taxon \( t_i \) must be estimated. Unfortunately in most practical applications they cannot be estimated directly from the records of the sample. To obtain useful estimates of these probabilities directly from the sample data, for any given result pattern \( R \), would require at least a few cases in the sample records of exactly similar result patterns. Since the number of possible result patterns is likely to be very large (e.g. for 20 two-state tests the number would be \( 2^{20} \approx 10^6 \)) it is unlikely that this would be achieved. To estimate the probability of a result pattern it is necessary to make some assumptions about the form of the distribution of the result patterns in the various taxa. The usual assumption is that the individual test results which make up the result pattern \( R \) are independent in each taxon. Now the probability of a number of independent results is the product of their individual probabilities; if the result pattern \( R \) consists of individual test results \( r_1 \ldots r_n \) then

\[
P(R | t_i) = P(r_1, r_2, \ldots r_n | t_i) = P(r_1 | t_i) P(r_2 | t_i) \ldots P(r_n | t_i)
\]

if the \( r_j \) are independent given the taxon \( t_i \). All the members of the sample which belong to the taxon and have been examined in the test are available for estimating the probabilities of the
individual test results, \( P(r_j | t_i) \), so reasonable estimates should be possible. The method of identification/diagnosis using Bayes's theorem with an assumption of independent tests has been termed method \( L \) (Anderson & Boyle, 1968). It has been considered by Dickey (1968) that it is the assumption of independence of tests, rather than the use of Bayes's theorem as such, which is the most serious objection to this method.

As an alternative to assuming the independence of test results, Fisher's linear discriminant function (Kendall & Stuart, 1968) can be used. Here the assumption is that the test results have a multivariate normal distribution with the same covariance matrix in each taxon. This method may be much more realistic if the tests are quantitative in nature, but for qualitative tests method \( L \) seems to be preferred (Anderson & Boyle, 1968). More recently some methods have been proposed which depend on far less restrictive assumptions in estimating the probabilities of the test results (Hills, 1967; Dickey, 1968). The probability of a pattern of test results for each taxon is estimated from the sample data in the neighbourhood of the test results – that is, those sample strains whose results differ from the pattern in question in not more than a given number of tests. At present these methods appear to be still in process of development (Card & Good, 1970).

A problem not so far discussed is the requirement that the \( q_i \) of equation (1) and hence the taxa or diseases in the application of Bayes's theorem are exclusive and exhaustive. In medical diagnosis a frequent problem is that the diseases are not exclusive; a patient may have two or more diseases at the same time. In bacteriological identification the taxa can be made exclusive but in practice it may be difficult to make them exhaustive, i.e. to cover all possibilities, even when the field of application is defined (Lapage et al. 1973).

**APPLICATION OF THE IDENTIFICATION METHOD**

The method adopted for use in the trial of computer-aided identification of bacteria is based on Bayes's theorem, equation (2). The tests are assumed to be independent within each taxon (method \( L \)) and the prior probabilities of the taxa are taken as equal. Equation (2) then becomes

\[
P(t_i | R) = \frac{P(R | t_i)}{\sum_t P(R | t_i)}
\]

(4)

and

\[
P(R | t_i) = P(r_1, r_2, \ldots | t_i) = P(r_1 | t_i) P(r_2 | t_i) \ldots,
\]

(5)

where \( P(t_i | R) \) is the probability of taxon \( t_i \) given the test results \( R \) and \( P(r_j | t_i) \) is the probability of one of the individual results given taxon \( t_i \). Most of the tests which were to be used were of the two-state type and for such tests the probability of a negative result is one minus the probability of a positive result. Thus a matrix giving the probability of a positive result for each of the taxa in each of the tests is all the information required to calculate by equations (4) and (5) the posterior probabilities of all the taxa for any set of test results.

