Computers in bacteriology

P. H. A. SNEATH

*From the Medical Research Council Microbial Systematics Research Unit,
University of Leicester, Leicester*

Numerical methods are now widely used in the classification of bacteria, and they are also being explored in the field of identification. Identification is the major interest for clinical work, but it should be emphasized that good taxonomies are a prerequisite for satisfactory schemes of identification or diagnosis. Much remains to be done in the taxonomy of bacteria, and for this reason a brief summary of numerical methods of classification will be given before considering diagnosis. First, however, some comments will be made on the collection of laboratory data.

COLLECTION OF DATA

One of the special problems in bacteriology is that the time and effort required to obtain a single piece of information about a specimen (a bacterial strain or a clinical sample) are high compared with those for a higher organism like an insect. A brief glance at a fly shows that it has two wings, six legs, and so on. But to determine whether a bacterium is indole positive or indole negative normally requires cultivation in a tube of an appropriate medium followed by chemical testing some hours later. Since a tube of sterile medium costs quite a lot (perhaps 4d, *ie*, 1.67 new pence, most of it due to preparation costs) there is scope for methods that reduce time and money, although there will still be the costs of skilled labour and equipment for setting up tests and reading them. High costs are of course found in many areas of clinical investigation, and provide much of the impetus toward the current efforts in automation.

Savings in cost can be made in various ways. Thus a divided Petri dish with 25 sterile compartments has been developed in this Unit, which allows many bacteriological tests to be made quickly and cheaply using a multiple inoculator (Sneath and Stevens, 1967). Each well in this dish costs something like 1d (about 0.4 new pence) of which about half is the cost of medium plus dish and half due to preparation. Also, with this dish one can fill and inoculate about 500 tests per hour. Similarly there

are micro-methods that greatly shorten the time for performing tests (see Hartman, 1968, for a comprehensive review of these micro-methods).

The use of methods like this still requires a good deal of ingenuity if they are to be applied to clinical bacteriology. They are at present best suited to the collection of data for classification of bacteria (*eg*, Lovelace and Colwell, 1968), or to screening large numbers of specimens during an epidemic. They could, however, be modified for routine use. For example, each well in a divided dish could contain a different medium, chosen so as to offer a range of the most useful tests for identification; each bacterium could then be inoculated into the wells of one dish, and reading would be quick and simple. The reading itself might perhaps be automated by photoelectric methods.

Of course in all such modifications of standard methods it is extremely important that the results should be highly reliable. This does not seem an insuperable problem, though experience will be needed to adapt them. The biggest obstacle to employing automatic analysers of the kind used in biochemical laboratories is the need to achieve rigorous standards of sterility, and progress in this respect is likely to be rather slow.

CLASSIFICATION

Bacteria have traditionally been classified by considering relatively few characters of the organisms, sometimes from tables of data (often incomplete), and giving great prominence to certain of the characters. Numerical methods developed in the last few years are intended to handle large tables of data (as complete as possible) without giving undue importance to any of the characters. Since it is difficult to make sense of large tables of data, computer methods are used to process them, and an outline of these is given below. Fuller descriptions of numerical taxonomy will be found in various publications (*eg*, Sokal and Sneath, 1963; Sneath, 1962; Skerman, 1967).

A distinction should first be made between

classification and identification. In the former collections of bacteria are grouped into taxonomic groups such as genera or species. In the latter an unknown strain is identified with a previously known taxonomic group. Although these two processes can be to some extent performed simultaneously (or by repeated modification of preliminary classifications), it is convenient to distinguish them because of the logical differences between these two procedures. Classification is generally of the kind loosely termed 'natural', and, in more technical language, 'polythetic'. In such classifications there are not necessarily any constant characters in a given taxonomic group, though fortunately there are usually a number of constant characters in practice. These constant characters are, however, not chosen *a priori* but emerge from the analysis of a large set of characters whose constancy within groups is not known initially for the obvious reason that the groups themselves are not known at this stage.

