The Application of Computers to Taxonomy

By P. H. A. Sneath

National Institute for Medical Research, Mill Hill, London, N.W. 7

SUMMARY: A method is described for handling large quantities of taxonomic data by an electronic computer so as to yield the outline of a classification based on equally weighted features. This enables Similarity to be expressed numerically, and would allow taxonomic rank to be measured in terms of it. An example in bacteria is given, and the results compared with the conventional classification. The method is to count the number of similar and of dissimilar features between strains and to sort the strains into groups whose members have a high percentage of similarities.

The classification of bacteria has reached a point at which there is a need for some mechanical aid to sorting the data. This is because it is difficult to remember the reactions of all the strains in all the tests and therefore one cannot easily separate the strains into valid taxonomic groups. In a discussion on bacterial taxonomy (Sneath, 1957b) four main conclusions were reached. First, that the ideal classification is the one which has the greatest content of information. Secondly, that overall similarity was the basic concept of such ideal classifications and that it is measured in terms of the number of similar features possessed by two organisms. Thirdly, that every feature should have equal weight. Fourthly, that division into taxonomic groups is made upon correlated features. Classifications based on the last three principles (which may for brevity be called Adansonian classifications) will contain the greatest content of information.

A development of these principles is that overall similarity may be estimated numerically, and in this paper an attempt is made to do this by using an electronic computer, and to suggest how overall similarity might be used as a criterion of taxonomic rank. The method has been briefly described in a preliminary communication (Sneath, 1957a). The first step is to convert one's data into a table of features which are scored as present or absent; for this the term 'feature' must be defined and a method of scoring devised. Secondly the table must be analysed so as to give a numerical value for overall similarity. For each of these steps there are a number of alternatives which are equally valid logically, but in general less convenient than the methods used in this paper.

In order to visualize the position the following statistical model may be made. If one considers an individual as possessing a very large number of features (all of which are observable with the techniques at present available) one could divide the features of two such individuals into three classes: (a) those features possessed by the first but not by the second individual, (b) those possessed by both individuals, and (c) those possessed by the second but not by the first individual. The numbers in the classes could be counted,
and shown diagrammatically as below where the numbers of features of the three classes is indicated by the distances, $a$, $b$, and $c$ respectively. In this model every feature has equal weight.

Overall similarity is indicated by the degree to which the two solid lines overlap. It may be represented by the fraction $b/(a+b+c)$, and could theoretically vary from 1.0 where the individuals are identical in all features to practically 0 where very different kinds of living creatures are compared. This ratio may be called Similarity, or $S_t$, which is the measure of overall similarity which is adopted in this paper. It is of course clear that the exact value of $S_t$ will depend on the way in which the term 'feature' is defined, but if the same definition is employed consistently, the values will be comparable. $S_t$ therefore can be looked upon as the best measure of overall similarity which could be obtained with present-day techniques.

Since in practice one cannot do all the possible tests, one must obtain an estimate, $S$, of $S_t$ by taking a sample of observed features and sorting them into classes (a), (b) and (c). Two questions here arise: (1) can one assume that the larger the sample the better this estimate will be, and (2) can one assume that the $S$ values will be fairly stable despite the discovery of new features by techniques which will only be devised in the future? The validity of these assumptions is discussed later, but one may deduce from past experience that the answer to both questions is yes. If it were not so one would find that new discoveries overturned the old classifications, and this does not, I think, generally happen with Adansonian classifications—in fact it is one of the points in favour of them. The relation of one $S$ value to another would be more stable than the $S$ values themselves, as is mentioned in the discussion. Although we cannot obtain absolute and stable values of $S$, but only approximations, this only amounts to saying that a classification must be made with the data which are available, which is true for any method, whether Adansonian or not.

In the method of analysis used here the only features employed are positive features. No account is taken of negative features, i.e. features which are absent from all the organisms under study. This is desirable since the class of absent features is almost infinite, and one cannot make a logical selection of them. Although this does mean that some negative data are not analysed, it does not reduce the information in the classification. For instance, if one states that strains of *Escherichia coli* are indole-positive, and that strains of typhoid bacilli differ from them in the indole reaction, then it is implied that typhoid bacilli are indole-negative.

This paper is concerned with $S$. The other observable ratio, of class (a) to
class (c) is of less interest, but it might give an estimate of the relative size of the genome (or more strictly of the sum of possessed features or 'phenome'), as measured by the ratio \((a + b)/(b + c)\).

**Symbols**

Individuals, or in bacteria individual strains, are indicated by numbers. Groups of individuals or groups of strains are indicated by letters. Subscripts are used to indicate the individuals or the groups which are referred to. A single subscript indicates a value within a group; two subscripts separated by a full stop indicate a value which compares two individuals or two groups.

\(i\) = the number of individuals, or in bacteria the number of strains.

\(m\) = the number of positive features under analysis which are possessed by at least one individual or strain.

For a particular pair of individuals the following symbols are used:

\(n_8\) = the number of positive features possessed by both individuals. It does not include the number of features which are not possessed by either individual.

\(n_d\) = the number of features possessed by the first individual but not by the second plus the number possessed by the second but not by the first.

\(n_f = n_a + n_d\).

\(S\) = Similarity, and is \(n_d/n_f\).

\(D\) = Dissimilarity, and is \(n_d/n_f\), or \(1 - S\).

The symbols \(S\) and \(D\) may be given as decimal fractions or as percentages. The pair of strains referred to is indicated by a double subscript, e.g. \(S_{1,2}\).

\(S\) and \(D\) also stand for Mean Similarity and Mean Dissimilarity when they refer to groups of individuals. If the mean is that within a group there are two ways of obtaining it, as is illustrated by Table 1. The first method gives the 'Square Mean' and is obtained by averaging all the values in an \(i \times i\) table such as Table 1, including the values for an individual compared with itself. The Square Mean \(S_a\) from Table 1 is

\[
(4 \times 1.0) + (2 \times 0.9) + (2 \times 0.8) + (8 \times 0.5))/16 = 0.712.
\]

The second method gives the 'Triangle Mean' and excludes self-comparisons and also the redundant repetitions (e.g. if \(S_{1,2}\) is counted then \(S_{2,1}\) is excluded since it is the same thing) and is the average of the \(S\) values in the outlined triangle. In this example, Triangle Mean \(S_t\) is

\[
(0.9 + 0.8) + (4 \times 0.5))/6 = 0.617.
\]

When the mean is that between strains or between groups there is only one way of averaging the values. In this example \(S_{1,2}\) is 0.9, while \(S_{b,c}\) is \((4 \times 0.5)/4 = 0.5\). The 'Square Mean' is the more useful as it is additive in certain circumstances, and is distinguished here by the prefix capital \(\gamma\), e.g. \(\gamma S\). The 'Triangle Mean' here has the prefix capital \(\delta\), e.g. \(\Delta S\). \(\Delta S\) is lower
than $\Gamma S$ but approaches it when $i$ becomes large. $\Delta S$ is related to $\Gamma S$ by the formulae

$$\Delta S = \frac{\Gamma S_i - 1}{(i-1)}'$$

$$\Gamma S = \frac{\Delta S(i-1)+1}{i},$$

$\Delta D$ and $\Gamma D$ show complementary properties, which may be derived from the relation $D = 1 - S$.