The prior probabilities of the taxa were set equal because at the outset there was little indication as to how the taxa would be represented among the strains submitted for identification by such a scheme. As discussed above the values of the posterior probabilities obtained when the prior probabilities are set equal are proportional to the likelihoods of the taxa which were used for bacterial identification by Dybowski & Franklin (1968). In fact in all documents produced by our system the values were referred to as the 'identification scores' of the taxa and users of the system were not encouraged to think of them as probabilities. However, at some stages in the application of the method it is convenient to regard these scores as probabilities; the points at which this has been done will be indicated.
The construction of the matrix, the basic data required by the method, for the present application is described in detail by Lapage et al. (1973) and Bascomb, Lapage, Curtis & Willcox (1973). As well as records of individual strains information from the literature was used, the final values being a subjective assessment of the available data. The possibility of using information other than complete records of sample strains seems to be an advantage peculiar to method L.

Having found the identification scores for a given set of test results, a definite identification is suggested only if one of the scores exceeds a threshold ‘identification level’; the level currently used is 0.999. Considering the identification scores as probabilities, an identification level of 0.999 should give an error rate, in the long run, of not greater than one in a thousand and this was the basis on which the level was originally set. The results of a trial of the identification method on over 1000 strains (Lapage et al. 1973) confirmed the suitability of this level. A procedure suggested by practical experience and incorporated into the identification method is to allow identification to certain specified composite taxa comprising two or more of the individual taxa, e.g. Escherichia coli/Alkalescens–Dispar. Identification to a composite taxon is suggested if the sum of the identification scores of the component taxa exceeds the identification level, a procedure which treats the scores as probabilities and was justified by the results of the trial.

Some problems which arose when the identification method was applied to a matrix for Gram-negative rods are now described with the procedures which were developed to overcome them.

**Limits on matrix entries.** The entries in the matrix were based on records of sample organisms and reports in the literature. A problem arises when all reported organisms of a certain taxon gave positive (or negative) results in a particular test. If such elements were entered as one (or zero), an organism giving a test result never before observed in a taxon could never be identified as a member of that taxon whatever its reactions in the remaining tests, since in equation (5) \( P(R|t_i) = 0 \) if any \( P(r_j|t_i) = 0 \). This is undesirable because any previous sampling cannot have covered all the possible reactions for a particular taxon and furthermore the test result in question might have resulted from an error in testing. The problem can be avoided by setting upper and lower limits of 0.99 and 0.01 respectively to matrix entries. The initial choice of these limits was to some extent arbitrary but was based on a bacteriological assessment of the weight to be attached to aberrant test results. A more formal basis for the limits is provided by the data error approach (see below), which follows the reasoning that apparently unexpected reactions may be the results of errors in testing.

The problem does not arise if the matrix entries are based entirely on sample results and Laplace’s law of succession is accepted. If in a sample of \( n \) organisms \( m \) are found to be positive in a test, the estimated probability, \( P \), that the next organism will be positive is given by Laplace’s law of succession (Good, 1965) as \( P = (m+1)/(n+2) \). Thus if 100 organisms had all been found positive the appropriate matrix entry would be very nearly 0.99.

The values chosen for the limits on matrix entries are crucial to some aspects of the performance of the identification method. The data errors approach suggests that limits of 0.99 and 0.01 are realistic but the method as a whole can only be justified by the successful results of the trial application (Bascomb et al. 1973).

