JOURNAL OF BACTERIOLOGY, May, 1965
Vol. 89, No. 5
Copyright © 1965 American Society for Microbiology
Printed in U.S.A.

Automatic Classification of Staphylococci by Principal-Component Analysis and a Gradient Method[1]

L. R. HILL, L. G. SILVESTRI, P. IHM, G. FARCHI, AND P. LANCIANI

Progetto Sistematica Actinomiceti, Università Statale, Milano, Italy, CETIS, Euratom, Ispra, Italy, and Laboratori di Fisica, Istituto Superiore di Sanità, Rome, Italy

Received for publication 11 January 1965

ABSTRACT

HILL, L. R. (Università Statale, Milano, Italy), L. G. SILVESTRI, P. IHM, G. FARCHI, AND P. LANCIANI. Automatic classification of staphylococci by principal-component analysis and a gradient method. J. Bacteriol. **89**:1393–1401. 1965.—Forty-nine strains from the species *Staphylococcus aureus*, *S. saprophyticus*, *S. lactis*, *S. afermentans*, and *S. roseus* were submitted to different taxometric analyses; clustering was performed by single linkage, by the unweighted pair group method, and by principal-component analysis followed by a gradient method. Results were substantially the same with all methods. All *S. aureus* clustered together, sharply separated from *S. roseus* and *S. afermentans*; *S. lactis* and *S. saprophyticus* fell between, with the latter nearer to *S. aureus*. The main purpose of this study was to introduce a new taxometric technique, based on principal-component analysis followed by a gradient method, and to compare it with some other methods in current use. Advantages of the new method are complete automation and therefore greater objectivity, execution of the clustering in a space of reduced dimensions in which different characters have different weights, easy recognition of taxonomically important characters, and opportunity for representing clusters in three-dimensional models; the principal disadvantage is the need for large computer facilities.

The main purpose of taxometrics, or numerical taxonomy, is the construction of a classification which is as objective as possible and is removed as much as possible from the subjective judgment of the taxonomist. This purpose justifies the assumption of equal weight of characters (Sneath, 1957*a*, *b*), and of equal weight of both positive and negative attributes (Hill et al., 1961; Beers and Lockhart, 1962); it also justifies the search for automatic methods of clustering the organisms, such as that proposed by Rogers and Tanimoto (1960) and by Silvestri et al. (1962), and that utilized by Sokal and Michener (1958). Finally, it justifies the present paper, which describes a further and more advanced attempt towards complete elimination of human bias.

The first taxometric methods available required a certain amount of constant feedback from machine to man and vice versa, a fact that, if it has the advantage of permitting control of the process by the expert taxonomist, also has the drawback of allowing infiltration of the results with his subjective bias and his sharing of the common endowment of taxonomic tradition. In the method of clustering organisms into groups by simple average linkage based on the utilization of shaded triangular diagrams of per cent similarity, *S* (or per cent matching, *M*) coefficients (Sneath, 1957*a*, *b*; Gilardi et al., 1960; Hill et al., 1961), it sometimes happens that particular strains become obviously misplaced, for the ordering in these triangles is a bidimensional one, requiring subjective visual inspection of the diagram for correction. In the method based on searching for nodes in a multidimensional space, such as that applied by Silvestri et al. (1962), the selection of the "second nodes," which determine the radii of the "spheres" or clusters to be isolated, is not completely objective. Rogers and Tanimoto (1960), however, proposed a "reasonable measure of inhomogeneity" which should result in a more objective determination of the clusters, but we are not aware that studies utilizing it have appeared.

To realize an automatic method of clustering, it was necessary to work with organisms localized in a space having a small number of dimensions. To reduce the original *n* dimensions (where *n* is

[1] Publication no. 15 of the Progetto Sistematica Actinomiceti.

the number of characters) to a smaller number, we have used a variety of factor analysis which is known as "principal component analysis." Applications of some kinds of factor analysis to taxometric problems were made previously by Schuessler and Driver (1956), Driver and Schuessler (1957), and Stroud (1953) in anthropology, by Braffort and Ihm (1960) in linguistics, by Rohlf and Sokal (1962) in entomology, and by Defayolle and Colobert (1962) in bacteriology.

For the clustering of organisms, a "gradient method" was used. This method was first used by Schnell (1964*a*, *b*) as part of a very general program (CLAUTO) which also uses linear algebraic methods described by Ihm (1962, *in press*), and some statistical tests. Schnell starts with the hypothesis of a division of the sample into g groups which he finds automatically. Instead of this, we extend the hypothesis to groups, subgroups, etc., and try to obtain with Schnell's gradient method the entire taxonomic dendrogram. This method, combined with principal-component analysis, was applied to the same data which had been previously analyzed by one of us (Hill, 1959) by the triangular shaded diagram method with per cent S. To allow a more complete comparison of the results, the data have been analyzed also with the triangular diagram method with per cent M, and, by courtesy of F. J. Rohlf, by the "unweighted pair group method" (Sokal and Sneath, 1963) also with per cent M coefficients.

