Test Reproducibility in Relation to Identification

P. H. A. SNEATH

Medical Research Council Microbial Systematics Unit, University of Leicester, Leicester, England

Test errors are much greater than usually thought and their investigation requires good statistical design. Analysis of variance of blind randomized trials is especially valuable. For representative selections of tests currently used in bacterial taxonomy and identification, the discrepancies within one laboratory are usually less than 4%; with care they can be reduced to 2% or less. Between-laboratory variation is much greater; discrepancies of about 8% are usual under routine conditions, and even 15% is common. The effect of errors on identification is due to (a) error in the reference descriptions of taxa, (b) error in the description of the unknown strain to be identified. Simultaneous polythetic methods (e.g., using matching coefficients or analogues of these) are robust to both types of error, and misidentifications are usually minor in degree, i.e., the unknown is allocated to a taxon close to the correct one. The rate of failure is expected to be greatest when the reference descriptions are based on the tightest clusters. Sequential identification methods are especially sensitive to errors because the unknown may then be allocated to a taxon far from the correct one. Monothetic sequential keys are sensitive to errors of type (b), whereas type (a) error is usually screened out during construction of the key.

Until recent years the accuracy and reproducibility of tests used in microbiology had received little attention, although a few reports did give a good deal of attention to this when new testing methods were first devised. For example, Williams and Rippon (41) included a main section on this theme in their early work on staphylococcal phage-typing. But lately, with the current move toward closer standardization of testing methods of various kinds, there has been renewed interest in this and in the quality control of microbiological techniques (e.g., see 3, 10, 35), a trend illustrated also by a symposium at this Congress on standardization of antibiotic sensitivity tests. This trend will be accelerated by pressures toward better descriptions in systematic bacteriology and toward automation and computer identification in diagnostic microbiology. It is therefore appropriate that study should be given to the effect of erroneous results on the processes of microbial identification.

The science of identification has not advanced very far, nor has the examination of errors in relation to numerical methods of classification. We have at present, therefore, rather little experience to draw upon, and in particular there has been almost no work on the statistical aspects of test reproducibility and error. It is thus necessary to extrapolate from rather slim foundations, but I hope that this will at least help to outline the areas where future work is particularly needed.

It is important, in discussing the extent of test reproducibility, and the sources of variation, to emphasize that in most microbiological work the problem is closer to one of quality control than to an absolute standard of scientific accuracy. With almost all tests the sensitivity of the technique has a pronounced influence on the results, so that the common plan of positive and negative controls is seldom of much help: in practice it is weakly reacting strains that often give trouble, rather than complete failure of the method. There is also a need to get away from viewing a test as positive or negative in absolute terms, because we can seldom be sure whether a strain entirely lacks a metabolic activity. When histograms of reactions or zone sizes are made, it is found that many tests do not show the sharply bimodal behavior that may be expected (Fig. 1). Sometimes the slow development of positive results over time introduces difficulties (39). Again, it is easy to become confused over what is being observed. Microbiological tests are empirical, so that detectable acidity from glucose, for example, is not equivalent to ability to metabolize glucose.

Recent experiences of several groups of workers have indicated that the extent of test errors is much greater than has commonly been supposed (10, 16, 17, 29). Much information has recently come available as the result of work by a *Pseudomonas* Working Party, whose report is now awaiting publication (P. H. A. Sneath and V.

FIG. 1. Histograms of the frequency of different grades of reaction in tests on a set of strains examined in a single laboratory. The examples are chosen to illustrate tests with a bimodal response, with weaker but distinct bimodality, and with little tendency to show two modes. They illustrate (a) tests where the results are recorded in intensities (—, +, ++, +++, +++++), and also (b) those where the width is scored. From unpublished data of the *Pseudomonas* Working Party on a set of strains of *Pseudomonadaceae* incubated at 25 C.

G. Collins, manuscript in preparation), and some of whose findings were briefly cited by Sneath and Johnson (29). The experiences of the Working Party have shown the critical importance in such work of a good experimental design. My own views are that such collaborative work (which is difficult, but which is essential for good standardization of test methods) should be placed on a firm statistical basis.

Great efforts should be made to prescribe in detail the techniques; exact criteria should be given for scoring the results and these should be rigidly adhered to; if possible, agreement should be reached in advance on the quantity of data needed to evaluate a test, and considerable

attempts must be made to obtain complete sets of data, otherwise much of the work may be wasted. There is merit in reaching agreement not only on anonymity about the results obtained by individuals or laboratories, but also in agreement not to withdraw data even if some findings are unpalatable, a course advocated by the International Working Group on Mycobacterial Taxonomy (40).

One lesson from the *Pseudomonas* Working Party mentioned above is the importance of replicates as a check on the consistency of the results, and I personally believe that for future studies it is essential to use randomized duplicate sets of strains whose identities are not revealed until the completion of the work. This is needed to avoid the many subjective factors that creep in and to give an unbiased estimate of the reproducibility. Analysis of variance is a particularly valuable technique to assist in finding the sources of variation and thus to expedite the design of experiments to test the influence of various factors; its use is described in detail in Sneath and Johnson (29) and by Sneath and Collins (manuscript in preparation).

Some data are available for representative selections of tests currently used in bacterial taxonomy and identification. For ease of exposition in the following discussion, the results are treated by appropriate scoring (consistent with usage in numerical taxonomy) as positive or negative, and the test variance for test i , s^2_i (29) has been used to estimate the reproducibility. From the variances the percentage of disagreements, p_i , have been obtained, which are easier to understand and which differ but little from s^2_i for small values of these quantities. The distribution of p_i is generally skewed to the right, however, so that the use of s^2_i in obtaining the average disagreement over the tests, p , (formula 11 of Sneath and Johnson, reference 29) is recommended.

For replicates within one laboratory, p is usually less than 4%, although some tests give higher or lower values. In the *Pseudomonas* Working Party data cited by Sneath and Johnson (29), the mean for replicates within laboratories is about 2.9%, with individual values from almost zero to about 12%. The data of Sneath and Johnson on *Bordetella* and allied bacteria gave within one laboratory p of about 2%. Bergan and Lystad (3) report a similar value for phage sensitivities. Snell and Lapage (32) report a rate of 5% for carbon source utilization tests within one laboratory. Recent findings in our laboratory by D. Jones, B. J. Wilkinson, and R. K. A. Feltham, on a number of gram-positive and gram-negative bacteria, are of similar order of magnitude, e.g.,

about 3% in a study of *Listeria*, about 2.5% with enterobacteria, and about 2.4% with coryneform bacteria. We have noted that care must be taken to make an unbiased selection of strains and tests when investigating this question, because otherwise there is a tendency to choose the more difficult tests and the more troublesome strains when performing checks and replicates, and this makes the average discrepancies apparently much greater.

The variation between laboratories is much greater than that within laboratories. The *Pseudomonas* Working Party data cited by Sneath and Johnson (29) show a mean p_i of about 10%, ranging from zero to over 20%. Other values from the literature cited there are of a similar magnitude, sometimes averaging almost 20%, as in that quoted by Taylor et al. (36) for streptomycetes. Perhaps the most extensive results are those of Lapage and his colleagues (16, 17), which show averages of p_i of about 4% in large series of tests in which special attention was paid to reproducibility. The antibiotic sensitivities reported by Ericsson and Sherris (10) show averages of 6 to 10% of disagreements, depending somewhat on the method of scoring.

Other comparisons are still scanty, but differences between different times of reading of tests (29, Sneath and Collins, manuscript in preparation) may be large, showing that some tests are very sensitive to length of incubation. One would expect this to be particularly pronounced for tests that are routinely read at a time when the proportion of positive results is increasing swiftly; under such conditions, small changes in temperature, aeration, and the like, by affecting the growth rate, could well account for many of the discrepancies.

Experimental error is quite significant in a number of other techniques, as is well known, for example, with electrophoretic patterns, but no detailed quantitative studies appear to have been yet made. Preliminary experience of R. K. A. Feltham in our laboratory is that, under carefully controlled conditions, one can reduce discrepancies to 1 to 2%. In pyrolysis-gas-chromatography, replicates agree to within a few percent, but differences in the columns used for gas chromatography can cause pronounced effects (7, 24). From available evidence (8), experimental error in deoxyribonucleic acid pairing seems to have a standard deviation of several percent.