It is usual to give each character equal weight when constructing a classification, which should be contrasted with the weighting of characters that is appropriate in identification. This point has aroused some controversy, but the basic criticism of character weighting for classification is that it is extremely difficult to find any logical and consistent grounds for allocating weights before the groups are known. For certain special purposes one may of course use frankly arbitrary classifications, such as phage types, which do not necessarily agree with the orthodox taxonomic groups; these are not further considered here, because the selection of optimal divisions for special purposes has not received much careful study.

Numerical taxonomy relies on a large number of characters of the bacteria, and the choice of these requires consideration. The aim is to try to obtain a representative sample of all the attributes of the organisms; this is not strictly a random sample, but should aim towards it. It is unsafe to rely on only one class of characters, *eg*, serological or carbohydrate fermentations, because they may not always correlate well with other sorts of characters. Nevertheless such discordance has usually been fairly small, so that a set of characters covering the main classes of attributes—morphological, physiological, biochemical, etc—is generally quite reliable.

It is, however, important to use a large number of characters, because if too few are used the resemblances between bacteria are too uncertain for satisfactory work. About 100 characters are desirable, and more if possible.

The next step is to convert the characters into

suitable coded form. With most bacteriological data these can be conveniently scored as plus (+) and minus (−) (usually represented for computation as 1 and 0), while missing entries are marked NC, meaning that no comparison is to be made when this entry is considered in the computer program. Certain quantitative or semiquantitative characters may need special coding (see Sneath, 1962; Lockhart, 1964), as may a few qualitative characters. The resulting table of strains versus characters is fed into the computer and the computer program carries out the following steps.

1 The resemblance between each strain and every other is calculated, commonly as the percentage of agreements between the two arrays of character values. This gives a chequerboard table, in which the percentage similarity between each pair of strains is recorded. A similarity value, for example, between two strains that score respectively + + − − + and + − − + + might be given as three agreements out of five, or 60%.

2 The computer next sorts through the similarity table and groups together the strains with the highest mutual resemblances, giving groups of strains that are highly similar to each other. For example, if strains A and B are 95% similar, while A and C, and B and C, are only 60% similar, then AB would be a small group (in this instance of only two members). Similarly C and D might be 88% similar, but all similarities between them and A or B might be low, around 60%. We would then have two separate groups, AB and CD.

3 This grouping process is repeated so as to bring together the most similar groups, and groups of groups, until all the bacteria have fused into a single large group.

4 The computer then displays these groups in one of several ways. The commonest is a taxonomic tree, or dendrogram, which summarizes the groupings in a convenient form. We might show the four strains A, B, C, D in our example:

These dendrograms can be used to decide what groups should be recognized in a formal taxonomy, although it should be noted that the choice of similarity levels to represent species, genera, etc, is largely arbitrary at present. It should also be noted that there are commonly some aberrant strains that

do not fit neatly into any group, though usually they are a minority.

5 The computer can then be asked to list the more constant characters for each group, as these are the most suitable for use in schemes for identification.

There are various numerical methods for these computer steps, but in general they give very similar results when applied to bacteria, although improvements in the methods are constantly being made. Numerical methods are usually very successful with bacteria (for reviews see Floodgate, 1961; Sneath, 1964) and have also been applied to other fields, *eg*, clinical medicine (see Baron and Fraser, 1968). They may also have a part to play in such procedures as the automatic scanning of cervical smears.

IDENTIFICATION

There are two well-known ways of identifying a bacterium against recorded data. The first is by means of a dichotomous key. The second is by comparing the results of a number of tests with diagnostic tables that list the characters of the groups. These are examples of two basic strategies for identification, which are respectively the 'sequential' and the 'simultaneous'. In the first a single character is used at a time and a sequence of these is followed through. In the second a number of characters are considered simultaneously. Of course many schemes make use of both principles. A key may have several tests listed at each dichotomy, although then the user may be puzzled if his unknown agrees only with some of them. A diagnostic table, too, may be broken into successive stages, so that the user moves from a first-stage table to a second-stage table on a smaller set of bacteria, and so on.