Table 1

$S$ values for strains 1—4, which together form group a. The group is divisible into subgroup b which consists of strains 1 and 2, and subgroup c consisting of strains 3 and 4.

<table>
<thead>
<tr>
<th>Strains</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.9</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>0.5</td>
<td>1.0</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
<td>0.5</td>
<td>0.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**Definition of the term ‘feature’**

If one accepts the thesis that the ideal scientific classification is the one which contains in its divisions the greatest amount of information, then one may logically define the word feature in terms of the number of statements which can be made about an organism. The working definition adopted here is that one feature is a thing about which one statement can be made, but that not all features can be counted when creating a classification.

To be counted the statement must be based on observed data and not merely assumed, otherwise it leads to circular reasoning of the kind ‘I assume this creature has a mammalian heart and therefore the nature of its heart supports my contention that it is a mammal’.

To be counted the statement must add new information about the organism. Therefore it must not be: (1) A tautology. (2) True by definition, e.g. ‘penicillin is the substrate of penicillinase.’ (3) True because it describes a property which is known to be obligatory, e.g. ‘haemoglobin is red’. (4) A redundant statement which is implied by statements listed previously, e.g. ‘growth occurs at 37°’ is redundant if one has listed ‘growth occurs from 12° to 44°’. (5) A combination of these.

Only positive features can be compared. A positive feature is defined as one which we must assume to be present in order to explain an observation. A negative feature is one for which we have no evidence, i.e. the observations may be explained by the absence of a feature. In bacteriology a negative feature will generally be one which gives the same result as a control test omitting the bacterium; indole-negative, for instance, is a negative feature.
Computers in taxonomy

since it may mean that the bacterium possesses no genes or enzymes which are in any way concerned with indole formation. There are few features which cannot be logically sorted into these two categories. The most difficult are those describing the toxicity of drugs, etc., in which several logical alternatives are possible: these are discussed in the appendix. If two organisms both possess the same positive feature this counts as one similarity; if one possesses it and the other does not this counts as one difference; if neither possesses it this does not count at all. This is logical, since it is not pertinent to count 'lack of feathers' as a feature when comparing bacteria with one another, but it would be pertinent in comparing a bacterium with a bird. If one once admits comparisons between negative features one does not know where to stop.

Reduction of bias

Although the observed features of an organism cannot be proved to be unbiased with respect to \( S \), one should not introduce deliberate additional bias by using a test which was selected because it readily separates the group of organisms under study. Neglect of this could (in theory) lead to ridiculous results. For example, if one were obsessed with the notion that the typhoid bacillus was in no way related to the colon bacillus one could spend one's life looking for new tests which separated them, and one might obtain a few such tests by trying out a large number of new tests. An analysis with the selected tests would show no similarity at all between the two bacteria. Clearly, one should either exclude all the tests or include all of them. However, there is no reason for assuming that the selected tests would necessarily be biased when studying a distant family such as the micrococci. Usually there is no record in the literature of the tests which were tried out to discover a test of this kind. One should therefore exclude a test which was devised in order to divide the group under study, or whose sensitivity was deliberately adjusted to this end, but may employ it when studying other bacteria. The same objection applies to tests which were devised in order to unify the group under study: e.g. the urease test adjusted so as to be positive only with Proteus strains is inadmissible when studying Proteus. The number of such tests is not great, and their inclusion might not in practice matter very greatly. The number of tests used must be adequate. The accuracy of \( S \) is roughly indicated, though in general underestimated, by its standard deviation \( \sqrt{(S \times D)/n_r} \). Ideally one should employ all the unbiased tests one can, but if one must make a selection of them, it seems wisest to pick a random selection, taken from many different classes such as morphology, carbohydrate fermentation, etc.

The programme of analysis

A simple example of the way in which the value for \( S \) is obtained is shown below.

Three bacteria have the following 'IMVIC' reactions:

<table>
<thead>
<tr>
<th>Strain</th>
<th>Indole</th>
<th>Methyl Red</th>
<th>Voges-Proskauer</th>
<th>Citrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain 1</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Strain 2</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Strain 3</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
On comparing a pair of strains one first lists the number of tests in which both are positive. Thus strains 1 and 2 are both positive in two tests, while strains 1 and 3 are both positive in the Methyl Red test only. These comparisons are made for all possible ways of pairing the strains, and the resulting table of \( n_p \) values is as follows. Note that where both strains are negative, this is not included in finding \( n_p \).

<table>
<thead>
<tr>
<th></th>
<th>Strain 1</th>
<th>Strain 2</th>
<th>Strain 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain 1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Strain 2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Strain 3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

One then compares the strains and lists the number of tests in which one strain is positive and the other strain is negative. Thus, strains 1 and 3 show two differences (the indole and the citrate tests). These figures give the table of \( n_d \) values shown below:

<table>
<thead>
<tr>
<th></th>
<th>Strain 1</th>
<th>Strain 2</th>
<th>Strain 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Strain 2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Strain 3</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

The value of \( n_d/(n_p+n_d) \) can then be obtained for each way of pairing the strains, giving the table of \( S \) values.

<table>
<thead>
<tr>
<th></th>
<th>Strain 1</th>
<th>Strain 2</th>
<th>Strain 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain 1</td>
<td>2/2</td>
<td>2/2</td>
<td>1/3</td>
</tr>
<tr>
<td>Strain 2</td>
<td>2/2</td>
<td>2/2</td>
<td>1/3</td>
</tr>
<tr>
<td>Strain 3</td>
<td>1/3</td>
<td>1/3</td>
<td>2/2</td>
</tr>
</tbody>
</table>

The \( S \) values are conveniently expressed as percentages:

<table>
<thead>
<tr>
<th></th>
<th>Strain 1</th>
<th>Strain 2</th>
<th>Strain 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain 1</td>
<td>100</td>
<td>100</td>
<td>33.3</td>
</tr>
<tr>
<td>Strain 2</td>
<td>100</td>
<td>100</td>
<td>33.3</td>
</tr>
<tr>
<td>Strain 3</td>
<td>33.3</td>
<td>33.3</td>
<td>100</td>
</tr>
</tbody>
</table>

Part of the tables of \( n_p \), \( n_d \) and \( S \) is redundant, and may be omitted, since it is obvious that any strain when compared with itself will have \( S = 100 \% \) while the upper right-hand part of the table is a 'mirror image' of the lower left-hand part. Of course no useful results can be expected from considering only four tests, but the same method is used for larger numbers of features.

**Preparation of data.** The data are tabulated with features in the columns and individual strains in the rows (an \( i \times m \) table). Strains are scored as + indicating presence of the feature, or as − indicating its absence. Where no datum is available the space is marked NC (standing for 'No comparison').