The effect of test errors on numerical taxonomy is to degrade the taxonomic structure, and the results for the commonly-used simple matching coefficient, S_{SM} , are given by Sneath and Johnson (29). If the average p_i becomes more than about 10%, the taxonomic structure is

severely damaged, and when p is over 20% very little of the original structure can be recovered even if the operational taxonomic units (OTUs) were in tight clusters (Fig. 2). The effects of error on taxonomic distance (which may also be of some application in microbiology) are described by Sneath and Sokal (31, p. 168).

Identification is a process where an unknown OTU, u , is compared with a series of taxa, q in number, in the hope of deciding which taxon it belongs to. A comprehensive discussion of this area is afforded by Sneath and Sokal (31, p. 381–408). There are two ways in which error in descriptions can affect an identification system. There may be (a) error in the reference description of the taxa, as when the characters have been incorrectly determined, and there may be (b) error in the description of the unknown OTU that is to be identified. The former is error in the identification matrix, the latter in the vector u in the symbolism of Sneath and Sokal (32, p. 382). Both types of error can be due to test error, but they can also, of course, occur for other reasons, such as mistakes in punching or copying, or misprints in a key, or inadequacies in the taxonomy on which the taxa were based. These types of error are seldom distinguished although they have quite different implications. Rhoden et al. (20) and Tomfrohde et al. (37), in discussing the performance of certain novel identification systems for enterobacteria, do not pursue the reasons that poor identification was found with the *Enterobacter-Serratia* group, although it seems likely that a major reason is the inadequate definition of the several species within this group. In one series of 12 misidentifications, 7 were due to the strains being atypical, not to the tests being in error. In other words the reference descriptions of the taxa did not accommodate the atypical strains. If such strains are to be excluded from the taxa, their failure to identify cannot be regarded as misidentification, but if they are to be included in these taxa the reference descriptions need correcting; this then is error of type (a). Lapage and his colleagues (1, 16, 17) have noted that inadequate descriptions are probably a main cause of failure in computer identification of bacteria, and though much of this will be due to poor taxonomic knowledge, some will be caused by error-prone testing methods employed by the designer or the user of the system. Error of type (a) is particularly important because it is hard to correct, for it requires republication of amended keys or tables, although it is easier to update the reference descriptions in computer-based systems.

Little is known about the statistics and information structure of most identification systems. The effect of errors on an identification system

FIG. 2. The effect of test errors on taxonomic structure. The figure illustrates the way in which the introduction of increasing test error, p , degrades the taxonomic structure, which consists of two tight clusters and two less closely related OTUs. From the example in Sneath and Johnson (29) represented as a shaded similarity matrix.

depends on the model that is used to construct the system. A full discussion of such models is given by Sneath and Sokal (31), and briefer accounts by Sneath (27, 28). The main forms are the sequential and the simultaneous. The sequential methods are those where successive characters are considered in turn until only one taxon remains. They include the usual diagnostic keys, and some related multiple-entry keys and "peek-a-boo" systems. An example is the system of Rypka et al. (22) and Rypka and Babb (21). The simultaneous methods are those where the unknown is compared on many characters simultaneously against the q taxa to obtain an identification at one step; they include the simultaneous keys and synoptic tables, the distance and taxon-radius models, and related probabilistic systems such as that of Lapage and his colleagues (1, 16, 17, 42) based on the work of Dybowski and Franklin (9), as well as classical discriminant functions. A good discussion of many of these is given by Sebestyen (23). The simultaneous methods are polythetic and can be visualized as ones where the unknown u is matched by some similarity measure to the q reference taxa in turn.

The choice of model depends on assumptions on the phenetic structure of the organisms. If all OTUs fall into tight well-separated clusters, then all the usual methods are available. But if they do not form tidy assemblages, then the diagnostic keys are less useful. There is evidence (28) that in bacteria there are some scattered aberrant strains between the denser clusters, which makes the simultaneous polythetic methods particularly suitable. These can often be represented as taxon-radius models (Fig. 3), following the concept of Gyllenberg (13), in which the unknown u is identified with a taxon J if it lies closer to the center of J than to any other taxon and also lies within the critical radius r_J of taxon J . The effect of character correlations in bacteria is probably small enough to ignore, though this has not been investigated in detail. The probabilistic systems can be partly represented in such a model by treating the distances for each character as logarithms (27). A new possibility, suggested to me by the way in which different tests become positive over time, is to treat such taxon-radius models as modified over time, so that the volume occupied by a taxon, instead of being a hypersphere, would be an elongated shape, with

FIG. 3. Identification in a taxon-radius model. The figure shows two taxa **J** and **K** represented by a number of OTUs whose phenetic position in multidimensional space has been shown schematically in two dimensions. The upper part of the figure illustrates the way in which three unknown strains, u_1 , u_2 , and u_3 , are examined to see whether they lie within a taxon by testing whether they are within the critical radius, r_J for taxon **J**, or r_K for taxon **K**. It is seen that u_1 lies within taxon **J**, and is identified as a member of this taxon. The unknown u_2 is outside taxon **K** but might be considered as an intermediate between **J** and **K**. The unknown u_3 evidently belongs to neither taxon, and may be a representative of some taxon that is not included in the scheme, or perhaps it may simply be an aberrant organism. The lower part of the figure illustrates the way the critical radius r_J has been found, by determining the distribution of frequency, f , of distances of the OTUs from the centroid c_J of taxon **J**. The mean distance $\bar{d}_{cJ}$ is shown by the broken circle, and by adding to it k times the standard deviation of these distances, $s_{d_{cJ}}$, one obtains the radius r_J . The mean is not identical to the mode because of the skewness of the curve. The figure also shows on the right the frequency distribution of members of taxon **K** from the centroid of **J**, and the small extent of overlap between the two distributions. For calculating r_K , a separate frequency distribution measured from the centroid of taxon **K** would be employed.

successive portions representing the hypersphere for successive times of readings: identification would require u to fall within one such volume and sufficiently close to the curve of its centroid-time loci.

Another general point in relation to the models is the number of characters required, m , to identify the q taxa. This is still poorly understood,

because although the theoretical minimum relation of m to q is given by $q = 2^m$, in practice many more characters are required to get sufficiently certain identifications. The small amount of available information suggests that the ratio m/q is usually 1.0 or somewhat greater. This seems true both for keys and synoptic tables for bacteria (2, 27, 31). My colleague A. Chater has investigated these ratios in botanical keys and finds the ratio is usually 1.0 to 1.3. The polythetic methods appear to be able to work with lower m/q ratios: thus for the system of Lapage et al. (16), it is about 0.5. The implication of this ratio for errors is that the smaller the number m , the more prone an error will be to cause a misidentification. Although it may seem that keys have the advantage here, it will be seen that they suffer from a danger of a different and more serious kind.

The number of tests employed in an identification scheme, m , is usually a selection from the totality of tests or characters studied, whose number is symbolized by n in numerical taxonomic work. The reduction from n to m offers the opportunity to reduce the errors by eliminating the less reproducible tests. If n is very large, then one would expect that it would be easy to choose enough tests that possessed both high constancy within taxa (a prerequisite for diagnostic characters) and also good reproducibility. In practice this often appears to be difficult, although systematic studies of this do not seem to have been made in microbiology. Nevertheless, we would hope that by such selection the p_i values for the diagnostic tests could be reduced from the levels cited earlier, even without improvements in standardization of techniques. A discussion of methods for choosing these tests is given by Sneath and Sokal (31, p. 384), and a method that takes into account test error has also been proposed (29). The reduction in the number of tests cannot be carried too far, however. This is not only because there may then be insufficient tests to distinguish all the taxa. It is also because the elimination of even unreliable tests may reduce the overall performance of the system; it has been shown (29) that this procedure may increase the uncertainty of estimates of similarity coefficients, and this in turn implies that similar behavior would be expected with polythetic identification schemes.