Keys have certain advantages. They are compact and easy to use. However, they suffer from disadvantages. It may not be possible to find mutually exclusive and constant characters for the groups. Also, if an error is made somewhere in the keying-out process, the user may be led a long way from the correct answer. These are particularly troublesome with bacteria, where completely constant characters may be hard to find, and where it is easy to make technical errors in performing the tests. These disadvantages are less with diagnostic tables, because it is not difficult to see that one or two tests are atypical as the user compares the unknown with the columns of the table. It will be seldom worth considering the use of a key stored in a computer though the use of a computer to generate a printed key from a given set of data may be worthwhile.

Tables, however, are well suited to computer use, as the machine can store large tables and can very quickly match the unknown against all the alternatives. It can also make some estimate of the probability of correct identification, which is particularly important in clinical work.

The basic form of a tabular system of identification can be illustrated by considering a simplified example. Below are shown three groups of bacteria, together with an unknown strain which it is desired to identify.

Character	<i>Klebsiella</i>	<i>Escherichia</i>	<i>Salmonella</i>	Unknown, u
Gram reaction	—	—	—	—
Lactose fermented	+	+	—	—
Indole	—	+	—	—
MR	—	+	+	+
H ₂ S	—	—	+	+

In this example characters are taken as being constant in the three genera. In practice the user has to make allowances for the variability of characters within taxonomic groups (as described below, note that *Klebsiella*, for example, is sometimes MR+) but the main principle is clear. We compare the unknown with each group in turn and choose the best matching group as the correct identification. In this example the unknown is clearly *Salmonella*. We may, however, find that the unknown does not match well with any group and perfect matching is rare if there are a lot of characters. This may be because we have a strain that is atypical in some respects, or one that is intermediate between two groups, or perhaps our unknown belongs to a group far from any that we have in our table. Cowan and Steel (1965) give extensive tables for bacteria of medical interest.

We can improve the power of our scheme by calculating the degree of match between our unknown and the various groups, and this leads naturally to computer methods. The match values are now virtually a set of similarity coefficients between the unknown and the various groups in the table, or some central representative of each group.

There are several ways in which the computation can be done. Thus Gyllenberg (1965) treats the similarities (or rather the dissimilarities) as distances between the unknown and each group; identification is made with the nearest group, provided it is close enough upon various chosen criteria. Rypka, Clapper, Bowen, and Babb (1967) have described another computer technique.

We shall here take as an example the broad outline of a method proposed by Dybowski and Franklin (1968). They use a table in the computer

which contains, instead of plus or minus values, the frequency of positives of each test in the various groups. Because there is always a small chance that an atypical result may be obtained with the unknown, or that a mistake has been made in testing, the entries for constant characters are never 0% or 100%, but are set to figures such as 5% or 95%. This prevents a mismatch on one test from completely excluding a group in the process described below. They then argue as follows. If a group has say 20% of positives in a given test, then the chance that an unknown which scores plus really belongs to that group is taken as only 20%, *ie*, 1/5. If it scores minus, then the chance is taken as 80%, or 4/5. Further, if a second test is considered, in which the group has 75% of positives, then the likelihood that the unknown belongs to it is 75% if it scores plus and 25% if it scores minus. But suppose the unknown scores plus, minus, on the two tests. The chance on the first is 0.2 and on the second 0.25. They take the likelihood considering both tests to be $0.2 \times 0.25 = 0.05$, or only 1/20. In this manner the individual probabilities are multiplied together for as many tests as are available.

As the number of tests is increased the joint likelihood becomes increasingly small for a misidentification. For a correct identification it falls more slowly. We therefore need some expression that takes this into account. One of the most useful is to compare the observed likelihood with the maximum possible for a strain of the given group, that is, a strain that possesses the commonest states for every character. Thus, for our example above, the best we could get was a stain scoring minus on the first and plus on the second test, whose likelihood would be $0.80 \times 0.75 = 0.60$. The unknown which scored plus, minus, would be compared with this maximum, as $0.05/0.60 = 0.083$. This is the model likelihood fraction, or MLF, and approaches one for the best possible identifications. When, for example, it is 0.95 we may take this as indicating roughly a 95% probability that the identification is correct. At some chosen level the computer can be instructed to print the presumptive identification, together with the next best candidate and its probability.