**Sorting of quantitative data.** It is difficult to score quantitative data in an entirely logical fashion, and an empirical method has to be used. The basic principle is to assume the existence of only such features as are necessary to explain the experimental findings. The quantitative results are first inspected to see how many steps can be clearly distinguished in the material under study by the experimental method, so that one can be confident that there is a real difference between strains in one category and strains in another category, after allowing for biological and experimental error. Strains which are intermediate between two categories, which will occur if there is a con-
Computers in taxonomy

A continuous curve of variation, must perforce be allocated arbitrarily, but one trusts to the large number of features under analysis to prevent systematic errors from this cause. One must now find the numbers of similarities and differences which are implied by the categories. To do this, it is first necessary to make one of two assumptions, either that features are additive in their properties (which would seem to be a basic assumption in using correlation coefficients), or that they are not. Unless one can decide between these alternatives from the data, which will be very rarely possible, this choice is necessarily an arbitrary one. I feel that it is best to assume that features are not additive, and this leads to the following method of scoring.

Each step is assumed to be under the control of a feature which determines the quantity of the characteristic, and which is different from the features determining the quantity of the other steps. However, all those strains which give a detectable reaction at all resemble one another in this respect. Each such strain must therefore possess a minimum of two features (or two subfeatures, which by the working definition of feature, may be regarded as distinct features)—one allowing expression of the characteristic, and one determining the quantity. An example is as follows. Five strains produce respectively no detectable indole, 5, 10, 10.5 and 20 μg./ml. of indole under defined conditions. The difference between 10 and 10-5 μg./ml. is considered to be of no significance because of the errors of the test. The first strain is allotted no features. Strains 2–5 are allotted a common feature, A, for indole-production, and are allotted features B, C and D, for the amount of indole which is produced. The constitution of the strains is provisionally written as follows:

<table>
<thead>
<tr>
<th>Strain</th>
<th>Constitution</th>
<th>(μg./ml.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(producing no indole)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>AB (producing 5 μg./ml.)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AC (producing 10–10.5 μg./ml.)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AC (producing 10–10.5 μg./ml.)</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AD (producing 20 μg./ml.)</td>
<td></td>
</tr>
</tbody>
</table>

One has, however, no reason for assuming the existence of feature B, since one could not distinguish, in this instance, between a strain which was AB and a strain which was A. One therefore deletes feature B. We have also assumed that the effect of features is not additive, and it is therefore impossible to distinguish between AD and ACD as the constitution for strain 5; that is, the feature C cannot be scored either as positive or negative in strain 5, and must be scored NC. Also, strain 1 does not, by definition, possess feature A, and therefore it is meaningless to ask whether it has features C or D, so that these two features must be scored NC. The strains are then coded as follows:

<table>
<thead>
<tr>
<th>Features</th>
<th>Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>+</td>
</tr>
</tbody>
</table>
Although C is not detectable in strain 5, yet one can say that D is absent from strains 2-4, because of the way we have defined the features. Similarly, both C and D are absent from strain 2.

This scheme (which is referred to in the Appendix as method C) fulfils two necessary conditions. First, that the $S$ values between a pair of strains are the same however many steps are distinguished: this can readily be seen by extending the steps (so that a sixth strain, for instance, producing 50 µg./ml. indole, this quantity being controlled by a new feature E, would score +, NC, NC, + against the features A, C, D, and E respectively). Secondly, that when the test is turned into a purely qualitative test, the strains will score in the same manner as they would have if the test had never been made a quantitative test. For instance, in the above example, if the division between indole negative and indole-positive lies at 7.5 µg./ml. when using a qualitative indole reagent, then the strains score as follows:

<table>
<thead>
<tr>
<th>Indole</th>
<th>Indole</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strain 1</td>
<td>–</td>
</tr>
<tr>
<td>Strain 2</td>
<td>–</td>
</tr>
<tr>
<td>Strain 3</td>
<td>+</td>
</tr>
<tr>
<td>Strain 4</td>
<td>+</td>
</tr>
<tr>
<td>Strain 5</td>
<td>+</td>
</tr>
</tbody>
</table>

This is because we are now only able to observe the presence of AC or of AD, and we have therefore no warrant for assuming the separate existence of feature A, and we cannot distinguish between features C and D: the feature ‘indole’ thus means ‘C or D’. This, the simplest way of scoring a test, is referred to in the Appendix as method A.

There are a few tests in which one should retain the feature called B in the above example. This is where there cannot be a strain which is A negative—for instance in coding the length of a bacterium. This is referred to in the Appendix as method B.

Conversion of data. The data in the large $i \times m$ table are now punched on to teleprinter tape or cards in the manner required for the computer which is used.

Method of analysis. The method given below is converted into appropriate instructions for the computer, and then forms the ‘computer programme’, which is fed in with the data. The steps are:

1. The $i \times m$ table is stored.
2. Each pair of strains in all possible combinations is compared in turn, and the number ($n_s$) of similar features in which both strains of the pair score +, and the number ($n_d$) of different features in which one strain scores + and the other – are counted. The machine makes no comparison of a – with a –, nor of either a + or a – with a space marked NC. The value of $S$ is calculated (being $n_s/(n_s+n_d)$) for each pair.
3. The $S$ values are tabulated in an $i \times i$ table, and this table may be printed for checking. It need not be the full table, but can be the ‘non-coincidental triangle’ referred to under Triangle Mean in the section on symbols.
4. The $i \times m$ table is forgotten, and only the table of $S$ values is used.
(5) The machine sorts the strains into groups, working in small steps of $S$. Any pair of strains with $S$ values of over 0.99 are placed together, and these groups (each consisting of almost identical strains) are printed.

(6) The machine repeats step 5, working at $S$ values of over 0.98, and adds any strains (which have an $S$ value to any strain already grouped and printed of more than 0.98) to the earlier groups, and prints them. It also creates and prints any new groups it finds.

(7) The machine repeats step 6 for $S$ values of 0.97, 0.96 and so on until all strains have been grouped and printed, when it is instructed to stop.

(8) The results indicate the groups which are present and roughly their relationships. One might for instance obtain a record:

<table>
<thead>
<tr>
<th>$S$</th>
<th>Strains</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.99</td>
<td>1, 2</td>
</tr>
<tr>
<td>0.98</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>0.97</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>0.80</td>
<td>1, 2, 3, 4, 5</td>
</tr>
</tbody>
</table>

This record shows that strains 1, 2 and 3 form a group (a), and that strains 4 and 5 form another group (b). The two groups are not very similar, since they do not fuse into one until the $S$ value falls to 0.8.

(9) The machine may now be asked to calculate intragroup and intergroup Mean Similarity values and to print the $S$ value table in a rearranged form. The Mean $S$ values give a truer picture than those indicated in step 8, since the machine is sorting strains by the smallest differences between strains and not by Mean values. The example above might yield the better values

$$\text{Mean } S_a = 0.988$$
$$\text{Mean } S_b = 0.981$$
$$\text{Mean } S_{a,b} = 0.781$$

$S_a$ and $S_b$ will be either $\Gamma S$ or $\Delta S$ values according to the instructions.