Before discussing the sensitivity to errors of different identification systems, it should be noted that mistakes in identification are of three kinds: (1) the unknown u may be incorrectly rejected as a member of taxon **J** when it should have been identified with it; (2) u is identified as belonging to **J** when in fact it belongs to another taxon in

the system, **K**; (3) **u** is identified with **J** (or some other taxon) when it belongs to a taxon not represented at all in the system (or is an aberrant isolated form), so that it should have been recorded as unidentifiable. The extent to which experimental test error leads to these different mistakes depends on the model used for identification, but generalizations about type (3) mistakes are difficult, because their frequency depends primarily on the comprehensiveness of the identification system and on the taxa from which **u** is most often chosen. The development of identification as a science also requires the development of some measure of the *degree* of misidentification. Two bases suggest themselves for this: the phenetic discrepancy or the cost to the user. The former could be expressed as the distance between the correct taxon and the taxon to which **u** was wrongly allocated (e.g., the phenetic dissimilarity between the centroids of the taxa, or their cophenetic distance is a phenogram). This would be a measure that is primarily of taxonomic significance. The latter criterion would be primarily of economic significance, and would employ some measure of the cost to the user of making a given mistake (e.g., misidentifying a strain of the typhoid bacillus as *Citrobacter freundii* would have a higher cost than the much grosser taxonomic error of misidentifying an *Aeromonas* strain as *C. freundii*). The failings of the system would then be expressed as the sum of the phenetic distances or the sum of the costs when a large sample of OTUs was identified.

Simultaneous polythetic methods, e.g., those using matching coefficients (5, 6), taxonomic distances, or analogues of these, are robust to test error of both type (a) and type (b), and misidentifications are usually minor in degree, i.e., the unknown is allocated as a rule to a taxon near to the correct one. This behavior can be readily understood by reference to taxon-radius models. The distance from **u** to the centroid of a taxon **J** is analogous to a dissimilarity coefficient in numerical taxonomy (the complement of a similarity coefficient). A few errors, when the number of characters is large, have little effect on similarities and dissimilarities. If **u**, therefore, should lie within the critical radius r_J , then a few test errors in **u** will seldom move it to an apparent position outside the taxon. This is test error of type (b), but a similar principle holds for test error of type (a); the reference description of taxon **J** is a formalized version of the position and radius of the taxon, and again small amounts of error will produce only small changes in the estimated centroid and radius. Type (a) error may creep in because polythetic systems seldom demand very

high constancy of character states with taxa, so that small amounts of error may be overlooked.

We can also expect that type (a) and type (b) error will be most serious if the taxon is an extremely compact cluster (i.e., with a very small radius), because it is the most compact clusters that will be susceptible to small amounts of error (29), perhaps leading to fusion with adjoining clusters. Also, of course, a small discrepancy in the position of **u** will more readily move it out of the cluster if the cluster radius is small.

The quantitative relations between p and misidentification rates remain to be worked out for such models. From results given by Sneath and Johnson (29) for the simple matching coefficient, S_{SM} , and the relation between taxonomic distance and that coefficient, i.e., $d = \sqrt{1 - S_{SM}}$, one can develop several approximations, whose derivations are given in the appendix to this paper. There are general grounds for believing that the distribution of OTUs of a taxon in phenetic space will be approximately multivariate normal, if the taxon is homogeneous, both for bacteria and other organisms: reasons are given by Sneath (28) and Sneath and Sokal (31, p. 197, 307). This is not strictly true for 1, 0 data, as commonly used in bacteriology, but it still offers a useful approximation. This is because for highly multivariate spaces the distances of OTUs from the centroid of a cluster will follow a bell-shaped curve, which, although theoretically not Gaussian, is likely to be sufficiently close to a normal curve for many practical purposes.

If correlations between characters are small, one can calculate the proportion of OTUs (strains) that lie within a taxon radius r_J when this is given in terms of the mean $\bar{d}_{cJ}$ and standard deviation $s_{d_{cJ}}$ of the taxonomic distances of strains from the centroid, **c**, of the taxon **J** (Appendix, formula 1). By this means one can calculate a radius that includes a given proportion of strains, $P(k)$, by appropriate choice of the constant k . The movement of an OTU from its correct position in phenetic hyperspace will be close to $\sqrt{p}$ when the mean test error is p , and this will increase the mean distance to $\bar{d}'_{cJ}$. The standard deviation will also be somewhat increased to $s'_{d_{cJ}}$. The Appendix gives approximations for these quantities if p is known (formulas 2 to 11). One can then evaluate the new value, k' , that corresponds to the radius r_J after the OTUs have been perturbed by the test error, and thus find $P(k')$.

Then $P(k) - P(k')$ will be the expected proportion of OTUs that will move from within the radius r_J to outside it. This will also be the expected proportion of extra misidentifications

FIG. 4. The effect of test error on identification in a taxon-radius model. The normal curves of distances from the centroid are shown, with the critical radius r_J chosen by setting k to 2.33 so that $P(k)$ is 0.99. The result of adding error, p , of 1, 2, 5, and 10% is shown, and the dark portion of the curve shows the proportion of OTUs that would not be identified (i.e., $1 - P(k')$ is shown dark, and these values are respectively 1.0, 7.8, 18.6, 55.4, and 91.1%). Based on the *Yersinia pseudotuberculosis* cluster mentioned in the text. The histograms show the actual frequencies of strains, f , both without error ($p = 0$) and after Monte Carlo perturbation to simulate increasing error. It is seen that the agreement with the theoretical distribution is reasonably good.

(at the critical level set by k , and hence r_J) of unknowns belonging to taxon J due to the introduction of the error p .

An illustrative example (Fig. 4) is given, based on a cluster of 34 *Yersinia pseudotuberculosis* strains with 49 characters (28, 34). The mean

distance, $\bar{d}_{cj}$, of strains from the centroid is 0.1382 with standard deviation, s_{dcj} , of 0.0463. If k is set to 2.33 [giving $P(k)$ of 0.99], then 99% of OTUs should lie within the radius $r_J = 0.1382 + 2.33 \times 0.0463 = 0.2461$. If one adds an error of $p = 0.05$, then using the approximations given in the Appendix, the mean distance will increase to $\bar{d}'_{cj} = 0.2547$ and the standard deviation to $s'_{dcj} = 0.0632$. Substituting these values into the formula for r_J (Appendix, formula 1) with r_J retained at 0.2461, gives k' as $(r_J - \bar{d}'_{cj})/s'_{dcj}$ which is -0.1361 . This corresponds to $P(k')$ of 0.4459. Therefore, almost 55% of u's (namely, 99% - 44.59%) that should be identified as members of J will no longer be identified, in addition to the 1% that fail to identify when no error is present. Figure 4 shows the effect of increasing error on the theoretical normal distribution of distances from the centroid in this example.

In practical schemes more sophisticated criteria will be used, but Fig. 4 illustrates the principle involved. A small amount of error will shift the distribution only a little to the right, so that only a small proportion of OTUs will cross the line set by r_J (in addition to the small proportion that already lie outside this radius, determined by the choice of k). As error increases, a growing proportion will appear to be outside the cluster, until the crest of the curve is pushed past r_J which occurs in this example when p is about 0.05. With yet greater p , almost all OTUs come to lie outside the cluster radius.

The curve of the proportion $P(k')$ plotted against p follows a falling sigmoid shape (Fig. 5). It will be appreciated that this general behavior of a sigmoid curve will occur with any plausible bell-shaped distribution, even if not Gaussian. The figure also confirms the deduction that tight clusters will be most affected by error, a deduction evident from the consideration that if an unknown is displaced about $\sqrt{p}$ from its correct position by error p , then it is very likely to move outside the cluster when r_J is of the order of $\sqrt{p}$ or less. Figure 5 also shows the curves for a cluster of 23 strains of *Yersinia enterocolitica* with 49 characters from the study of Stevens (34), a cluster that is less compact than *Y. pseudotuberculosis*, with $\bar{d}_{cj}$ of 0.2204 and s_{dcj} of 0.0495. Whatever value is chosen for k , the percentage of identifications falls more swiftly in the more compact cluster. In addition, the figure illustrates the way in which the loss of OTUs is proportionately less for any given p if k is high and r_J is therefore correspondingly larger. It must be remembered, however, that r_J cannot be increased indefinitely, because this will lead ultimately to

FIG. 5. The proportion $P(k')$ for the *Yersinia* clusters plotted against test error p for various values of k . The horizontal axis shows p on a square root scale for convenience. The solid lines are *Yersinia pseudotuberculosis*, the broken lines are *Y. enterocolitica*.

overlapping of clusters in the diagnostic system, so that large values of r_j correspond to less stringent criteria for identification. A similar falling sigmoid curve is seen in the proportion of correct identifications in the diagnostic system of Lapage et al. (17) when criteria are made more stringent.