Furthermore, we can ask for the next tests to perform, *ie*, those that will be most efficient in distinguishing between the candidates, or in raising the probability to an acceptable level. On obtaining the results on these we can repeat the identification in the hope of clinching the matter.

The original scheme of Dybowski and Franklin ran into various difficulties. It does of course assume that the joint probability can be obtained as above, and therefore that there is not a significant degree

of correlation between characters. There will certainly be some correlation, but this does not seem to be a serious cause of trouble. Recent experience at the National Collection of Type Cultures with a modified form of this method is much more promising. Drs S. P. Lapage, S. Bascomb, W. R. Wilcox, and M. A. Curtis presented this method at a demonstration meeting of the Society for Applied Bacteriology in October 1968, and I am indebted to them for an outline of their findings. They use a set of 50 tests, but 30 are usually sufficient to identify the 60 groups covering the general run of Gram-negative rods of medical interest that grow aerobically on nutrient agar. If there are only 15 tests then the number of doubtful results rises steeply. This figure of 30 is of some interest, because we would like to know whether it is near the theoretical minimum.

At first sight it would seem that about six positive or negative tests would separate the groups, but of course not all the 64 combinations of six tests correspond to actual groups. It is worth noting that in conventional keys for higher organisms the number of characters, m , required to identify T groups is usually such that m about equals T , so that the ratio m/T is around 1.0. In the tables of Cowan and Steel (1965), which must represent about the most efficient scheme for bacteria without a computer, m/T is on the average 1.2, and is never much less than 1, even for the best known groups.

The ratio obtained by the workers at the National Collection of Type Cultures is around 0.5, and if this should be generally true for computer methods it is a very significant saving in time and effort. However, they did not obtain more than about 50% of correct identifications (at the rather stringent probability level of 99.9%) on the first runs when a free choice of 20 out of 50 tests was allowed to the sender, so a value of $m = 20$ is probably too low. They believe, however, that it may be possible to reduce the 50 tests to about 30 with improved knowledge. (An example of how to measure the diagnostic power of a new test is given in Lapage and Bascomb, 1968.) With $m = 30$ this would still represent a saving of about half the usual number of tests. At all events it seems very probable that we cannot expect to reduce this number much below 30, however powerful the diagnostic method. They believe that some gains are possible by selecting 20 out of the 30 according to the source of the specimen; an opportunity to test this hypothesis has not yet arisen.

Lapage and his colleagues found that with 35 tests they obtained around 80% of identifications (at the 99.9% level), and that increasing this number did not raise the correct identifications very much.

Sometimes a second or third run was needed, but it may be possible to avoid this with better selection of tests. It should be noted that even if only 80% of strains can be uniquely identified as to group this does not mean that the other 20% are badly misidentified, for they are usually allocated to the correct general area. Also, these figures are based on field strains which are largely unselected. (It is not always possible to exclude some selection, and in fact there was rather a high proportion of 'difficult' strains.) When culture collection strains are used, again with some 'difficult' strains, the acceptable identifications rise to over 92%. This itself indicates one possible explanation; there may in fact be around 20% of field strains that are not members of recognized groups but are either intermediate forms or aberrant forms. Certainly there must be some new groups that need defining, but these observations may provide support for the suggestion (Sneath, 1968b) that the pattern of variation in bacteria exhibits many scattered strains around and between the denser clusters that are commonly recognized as distinct groups.

Another point brought out by the study of Lapage and his colleagues is that some tests are very much worse than others as far as reproducibility is concerned. Thus H_2S , gelatin hydrolysis, and nitrate reduction are much less repeatable than acid from glucose or phenylalanine deamination. On the whole, however, the tests are fairly reliable, for there was only about an average of 6.2% of disagreements on repeating the tests on a second occasion, and I have the impression that this is quite good for routine testing methods. Unpublished work of the *Pseudomonas* Working Party of the Society for General Microbiology supports this conclusion.