These results may now be converted into the more familiar and understandable form of a taxonomic tree. We might for instance conclude that Groups (a) and (b) are sufficiently similar (as indicated by $S_{a,b}$) to belong to one genus, but that they are different species, viz:

Genus X

Species Y (Group a)  Species Z (Group b)

It must be pointed out that it is impossible to summarize an $S$ value table without loss of some information. This is because the $S$ values are quantities in multidimensional space, and the sorting procedure is forming hierarchical groups in a few dimensions. Nevertheless, it is impossible to remember tables of $S$ values, and if anyone else is to understand or to use the classification, these hierarchies must be formed, and one must strike a balance between convenience and accuracy.

This procedure clearly measures overall similarity based on features of equal weight: in addition, the divisions are based on correlated features.
Results of the worked example given in the appendix

The appendix contains the details of the analysis of a study of 45 strains of bacteria. Thirty-eight of these were strains of the genus *Chromobacterium*, which had previously been found to belong to two groups, a mesophilic group (18 strains) and a psychrophilic group (20 strains) (Sneath, 1956b). The remaining strains were not of this genus but were included for comparison.

The sorting procedure (steps 5-8 above) correctly sorted the strains into the expected groups as shown below.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group a</td>
<td>18 mesophilic strains of <em>Chromobacterium</em></td>
</tr>
<tr>
<td>Group b</td>
<td>20 psychrophilic strains of <em>Chromobacterium</em></td>
</tr>
<tr>
<td>Group c</td>
<td>2 strains of <em>Serratia marcescens</em></td>
</tr>
<tr>
<td>Group d</td>
<td>2 strains of <em>Pseudomonas aeruginosa</em></td>
</tr>
<tr>
<td>Group e</td>
<td>1 strain of 'Chromobacterium viscosum'</td>
</tr>
<tr>
<td>Group f</td>
<td>1 strain of 'Chromobacterium iodinum'</td>
</tr>
<tr>
<td>Group g</td>
<td>1 strain of 'Chromobacterium ianthinum'</td>
</tr>
</tbody>
</table>

This may be seen from Fig. 1, which is a graphical representation of the sorting procedure. It is seen that at the level of $S=100\%$, no groups are formed (or alternatively, one might say that every strain formed a separate group containing only itself). At $S=95\%$ the two strains of *Serratia* form a group. At $S=98\%$ two of the mesophilic strains, MK and DK, form a group to which other mesophilic strains are added as $S$ decreases. Similarly, at $S=92\%$ the psychrophilic group begins to form. Strain FH is evidently an aberrant strain of group $a$, and strain NT an aberrant strain of group $b$, since they are added at low values of $S$.

The groups $a$ and $b$ remain distinct until $S=64\%$, when they fuse together by virtue of the pair SH.HD (which has $S=64\%$), the highest value of $S$ for any comparison of a mesophil with a psychrophil. A little below this, groups $c$ and $d$ fuse with groups $a$ and $b$, as does the group $g$ represented by only the single strain TI. The two diphtheroid-like organisms of groups $e$ and $f$ only fuse with the rest at the level of $S=44\%$; this is expected, since they appear to be distantly related to the other bacteria studied. They are, in addition, not very similar to one another, since for the comparison GR to TE the $S$ value is $65\%$. A cleft between two groups in Fig. 1 shows that the groups are completely distinct at $S$ levels above that of the bottom of the cleft, since the bottom of the cleft indicates the highest $S$ value found in comparing strains of one group with strains of the other.

The full table of $S$ values (Step 3 above) was rearranged, so that the strains were in the same order as they are in Fig. 1: this table is shown in the Appendix (Table 7 and diagrammatically in Fig. 4). This full table does not force any major modifications of the conclusions reached from Fig. 1. Only minor points emerge—one might for instance judge that Fig. 1 over-emphasizes the aberrancy of strain FH and under-emphasizes the aberrancy of strain AM.

A few subgroups may be noted—thus strains MK, DK, TB, TA, TV, and LG and strains RT, SH, and LW form subgroups of group $a$, and similar sub-
Fig. 1. Diagram of the results of the sorting procedure on the forty-five strains used in the worked example.

Table 2. Mean % of $S$ values for the groups a–g

The $S$ values in the squares are those of the combination implied from the groups tabulated in the row and column. Thus $S_{a,c}$ is 40.1%. The value for a group compared with itself is the $T$ value.

<table>
<thead>
<tr>
<th>Group</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>77.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>51.5</td>
<td>75.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>54.9</td>
<td>40.1</td>
<td>97.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>52.3</td>
<td>52.7</td>
<td>54.0</td>
<td>91.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>40.0</td>
<td>32.4</td>
<td>32.5</td>
<td>39.0</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>34.3</td>
<td>30.0</td>
<td>29.5</td>
<td>37.0</td>
<td>63.0</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>52.4</td>
<td>52.5</td>
<td>44.0</td>
<td>61.5</td>
<td>38.0</td>
<td>42.0</td>
<td>100</td>
</tr>
</tbody>
</table>
groups are seen in group b. One might consider these to be rather indistinct varieties or subspecies. The suggested neotype strains (Sneath 1956a, b) MK and HB are typical members of their group.

The sorting procedure is clearly making sense and not nonsense, but there is one important difference between these results and the conventional classification. One sees from the mean $S$ values for the groups shown in Table 2 (and it is also apparent in Fig. 1) that the groups a, b, c and d appear to be about equally related to one another. This is shown diagrammatically below:

That is to say, the two groups of *Chromobacterium*, groups a and b which have been considered by most authors including myself (Sneath, 1956a, b) to constitute two species within one genus, are as distantly related as *Pseudomonas* is to *Serratia* which are usually regarded as belonging to different families (the Pseudomonadaceae and Enterobacteriaceae respectively). Leifson (1956) has already pointed out that the species of *Chromobacterium* are very different, and has suggested that they might be best raised to generic rank. However, one could also argue that the families Pseudomonadaceae and Enterobacteriaceae should be reduced to the rank of genera. These two families may not be as distinct as is commonly assumed, for there are an increasing number of reports of organisms which are intermediate between the typical pseudomonads and the typical Enterobacteriaceae (e.g., Leifson & Hugh, 1958; Crawford, 1954; Asai, Okumura & Tsunoda, 1956; Sreenivasan & Venkataraman, 1956), and the two families are now beginning to look like arbitrary categories based only upon whether the flagella are polar or peritrichous. This feature itself is less simple than was thought, since Leifson (1956) and Sneath (1956c) have shown that *Chromobacterium* sp. may have both polar and peritrichate flagella.

It seems to me that bacteriologists are apt to assume that two bacteria which differ in a striking feature such as their pigment are also different in many other uninvestigated features, and that this is the reason why *Chromobacterium*, *Pseudomonas* and *Serratia* are considered to be distantly related. But this assumption is quite unproved. It is possible that a variant of *Pseudomonas* which produces violacein, or a strain of *Chromobacterium* which produces fluorescein, might be discovered, which would then suggest that the two genera were very closely related. Similarly, the tacit assumption that the type of flagellation reflects a fundamental difference between the Enterobacteriaceae and the Pseudomonadaceae (or more accurately that there are many features correlated with flagellation), is unproved, and is now looking unsafe.