We still lack information on the distribution of radii of clusters representing bacterial species, but it is well known in numerical taxonomy that species clusters are commonly formed in the 80 to 90% S region, so that their effective $\bar{d}_{cj}$ is commonly 0.2 to 0.3, using the relation between mean intracluster distances and centroid-to-OTU distances (Appendix, formulas 21 to 23). In practice, rather tighter clusters can usually be obtained by selecting the more constant tests for identification schemes, so $\bar{d}_{cj}$ may often be around 0.1. It is thus seen that, if values of p approach levels like 0.2, this will lead to serious trouble with identification, in the same way as levels of this magnitude degrade the taxonomic structure severely. Thus, errors of the magnitude reported by Taylor et al. (36) would preclude good diagnostic work. However, small levels of p , such as 0.01, will not lead to serious difficulties (Fig. 5), illustrating the robustness of polythetic systems to small disturbances.

Despite certain drawbacks which will be mentioned later, the model discussed above offers

advantages: it provides a conceptually simple way of visualizing diagnostic problems; it allows rough estimates of cluster parameters to be obtained from crude data; it can be developed into more sophisticated forms if required. As shown in the Appendix (formulas 12 to 24), one can estimate a good deal from the percent frequencies of positive states of characters, P_i , among strains of a taxon or cluster, and even some indication of the separateness of taxa. It also permits some estimate of the effect of the number of tests upon the reliability of identification. These last two points are discussed in turn below.

The series of values of P_i gives the centroid of a taxon in n -space. The mean squared distance from the centroid $\bar{d}_{cj}^2$ is also readily obtained (formula 14), and, as has been noted, this is one measure of cluster radius. The standard deviation of $\bar{d}_{cj}^2$ is not obtained thus simply, because it is affected by character correlations, but a rough estimate can be found (formula 15). Even if only the fraction of variable characters is given, one may get some indication because of the surprisingly uniform nature of the distribution of P_i in practice (formulas 16 to 18).

For any single cluster, the frequency curve of P_i is uneven, but if frequency distributions from several clusters are pooled, the curves become smoothed. It can then be seen that the distribution of P_i within homogeneous taxa (i.e., those without internal subgroups) is usually strongly U-shaped (Fig. 6). This seems true both for ordinary series of tests and also for more restricted sets, for example carbon utilization tests, as shown by data of Stanier, Palleroni, and Doudoroff (33) and Redfearn, Palleroni, and Stanier (19). But, apart from the high frequencies close to $p_i = 0$ and 1, the curves are almost flat, so the curves are trench-shaped. Although in theory even homogeneous clusters can show low values near 0 and 1, this is likely to be due to the exclusion of constant characters as is usual in a numerical taxonomic study. In heterogeneous clusters, however (i.e., those containing quite distinct subgroups), the frequencies at the extremes are less, and the shoulders near these extremes seem higher, so the curves have more of a U-shape or V-shape, culminating in extremely heterogeneous clusters where scarcely any characters are constant (see Fig. 6b and c).

No particular tendency has been noted in the examples so far examined for the frequencies at 0 to be higher or lower than those at 1, and this would vary with metabolic vigor (26) of the group. But test error will affect the shape of the curves, because it will shift some of the constant characters into the shoulders near 0 and 1, thus

FIG. 6. Frequency distributions of P_i from various taxonomic studies, obtained by pooling histograms from sets of comparable clusters. (a) Solid line, 8 clusters of gram-negative bacteria (15, 30); broken line, 70 clusters of gram-negative bacteria from the identification matrices of Bascomb et al. (1). The peak at $P_i = 0.5$ in the latter appears to be due to the need to allocate in poorly-known taxa the frequency of 50% to a number of tests, to insure a value that is effectively neutral in their identification system. (b) Solid line, 8 clusters of the *Mycobacterium rhodochrous* group (12); broken line, the *M. rhodochrous* clusters treated as one heterogeneous group. (c) Solid line, 6 clusters of gram-positive bacteria (B. J. Wilkinson, unpublished data); broken line, 8 homogeneous groups of pseudomonads (19, 33); open circles and dotted line, the pseudomonad data pooled to give one heterogeneous group. The frequency polygons show the percent of constant tests at $P_i = 0$ and $P_i = 1$, and the variable tests are grouped into bands so as to make the polygons symmetrical about $P_i = 0.5$.

raising these. In the region near $P_i = 0.5$, the error will have little effect except to make the central portion even more level. The test error p in the data from Johnson and Sneath (15) and Sneath and Johnson (30) was about 2%. If one assumes a Poisson distribution of erroneous results, this would convert about 1.9% of constant tests to apparent P_i of 0.01 and 0.99, and a very few to 0.02 and 0.98. These figures account in part for the shoulders seen in the curve in Fig. 6a. Therefore, it may be reasonably safe to assume that with homogeneous clusters, if one excludes the constant characters, one can treat the remainder as having a uniform flat distribution. One can then estimate the mean and standard deviation of the squared distances by the Appendix formulas 16 and 17, and from these make an estimate of $\bar{d}_{cj}$ and s_{dcj} from formulas 8 to 11. The true values will be higher because of character correlations, but some examples are close to expectation (see Appendix). The contribution to the variance (formula 13) varies strongly with P_i , as shown in Fig. 7, from which it can be seen that it is zero at $P_i = 0, 0.5$, and 1. The standard deviation is even more peaked, but it is the vari-

ance that is additive over characters. In contrast, the average contributions to distance (Fig. 7) are greatest at $P_i = 0.5$.

One can also calculate the distance between the centroids of two clusters L and M (Appendix, formulas 24 to 29) from the P_i values or from the mean inter- and intragroup similarities. This may give some indication of whether the clusters overlap. Although also dependent on the degree and direction of character correlations, there is probably some overlap if the distance between the centroids is less than the sum of the radii of the clusters (at some chosen value of k). If the distance between the centroids is much greater, then overlap is correspondingly less probable. The extent of the overlap is best calculated directly from observed distances of OTUs from the centroids (28) or from other central points (14, 38). This is illustrated in Fig. 3. It would seem unsafe to rely on the approximations given in the Appendix for this calculation.

The taxon-radius model also has the important advantage that, unlike most identification schemes, it can readily take into account the effect of a reduced number of tests, m . Thus it

would seem reasonable to regard an identification based on 50 tests to be more secure than one based on only 5. I have not derived the detailed statistics, but a rough method is based on the binomial approximation to the standard error of sampling of d^2 . If an unknown u lies within the chosen radius r_J , one may ask what is the expectation that its observed distance from the centroid, d_{cu} , could be as great as r_J due to accidents of sampling at random a smaller number of tests. We may assume that the sampling error of the larger number of tests has already been effectively incorporated by the choice of the radius r_J . An example is given in the Appendix. Conversely, if a new strain is to remain within the radius r_J when tested with a different random sample of tests, one can estimate (at some chosen confidence level set by k) the maximal distance $d(k, m)$ it should be from the centroid on the first set of tests.

The model just discussed does not take into account correlations between characters, but this disadvantage may be more apparent than real for 1, 0 data (reference 31, p. 406). Even full discriminant analysis is not suitable for a large array of taxa if the variances and covariances of the taxa are very different. A more serious drawback is that it gives relatively little differential weight to the more constant characters: the contribution to d^2_{cj} of a mismatch on an almost constant character is only four times that on a character with P_i close to 0.5. It can be converted into a model that allocates heavier weights, and more closely parallels the probabilistic models (1, 9, 16, 17, 42), but has a less explicitly Bayesian form. This can be done by scaling the contribution toward d^2_{cj} by the character variance (Appendix, formulas 32 to 34). This increases the differential weights of the characters greatly, but some provision must be made for preventing the effect of a mismatch on an almost constant character from becoming indeterminately large (as is needed also in the probabilistic models), and an estimate of p_i based on Laplace's law of succession as cited by Good (11) might be appropriate (Appendix, formulas 35 and 36).