**Unknown matrix entries.** Any general application of a probabilistic identification method to bacteriology is likely to meet the problem of a complete lack of data on the performance of some taxa in some tests. An analogous problem arises in the treatment of linked tests (see below); the procedure developed is applicable to either case.
Suppose the matrix element required to obtain the probability of an individual test result, $P(r_j | t_i)$ of equation (9), is unknown. In the calculation a value for the element is assumed so as to give the maximum value for $P(r_j | t_i)$. Thus, if the test result in question is positive a value of 0.99 is assumed, if the result is negative a value of 0.01; in either case $P(r_j | t_i) = 0.99$. Values of 0.99 and 0.01 are taken for consistency with the limits on matrix entries discussed above but if these limits were not being applied values of 1 and 0 giving $P(r_j | t_i) = 1$ could be used. If, having applied this procedure to all unknown matrix entries for the given set of test results, the identification score of a taxon exceeds the identification level a tentative identification to this taxon is made. Then if it turns out that there are no unknown matrix entries for this taxon in the tests considered the identification will hold because the values assumed for the unknown entries of the other taxa were the most favourable to those taxa and thus the least favourable to the identifying taxon. If, on the other hand, there are one or more unknown matrix elements in the scoring of the tentatively identified taxon the identification cannot be accepted because values of matrix entries favourable to that taxon have been assumed. The identification scores are then recalculated without assuming values favourable to the taxon (a process termed ‘rescoring’) and the identification is only accepted if the rescore of the same taxon still exceeds the identification level. Two possible strategies are available for rescoring. In the ‘lenient strategy’ the test results for which the matrix entries are unknown for the tentatively identified taxon are simply ignored, for all taxa, in the rescoring, the scoring being based on the remaining test results. In the ‘stringent strategy’ values are assumed for these matrix entries which are the least favourable to the tentatively identified taxon. The stringent strategy requires a previous setting of limits to matrix entries; otherwise the least favourable assumption would give $P(r_j | t_i) = 0$. For unknown matrix entries the lenient strategy would seem to be the most suitable but for the linked tests problem the stringent strategy is applicable.

In the full trial of the identification method there were no unknown matrix entries and rescoring was only required in the treatment of linked tests (see below). The stringent strategy was used and in the trial on over 1000 strains rescoring affected the identification of only one strain.

**Linked tests.** The identification method assumes that the test results are independent given the taxa so that the probability of a set of results is the product of the probabilities of the individual results. For some of the tests in the present matrix it is known that this is not true and the test results are governed by a simple logical relationship valid for all taxa. For example, if a strain does not grow at 37 °C the results for motility, methyl red and Voges–Proskauer at 37 °C must all be negative. These linked tests are allowed for in the calculation of the probabilities of the test results given the taxa.

Equation (5) gives $P(r_1, r_2, \ldots, t_i) = P(r_1 | t_i) P(r_2 | r_1, t_i) \ldots$, valid only if the $r_j$ are independent given $t_i$. Now if results $r_1, r_2, r_3, r_4$, say, are not independent of each other but are independent of the remaining results, then

$$P(r_1, r_2, r_3, r_4, t_i) = P(r_1 | t_i) P(r_2 | r_1, t_i) P(r_3 | r_2, r_1, t_i) P(r_4 | r_3, r_2, r_1, t_i) \ldots$$

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

$$P(r_1, r_2, r_3, r_4 | t_i) = P(r_1 | t_i) P(r_2 | r_1, t_i) P(r_3 | r_2, r_1, t_i) P(r_4 | r_3, r_2, r_1, t_i),$$

where, for example, $P(r_4 | r_1, r_2, r_3, t_i)$ is the probability of result $r_4$ given taxon $t_i$ and having observed results $r_1, r_2$ and $r_3$. For independent tests terms such as $P(r_4 | r_1, r_2, r_3, t_i)$ equal
Theory of identification

$P(r_i | t_i)$ leading to equation (5). For dependent tests this is not so and such terms must be evaluated.

A result $r_j$ is taken as representing one of the two possible results of the test $x_j$, the possible results being positive, $+$, and negative, $-$. The condition ‘given taxon $t_i$’ will be assumed and $t_i$ is omitted from the following equations. The tests $x_1, x_2, x_3, x_4$ are taken to be linked such that a negative result in $x_1$ implies a negative result in $x_2, x_3$ and $x_4$; it follows that a positive result in any of $x_2, x_3$ and $x_4$ implies a positive result in $x_1$.

For the term $P(r_2 | r_1)$, considering all possible test results, it is necessary to find $P(+_2 | +_1)$, $P(-_2 | +_1)$, $P(+_2 | -_1)$ and $P(-_2 | -_1)$. Now a negative result in $x_1$ implies a negative result in $x_2, x_3$ and $x_4$, so

$$P(+_2 | -_1) = 0$$

and

$$P(-_2 | -_1) = 1.$$  

Also

$$P(+_2) = P(+_1, +_2) + P(-_1, +_2),$$

where $P(+_1, +_2)$ is the probability of $+_1$ and $+_2$, equals $P(+_1)P(+_2 | +_1)$, so since $P(+_2 | -_1) = 0$,