The example given here may not be a very good test of the method, since
sharply defined groups of bacteria were used, and the least one would expect is that the method should separate them. The example does, however, suggest that the method does not disrupt sound taxa.

DISCUSSION

Probably the most controversial point in the method described here is the dictum that every feature should have equal weight. This is, however, not a matter for experimental proof; it is for taxonomists to decide whether they want the kind of classification which is given by equally weighted features. The viewpoint taken here is that it is impossible in practice to decide on relative weight, and this avoids much sterile argument.

The definition of the term feature is a makeshift. It may be possible in the future to re-define it in terms of genes and perhaps nucleotides; this will not effect the basic concepts of Adansonian methods but may simplify them.

The $S$ values are analogous in some respects to the 'coefficient of racial likeness' which has been developed in anthropology by Karl Pearson and others (see discussion by Rao, 1948). In the field of entomology Sturtevant (1939, 1942) made an analogous approach to the problem of classification, and Dr R. R. Sokal informs me that he has developed a similar method (Sokal, 1956); Sokal & Michener, 1957a, b).

Stability of $S$ values

Reference was made earlier to the question of whether one obtains a better estimate of $S_1$ as one increases the number of tests under analysis. This is not, as at first sight, a simple question. The tests may be biased because they contain a high or a low proportion of tests in class (b) (see p. 201), and this can occur by the accidents of history. For example a school of workers might develop many simple tests for biosynthetic pathways of the amino acids, and in a group of prototrophic bacteria most strains would be positive in every test. This would raise the $S$ values in the entire group. A further development might reduce the $S$ values again (e.g. by studying the antigenic differences of the enzymes concerned). One cannot prove that either estimate was the better, although by definition $S$ will equal $S_1$ when all possible tests are performed. This means in practice that one cannot extrapolate an observed tendency to obtain $S_1$. Similarly, new technical developments will cause shifts in $S$ values and one cannot extrapolate to obtain $S_1$ values that might be obtained with tests that will be invented in the future. The $S$ values are useful nevertheless because they are the best estimates available at the time.

It would be inconvenient if $S$ values should fluctuate too widely. However we would expect some constancy in the relations between $S$ values from different analyses since it is usual to find that organisms which are similar in a given set of features are also similar in other features. If this were not true, one would observe individuals with all possible combinations of features—birds with fur and such-like chimeras. One does not observe this even in bacteria (though this has been disputed) since we do not find streptococci
causing tetanus nor salmonellas causing diphtheria: it is also evident that some of the possible combinations would cause inviability.

Thus, if one had three strains of typhoid bacillus, strains 1–3, and analysed them using a hundred current tests one might obtain the following S values—$S_{1.2} 90\%$, $S_{2.3} 80\%$, $S_{1.3} 72\%$. Now if new tests were devised for a hundred biosynthetic enzymes, and a new analysis was made including these, one might observe that strain 1 had all of these. If this were so then I think that one would find that strains 2 and 3 also possessed all the enzymes, or the great majority of them. The second analysis would then give $S_{1.2} 95\%$, $S_{2.3} 90\%$ and $S_{1.3} 86\%$. In this particular instance the relative D values are the same, since $D_{1.2}$, $D_{2.3}$ and $D_{1.3}$ are respectively 10, 20 and 28% on the first analysis and 5, 10 and 14% on the second analysis. This might be commonly true in practice, but when it was not, one would still be able to calibrate one scale of S values in terms of another with reasonable accuracy provided there were several strains which had been studied in both analyses.

**Taxonomic rank**

If one bases taxonomy upon the concept of overall similarity it is illogical to base taxonomic rank upon other criteria. The rank of species is commonly defined in terms of interbreeding, geographical distribution, divisibility into subspecies, etc., but these criteria are usually not applicable to higher ranks of taxa. Similarity is applicable at all levels of rank. It is true that the use of similarity to define taxonomic rank would cause some changes. For instance, a group which is presently called a species might become a genus, but this is only a different way of stating the same thing; instead of being a species (i.e. a group showing free interbreeding) which is very variable in its features it would now be called a genus (i.e. a group showing much variability) which shows free interbreeding. In bacteriology our criteria for taxonomic rank are so poor that a great simplification would result from basing rank upon similarity, as expressed by S. By ignoring comparisons between negative features, the S values for most bacteria will not be altered by the discovery of bacteria with very unusual positive features, nor will they be affected by positive features which are absent throughout bacteria, such as feathers.

In order to use S values as criteria of rank one must select a few type strains as reference points for the ranks, and one can then decide the rank of other strains and taxa. Thus if one takes a strain of *Bacillus anthracis* and a strain of *Bacillus subtilis* and says that they belong to the same genus but to different species one can decide from the S values whether other strains should be regarded as species, subspecies or varieties. These type strains would then be used in subsequent analyses as reference points. This would provide a practical solution to the problem of reaching general agreement on the criteria of taxonomic rank in bacteria. Clearly one of the first steps is to make an analysis using a few strains from each family of bacteria (including wherever possible the type strain of the nomenclatural type.). This would provide a skeleton on which to build other analyses, and would hasten the international agreement of type strains of the better-known bacteria, a reform which is overdue.
Computers in taxonomy

Practical applications in bacteriology

The massive data now accumulating in the records of reference laboratories will never be adequately analysed without mechanical aid. These data can be analysed in a 'scientific' manner if the machine is correctly instructed. It is true that the data must be copied into large tables or typed on to computer tape or cards, but this is a necessary preliminary to any form of analysis, and there is no reason why they should not be compiled piecemeal and then joined together in the right order. Once a machine is programmed to the problem it can analyse subsequent examples without delay, and the computer programme can be readily translated into the code required for a different computer. The only limitations are where the problem is basically different, or where the data are too numerous for the machine to handle. If the number of tests is small then the data will not be worth analysis however numerous are the strains, but bacteriologists could build up suitable data by taking a random selection of their strains for a bigger testing routine.

Antigenic analysis would provide a great number of possible features. The data must be codified in terms of antigens whose presence is warranted by the serology, i.e. numbered or lettered antigens. These may however be biased features, since in many instances the antisera have been cross-absorbed so as to make them efficient at revealing antigenic differences within a particular taxon. It is not permissible to include features which are just thought to be likely, e.g. R antigens in Salmonella are not demonstrated in routine work, and one is not entitled to assume their presence in all the strains. This rule is important, since this method of analysis is intended to give the best estimate of overall similarity which is possible from the data available, and therefore guesswork must be excluded. Admittedly, a shrewd taxonomist might make a better estimate by guessing, but much of present-day bacterial taxonomy is an art based on guesswork, conscious or unconscious, and the results have not been encouraging.

One other point requires emphasis. The analysis is intended to produce taxonomic groups and not to identify the groups. It is therefore not necessary to employ tests which are standardized for use with all bacteria, though they must of course be standardized throughout the group under study. Different programmes could employ very diverse selections of features and still yield useful values of S. The tests most useful for identification may be obtained by inspection of the data and included in the description of the groups which were distinguished by the analysis.