TABLE 1. *List of species used, named according to the scheme of Shaw et al. (1951)*

No.	Name	NCTC no.
1	*S. aureus*	4136
2	*S. aureus*	4163
3	*S. aureus*	6571
4	*S. aureus*	8532
5–20	*S. aureus*	—
21	*S. saprophyticus*	7292
22	*S. saprophyticus*	7604
23	*S. saprophyticus*	7612
24–28	*S. saprophyticus*	—
29	*S. lactis*	189
30	*S. lactis*	1630
31	*S. lactis*	7564
32	*S. lactis*	7944
33–37	*S. lactis*	—
38	*S. roseus*	7511
39	*S. roseus*	7512
40	*S. roseus*	7514
41	*S. roseus*	7523
42	*S. roseus*	7528
43	*S. roseus*	7738
44	*S. roseus*	—
45	*S. afermentans*	196
46	*S. afermentans*	2665
47	*S. afermentans*	3874
48	*S. afermentans*	7563
49	*S. afermentans*	—

MATERIALS AND METHODS

Data for analysis, per cent S, per cent M. Experimental methods were described previously (Hill, 1959). The strains of staphylococci used (Table 1) were provisionally identified according to the scheme of Shaw, Stitt, and Cowan (1951) unless received already named; they comprised 20 strains of *S. aureus*, 8 strains of *S. saprophyticus*, 9 strains of *S. lactis*, 7 strains of *S. roseus*, and 5 strains of *S. afermentans*. The 49 strains ($N = 49$) had been scored for 80 features ($n = 80$) and the $N \times n$ matrix contained plus, minus, and not counted (NC) attributes. The same $N \times n$ matrix was used for all four taxometric methods.

Between all pairs of strains, per cent S coefficients were calculated according to the formula

$$S = \frac{n_s}{n_s + n_d}$$

and per cent M coefficients according to the formula (Sokal and Sneath, 1963)

$$M = \frac{n_c}{n_c + n_d}$$

Since the per cent S coefficients had been computed by hand for the earlier paper, and as the computer program available for the calculation of per cent M coefficients also calculated per cent S, the opportunity was taken during the present work to check the earlier per cent S results. Groups were then sorted from these per cent S and, separately, per cent M coefficients by ordering in triangular shaded diagrams. Intra- and intergroup mean values were estimated from rearranged $N \times N$ matrices, and taxonomic dendrograms were constructed from the values (Fig. 1 and 2).

Mathematical method. Details of the mathematics involved were published elsewhere (Ihm, *in press*). With reference to principal-component analysis, several textbooks may be consulted (e.g., Cattell, 1952; Brambilla, 1959; Harman, 1960). Here, the method will be described briefly, only to introduce the geometric model upon which our method is based, so that the taxonomist can understand the principles. The method is carried out in two successive steps: principal-component analysis and clustering. Principal-component analysis is used simply to reduce the number of dimensions (characters) to only five or six, that is, the maximal number with which the clustering program can economically operate. Principal-component analysis is the most efficient way of reducing dimensions, ensuring that the inevitable loss of information is the minimum mathematically possible. If the original dimensions (i.e.,

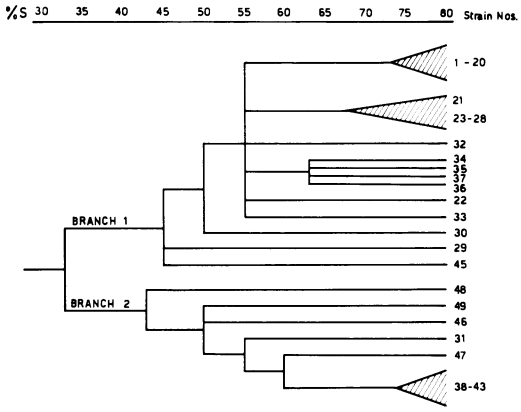

FIG. 1. *Dendrogram based on per cent S coefficients.*

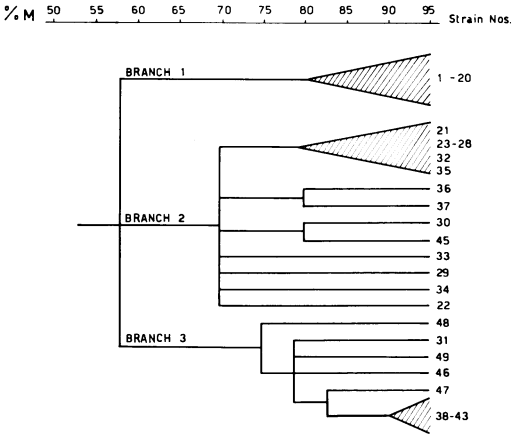

FIG. 2. *Dendrogram based on per cent M coefficients.*

characters) were only a few (five or six), then the clustering method could be applied directly to the data. However, the use of only a few characters would be in contradiction with the Adansonian concepts of overall similarity estimated over a wide range of characters.