Conditional probability methods of the kind described above are potentially powerful, but they have some weaknesses. If there is one bad mismatch, *eg*, a plus when the table contains 5%, this will pull down the probability measure a good deal. Yet an aberrancy in one character perhaps should not be treated so harshly. It may be noted that it is multiplying the probabilities which is responsible. With distance models the discrepancies are summed, and the effect of one aberrancy is less marked. Multiplication is equivalent to summing logarithms, so it might be worth considering, too, the summation of other transformations that have a less drastic effect than logarithms. All identification methods will attempt to identify an unknown even if it comes from outside the reference set of organisms, but this should be readily detected in either method.

It is also possible that some of the difficulty may lie in systematic errors that are due to different

times of reading or different growth rates of the bacteria. Some suggestions have been made (Sneath, 1968a) to overcome these effects, and one of the 'pattern' coefficients proposed there may be of use in improving methods of computer identification. Thus, the coefficient D_P might be treated as a distance in distance models, and may give a worthwhile gain in correcting for systemic errors of the kind mentioned above. For example, a slowly growing strain of a group would not appear so different from the reference description of that group as it would if the more usual coefficient D_T were used. To do this, however, D_P would have to be calculated using non-integral values in the 2×2 table, because the reference description would in general contain non-integral values for the characters. This has no very obvious statistical justification, and would have to be justified empirically. The correction obtained by using D_P is greatest (in general) when the reference description has most of its entries near 1 or 0. If they are all close to 0.5 then any coefficient is apt to be misleading.

However, D_P is not directly applicable to conditional probability models, and the principle would need modification, and empirical justification, using logarithmic transformations. This could be done on the following lines. A 2×2 table is set up with the cells labelled *a*, *b*, *d*, *c* in the usual way. Let the logarithms of the frequencies in the reference group be symbolized by *e* for the 1 state and *f* for the 0 state. Then for each character of the unknown *u* the following scoring is used.

If *u* scores 1, then if 1 is the commoner state in the group add *e* to *a*, otherwise add *e* to *c* and *f* to *b*.

If *u* scores 0, then if 0 is the commoner state add *f* to *d*, otherwise add *e* to *c* and *f* to *b*.

The logarithm of the corrected likelihood is then taken as $a + b + c + d$ minus the absolute difference between *b* and *c*.

This provides substantial correction for growth rates, etc, though it is not quite complete, and this is probably desirable. It amounts to excluding a large part of the adverse probabilities due to *u* having more or fewer positive states than the reference group. The differences due to pattern of tests is some what accentuated. It would be necessary to use only characters that can properly be considered positive rather than negative.

Looking to the future, we may ask how computers are likely to be used in clinical bacteriology laboratories. Probably there will be two main ways. Where time is not so important, particularly for difficult problems, the computer might be used to process data sent by post. It would have in store more information than would be readily available

in a laboratory. But quick access is extremely important, both to obtain an answer the same day and also to tell the laboratory what further tests are suggested. For this it would be necessary to have an on-line system, perhaps by a Telex installation, because telephone enquiries would not be practicable except in an emergency.

Another alternative would be to have small computers in the laboratory, and as the cost of desk computers becomes less it may be possible to have one of these actually in the laboratory. Existing desk computers have too small a store. There may be a place for special purpose computers (perhaps of analog type) in the more distant future. There is already interest in machines specially made for identification. Olds (1966) has described one in which light shines through superimposed punched cards in such a way that it indicates those bacteria compatible with the unknown, and he showed an improved model at the Society for Applied Bacteriology demonstration meeting in October 1968. Aberrant test results, however, may cause trouble. The sort of device described by Cowan and Steel (1960) for use with diagnostic tables might be modified to give semiautomatic identification. We should, then, anticipate that routine automatic identification will become possible before very long.