Nevertheless, character weighting will not greatly affect the main conclusions of the effect of test errors. It is true that error in the more constant characters of a taxon will be more serious than errors in the less constant ones, and this is an important generalization for any particular trial of u against a given taxon. Taken over the whole system, however, it is likely that these constant characters will be variable in other taxa. Therefore, the bias in the estimates for some characters and taxa will tend to average out when the whole system is considered. This argument

FIG. 7. Values of mean distances and their variances for a single character i for P_i from 0 to 1. The upper part of the figure shows the mean distance from the centroid, which equals s_i , and the mean squared distance, s_i^2 . The lower part shows the standard deviation ($s_i^2 - 4s_i^4$) and the variance $s_i^2 - 4s_i^4$ (see text).

applies to probabilistic models as well. Possibly better predications of $P(k')$ could be obtained by taking into account the overall separation values of characters (for discussion see reference 31, p. 363).

Turning to the types of mistake that are likely as a result of error, we may assume that the majority of OTUs belong to separate clusters, so that with polythetic systems error is most likely to move the apparent position of u into empty phenetic space between clusters, leading to mistakes of type (a). The chance that u will enter another cluster will be small, so type (b) mistakes should be few.

We have little experience of the quantitative effect of errors on probabilistic methods of the kind used by Lapage and his colleagues (1, 16, 17, 42). The general behavior is likely to be fairly similar to the taxon-radius model scaled for character variances, because of the analogy between the negative logarithm of the probability and a taxonomic distance. The detailed statistics are now being developed (42). However, in one respect these models may be more sensitive to errors on single characters than the unscaled taxon-radius model: a mismatch on a constant character corresponds to a large distance and to a low probability of correct identification, because of the consequence of division of multiplication by a number close to zero. This may, therefore, be apt to displace an unknown more drastically than in the other model and thus to

produce a rather high rate of mistakes of type (1) due to a single test aberrancy. Working schemes do in fact make provisions for reducing the size of this effect (42).

Monothetic sequential identification schemes, in particular diagnostic keys, respond in a very different way to error. Error of type (a), i.e., in the reference description, is however commonly screened out during construction of the system, because it will usually show up as characters that are inconstant within taxa. These are consequently unsuitable as diagnostic characters and are more stringently excluded in monothetic systems than in polythetic ones. If type (a) error is left in, then it has serious results because it will be apt to divert identification down an entirely wrong pathway in a key. Type (b) error, i.e., error in the unknown, is also serious because it too will divert identification along the wrong path. Most monothetic systems make little provision for checking or retrieving mistakes; indeed, if they do make such provision they usually begin to assume the guise of simultaneous polythetic methods. Mistakes are also more apt to be serious in monothetic than in polythetic systems because the unknown may be identified with a taxon far from the correct one. It should be pointed out, however, that the average degree of misplacement in monothetic keys is not very high. The values depend partly upon the branching pattern of the key (only dichotomies are considered for simplicity). If each branch again branches repeatedly (a symmetrical and short key), then the mean misplacement is to a branch about two levels removed from the correct one. If the key is asymmetric, throwing off successive unbranched side-shoots, the mean misplacement is higher, approaching $\frac{1}{2}q$ if the levels are given sequential values for estimating misplacement.

Monothetic systems that are not explicitly hierarchic, such as the phage-typing schemes, are similarly sensitive to error of type (b). The requirement here is that as many strains as possible should be allocated to a definite class (e.g., a unique phage type). This in turn requires that the proportions of strains that fall into each class should be as nearly equal as possible because classes containing very large or very small percentages of strains are inefficient. One measure of the value of a given scheme with q classes, of proportionate frequency f , is that for which the information statistic $H = -\sum^q f \ln f$ is highest. A more direct measure, related to j_x , the probability of allelic identity in genetics (18), is $\sum^q f^2$. This is the probability that two strains taken at random will fall into the same class, and the scheme with lowest j_x is the most efficient. The most efficient schemes are, however, likely to be

the ones that are most subject to error of type (b), because it is the most valuable tests (that divide a group into classes of equal size) that will yield the most mistakes.

The mistakes that will usually occur in monothetic systems will be of types (1) and (2) combined. This is because the key structure generally shifts an unknown to an incorrect branch, resulting in identification with the wrong taxon, and at the same time it generates a type (1) mistake by removing it from the right taxon. Few systems of this kind make provision for unidentifiable OTUs.

Little experience has accumulated in microbiology with multiple-entry keys, commonly implemented by "peek-a-boo" systems (43). In these, cards—one per character and punched in locations representing taxa—are chosen according to the characters of the unknown and are superimposed until only one hole remains: this indicates the taxon with which u is identified. The information structure of these is poorly understood. Ideally one would wish each card to exclude half the remaining taxa, because a 50% reduction is the most economic use of presence-absence characters, and would lead to the ideal relationship $q = 2^m$. In practice some characters are variable, often leading to two or more cards per character, and the reduction of twofold with each card cannot be achieved. In such systems the effect of error is hard to forecast because on the one hand it may cover up the correct taxon and on the other hand it may open up incorrect taxa as possibilities which, however, may then be excluded again by later cards. One would expect them to behave more like monothetic keys than like polythetic systems because of the reliance on largely monothetic strategy in excluding incorrect taxa. Like keys, they are apt to generate type (1) plus type (2) mistakes by moving the unknown to an incorrect taxon. However, failure to identify at all can easily occur and is made obvious to the user. For this reason there is an inbuilt element of safety which warns the user of the danger of misidentification. It is likely, however, that type (a) error will have relatively little effect because it will lead to inconstant characters that will usually be excluded in building the system; alternatively the system will permit either state of a character, and this is relatively innocuous because it will lead to no decision on that character.

In conclusion, it can be seen that test reproducibility has implications for many fields in microbiology. It points to the urgent need to standardize methods if world-wide diagnostic systems or data banks are to prove worthwhile. It raises problems for collaborative studies on taxonomic groups and poses other problems in

the free exchange of standard media that may be unavailable in many parts of the world. We should not be too pessimistic about the prospects of improvement: thus, in antibiotic sensitivity tests, some remarkable consistency has been obtained in standardizing zone sizes to a standard deviation of less than 1 mm for a mean of about 20 mm (3). Some tests have been very reliable in most comparative studies on testing methods. We need to discover more tests of this kind as well as to improve the less reproducible ones. Such work will aid not only in bringing about the adoption of minimal standards for descriptions of bacteria, the better classification of bacteria, and the profitable storage and use of extensive microbiological data but also in improving the speed and certainty of identification of microbes, one of the most important activities of microbiologists.

APPENDIX

If we assume that the distribution of strains of a taxon **J** in a phenetic hyperspace of 1, 0 characters is sufficiently close to being multivariate normal, without strong correlation between characters, then the proportion of OTUs lying within a radius of

$$r_J = \bar{d}_{cj} + k s_{d_{cj}} \quad (1)$$

of the centroid **c** will be the one-tailed normal probability integral $P(k)$ for k standard deviations, where $\bar{d}_{cj}$ is the mean taxonomic distance of OTUs from the centroid, and $s_{d_{cj}}$ is the standard deviation of these distances.