Principal-component analysis. Each individual strain can be represented by a point, or vector, in an n-dimensional real space, R^n, where n is the number of binary (1 and 0, or + and −) characters. The n axes of this space are orthogonal one to the other. For example, if a set of organisms is studied for only one character with two attributes (0, 1), they can be represented at the ends of a unit segment (Fig. 3, a). Considering two characters simultaneously, the organisms could be imagined placed at the corners of a square with unit sides (Fig. 3, b). With three characters, they would be placed on a cube (Fig. 3, c); and with more than three characters, on hypercubes, which cannot be represented on paper but can be described algebraically and hence can be dealt with by computers. In the complete program, a special allowance could be made for the coding of characters having at least one NC score in the original $N \times n$ matrix. For the sake of simplicity, however, NC were counted as negative scores in the present example.

Of course, if all eight possible combinations of the three characters exist, then each of the eight corners of the cube would be occupied by an organism. And, analogously, if all the 2^{80} possible combinations of the 80 characters of the present study existed, the organisms would be located on the corners of a 2^{80}-dimensional hypercube. It is highly improbable that all the possible combinations exist in nature, because some of them have been eliminated by natural selection. This means geometrically that some of the corners of the hypercube will be vacant or, in other words, the n-dimensional space will be more densely occupied in some regions than in others. To this point dispersion could be fitted a hyperellipsoid, rather than a hypersphere, as would be the case if all the corners were occupied. The hyperellipsoid will have its own axes, which are the principal components, and these will have different lengths.

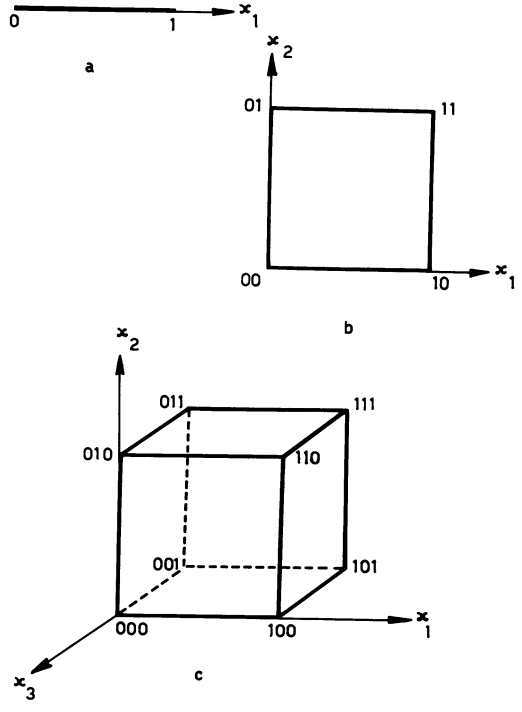

FIG. 3. *Representation of organisms as points in real space. Each binary character (x_1, x_2, x_3) is represented by a segment of unit length orthogonal to the others, and each organism, as a point situated at the extremes of these segments, according to the score, 0 or 1, for each character. In (a) are represented organisms for which one character only; in (b) two and in (c) three characters were taken into consideration.*

The axes can be determined by solving the eigenvalue problem of the covariance matrix. Thus, the localization of the points, which before were specified in relation to the n axes of R^n, becomes possible in relation to the newly found axes, each of which is not a unit character but a linear combination of all characters.

A bidimensional example may be useful to explain the purpose of determining the new axes, by means of reference to a more familiar problem. Suppose we have a set of i points determined by two variables x and y, e.g., i bacteria for which length x and breadth y have been measured. As is well known, the regression of x on y and of y on x can be determined, as well as the orthogonal regression z. The latter has the property of being that line for which the sum of the squares of the distances of the points from it is minimal. This orthogonal regression would be called the first principal component of the set of input data in factor analysis.

The projections of the i points onto this orthogonal regression (first principal component) can be utilized to describe the set of points in one dimension z instead of in the original two dimensions x and y. Of course, z is not a character in the sense that it could be directly measured, but it is derived by calculation from the original measures. In a similar way (through so-called principal-component analysis), the first component can be also found when the original dimensions are more than two. The further principal components (second, third, etc.) are those lines orthogonal to the previous ones, which also minimize the sum of squares of the distances.

If the total covariance matrix of the sample is known, the total marginal variance can be calculated in each direction OU, which is the total variance of the projection of the sample on an axis determined by the segment going from point O to point U. A set of p such axes, which may be considered as orthogonal, spans a p-dimensional (hyper-) plane E^p in the R^n. The objective is to determine E^p, i.e., the directions OU_1, OU_2, $\cdots$ OU_p such that the sum of the marginal variances in these directions is as great as possible. Then the points of the sample can be projected onto E^p, which is itself a p-dimensional real space, and the clustering can be carried out on this p-dimensional space. The reason for using a reduced-dimensional space is that, if there were g groups, a $(g-1)$-dimensional (hyper-) plane E^{g-1} could be fitted to the g centers of gravity of each group, and all information about these centers would be contained in E^{g-1}. The between-group variance in any direction orthogonal to this plane is equal to zero. If there is now a (hyper-) spherical dispersion of the points of a group around their group center of gravity, the marginal within-group variance is constantly equal in all directions. Since the total variance is composed of within- and between-group variance, and the latter is unequal to zero in every direction of E^{g-1}, the maxima of the total marginal variance determine E^{g-1}. If one projects the points perpendicularly onto E^{g-1}, all information about the groups will be contained in E^{g-1}, and, because of the supposed within-group dispersion, two clearly separated groups in the R^n will also be clearly separated in the projection on E^{g-1}.