$$P(+_2) = P(+_1) P(+_2 | +_1),$$

so

$$P(+_2 | +_1) = \frac{P(+_2)}{P(+_1)},$$

and since

$$P(-_2 | +_1) = 1 - P(+_2 | +_1),$$

$$P(-_2 | +_1) = 1 - \frac{P(+_2)}{P(+_1)}.$$  

$P(+_1)$ and $P(+_2)$ are just the appropriate matrix entries and all the required probabilities can be found from (7), (8) and (9). The results $-_4$, $+_4$ given a zero probability by (6) are detected as ‘impossible’ by an earlier editing procedure.

Terms $P(r_3 | r_1, r_2)$ and $P(r_4 | r_1, r_2, r_3)$ are similarly evaluated.

Since $-_1$ implies $-_3$ and $-_4$

$$P(+_3 | -_1, -_2) = P(+_4 | -_1, -_2, -_3) = 0$$

and

$$P(-_3 | -_1, -_2) = P(-_4 | -_1, -_2, -_3) = 1.$$  

It is assumed that given a positive result in $x_1$, tests $x_2, x_3$ and $x_4$ are independent, so

$$P(+_3 | +_1, r_2) = P(+_3 | +_1) = \frac{P(+_3)}{P(+_1)}$$

and

$$P(+_4 | +_1, r_2, r_3) = P(+_4 | +_1) = \frac{P(+_4)}{P(+_1)}$$

by analogy with (8) and

$$P(-_3 | +_1, r_2) = P(-_3 | +_1) = 1 - \frac{P(+_3)}{P(+_1)}$$

and

$$P(-_4 | +_1, r_2, r_3) = P(-_4 | +_1) = 1 - \frac{P(+_4)}{P(+_1)}$$

by analogy with (9).

The logical linkage examined above was the only one which occurred in the present matrix though any logical expression could be similarly analysed. It was assumed above that all of
the tests \( x_1 \), \( x_2 \), \( x_3 \), \( x_4 \) had been carried out. So long as a result is available for \( x_1 \) any of the other tests can be omitted and the equations will hold. If \( x_1 \) is not carried out the expressions for the conditional probabilities for \( x_3 \) and \( x_4 \) become more complex. The evaluation of these probabilities was avoided in the present application by assuming that \( x_1 \) was positive if two or more of \( x_2 \), \( x_3 \) and \( x_4 \) had been carried out and \( x_1 \) had not. This seemed a reasonable assumption for the tests involved.

Consider the calculation of the probability of the results \( +_1, -_2 \) for linked tests. Using equation (9),

\[
P(+_1, -_2) = P(+_1) P(-_2 | +_1) = P(+_1) \left[ 1 - \frac{P(+_2)}{P(+_1)} \right],
\]

and if \( P(+_2) = P(+_1) \) then \( P(+_1, -_2) = 0 \), which is unacceptable for the reasons given under limits on matrix entries, above. In the present matrix this situation occurs quite frequently; in all cases \( P(+_2) = P(+_1) = 0.01 \) or 0.99. The problem arises because of the approximate nature of the matrix entries, especially for probabilities near one and zero where the limits on matrix entries are applied. Reasonable results can be obtained by assuming that a matrix 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 dealing with unknown matrix elements the most favourable values are first assumed; for \( P(+_2) \) and \( P(+_1) \) entered as 0.99 assume \( P(+_1) = 1 \), \( P(+_2) = 0.99 \), giving \( P(-_2 | +_1) = 0.01 \), and for \( P(+_2) \) and \( P(+_1) \) entered as 0.01 assume \( P(+_1) = 0.01 \), \( P(+_2) = 0 \) giving \( P(-_2 | +_1) = 1 \). Because values for probabilities have been assumed, rescoring will sometimes be necessary. In practice rescoring should be required infrequently because the problem only occurs when unexpected results, \( +_1 \) where \( P(+_1) = 0.01 \) and \( -_2 \) where \( P(+_2) = 0.01 \) are evaluated. The lenient strategy for rescoring is not appropriate here because the results in question are known to be unexpected for the tentatively identified taxon and ignoring them could increase the score of that taxon. On the other hand the stringent strategy is not immediately applicable as the least favourable assumptions give \( P(+_2) = P(+_1) \) again and \( P(-_2 | +_1) = 0 \). The data errors approach (see below) suggests appropriate values for the probabilities in the rescoring of these test results.