If the simple matching coefficient, S_{SM} , is used, one can derive the mean and standard deviation of d_{cj} after error p has been added (where p is the mean of p_i over the n characters). This is not the same as the effect of error on similarities (or distances) between actual OTUs (29). Instead, the distance is measured from the centroid, **c**, of a cluster to an OTU **j** that carries error on the assumption that the centroid is not itself affected by the error. If a character state of an OTU is changed from 1 to 0 then δ^2_i , the increase in squared distance from the centroid is $2P_i - 1$, and if an OTU character state is changed from 0 to 1, it is $1 - 2P_i$. Therefore

$$\delta^2_i = (1 - 2P_i)(1 - 2X) \quad (2)$$

where X is the original state and P_i is the proportion of 1 states in character i . If one treats the cluster as having a mean value of P_i of $\bar{P}$, then (see equation 12 below) the mean of d^2_{cj} is $\bar{P}(1 - \bar{P})$, and for a typical OTU of that cluster the expected number of changes (due to error) from 0 to 1 will be $p(1 - \bar{P})$ and of 1 to 0 will be $p\bar{P}$. This can be treated approximately as a binomial

distribution whose mean is p and whose variance is $p(1 - p)/n$, where n is the number of characters. However, the average magnitude of the changes is $1 - 2\bar{P}$ or $2\bar{P} - 1$, and by appropriately scaling the expected number of each change one obtains

$$\bar{d}^2_{cj} = E(d^2_{cj}) = \bar{d}_{cj} + p(1 - 4\bar{d}^2_{cj}) \quad (3)$$

The increase is seen to be $p(1 - 4\bar{d}^2_{cj})$. The corresponding standard deviation is

$$SD(d^2_{cj} - \bar{d}^2_{cj}) = \sqrt{[p(1 - p)(1 - 4\bar{d}^2_{cj})/n]} \quad (4)$$

The additional variance from equation 4 must be added to the original observed variance of the cluster. One therefore obtains

$$s'^2_{d^2_{cj}} = SE(d^2_{cj}) = \sqrt{[s^2_{d^2_{cj}} + s^2_{d'^2_{cj}}]} \quad (5)$$

The other quantities in equations 3 and 5 are

$$\bar{d}^2_{cj} = \sum_{t_J} d^2_{cj}/t_J \quad (6)$$

$$s^2_{d^2_{cj}} = \sum_{t_J} (d^2_{cj} - \bar{d}^2_{cj})^2/(t_J - 1) \quad (7)$$

where there are t_J OTUs in cluster **J**. It seems preferable to correct the denominator as shown to $(t_J - 1)$ to give the degrees of freedom.

Although these approximations are not exact, they give values quite close to those obtained by Monte Carlo trials for the region of interest (i.e., where $\bar{d}^2_{cj}$ and p are small to moderate). It is in theory possible for the variance to decrease on adding error, because when p is near 0.5 the variance becomes close to $1/4n$, the value for equal numbers of 0 and 1 states distributed randomly, but this is unlikely to be of consequence in practical situations.

Equations 3 and 5 therefore give estimates of the new distribution of d^2_{cj} . However, the squared distances are apt to be skewed toward higher values, and better normal distributions are obtained from unsquared distances that correspond to $\sqrt{(1 - S_{SM})}$. Thus, for the clusters of *Y. pseudotuberculosis* and *Y. enterocolitica* used as an illustration, the coefficients of skewness, β_1 , were 1.42 and 0.82, respectively, for d^2 , but 0.70 and 0.29 for d . The first two are significant at the 0.05 level, but not the second two. Kurtosis is less marked; of the coefficients of kurtosis, β_2 , only that for d^2 with *Y. enterocolitica* was significant. The transformation of d^2 into d is not straightforward, however. The simplest method is to calculate estimates of $\bar{d}_{cj}$ and $s_{d_{cj}}$ from $\bar{d}^2_{cj}$ and $s_{d^2_{cj}}$ as follows:

$$\hat{d}_{cj} = \sqrt{[(\bar{d}^2_{cj}) - s^2_{d^2_{cj}}/4(\bar{d}^2_{cj})]} \quad (8)$$

(the parentheses have been placed about $\bar{d}^2_{cj}$ to em-

phasize that this is the mean of the squared distances, not the square of the mean distances mentioned below).

$$\hat{s}_{dej} = \sqrt{(s_{d^2ej}^2/[4(\bar{d}_{ej}^2) - s_{d^2ej}^2/4(\bar{d}_{ej}^2)])} \quad (9)$$

It may be preferable to start with the observed mean $\bar{d}_{ej}$ and its standard deviation s_{dej} , and to add estimates of error due to p . There are several methods of doing this, but the one that seems most satisfactory and which has been used in the examples, is as follows:

$$\hat{\bar{d}}_{ej} = \sqrt{[(\bar{d}_{ej})^2 + p - 4p((\bar{d}_{ej})^2 + s_{dej}^2)]} \quad (10)$$

$$\hat{s}_{dej} = \sqrt{[s_{d^2ej}^2/4((\hat{\bar{d}}_{ej})^2 + s_{dej}^2)]} \quad (11)$$

If equations 10 and 11 are used, the values without error, $\bar{d}_{ej} = \sum d_{ej}/t_J$ and $s_{dej} = \sqrt{[\sum (d_{ej} - \bar{d}_{ej})^2/(t_J - 1)]}$ are found by observation, and then squared. The value for $s_{d^2ej}^2$ is obtained from equation 5 and that for $\hat{\bar{d}}_{ej}$ is obtained from equation 10. When $p = 0$, the mean and standard deviation are identical to the observed ones.

We next turn to estimates of parameters of a cluster obtainable from values of P_i , the proportions within a single taxon of the 1 states of the various characters. The P_i values themselves represent the centroid. The variance of the character i , s_i^2 , is $P_i(1 - P_i)$, and this is also the mean squared distance of OTUs from the centroid on the i axis, so that

$$\bar{d}_{ej,i}^2 = P_i(1 - P_i) \quad (12)$$

The mean unsquared distance is twice this, but this gives a Manhattan metric, not a Euclidean metric, so it is not considered further. The variance of $\bar{d}_{ej,i}^2$ is

$$\text{var}(\bar{d}_{ej,i}^2) = s_i^2 - 4s_i^4 = P_i(1 - P_i)(1 - 2P_i)^2 \quad (13)$$

This is zero at $P_i = 0, 0.5$, and 1.0 and maximal at $P_i = (2 \pm \sqrt{2})/4$, i.e., at $P_i = 0.1465$ or at 0.8536 , when the variance is $1/6$. The variance for $\bar{d}_{ej,i}$ is equal to that for $\bar{d}_{ej,i}^2$, surprisingly.

Character correlations do not affect mean distances, so that

$$\bar{d}_{ej}^2 = \sum_i P_i(1 - P_i)/n = \sum_j t_J d_{ej}^2/t_J \quad (14)$$

The variance is affected by correlations, but if these are all zero, then

$$s_{d^2ej}^2 = \sqrt{(\sum \text{var}(\bar{d}_{ej,i}^2)/n^2)} \quad (15)$$

In many cases the character correlations are sufficiently

low that equation 15 gives a reasonable estimate. For example, for the *Y. pseudotuberculosis* and *Y. enterocolitica* clusters used as illustrations, the observed standard deviation and that calculated from equation 15 was 0.0147 and 0.0127 for the former species and 0.0253 and 0.0195 for the latter species. If the distribution of P_i is uniform between 0 and 1, the expected values of $\bar{d}_{ej}^2$ and $s_{d^2ej}^2$ are obtained by integration and are $1/6$ and $1/\sqrt{30}$, respectively. It has been noted in the text that the distribution of P_i is approximately uniform for the variable characters, and it is therefore possible to obtain very rough estimates from the number of variable characters, n_v , as follows

$$\bar{d}_{ej}^2 = n_v/6n \quad (16)$$

$$s_{d^2ej}^2 = \sqrt{(n_v/30n^2)} \quad (17)$$

In the two *Yersinia* clusters the observed and the estimated squared distances were 0.0212 and 0.0408 for *Y. pseudotuberculosis*, whereas for *Y. enterocolitica* they were 0.0509 and 0.0748, respectively. The corresponding standard deviations for *Y. pseudotuberculosis* were 0.0147 and 0.0138, and for *Y. enterocolitica* they were 0.0235 and 0.0175. Considering the weakness of the assumptions, the agreement is quite good.

For series of clusters from the literature, the estimates given by equations 14 and 15 are again reasonably close to those from equations 16 and 17. Thus, for the eight clusters from Goodfellow (12), the average of s_i^2 over the n_v variable characters was 0.1427 with standard deviation 0.0495, and the average of $s_i^2 - 4s_i^4$ was 0.0365 with standard deviation 0.0138; for the eight clusters from data of Stanier and his colleagues (19, 33), they were respectively 0.1233 (SD 0.0176) and 0.0389 (SD 0.0033). The values expected from equations 16 and 17 are 0.1667 and 0.0333. Even quite heterogeneous clusters give similarly good estimates. Thus the P_i values for the clusters from Goodfellow pooled into one heterogeneous group give $s_i^2 = 0.1325$ and $s_i^2 - 4s_i^4 = 0.0371$ averaged over the variable characters, whereas the corresponding figures from the pooled clusters from the data of Stanier and his colleagues are 0.1734 and 0.0287.