Finding the directions OU_1, OU_2, $\cdots$ OU_{g-1} is the same problem as that of finding the $g - 1$ principal components in factor analysis. To the direction OU_i belongs a so-called eigenvalue (or characteristic root) λ_i, which is equal to the total marginal variance in the OU_i direction and is maximal for the principal components (Rao, 1952). If the eigenvalues are determined in decreasing order of magnitude, λ_g, λ_{g+1}, $\cdots$ λ_n become constantly a multiple of s^2, or approximately the same because of sampling fluctuations, so that the number of dimensions necessary can be determined from the behavior of the λ values.

Twenty principal components were extracted from the present data (Fig. 4). After the second or third eigenvalue, the remaining ones decrease very slowly. It can be assumed that, after the flexus, the remaining eigenvalues represent within-group variance and that, in the present case, only two or three axes (dimensions) are sufficient to represent the variance between groups. It is appreciated that placing reliance on such diagrams to determine the number of relevant axes is empirical, but it is the only course available at present. It should be mentioned that, whereas taking too few dimensions may lead to spurious overlapping of groups and difficult separation, the taking of too many, on the other hand, leads to no inconveniences other than lengthening calculations. In the present study, we used the first five dimensions for clustering. Since an R technique

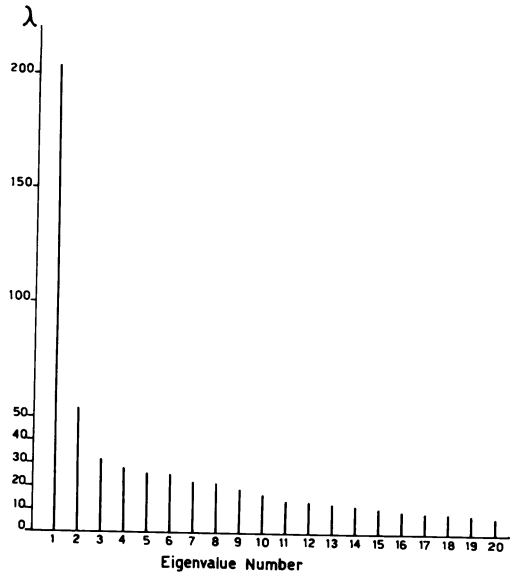

FIG. 4. *"Eigenvalues" (λ) of the first 20 principal components extracted from the present data. The eigenvalue of each principal component is proportional to the fraction of the total variance accounted for by each component.*

was used (covariance between characters), the result is a localization of characters in the reduced spaced spanned by the axes found. But since an organism is specified by its characters, then the projection of, for instance, the ith organism on the first axis can be found by summing the loadings on the first axis of all those characters which are scored 1 for the ith organism. This is done for all axes, and for all organisms.

Clustering. Having reduced the n dimensions to a more manageable number with minimal loss of between-group variance, or, in other words, having projected the N points from an n-dimensional space into, e.g., a five-dimensional one, we could proceed to the clustering method which is based on normal density functions, coupled with a gradient method to locate the centers of gravity of the zones where there is maximal gathering of points. Each point (representing an organism) is considered as the mean of a normal density function, of which the standard deviation, σ, can be fixed arbitrarily but is the same for all N functions. The N density functions are then added together. With a sufficiently small value of σ, the sum of the N normal density functions will have as many maxima as there are points, i.e., N maxima. If a larger σ value is selected, then the summed function will have fewer than N maxima, as in Fig. 5 in which nine points are ranged in one dimension (only to simplify the graphical representation), showing three maxima. With a sufficiently larger σ, the summed function will have only one maximum, as in Fig. 6. The same technique can be applied simultaneously in more than one dimension.

It is evident that, for a given value of σ, the maxima will lie in regions of maximal clustering of points and, therefore, the existence of maxima is an indication of clusters. The next step is to localize the maxima and to identify the points lying under the slopes of each particular maximum. This determination is made by a gradient method (Schnell, 1964*a*, *b*). In the one-dimensional example in Fig. 5, the points can be imagined as projected onto the summed function. Points 1, 2, 3 lie on the slopes of maximum I; points 4, 5, 6, 7, on those of maximum II; and points 8, 9, on the slopes of maximum III. The direction of maximal slope is determined for each point. Each point is moved a small step in the direction of maximal slope. The point is again moved, and the process is repeated until the slope is null; i.e., the point has reached the maximum and it can "climb" no farther. This method is therefore also called a "climbing" method or a "steepest ascent" method. All points coming to the same maximum are considered as belonging to the same group.

It is evident that different groups result at different values of σ and subsequently merge together as σ is increased. This permits the construction of an automatic dendrogram by the computer. As σ increases, the number of groups decreases (Fig. 7). The machine prints out the smallest σ value for which the number of groups, g, is a little less than or equal to N, and the members of the corresponding groups. The value of σ is then increased, and the computer repeats the clustering method, printing out the new σ value and the corresponding groups. This is repeated until a value of σ is reached for which only one group results. Thus, a dendrogram is obtained in which the scale is expressed in units relative to the smallest σ used (Fig. 8). In the present example, the rate of increase was the addition of 2 σ at each iteration.