The method suggested in this paper is applicable to any group of living creatures. Although the data would often be less easy to codify there is no difference in essentials. In higher organisms one can form, by looking at morphology, an intuitive idea of the overall similarity, which makes their classification much easier than that of bacteria, but one could use the computer to obtain numerical values if one wished. For instance a collection of thrushes, blackbirds and ostriches would be correctly sorted by the computer into these three groups; precision could be given to the vague statements
that the first two are very similar birds but ostriches are very different by estimating Mean Similarities, such as for thrushes versus blackbirds $S_{t,b} = 0.92$, for thrushes versus ostriches $S_{t,o} = 0.62$ and for blackbirds versus ostriches $S_{b,o} = 0.68$.

For bacteria and viruses the method could be widely used, but it would mean that workers would have to study a few strains at a time very thoroughly and build up the classification piece by piece. It would be ‘classification from below’ (White, 1937), and though slow it would, I believe, be sure.

I am particularly indebted to Drs J. Mandelstam and S. T. Cowan for very helpful discussion and comments, and to Mr R. G. Mills of Elliott Brothers for much help with the computer programme.

REFERENCES


(Received 26 February 1957)
APPENDIX

Details of analysis of the example

The data are from protocols used in a study of the genus *Chromobacterium* and most have been already published, together with details of the techniques (Sneath, 1956b). In some tests the reactions of every strain were not given, but they can be obtained from Figs. 2 and 3 and the details given below.

The data were entered in a table of strains versus features (an \(i \times m\) table) in the form of entries of +, — or NC. This table in a modified form (see below) is given in Figs. 2 and 3. Features which were negative with all of the 45 strains were omitted, and a few data which could not be easily scored, such as the source of the strains, were also omitted. Conscious bias was avoided as far as possible by the following procedures. (1) Any test which had been devised, or selected by trial, or adjusted in sensitivity, because it readily separated or united some of the taxa under study was excluded. (2) Standard techniques were used wherever possible (e.g. methylene-blue reduction). (3) Where observations were made at various times (e.g. gelatin liquefaction) the scoring was based, when possible, on the time required to obtain a definite positive result. (4) Where observations were made at different times or by different techniques but it was not possible to score them as in (3) (e.g. the size of cells after 18 and 96 hr. incubation are dependent on one another to some extent, but one cannot tell to what extent), then one time or one technique was selected at random by coin-spinning, and only these data were used. All tests were performed at 25°C unless otherwise stated.

One hundred and five features were used. Quantitative characteristics were arbitrarily divided into a few steps for scoring. The letters after the feature numbers refer to the method of scoring described on p. 208.

*Morphological characteristics (features 1–24)*

All are from nutrient agar plate cultures after 18 hr. incubation.

Length of organisms. Features 1–5(B) scored by method B as shown below:

<table>
<thead>
<tr>
<th>Feature</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 2(\mu)</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2 to 2.9(\mu)</td>
<td>+</td>
<td>NC</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 to 3.9(\mu)</td>
<td>+</td>
<td>NC</td>
<td>NC</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Over 4(\mu)</td>
<td>+</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>+</td>
</tr>
</tbody>
</table>

Breadth of organisms. Features 6–9(B). Scored in the same way as for length of organisms, with three steps (up to 0.7\(\mu\), 0.7–0.9\(\mu\), over 0.9\(\mu\)).

Rounded ends. Feature 10(A).

Bulged or clubbed ends. Feature 11(A).

Chains of organisms readily seen. Feature 12(A).

Palisade arrangement of organisms. Feature 13(A).

Marked curvature of rods (if length was less than 2\(\mu\) this could not be estimated and was scored NC). Feature 14(A).
Pleomorphic forms prominent. Feature 15 (A).
Presence of polar flagella. Feature 16(A).
Presence of lateral flagella. Feature 17(A).
Tufts of polar flagella (i.e. lophotrichous arrangement). Feature 18(A). If there were no polar flagella this was scored NC.
Prominent barred or bipolar staining. Feature 19(A).
Gram-positive staining. Feature 20 (A). (Gram-positive granules were ignored.)
Capsule visible by Hiss's stain. Feature 21(A).
Microscopic fat in the organisms by Burdon's stain. Features 22–23(C).
This was scored as below:

<table>
<thead>
<tr>
<th>Feature</th>
<th>22</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat scanty or none</td>
<td>-</td>
<td>NC</td>
</tr>
<tr>
<td>Moderate fat</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Abundant fat</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Metachromatic granules (i.e. blue granules) clearly visible by Neisser's stain. Feature 24(A).

Cultural characteristics (features 25–40)

I. Nutrient agar plate cultures.

Gelatinous consistency of colonies after 7 days' growth. Features 25–26(C).
Graded into none, slight, and marked.
Opaque colonies after 2 days' growth. Feature 27 (A) (transparent or semi-transparent colonies scored as -).
Degree of production of the pigment violacein after 7 days. Features 28–29 (C). The scoring was by method C modified as follows since 'detectable violacein' is scored under Feature 68.

<table>
<thead>
<tr>
<th>Features</th>
<th>28</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moderate violacein</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Abundant violacein</td>
<td>NC</td>
<td>+</td>
</tr>
</tbody>
</table>

II. Colony form in gelatin plate surface cultures after 4 days at 20° (the scoring is arbitrary)

Smooth Surface. Feature 30(A).
Lobate or erose edge. Feature 31 (A).

III. Other media

Rate of liquefaction in gelatin stab cultures at 20°. Features 32–35(C).
Divided into five steps—no liquefaction in 14 days, liquefaction within 14 days, within 8 days, within 5 days and within 2 days. The last four steps correspond to features 32, 33, 34, 35 respectively; the first step is scored as negative for feature 32 and NC for the others.
Degree of digestion of Löffler serum slope in 6 days. Features 36–37(C).
Divided into slight or none, moderate and marked.
Computers in taxonomy

Ability to grow on potato slope. Feature 88(A).
Degree of violet pigment on potato slope after 14 days. Features 39–40(C).
Divided into none, slight and marked.

Resistance tests (features 41–47)

The coding of these tests is arbitrary, since there are a number of logical alternatives. An organism which is killed by a certain treatment must possess at least one characteristic (and one assumes only one if there is no evidence to the contrary) which is both essential to life and sensitive to the treatment. These two properties may be regarded as sub-features, and called $E$ and $S$: a sensitive organism is then $E + S +$. It seems most logical and convenient to consider them as positive features. A resistant organism may be assumed to lack one of the subfeatures, so that it is either $E - S$ NC, indicating that the sensitive characteristic is not essential, or $E$ NC $S -$ indicating that the characteristic may be essential but is not sensitive to the treatment. If one uses the first convention, then one can code the strains by method C, in which a number of different degrees of sensitivity are indicated by $S_1$, $S_2$, etc., and this seems the simplest method, e.g.