Alternatively, if the mean squared distance is known, but not the number of variable characters, one may obtain an estimate of $s_{d^2ej}^2$ on the assumption of a uniform distribution of P_i for variable characters as

$$s_{d^2ej}^2 = \sqrt{(\bar{d}_{ej}^2/5n^2)} \quad (18)$$

This may be of some assistance in roughly estimating cluster parameters if only inter- and intragroup average similarity coefficients are available.

The relation of intragroup similarity to distances from the centroid depends on the way the average similarity has been calculated. Two methods have been

used (25):

$$r\bar{S} = \frac{1}{t_J^2} \sum S_{jk} \quad (19)$$

$$\Delta\bar{S} = \frac{1}{\frac{1}{2}t_J(t_J - 1)} \sum S_{jk} \quad (20)$$

The former includes the self-comparisons in the principal diagonal. One can show (31, p. 292) that

$$\bar{d}_{ej}^2 = \frac{1}{2}(1 - r\bar{S}) = \frac{(t_J - 1)}{2t_J} (1 - \Delta\bar{S}) \quad (21)$$

One can readily obtain $\bar{S}$ or $\Delta\bar{S}$ from $\bar{d}_{ej}^2$ by the relations

$$r\bar{S} = 1 - 2\bar{d}_{ej}^2 \quad (22)$$

$$\Delta\bar{S} = \frac{1 - 2t_J \bar{d}_{ej}^2}{(t_J - 1)} \quad (23)$$

Turning next to relations between two clusters, **L** and **M**, with t_L and t_M OTUs, respectively, one can express similarities in the form of distances and partition the distances in the form of sums of squares about centroids. If the values of P_i are given for **L** and **M** separately, the squared taxonomic distance between the centroids is

$$d_{LM}^2 = \frac{1}{n} \sum (P_{iL} - P_{iM})^2 \quad (24)$$

If given the mean inter- and intragroup similarities, one can obtain by equations 22 or 23 the mean squared distances about the respective centroids, $\bar{d}_{ej,L}^2$ and $\bar{d}_{ek,M}^2$. The mean intergroup squared distance is

$$\bar{d}_{jL,kM}^2 = 1 - \bar{S}_{LM} = 1 - \frac{1}{t_L t_M} \sum S_{jL,kM} \quad (25)$$

The total sum of squares is

$$SS_T = \frac{1}{(t_L + t_M)} \cdot (t_L^2 \bar{d}_{ej,L}^2 + t_M^2 \bar{d}_{ek,M}^2 + t_L t_M \bar{d}_{jL,kM}^2) \quad (26)$$

This can be partitioned into the within groups sum of squares

$$SS_W = t_L \bar{d}_{ej,L}^2 + t_M \bar{d}_{ek,M}^2 \quad (27)$$

and the between groups sum of squares

$$SS_B = \frac{t_L t_M}{t_L + t_M} (\bar{d}_{jL,kM}^2 - \bar{d}_{ej,L}^2 - \bar{d}_{ek,M}^2) \quad (28)$$

The squared distance between the centroids of **L** and **M** is

$$d_{LM}^2 = \left(\frac{t_L + t_M}{t_L t_M} \right) SS_B \quad (29)$$

$$= \bar{d}_{jL,kM}^2 - \bar{d}_{ej,L}^2 - \bar{d}_{ek,M}^2$$

The distances of the centroids of **L** and **M** from the grand centroid are respectively $t_M/(t_L + t_M)$ and $t_L/(t_L + t_M)$ times the intercentroid distance.

The mean squared distance of OTUs of **L** from the centroid of **M** is

$$\bar{d}_{jL,eM}^2 = d_{LM}^2 + \bar{d}_{ej,L}^2 \quad (30)$$

and that for OTUs of **M** from the centroid of **L** is

$$\bar{d}_{eL,kM}^2 = d_{LM}^2 + \bar{d}_{ek,M}^2 \quad (31)$$

The effect of the number of tests, m , on the reliability of identification can be illustrated as follows. Suppose a strain of *Y. pseudotuberculosis* in the example given earlier was found to have a distance, d_{eu} , of 0.1850 from the centroid when only 9 tests chosen at random had been performed instead of the full 49. The standard error of sampling is approximately $\sqrt{[d^2(1 - d^2)/m]}$, which here is $\sqrt{(0.0331/9)} = 0.0606$. However, if the radius, r_J , has effectively made allowance for the sampling error of the full 49 characters, this error should be subtracted, otherwise the estimates will be too large. This gives the increase due to the reduced number of tests as $\sqrt{[(0.0331/9) - (0.0331/49)]} = 0.0548$. The radius $r_J = 0.2461$ is 1.1150 of these standard errors above the observed distance of 0.1850. There is thus an appreciable chance, about 13.24% [since $P(k)$ is only 0.8676 for $k = 1.1150$], that this strain would fall outside this radius if a different set of 9 tests had been chosen at random. If, however, one had performed 36 of the 49 tests, one would have an increased standard error of only 0.0156, and r_J is 3.9167 standard errors above 0.1850. $P(k)$ for $k = 3.9167$ is over 0.9999, so there is less than a 0.01% chance that the strain would fall outside r_J on a different selection of tests (indeed rather less, because the sampling would be with replacement).

The converse allows one to estimate the approximate distance from the centroid, $d(k, m)$, that a new strain must not exceed in order that (at some chosen confidence level) it would be expected to remain within the radius r_J if tested with a different random sample of tests. This is obtained by setting r_J to $d(k, m) + k[SE(d(k, m))]$, and is most readily calculated by successive approximations. In this example for $k = 2.33$ (the 99% confidence level), $d(k, m)$ is 0.1453 for $m = 9$ and 0.2058 for $m = 36$, after correcting as above for the sampling error of the 49 tests.

In the model in which the contributions to squared

distance from the centroid made by character i in cluster J are scaled by the variance, the contribution, $d^{*2}_{ejJ,i}$ is

$$d^{*2}_{ejJ,i} = d^2_{ejJ,i}/s^2_{iJ} \quad (32)$$

where $d^2_{ejJ,i}$ is the squared distance of j from the centroid of J on character axis i , and s^2_{iJ} is the variance of i in cluster J , that is $P_{iJ}(1 - P_{iJ})$. For a 0 state, $d^{*2}_{ejJ,i}$ is $(1 - P_{iJ})/P_{iJ}$, whereas for a 1 state it is $P_{iJ}/(1 - P_{iJ})$. The mean is 1. The variance is

$$\text{var}(d^{*2}_{ejJ,i}) = (1 - 4s^2_{iJ})/s^2_{iJ} \quad (33)$$

The squared distance, over all characters, of OTU j from the centroid of cluster J is then

$$d^{*2}_{ejJ} = \sum^n d^{*2}_{ejJ,i}/n \quad (34)$$

To prevent the variance s^2_{iJ} from becoming zero if a character is constant (leading to indeterminately large values of d^{*2} and its variance), an estimate of P_{iJ} based on Laplace's law of succession, as cited by Good (11), may be useful. This assumes that, after an observed sample of t_1 positive and t_0 negative results, the probability that the next strain would yield a positive result is

$$\hat{P}_{iJ} = (t_1 + 1)/(t_1 + t_0 + 2) \quad (35)$$

The corresponding variance is

$$\hat{s}^2_{iJ} = (t_1 + 1)(t_0 + 1)/(t_1 + t_0 + 2)^2 \quad (36)$$

and this will never be zero.

ACKNOWLEDGMENTS

I am indebted to the members of the *Pseudomonas* Working Party, to Dorothy Jones, M. Stevens, R. Johnson, R. K. A. Feltham, and Brenda J. Wilkinson for unpublished data on bacteria they have studied, and to M. J. Sackin for assistance with computing.