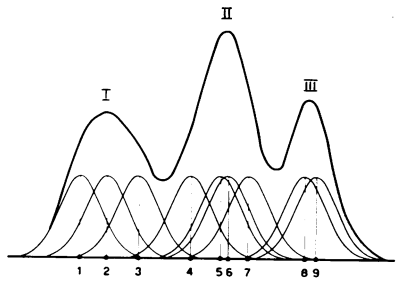

FIG. 5. *Projection of points (organisms) onto one dimension (corresponding to one principal component of the factor analysis) with normal density functions fitted to each point and with the sum of these. The summed function has three maxima. In the gradient method, the points can be projected onto the summed function and then each separately moved in direction of its maximal slope progressively until they reach a maximum. All points coming to the same maximum are considered as belonging to the same group (e.g., points 1, 2, 3 belong to group I; points 4, 5, 6, 7, to group II; and points 8, 9 belong to group III).*

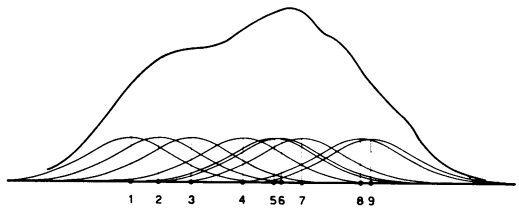

FIG. 6. *Projection similar to that in Fig. 5, with σ increased. The sum of the nine density functions now has only one maximum and, therefore, with this value of σ, all nine points form one group only.*

Representation of the results by physical models. Principal-component analysis is used to locate the strains in relationship to a restricted number of orthogonal axes. In many cases, the projection of the strains onto the first three axes may be sufficient to represent groups without spurious overlapping. This permits the construction of a physical three-dimensional model which can represent an interesting didactic and heuristic device. The computer has been programmed to print directly on paper the projection of the organisms in rela-

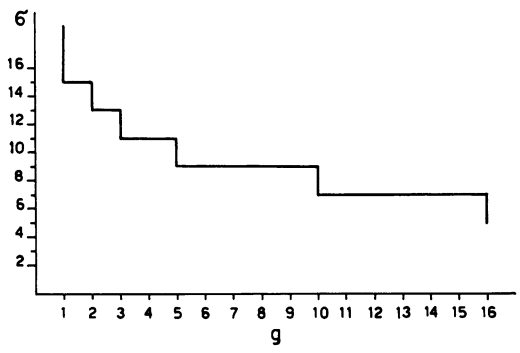

FIG. 7. *Histogram of the number g of groups formed in function of σ value used at each iteration in the present example.*

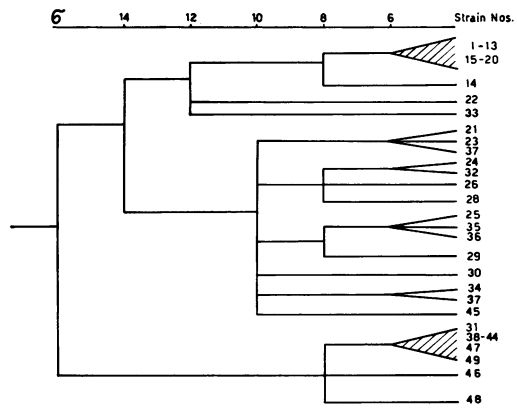

FIG. 8. *Dendrogram resulting from successive applications of the GRADIENT method of clustering, with different multiples of the smallest σ.*

tion to pairs of axes, the 1st and 2nd, then the 1st and 3rd; from such diagrams, models such as that shown in Fig. 9 can be rapidly constructed with colored balls and metal wire.

The first introduction of physical models in taxometrics was by Lysenko and Sneath (1959). Their method, however, can only work well either when the number of the characters is very small or when the characters themselves are correlated into not more than three groups, since the complements of the similarity indexes ($1 - S$) are calculated for an n-dimensional space. A physical model based on principal-component analysis can always be made, possibly without overlapping of the groups.

Execution of calculations. Per cent S and M coefficients were calculated with an IBM 650 computer at the Centro di Calcolo Numerico of the University of Genoa. Principal-component analysis and projection of the elements were carried out with an IBM 7090 at CETIS, Euratom, Ispra, Italy, with a program called "automatic classification by principal-component analysis," written by P. Ihm, H. Fangmeyer, and P. Schnell. The "gradient method" was carried out with an IBM 7040 at the Laboratorio di Fisica of the Istituto Superiore di Sanità, Rome, Italy, with a program called "gradient method clustering" (GRADIENT), written by G. Farchi and P. Lanciani. The "unweighted pair group method" (using also per cent M) was carried out by courtesy of F. J. Rohlf of University of California, Santa Barbara, with an IBM 7094 and a program called TAXON, written by Rohlf.

RESULTS

Sneath's method with the use of per cent S. Only a few errors were found in the earlier (Hill, 1959) hand-computed per cent S coefficients. The corrected dendrogram is presented in Fig. 1. It differs but little from that of the earlier paper. The taxonomic conclusions previously reached remain, therefore, unaltered. Since, with per cent S, *S. aureus* and *S. saprophyticus* together with the miscellaneous *S. lactis* strains appear more closely related to each other than they are to *S. roseus* and miscellaneous *S. afermentans* strains, then better nomenclature would be obtained by giving the generic names *Staphylococcus* and *Micrococcus* to branches 1 and 2, respectively, of the per cent S dendrogram.