**Multistate tests.** For two-state tests the \( P(r_j | t_i) \) of equation (5) can be obtained from a matrix giving the probability of a positive result for each of the taxa in each of the tests; the probability of a negative result is one minus the probability of a positive result. In general terms, for an \( n \)-state test, either \( n - 1 \) probabilities can be held in the matrix for each taxon, the probability of the \( n \)th result being calculated as required, or \( n \) probabilities can be held and referenced directly. For two-state tests the first approach halves the amount of computer store required and gives a more compact representation of the matrix outside the computer. For multistate \( (n > 2) \) tests the proportional saving of store is less and more computation is necessary to obtain the \( n \)th probability.

Because the majority of the tests in the present matrix are two-state the program uses the first approach but the second approach is adopted for multistate tests by the following procedure. An \( n \)-state test \( (n > 2) \) is represented in the matrix by \( n \) component tests. Any result of the test is presented as a positive result in the appropriate component test, the other component tests being scored 'not done'. Then the same procedure used for the probability calculation for the results of two-state tests will obtain the correct probability for multistate results.

**Data errors approach.** In the application of method \( L \) to the identification of bacteria it was found necessary to set a lower limit to the \( P(r_j | t_i) \) of equation (5) to prevent the exclusion of a taxon on the basis of one unexpected result. This was originally carried out by setting
upper and lower limits to the entries of the matrix from which the \( P(r | t) \) were obtained. Even with these limits zero probabilities were returned for some result patterns in linked tests. The data errors approach provides a unified basis from which methods for dealing with these problems can be derived. In particular it justifies the setting of limits to matrix entries and gives values for the probabilities in linked tests where the formulae developed returned values of zero.

It is known that errors can occur in the execution and reading of bacteriological tests and in transcribing results. It is assumed that any test result considered has a certain probability of being incorrect, \( P_m \), and this probability is taken as 0.01 for all tests. If \( P(+) \) is the 'true' probability of a positive result in a given test for a given taxon, and \( P_a(+) \) and \( P_a(-) \) are the equivalent apparent probabilities of positive and negative results allowing for errors in test results, then

\[
P_a(+) = P(+) (1 - P_m) + P(-) P_m
\]

for \( P_m = 0.01 \)

\[
P_a(+) = 0.98 \ P(+) + 0.01
\]

and

\[
P_a(-) = (1 - P(+) ) (1 - P_m) + P(+) P_m,
\]

for \( P_m = 0.01 \)

\[
P_a(-) = 0.99 - 0.98 \ P(+).
\]

If the formulae (10) and (11) were used in the identification method and \( P(+) \) obtained from the matrix would be no need to set limits to matrix entries since for \( P(+) = 0 \), \( P_a(+) = 0.01 \). The matrix entries are taken as estimates of the true probabilities \( P(+) \); the error rate \( P_m \) is taken to apply only to the results of strains to be identified; the results on which the matrix entries were based are assumed to be free of errors. For reasons discussed below formulae (10) and (11) were not adopted but it can be shown that the results obtained by these formulae differ only slightly from those given by the original method of setting limits of 0.99 and 0.01 to matrix entries. The original method can be considered as giving values \( P_t(+) \), where \( P_t(+) = P(+) \) for \( 0.99 > P(+) > 0.01 \), \( P_t(+) = 0.99 \) for \( P(+) \geq 0.99 \), \( P_t(+) = 0.01 \) for \( P(+) \leq 0.01 \); and \( P_t(-) \) which equals \( 1 - P_t(+) \). Then, for example: at \( P(+) = 0.5 \), \( P_t(+) = 0.5 \); at \( P(+) = 0.05 \), \( P_t(+) = 0.05 \), \( P_a(+) = 0.059 \); at \( P(+) = 0.01 \), \( P_t(+) = 0.01 \), \( P_a(+) = 0.0198 \); and at \( P(+) = 0 \), \( P_t(+) = 0.01 \), \( P_a(+) = 0.01 \). The greatest discrepancy between \( P(+) \) and \( P_a(+) \), a factor of about 2 at \( P(+) = 0.01 \), is negligible considering the approximate nature of the matrix entries as estimates of \( P(+) \).