LITERATURE CITED

- Bascomb, S., S. P. Lapage, M. A. Curtis, and W. R. Willcox. 1973. Identification of bacteria by computer: identification of reference strains. *J. Gen. Microbiol.* **77**:291-315.
- Berd, D. 1973. Laboratory identification of clinically important aerobic actinomycetes. *Appl. Microbiol.* **25**:665-681.
- Bergan, T., and A. Lystad. 1972. Reproducibility in bacteriophage sensitivity pattern of *Pseudomonas aeruginosa*. *Acta Pathol. Microbiol. Scand. B* **80**:345-350.
- Blazevic, D. J., M. H. Koepcke, and J. M. Matsen. 1972. Quality control testing with the disk antibiotic susceptibility test of Bauer-Kirby-Sherris-Turck. *Amer. J. Clin. Pathol.* **57**:592-597.
- Bogdanescu, V., and R. Racotta. 1967. Identification of mycobacteria by overall similarity analysis. *J. Gen. Microbiol.* **48**:111-126.
- Campbell, I. 1973. Computer identification of yeasts of the genus *Saccharomyces*. *J. Gen. Microbiol.* **77**:127-135.
- Carmichael, J. W., A. S. Sekhon, and L. Sigler. 1973. Classification of some dermatophytes by pyrolysis-gas-liquid chromatography. *Can. J. Microbiol.* **19**:403-407.
- De Ley, J. 1971. Hybridization of DNA, p. 311-329. In J. R. Norris and D. W. Ribbons (ed.), *Methods in microbiology*, vol. 5A. Academic Press, Inc., London.
- Dybowski, W., and D. A. Franklin. 1968. Conditional probability and the identification of bacteria: a pilot study. *J. Gen. Microbiol.* **54**:215-229.
- Ericsson, H. M., and J. C. Sherris. 1971. Antibiotic sensitivity testing. Report of an International Collaborative study. *Acta Pathol. Microbiol. Scand. B* **217**(Suppl.):1-90.
- Good, I. J. 1950. Probability and the weighing of evidence. Charles Griffin & Co., London.
- Goodfellow, M. 1971. Numerical taxonomy of some nocardioform bacteria. *J. Gen. Microbiol.* **69**:33-80.
- Gyllenberg, H. G. 1965. A model for computer identification of microorganisms. *J. Gen. Microbiol.* **39**:401-405.
- Hutchinson, M., K. I. Johnstone, and D. White. 1965. The taxonomy of certain thiobacilli. *J. Gen. Microbiol.* **41**:357-366.
- Johnson, R., and P. H. A. Sneath. 1973. Taxonomy of *Bordetella* and related organisms of the families *Achromobacteraceae*, *Brucellaceae*, and *Neisseriaceae*. *Int. J. Syst. Bacteriol.* **23**:381-404.
- Lapage, S. P., S. Bascomb, W. R. Willcox, and M. A. Curtis. 1970. Computer identification of bacteria, p. 1-22. In A. Baillie and R. J. Gilbert (ed.), *Automation, mechanization and data handling in microbiology*. Academic Press, Inc., London.
- Lapage, S. P., S. Bascomb, W. R. Willcox, and M. A. Curtis. 1973. Identification of bacteria by computer: general aspects and perspectives. *J. Gen. Microbiol.* **77**:273-290.
- Nei, M. 1972. Genetic distance between populations. *Amer. Natur.* **106**:283-292.
- Redfearn, M. S., N. J. Palleroni, and R. Y. Stanier. 1966. A comparative study of *Pseudomonas pseudomallei* and *Bacillus mallei*. *J. Gen. Microbiol.* **43**:293-313.
- Rhoden, D. L., K. M. Tomfohrde, P. B. Smith, and A. Balows. 1973. Auxotab—a device for identifying enteric bacteria. *Appl. Microbiol.* **25**:284-286.
- Rypka, E. W., and R. Babb. 1970. Automatic construction and use of an identification scheme. *Med. Res. Engineering* **9**(2):9-19.

22. Rypka, E. W., W. E. Clapper, I. G. Bowen, and R. Babb. 1967. A model for the identification of bacteria. *J. Gen. Microbiol.* **46**:407-424.
23. Sebestyen, G. S. 1962. Decision-making processes in pattern recognition. Macmillan Co., New York.
24. Sekhon, A. S., and J. W. Carmichael. 1973. Column variation affecting a pyrolysis-gas-liquid chromatographic study of strain variation in two species of *Nannizzia*. *Can. J. Microbiol.* **19**:409-411.
25. Sneath, P. H. A. 1957. The application of computers to taxonomy. *J. Gen. Microbiol.* **17**:201-226.
26. Sneath, P. H. A. 1968. Vigour and pattern in taxonomy. *J. Gen. Microbiol.* **54**:1-11.
27. Sneath, P. H. A. 1969. Computers in bacteriology. *J. Clin. Pathol.* **22**(suppl.)**3**:87-92.
28. Sneath, P. H. A. 1972. Computer taxonomy, p. 29-98. In J. R. Norris and D. W. Ribbons (ed.), *Methods in microbiology*, vol. 7A. Academic Press, Inc., London.
29. Sneath, P. H. A., and R. Johnson. 1972. The influence on numerical taxonomic similarities of errors in microbiological tests. *J. Gen. Microbiol.* **72**:377-392.
30. Sneath, P. H. A., and R. Johnson. 1973. Numerical taxonomy of *Haemophilus* and related bacteria. *Int. J. Syst. Bacteriol.* **23**:405-418.
31. Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman & Co., San Francisco.
32. Snell, J. J. S., and S. P. Lapage. 1973. Carbon source utilization tests as an aid to the classification of nonfermenting bacteria. *J. Gen. Microbiol.* **74**:9-20.
33. Stanier, R. Y., N. J. Palleroni, and M. Doudoroff. 1966. The aerobic pseudomonads: a taxonomic study. *J. Gen. Microbiol.* **43**:159-271.
34. Stevens, M. 1969. Development and use of multi-inoculation test methods for a taxonomic study. *J. Med. Lab. Technol.* **26**:253-263.
35. Stokes, E. J. 1968. Quality control in diagnostic bacteriology. *Proc. Roy. Soc. Med.* **61**:457-463.
36. Taylor, G. R., R. K. Guthrie, and E. B. Shirling. 1970. Serological characteristics of *Streptomyces* species using cell wall immunizing antigens. *Can. J. Microbiol.* **16**:107-115.
37. Tomfohrde, K. M., D. L. Rhoden, P. B. Smith, and A. Balows. 1973. Evaluation of the redesigned Enterotube—a system for the identification of *Enterobacteriaceae*. *Appl. Microbiol.* **25**:301-304.
38. Tsukamura, M. 1971. Some considerations on classification of mycobacteria. Definition of bacterial species by introduction of the concept of "hypothetical median or mean organism." *Jap. J. Tuberc.* **17**:18-30.
39. Wayne, L. G. 1967. Selection of characters for an Andansonian analysis of mycobacterial taxonomy. *J. Bacteriol.* **93**:1382-1391.
40. Wayne, L. G., T. M. Dietz, C. Gernez-Rieux, P. A. Jenkins, W. Kppler, G. P. Kubica, J. B. G. Kwapinski, G. Meissner, S. R. Pattyn, E. H. Runyon, K. H. Schrder, V. A. Silcox, A. Tacquet, M. Tsukamura, and E. Wolinsky. 1971. A co-operative numerical analysis of scotochromogenic slowly growing mycobacteria. *J. Gen. Microbiol.* **66**:255-271.
41. Williams, R. E. O., and J. E. Rippon. 1952. Bacteriophage typing of *Staphylococcus aureus*. *J. Hyg.* **50**:320-353.
42. Willcox, W. R., S. P. Lapage, S. Bascomb, and M. A. Curtis. 1973. Identification of bacteria by computer: theory and programming. *J. Gen. Microbiol.* **77**:317-330.
43. Yourassowsky, E., W. Hansen, M. Labbe, and J. Van Molle. 1965. Problmes taxonomiques. Orientation mechanographique du diagnostic des espces microbiennes. *Acta Clin. Belg.* **20**:279-285.