Sneath's method with per cent M, TAXON, and GRADIENT. All these three methods gave essentially similar results; the dendrograms are given in Fig. 2, 10, and 8, respectively. One group

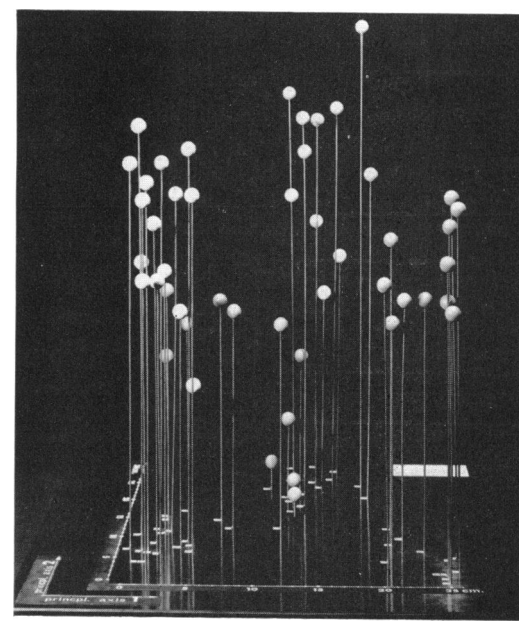

FIG. 9. *Three-dimensional model constructed from the projection diagrams furnished by the computer.*

comprises all the *S. roseus* and *S. afermentans* strains (excluding only strain 45 and including strain 31) well separated in all four methods. The constancy of this finding with different methods is evidence that this group of cocci is indeed taxonomically separated, confirming earlier conclusions. Strain 31 corresponds to the description of Shaw et al. (1951) for *S. lactis* as producing acid from glucose. However, it did not produce acid from the other glucids tested, and was, therefore, at best a "poor" fermenter.

The division between *S. aureus* and the miscellaneous group comprising *S. saprophyticus* and *S. lactis* strains is evident differently in different dendrograms. It is very sharp with per cent *M*, less so with GRADIENT method and TAXON, and least of all with per cent *S*. With the exception of per cent *M*, it is always less evident than the separation of the *S. roseus* and *S. afermentans* group.

Comparison with the classification of Shaw et al. Organisms named *S. aureus* according to the scheme of classification of Shaw et al. (1951) constitute a discrete taxometric phenon. The character chosen by them, coagulase possession, is a highly correlated character, as evidenced by its high loading in both first and second axes (Table 2). The sharp distinction between *S. saprophyticus* and *S. lactis* does not appear justified, particularly in view of the TAXON and GRADIENT dendrograms. The use of the Voges-Proskauer test to make such a division is suspected of creating artificial groups in their scheme. The character had relatively high loading only on the fourth axis. This finding illustrates the inherent danger of monothetic classifications. *S. lactis*, on the other hand, has also been found a nonhomogeneous group by Pohja (1960), Gregory and Mabbit (1957), and Baird-Parker (1963).

Also, the separation between *S. roseus* and *S. afermentans* appears dubious. A certain degree of separation of *S. roseus* can be seen with per cent *S*, but this is less evident in per cent *M* and TAXON, and almost nil with the GRADIENT method.

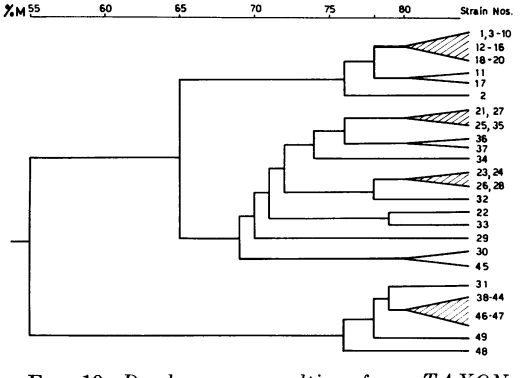

FIG. 10. *Dendrogram resulting from TAXON, applied to a matrix of per cent M.*

TABLE 2. *First and last five characters at both extremes of the first three axes**

Axis	No.	Feature	Eigenvector
Eigenvalue 1 = 204.3	76	Heat-sensitive, 3+	+0.168
	13	Optimal temp, 30 C	+0.138
	80	Phenol-sensitive, 3+	+0.113
	5	Diam, 0.9 to 1.0 μ	+0.109
	21	Pink pigment	+0.109
	64	Methylene blue reduced, 2+	−0.192
	18	Gold-yellow pigment, 1+	−0.197
	42	Phosphatase	−0.203
	45	One to four flocculation lines against staphylococcus antitoxin	−0.210
	32	Free coagulase, 1+	−0.211
Eigenvalue 2 = 55.1	23	White pigment	+0.362
	73	Growth with 15% NaCl	+0.259
	16	Smooth colonies	+0.220
	53	Acid from maltose	+0.213
	49	Acid from glucose	+0.208
	35	α-Lysin, 1+	−0.151
	45	One to four flocculation lines against staphylococcus antitoxin	−0.156
	21	Pink pigment	−0.164
	39	δ-Lysin	−0.165
	32	Free coagulase, 1+	−0.170
Eigenvalue 3 = 32.5	54	Acid from mannitol	+0.255
	47	Penicillinase, 1+	+0.235
	48	Penicillinase, 2+	+0.190
	59	Urease, 2+	+0.189
	25	Granular deposit in broth	+0.187
	24	Uniform turbidity in broth	−0.187
	69	Gelatin liquefaction 1+	−0.191
	67	Casein hydrolysis, 1+	−0.206
	62	$NO_3 \rightarrow NO_2$	−0.228
	75	Heat-sensitive, 2+	−0.267