Yet all the techniques so far discussed require the carrying out of a number of bacteriological tests. Is it likely that we might find some entirely different method that required only one operation by the user? Many suggestions have been made, such as infrared absorption spectra or gas chromatography, but none have yet proved very practicable. It is, I suppose, possible that we might exploit the specificity of serological techniques. Surface antigens are too often strain specific, so that others would have to be used. But one could not afford to do large numbers of serological tests on each isolate, so that in some way each antibody in a polyvalent serum would have to be labelled. The obvious way of doing this to exploit position on a surface, by binding antibodies in different patterns and detecting which spots take up the antigen. It will be a long time, though, before we can emulsify a colony in the magic x-reagent, and, on spreading it onto our special antibody slide, see its name outlined in pretty colours!

SUMMARY

Three aspects of automation in clinical bacteriology are discussed: rapid semiautomatic testing methods; numerical taxonomy of bacteria; and computer methods for identification of routine bacterial isolates. The discussion of this last topic is directed toward conditional probability models, in which the unknown isolate is compared with a data matrix of test results of various well defined groups of bacteria. Such methods are capable of giving a high proportion of correct identifications using a smaller set of tests than would otherwise be needed. They can also give some indication of the reliability of identification, together with the best additional tests for distinguishing between those bacterial groups that are most difficult to identify.

REFERENCES

- Baron, D. N., and Fraser, P. M. (1968). Medical applications of taxonomic methods. *Brit. med. Bull.*, **24**, 236-240.
- Cowan, S. T., and Steel, K. J. (1960). A device for the identification of microorganisms. *Lancet*, **1**, 117.
- , — (1965). *Manual for the Identification of Medical Bacteria*. Cambridge University Press, London.
- Dybowski, W., and Franklin, D. A. (1968). Conditional probability and the identification of bacteria: a pilot study. *J. gen. Microbiol.*, **54**, 215-229.
- Floodgate, G. D. (1961). Some remarks on the theoretical aspects of bacterial classification. *Bact. Rev.*, **26**, 277-291.
- Gyllenberg, H. G. (1965). A model for computer identification of micro-organisms. *J. gen. Microbiol.*, **39**, 401-405.
- Hartman, P. A. (1968). *Miniaturized Microbiological Methods*. Academic Press, New York.
- Lapage, S. P., and Bascomb, S. (1968). Use of selenite reduction in bacterial classification. *J. appl. Bact.*, **31**, 568-580.
- Lockhart, W. R. (1964). Scoring of data and group formation in quantitative taxonomy. *Devel. indust. Microbiol.*, **5**, 162-168.
- Lovelace, T. E., and Colwell, R. R. (1968). A multipoint inoculator for Petri dishes. *Appl. Microbiol.*, **16**, 944-945.
- Olds, R. J. (1966). An information sorter for identifying bacteria. In *Identification Methods for Microbiologists, Part A*, edited by B. M. Gibbs and F. A. Skinner. pp. 131-136. Academic Press, London.
- Rypka, E. W., Clapper, W. E., Bowen, I. G., and Babb, R. (1967). A model for the identification of bacteria. *J. gen. Microbiol.*, **46**, 407-424.
- Skerman, V. B. D. (1967). *A Guide to the Identification of the Genera of Bacteria*. 2nd ed. Williams and Wilkins, Baltimore.
- Sneath, P. H. A. (1962). The construction of taxonomic groups. *Symp. Soc. gen. Microbiol.*, **12**, 283-332.
- (1964). New approaches to bacterial taxonomy: use of computers. *Ann. Rev. Microbiol.*, **18**, 335-346.
- (1968a). Vigour and pattern in taxonomy. *J. gen. Microbiol.*, **54**, 1-11.
- (1968b). The future outline of bacterial classification. *Classification Society Bulletin Vol. 1* (no. 4), 28-45.
- and Stevens, M. (1967). A divided Petri dish for use with multipoint inoculators. *J. appl. Bact.*, **30**, 495-497.
- Sokal, R. R., and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman, San Francisco.