However, the setting of limits to matrix entries has the useful effect that when a probability is estimated to be above 0.99 or below 0.01 a more precise estimation is not required; this avoids the difficulties of estimating probabilities near to one and zero. It follows that the true probabilities of such test results are to some extent unknown. A matrix entry of 0.99 represents a probability, \( P(+) \), of between 1 and 0.99 and an entry of 0.01 a probability between 0.01 and 0. Following the approach adopted above for unknown matrix entries, values giving the highest \( P(r | t) \) are first assumed and then, for the tentatively identified taxon, values giving the lowest \( P(r | t) \) are taken. For example, if \( 1 \geq P(+) \geq 0.99 \): \( P_a(+)_{\text{max}} = 0.99 \) with \( P(+) = 1 \); \( P_a(+)_{\text{min}} = 0.9802 \) with \( P(+) = 0.99 \); \( P_a(-)_{\text{max}} = 0.0198 \) with \( P(+) = 0.99 \); \( P_a(-)_{\text{min}} = 0.01 \) with \( P(+) = 1 \). The differences between the maximum and minimum probabilities is clearly negligible for \( P_a(+) \) and the factor of about 2 for \( P_a(-) \) is considered to be negligible in this application. The rescoring of a tentatively identified taxon is not then required.

The above approach can be applied to the calculation of the probabilities of result patterns in linked tests. For each result pattern \( R \) an expression for \( P_a(R) \) in terms of \( P(+) \), \( P(+,-) \), etc., is found and compared with the expressions derived above. The effect of 0.99
and 0.01 matrix entries is considered by obtaining expressions for $P_a(R)_{\text{max}}$ and $P_a(R)_{\text{min}}$ for various situations involving such entries. When this is carried out for the linked tests occurring in the present matrix, up to four tests linked so that a negative result in the first implies a negative result in the remainder, it is found that: (i) a reasonable approximation, within a factor of four, to $P_a(R)$ and where appropriate $P_a(R)_{\text{max}}$ is always given by the previously developed straightforward method; (ii) the difference between $P_a(R)_{\text{max}}$ and $P_a(R)_{\text{min}}$ is always negligible unless an unexpected result, positive where the matrix entry is 0.01 or negative where the entry is 0.99, is considered. In these cases the ratio of $P_a(R)_{\text{max}}$ to $P_a(R)_{\text{min}}$ can be as high as $10^6$ and rescoring is required. (iii) When rescoring is required a reasonable approximation to $P_a(R)_{\text{min}}$ is given by $P_a(R') \times (0.01)^n$, where $R'$ is the result pattern excluding the unexpected results and $n$ is the number of such results.

In summary the standard method together with a simple procedure for rescoring unexpected results in linked tests provides reasonable approximations to the probabilities given by the data errors approach with probability of error of 0.01. It might be thought that if the data errors model was accepted the formulae derived from the model should be used in the identification method. However, there are objections to the approach and it seems reasonable to use it merely to suggest subsidiary procedures for situations in which the straightforward approach breaks down.

The main objection to the data errors approach is the difficulty of estimating the error rate, $P_m$. The use of a single value of $P_m$ for all tests and for both results obtained in the laboratory which constructed the matrix and results submitted by other laboratories is not entirely realistic. Sneath & Johnson (1972) found an average $P_m$ of about 2% with significant differences in reproducibility between the 63 tests in their study. If this error rate is assumed for testing in our laboratory, the rate of about 8% found for test result differences between our laboratory and those which send strains (Lapage et al. 1973) suggests a $P_m$ of about 6% for the sending laboratories. This higher value of $P_m$ does not imply that the testing within these laboratories is any less reproducible as it includes the effects of differences in media and testing methods between these laboratories and ours.