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
</tr>
<tr>
<td>Resistant strain</td>
</tr>
<tr>
<td>Moderately sensitive strain</td>
</tr>
<tr>
<td>Highly sensitive strain</td>
</tr>
</tbody>
</table>

Heat resistance. Features 41–48(C). Viable cells in one loopful of a broth culture after heating at 56° for 5 and 15 min. scored as follows:

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
</tr>
<tr>
<td>No viable cells after 5 min.</td>
</tr>
<tr>
<td>Viable cells after 5 min. but not after 15 min.</td>
</tr>
<tr>
<td>Viable cells after 15 min.</td>
</tr>
</tbody>
</table>

Phenol resistance: viable cells in one loopful of broth culture containing 1% (w/v) of phenol after 5 and 10 min. at 20°. Features 44–46(C) scored as below:

<table>
<thead>
<tr>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
</tr>
<tr>
<td>No viable cells after 5 min.</td>
</tr>
<tr>
<td>Viable cells after 5 min. but not after 10 min.</td>
</tr>
<tr>
<td>Viable cells after 10 min.</td>
</tr>
</tbody>
</table>

Penicillin sensitivity: growth inhibited for 2 days by 100 units/ml. of sodium benzylpenicillin in nutrient agar plates. Feature 47(A).

Metabolism and nutrition (features 48–70)

Anaerobic growth: visible growth after 4 days on anaerobic nutrient agar plate. Feature 48(A).
Upper temperature limit for growth: 7 days on moist nutrient agar slopes,
except for readings at 44° which were made after 4 days. Features 49–53(B). This test is scored by method B, since the characteristics concerned must all resemble one another in that they are all connected with growth at high temperatures. The scoring was as follows:

<table>
<thead>
<tr>
<th>Features</th>
<th>49</th>
<th>50</th>
<th>51</th>
<th>52</th>
<th>53</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visible growth at 25°</td>
<td>+</td>
<td>+</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
</tr>
<tr>
<td>Visible growth at 30°</td>
<td>+</td>
<td>NC</td>
<td>NC</td>
<td>+</td>
<td>NC</td>
</tr>
<tr>
<td>Visible growth at 35°</td>
<td>+</td>
<td>NC</td>
<td>NC</td>
<td>+</td>
<td>NC</td>
</tr>
<tr>
<td>Visible growth at 44°</td>
<td>+</td>
<td>NC</td>
<td>NC</td>
<td>NC</td>
<td>+</td>
</tr>
</tbody>
</table>

Lower temperature limit for growth: 7 days on moist nutrient agar slopes. Features 54–57(B). The scoring is by the same method as the preceding test. The features 54 and 55 to 57 describe respectively the common subfeature concerned with low temperature growth, and subfeatures for growth down to 16°, 10° and 4°.

Rate of citrate utilization: Simmon’s medium containing 0.01 % (w/v) yeast extract to supply growth factors. Features 58–59(C). The results were graded into none, slow and fast, according to whether the medium became blue in more than 5 days, within 5 days or within 2 days respectively.

Tolerance of NaCl. Features 60–61(C) growth on NaCl nutrient agar after 7 days. This test may be regarded as a toxicity test. Although one might postulate many features, since some highly halotolerant bacteria are known, the features used are only those implied from the results in this study. The scoring is as follows:

<table>
<thead>
<tr>
<th>Features</th>
<th>60</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>No growth on 3 % NaCl nutrient agar</td>
<td>-</td>
<td>NC</td>
</tr>
<tr>
<td>Growth on 3 % NaCl but not on 6.5 % NaCl agar</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Growth on 6.5 NaCl agar</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Production of detectable HCN in aerobic agar cultures. Feature 62(A). Production of pigments. Features 63–69(A). These features 63–69 describe the production of detectable amounts of the pigments violacein, prodigiosin, pyocyanin, fluorescein, the blue-violet ‘Chromobacterium viscosum’ pigment, iodonin and the magenta pigment of strain TI respectively under appropriate conditions. Some pigments would mask others (e.g. violacein might mask prodigiosin) and some NC entries are therefore given. In the bacteria studied here there is no point in allotting features for the conditions for pigmentation, so these are omitted.

Growth in serial subculture in Koser’s citrate medium. Feature 70(A). A positive reaction in this test presumably implies the possession of a large number of features allowing the biosynthesis of amino acids, nucleotides, etc. but as this was not investigated in this study it must be counted as a single feature.

Biochemical reactions (features 71–105) Ability to ferment glucose anaerobically in Hugh & Leifson medium within 7 days. Feature 71(A).
Figs. 2, 3. The table of strains versus features (i × m table) used for the worked example when printed on to X-ray film. The strains are tabulated in the rows (only alternate stains are lettered in) and the features in the columns. Symbols: positive value = circle, negative value = rectangle, NC value = blank, orientation marks = small dashes.  Fig. 2. Features 1–57.  Fig. 3. Features 58–105.
Ability to oxidize glucose in aerobic Hugh & Leifson medium within 7 days. Feature 72(A). When feature 71 is present it obscures the result and feature 72 is scored NC.

Definite acid production in peptone water carbohydrate media within 14 days. Features 78–84(A) for the carbohydrates fructose, mannose, trehalose, sucrose, starch, lactose, dulcitol, sorbitol, adonitol, melezitose, salicin and xylose respectively

Bulky rennin clot in litmus milk within 7 days. Feature 85(A).

Bleaching of the litmus in litmus milk within 7 days. Feature 86(A).

Table 3. *Part of the* $i \times i$ *table of* $n_i$ *values*

<table>
<thead>
<tr>
<th>Strains</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>29</th>
<th>30</th>
<th>31</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strains</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>37</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>39</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>26</td>
<td>25</td>
<td>22</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>25</td>
<td>24</td>
<td>21</td>
<td>37</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>26</td>
<td>27</td>
<td>25</td>
<td>34</td>
<td>36</td>
<td>41</td>
</tr>
</tbody>
</table>
Malonate utilization test positive in 2 days. Feature 108(A).
Gluconate test after 4 days. Features 104–105(C). Divided into negative, weak positive and strong positive on heating with Benedict’s solution.

The data after coding into features were entered into an \( i \times m \) table on squared paper, and they were analysed by two methods, an electronic computer and by a rapid manual method. This latter technique, although laborious, is suitable for small numbers of strains, and is described since it may be useful to some workers. Each positive value is covered with a disk of adhesive black paper; each negative value is covered with a rectangle of black paper covering the lower half of the square. NC entries are left blank. Strain numbers are marked in black ink with orientation marks at the ends of each row. After erasing pencil marks the table is printed photographically on to film (X-ray film of 35 x 43 cm. size is suitable) and two copies prepared. The table is shown in Figs. 2 and 3 in this modified form, and will enable the subsequent working to be checked. When the duplicate copies are placed over one another on an illuminated background, one can orientate one row on any other row, and count the number of transparent circles (where two positive values coincide) and the number of semi-circles (where a positive and a negative entry coincide), thus obtaining \( n_s \) and \( n_d \). Wherever NC occurs, it blots out the coincident symbol. The subsequent working is the same as that given below where an electronic computer was used.