* Characters were reordered according to their loadings (eigenvectors) from the highest positive value to the highest negative one on each axis.

Identification. As mentioned before, during the principal-component analysis, the computer furnishes lists of weights or loadings of the characters toward each axis, one list (eigenvector) per axis. These lists can be reordered from the highest positive loading to zero loading and then to the highest negative loading. The characters at the extremes of such reordered lists are the relatively most important ones in generating the classification.

The projection of the first two axes of the present principal-component analysis is sufficient

to separate the major groups found (*S. aureus*; *S. saprophyticus* and *S. lactis*; *S. roseus* and *S. afermentans*). At the extremes of the first axis are located *S. aureus* and *S. roseus*; hence, the most negatively loaded characters (with respect to this axis) are typical of *S. aureus*. In fact, particularly in the case of the better-known *S. aureus*, characters such as free coagulase, phosphatase, and golden yellow pigment were among the highest negatively loaded characters (Table 2). With reference to *S. roseus*, pink pigment appears (high positive loading), but unsuspected characters such as heat and phenol sensitivity are also revealed (Table 2).

Along the second axis, the organisms are ordered from *S. aureus*, together with *S. roseus* and *S. afermentans* strains, to *S. lactis* and *S. saprophyticus* strains. Again, the most negatively loaded characters are typical of the first two major groups of strains (Table 2, free coagulase again, δ-hemolysin, flocculation lines against staphylococcal antitoxin, α-hemolysin, and pink pigment), and the most positively loaded ones are typical of *S. lactis* and *S. saprophyticus* strains (Table 2, white pigment, growth with 15% NaCl, smooth colonies, acid from maltose and glucose).

DISCUSSION

The purpose of this paper has been not so much the elucidation of *Micrococcus-Staphylococcus* taxonomy per se, which would require the study of a larger number of strains, but rather the presentation and description of the GRADIENT method as a new taxometric method, the use of data regarding those organisms for its illustration, and, finally, the comparison of this method with others currently in use.

Our method, combining projection onto p dimensions and clustering, is evidently a much more complex method than the others used here. It presents the disadvantage of requiring good computing facilities. Final results are quite similar to the simpler method of Sneath (per cent M) and the TAXON method. Our method and TAXON both represent an improvement over Sneath's method (per cent S or M), in that both methods avoid the subjective visual reordering (often obligatory) of shaded triangle diagrams. Both Sneath's method (per cent S or M) and TAXON carry out clustering directly on similarity coefficients, between organisms, computed with all characters equally weighted, a practice which has given rise to vivacious controversy. Though groups, by these methods, can only emerge if correlated characters are present and these indeed determine the eventual taxonomic divisions made, this fact is not immediately apparent. These methods do not actually reveal which characters are the most important ones, this being probably the cause of controversy.

Out method, on the other hand, conducts first a procedure by which characters are weighted (and furnishes lists of these loadings), and then proceeds to clustering in a weighted-character space. For this reason, the method is less exposed to the criticisms raised against other Adansonian methods.

The GRADIENT method carries out clustering in a p-dimensional space ($p < n$), and, hence, there is loss of information; per cent S, per cent M, and TAXON methods are not amenable to the study of how much information is lost. With our method, on the other hand, the loss of information when passing from n to p dimensions is easily calculated. The total variance in the original data corresponds to the trace of the covariance matrix, and the eigenvalues, λ, associated with each axis determined, indicate how much variance is accounted for by each axis. Hence, the sum of the λ divided by the trace and multiplied by 100 gives the per cent of total variance accounted for by the axes used. In the present illustration, the trace was 636.94, and $\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5$ was 347.21; thus, $347.21/636.94 \times 100 = 54.5\%$ of the total variance was accounted for by the first five axes. Therefore, the loss of information was $100 - 54.5 = 45.5\%$. But this information is only within-group information and is without interest for group determination.

The use of the first three principal axes to construct physical three-dimensional models provides the best representation that can be visually realized in the sense that the loss of information is the minimum possible (Fig. 9).

Finally, the eigenvectors computed during the principal component program yield information of direct interest for identification. In another paper, we explained how taxometric results could be utilized to construct a diagnostic stochastic key (Hill and Silvestri, 1962; Möller, 1962). This same method could be applied to the present results. However, some diagnostic device could be based on the utilization of such information, represented by the character loadings obtained during the factor analysis. This utilization will be the subject of a further paper.