**TEST SELECTION**

If none of the scores of the taxa exceeds 0.999 the identification method does not suggest a definite identification. In such cases it is required to select, from among the available tests not already carried out, the most useful tests with which to continue the identification. An important factor in the design of the test selection method is the way in which the complete system – identification scores calculation followed if necessary by test selection and a further identification scores calculation and so on – is to be used. If only one test is suggested at each stage the fewest possible tests should be required for an identification because the situation is then reconsidered after each test result and the best test given the situation is selected. This procedure would be acceptable in fields in which the properties of a specimen can be observed without delay and provided continuous access to a computer was available, but the overall time required to identify a bacterial strain in this way would become very long. In bacteriology it seems preferable to present an ordered list of tests such that if all the tests listed were carried out a definite identification would probably result, but if only $n$ tests were required then the first $n$ of the list would be a useful choice. The bacteriologist could then choose to carry out all the suggested tests to obtain a rapid identification, or a few of them in the expectation of using fewer tests altogether after re-evaluating the results.

The current situation $C_0$ at any stage of an identification is represented by the probabilities
Theory of identification

of the taxa at that stage $P(t_1), P(t_2), \ldots$. The usefulness of a test $x_j$, not already carried out, is to be assessed. The probability of a positive result in $x_j$ given $C_0$ is $P(+j | C_0) = P(t_1) P(+j | t_1) + P(t_2) P(+j | t_2) + \ldots$. The addition of the result $+j$ will alter the probabilities of the taxa giving a new situation $C_1$ and similarly for a negative result $P(-j | C_0)$ and the resulting situation $C_2$ can be found. Now if a measure of the value of a given situation is available, $V(C)$, the expected utility of the test $x_j$ is $P(+j | C_0) V(C_1) + P(-j | C_0) V(C_2)$ and the test with the greatest expected utility can be selected. Gorry (1968) proceeded in this way using as $V(C)$ the reverse of the expected cost of accepting as final the diagnosis with the highest probability at $C$; all possible misdiagnoses had previously been allotted costs according to their seriousness. This might be a useful measure if the user of the system were able to force a definite diagnosis at any time, but Hills (1967) suggested that such measures do not lead to efficient test sequences. Good (1970) suggests several possible measures; for example, the negative entropy

$$V(C) = \sum_i P(t_i) \log P(t_i).$$

If more than one test is to be selected the above procedure can be followed, but now, for $m$ tests, there are $2^m$ patterns of results and the probability of each pattern and the value of the situation given by the pattern have to be found. If the number of tests required is fixed, all possible combinations of $m$ tests from those available can be evaluated but in practice the number of combinations becomes impossibly high. Alternatively, tests can be selected sequentially; first the test with the highest value alone is chosen, then the test which together with the first gives the highest combined value and so on. This reduces the amount of computation and has the advantage that an ordered list of tests is produced as suggested above. Even with sequential test selection, however, the amount of computation required can be very large. For, say, 10 tests there are $2^{10} = 1024$ possible result patterns and the complete identification scores calculation must be carried out on each pattern to obtain and evaluate the situations $C_1, C_2, \ldots C_{1024}$.

To apply test selection in the present problem and as a regular service, a method is required which involves much less computation than the general method outlined above. Two simplifying approximations are made: firstly instead of considering all the taxa and their current scores some of the taxa are taken as ‘possibles’ and are given equal weight when selecting tests, the remainder are ignored; secondly matrix entries are reduced from probabilities to a $+$ and $-$ form.

The possible taxa are found by taking taxa in order of their scores until the sum of the scores exceeds $0.999$. Treating the scores as probabilities, the true taxon would be among the possibles in $99.9\%$ of cases. A subsidiary matrix is then constructed of possible taxa by remaining tests. Entries in this matrix are derived from the original matrix by entering $+$ (always positive) for probabilities $\geq 85\%$, $-$ (always negative) for probabilities $\leq 15\%$ and $o$ (variable) for other probabilities.