The computer used was the Elliott 405 digital computer (Elliott Brothers, Boreham Wood, Herts). The computer programme for steps 1 to 9 is Bacterial Classification Mark I in the Elliott 405 computer library of programmes, and a copy of this programme is also deposited in the Science Museum Library, South Kensington, London, S.W. 7.

The computer worked through the steps given on p. 208. Some of the re-
Table 7. The full table of the $S$ values (%) from the worked example after re-arranging the strains by the sorting procedure

<table>
<thead>
<tr>
<th>Strain letters</th>
<th>Strain numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>MK</td>
<td>2</td>
</tr>
<tr>
<td>DK</td>
<td>15</td>
</tr>
<tr>
<td>TB</td>
<td>24</td>
</tr>
<tr>
<td>TA</td>
<td>23</td>
</tr>
<tr>
<td>TV</td>
<td>8</td>
</tr>
<tr>
<td>LG</td>
<td>12</td>
</tr>
<tr>
<td>SL</td>
<td>13</td>
</tr>
<tr>
<td>RV</td>
<td>14</td>
</tr>
<tr>
<td>RH</td>
<td>3</td>
</tr>
<tr>
<td>BN</td>
<td>4</td>
</tr>
<tr>
<td>MH</td>
<td>16</td>
</tr>
<tr>
<td>PT</td>
<td>26</td>
</tr>
<tr>
<td>MW</td>
<td>6</td>
</tr>
<tr>
<td>AM</td>
<td>5</td>
</tr>
<tr>
<td>LW</td>
<td>9</td>
</tr>
<tr>
<td>RT</td>
<td>11</td>
</tr>
<tr>
<td>SH</td>
<td>10</td>
</tr>
<tr>
<td>FH</td>
<td>1</td>
</tr>
<tr>
<td>HD</td>
<td>34</td>
</tr>
<tr>
<td>DA</td>
<td>21</td>
</tr>
<tr>
<td>NC</td>
<td>22</td>
</tr>
<tr>
<td>MC</td>
<td>27</td>
</tr>
<tr>
<td>MA</td>
<td>26</td>
</tr>
<tr>
<td>HA</td>
<td>29</td>
</tr>
<tr>
<td>RU</td>
<td>36</td>
</tr>
<tr>
<td>CA</td>
<td>30</td>
</tr>
<tr>
<td>IN</td>
<td>33</td>
</tr>
<tr>
<td>GA</td>
<td>32</td>
</tr>
<tr>
<td>HC</td>
<td>35</td>
</tr>
<tr>
<td>PB</td>
<td>37</td>
</tr>
<tr>
<td>HE</td>
<td>33</td>
</tr>
<tr>
<td>EA</td>
<td>17</td>
</tr>
<tr>
<td>EB</td>
<td>18</td>
</tr>
<tr>
<td>EC</td>
<td>19</td>
</tr>
<tr>
<td>HF</td>
<td>34</td>
</tr>
<tr>
<td>MB</td>
<td>36</td>
</tr>
<tr>
<td>NT</td>
<td>7</td>
</tr>
<tr>
<td>NCTC1377</td>
<td>39</td>
</tr>
<tr>
<td>NCTC9403</td>
<td>40</td>
</tr>
<tr>
<td>NCTC20000</td>
<td>41</td>
</tr>
<tr>
<td>NCTC6749</td>
<td>42</td>
</tr>
<tr>
<td>TI</td>
<td>45</td>
</tr>
<tr>
<td>GR</td>
<td>43</td>
</tr>
<tr>
<td>TE</td>
<td>44</td>
</tr>
</tbody>
</table>

The full values (%) from the worked example after re-arranging the strains by the sorting procedure.
results in the intermediate steps are given below. Part of the \( n_a \) and \( n_d \) values are tabulated in Tables 3 and 4, illustrating the results on three mesophilic strains (1, 2 and 3) and three psychrophilic strains (29, 30 and 31) of *Chromobacterium*. The figures for \( n_s \) for strains compared with themselves (i.e. the number of positive features possessed by the strains) seem to be characteristic of the group to which the strains belong. In Table 3 one can see that these figures are higher for strains 1–3 than for strains 29–31, and this was true generally for the other strains of the two groups. The \( S \) values were then calculated from the relation \( S = n_a/(n_a + n_d) \).

Fig. 4. Diagrammatic representation of the full \( S \)-value table of the worked example. The diagram is derived from Table 7 by shading the squares. The strains 39, 40, 41 and 42 are, respectively, NCTC 1377, NCTC 9493, NCTC 2000 and NCTC 6749.

The strains were then sorted into groups and the \( S \) table was rearranged in order to bring strains of a group adjacent to one another (steps 5–8). This may be done manually by the following procedure: the \( i \times i \) table of \( S \) values is converted into an \( S \) table of strain pairs. Tables 5 and 6 show this for strains 1, 2, 3, 29, 30 and 31. It is seen that high values of \( S \) are found when a...
mesophilic strain is compared with a mesophilic strain, or when two psychrophils are compared. For these six strains the sorting procedure yields the results:

<table>
<thead>
<tr>
<th>% S</th>
<th>29, 30</th>
<th>29, 30, 31</th>
</tr>
</thead>
<tbody>
<tr>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>29, 30</td>
<td>31</td>
</tr>
<tr>
<td>73</td>
<td>29, 30</td>
<td>31</td>
</tr>
<tr>
<td>71</td>
<td>29, 30</td>
<td>31</td>
</tr>
<tr>
<td>52</td>
<td>29, 30</td>
<td>31, 1, 2, 3</td>
</tr>
</tbody>
</table>

This shows the separation of these strains into two groups at $S = 71$ and the fusion of the groups at $S = 52$.

The result of the sorting is shown diagrammatically in Fig. 1, in which it is seen that the expected groups of bacteria have been correctly formed. The table of $S$ values was re-arranged to give the order of strains in Fig. 1, and this full table is shown in Table 7, to the nearest 1% (the uncertainties of scoring the data do not justify greater accuracy). In this table the strain letters used in the text and in Sneath (1956b) are added to the strain numbers used in the analysis for ease of comparison. The full table is needed to check the results of the sorting procedure, since this might occasionally fuse two groups too soon or too late owing to an aberrant $S$ value, and it is also needed for calculation of $D$ values and of mean $S$ values (step 9 and Table 2).

Overall similarity is a multidimensional quantity (one dimension for each feature) which has been expressed in three dimensions when given as an $i \times i$ table of $S$ values. Therefore only a small amount of condensation can be made without losing information. This is because the inherent restrictions are few: thus if $S_{1.2}$ is 90% and $S_{1.3}$ is 90% then $S_{2.3}$ cannot be below 80% provided there are no NC entries. But with NC entries and with more than a few strains this restriction becomes of little practical interest. It is possible without losing information to convert an $i \times i$ $S$ value table into a three-dimensional model, made of rods whose length is proportional to $S$, and this may be useful on occasion. It can also be shown as an $i \times i$ table whose squares are shaded to represent the approximate $S$ value, and the table from this example is shown in Fig. 4; the division between the groups is strikingly evident, together with some of the finer distinctions between one group and another.