In conclusion, all methods present advantages and disadvantages, of course. Final results are very similar with the different methods, which in itself is indicative of their statistical robustness. An eventual choice between alternative methods will, therefore, be determined by weighing the advantages of the greater "yield" of the GRADIENT method and the disadvantage that it

cannot be carried out with simpler computing facilities against the greater economy of the other methods with their smaller yield. The greater "yield" of our method consists of obtaining character loadings, evaluating losses of information, constructing physical models, and, finally, providing data useful for identification keys.

Another factor influencing the choice of method is represented by the original data itself. In the present example, one of the simpler methods would probably be preferred by most, but in more complex taxonomic situations our method may be more rewarding, in particular when the reordering of strains in triangular diagrams is not unequivocal.

ACKNOWLEDGMENTS

That part of this work which was contributed by Progetto Sistematica Actinomiceti was supported in part by Assegnazione no. 04/80/4/3802 of the Consiglio Nazionale delle Ricerche and by a grant from Lepetit S.p.A.

LITERATURE CITED

BAIRD-PARKER, A. C. 1963. A classification of micrococci and staphylococci based on physiological and biochemical tests. J. Gen. Microbiol. **30**:409–427.

BEERS, R. J., AND W. R. LOCKHART. 1962. Experimental methods in computer taxonomy. J. Gen. Microbiol. **28**:633–640.

BRAFFORT, P., AND P. IHM. 1960. Dépouillement et exploitation d'un échantillon linguistique. Rappt. CETIS no. 7.

BRAMBILLA, F. 1959. L'analisi dei fattori. Università Bocconi, Milan.

CATTELL, R. B. 1952. Factor analysis. Harper & Row, Publishers, Inc., New York.

DEFAYOLLE, M., AND L. COLOBERT. 1962. L'espéce *Streptococcus faecalis*. II. Etude de l'homogénéité par l'analyse factorielle. Ann. Inst. Pasteur **103**:505–522.

DRIVER, H. E., AND K. F. SCHUESSLER. 1957. Factor analysis of ethnographic data. Am. Anthropol. **59**:655–663.

GILARDI, E., L. R. HILL, M. TURRI, AND L. G. SILVESTRI. 1960. Quantitative methods in the systematics of Actinomycetales. I. Giorn. Microbiol. **8**:203–218.

GREGORY, M., AND L. A. MABBIT. 1957. The differentiation of bacterial species by paper chromatography. V. Preliminary examination of the micrococci. J. Appl. Bacteriol. **22**:307–316.

HARMAN, H. H. 1960. Modern factor analysis. Univ. Chicago Press, Chicago.

HILL, L. R. 1959. The Adansonian classification of staphylococci. J. Gen. Microbiol. **20**:277–283.

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

HILL, L. R., M. TURRI, E. GILARDI, AND L. G. SILVESTRI. 1961. Quantitative methods in the systematics of Actinomycetales. II. Giorn. Microbiol. **9**:56–72.

IHM, P. 1962. Methoden der Taxometrie. *In* V. F. Serbanesun [ed.], Information Retrieval, IBM Symposium, Blaricum, Holland.

LYSENKO, O., AND P. H. A. SNEATH. 1959. The use of models in bacterial classification. J. Gen. Microbiol. **20**:284–290.

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

POHJA, M. S. 1960. Micrococci in fermented meat products. Classification and description of 171 different strains. Acta Agral. Fennica Bull. 96.

RAO, C. R. 1952. Advanced statistical methods in biometric research. John Wiley & Sons, New York.

ROGERS, D. J., AND T. T. TANIMOTO. 1960. A computer program for classifying plants. Science **132**:1115–1118.

ROHLF, F. J., AND R. R. SOKAL. 1962. The description of taxonomic relationships by factor analysis. Systematic Zool. **11**:1–16.

SCHNELL, P. 1964*a*. Eine Methode zur Auffindung von Gruppen. Biomet. Z. **6**:47–48.

SCHNELL, P. 1964*b*. Diplomarbeit. Technische Hochschule, Darmstadt.

SCHUESSLER, R. F., AND H. E. DRIVER. 1956. A factor analysis of 16 primitive societies. Am. Sociol. Rev. **21**:493–499.

SHAW, C., J. M. STITT, AND S. T. COWAN. 1951. Staphylococci and their classification. J. Gen. Microbiol. **5**:1010–1023.

SILVESTRI, L. G., M. TURRI, L. R. HILL, AND E. GILARDI. 1962. A quantitative approach to the systematics of actinomycetes based on overall similarity. Symp. Soc. Gen. Microbiol. **12**:333–360.

SNEATH, P. H. A. 1957*a*. Some thoughts on bacterial classification. J. Gen. Microbiol. **17**:184–200.

SNEATH, P. H. A. 1957*b*. The application of computers to taxonomy. J. Gen. Microbiol. **17**:201–226.

SOKAL, R. R., AND C. D. MICHENER. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **38**:1409–1438.

SOKAL, R. R., AND P. H. A. SNEATH. 1963. Principles of numerical taxonomy. W. H. Freeman and Co., San Francisco.

STROUD, C. P. 1953. An application of factor analysis to the systematics of *Kalotermes*. Systematic Zool. **2**:76–92.