Methods previously developed for constructing diagnostic tables are available for analysing the subsidiary matrix. Diagnostic tables usually require that every pair of taxa is separated, that is there is at least one test in the table in which one of the pair is $+$ and the other $-$ . The methods attempt to find a set of tests which meets this requirement with the fewest tests. Willcox & Lapage (1972) described a non-sequential method which always constructs a table with the fewest possible tests but for the present application sequential methods are preferable as mentioned above. Sequential methods described by Gyllenberg (1963), Rypka et al. (1967) and Niemelä, Hopkins & Quadling (1968) choose tests one at a time, at each choice maximising some function of the resulting, partially complete, diagnostic table. Gyllenberg and
Table 1. Hypothetical entries in a subsidiary matrix for the component tests of a 3-state test: (a) examples of possible patterns, (b) the patterns as entered to allow the evaluation of the multistate test as a whole

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Tests</th>
<th>(a)</th>
<th>Tests</th>
<th>(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+</td>
<td>a</td>
<td>1</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>b</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>c</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>a</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>c</td>
<td>5</td>
<td>0</td>
</tr>
</tbody>
</table>

Rypka et al. simply maximize the total number of taxon pairs separated by the partial table, Niemelä et al. use a more complex function based on information theory, and some of the functions used for producing keys (Pankhurst, 1970; Gower & Barnett, 1971; Morse, 1971) might also be applicable.

It was decided to adopt Gyllenberg's method of test selection. This method is not affected by the presence of 0 entries in the subsidiary matrix and it allows the 'identification performance' (Willcox & Lapage, 1972) of the diagnostic table to be increased if required by specifying that more than one test is required to separate a pair of taxa. In the present application, although tests are selected as though for a diagnostic table, identification is carried out not by the logical matching process used in such tables but by a probability calculation. A consideration of the identification calculation suggests that, with an identification level of 0.999, if two tests are required to separate each taxon pair there is a reasonable chance that the discrimination assumed by the test selection model will be achieved.

The basic test selection procedure consists of the steps:

1. Take taxa as 'possibles' in order of their identification scores until the sum of the scores exceeds 0.999.
2. Take as available tests those not already carried out but exclude those tests whose results can be predicted from results already known, e.g. exclude motility at 37 °C if growth at 37 °C already found negative.
3. Form a subsidiary matrix of possible taxa by available tests, entries in the matrix are + if the original matrix probability \( \geq 0.85 \), – if \( \leq 0.15 \), 0 otherwise.
4. Form a taxon by taxon separation matrix and set all its elements to zero.
5. For each test, count the number of + and – entries for the test in the subsidiary matrix, \( n_+ \) and \( n_- \). Set the value of the test \( =n_+ \times n_- \). This is just a convenient way of finding how many pairs of taxa each test separates initially.
6. Find the test with the highest value, if more than one test has this value take the first encountered. If all tests have zero value the set is complete, test selection stops. If not, the test with the highest value, test X, is added to the list of tests selected; the value of X is set to zero.
7. For all pairs of taxa not already completely separated, find out whether test X separates the pair: that is, has an entry + in one taxon and – in the other. If so add one to the appropriate element of the separation matrix. If the separation matrix shows that there are now two tests separating this pair the pair is completely separated; inspect each test not already suggested and if the test also separates the pair reduce its value by one.
8. Return to step (6).
Theory of identification

An addition to the basic procedure allows for multistate tests. As mentioned above, these tests are represented by a set of component tests, one for each state. From the nature of these component tests the matrix entries for a given taxon must consist of just one + and the remainder −, or one or more 0 with or without some − (Table 1a). At stage 3 of the test selection these tests are entered so that, for each taxon, once a + entry is made the remaining entries are all 0 (Table 1b). The resulting entries are such that no more than one of the component tests can separate a given taxon pair. At step 6, before finding the test with the highest value, the values of the components of each multistate test are added together to give the value of the multistate test itself. If a multistate test is suggested, step 7 is repeated with each component test in turn as X.

We wish to thank the Public Health Laboratory Service for help and support, and the Department of Health and Social Security for the grant which made the work possible.

REFERENCES


