JOURNAL OF CLINICAL MICROBIOLOGY, Sept. 1988, p. 1745–1753
0095-1137/88/091745-09$02.00/0

Vol. 26, No. 9

Optimal Data Processing Procedure for Automatic Bacterial Identification by Gas-Liquid Chromatography of Cellular Fatty Acids

ERKKI EEROLA[1]* AND OLLI-PEKKA LEHTONEN[2]

Department of Medical Microbiology, Turku University,[1] and Central Laboratory, Clinical Microbiology, Turku University Central Hospital,[2] SF-20520 Turku, Finland

Received 1 March 1988/Accepted 17 May 1988

Gas-liquid chromatography of cellular fatty acids was used in automatic identification of clinical bacterial isolates. The intraspecies variation in the occurrence of fatty acids and the variation in the relative gas-liquid chromatography peak areas of different fatty acids were evaluated and compared with the relative peak areas of these acids. A new chromatogram comparison method involving the use of an exponential function was developed to adjust to data variation optimally. This method was compared with several previously published methods of correlation analysis with data from representative clinical bacteriological isolates. The efficacies of the methods in separating different bacterial species into distinct clusters were compared. The new exponential function method was superior to the others both in its ability to separate species into different clusters and in giving a greater degree of identity to strains within a proper cluster. The results indicate that the gas-liquid chromatography of bacterial cellular fatty acids can be used effectively in the identification of clinically isolated bacteria. However, the usefulness of the analysis depends on the comparison method used and on its ability to cope with data variations.

One of the main goals of clinical bacteriological analysis is accurate identification of bacterial isolates. This serves two purposes: taxonomic identification of species and comparisons between new and previous isolates. In addition to classical identification methods, based mainly on demonstrations of differences in the metabolic pathways of different species, methods have been developed which identify bacteria according to chemical compounds present in the bacterial cells (11). The main advantages of these methods are increased speed of analysis, since there is no need to cultivate the bacteria any further, and the possibility of identifying bacteria which are no longer viable. Several groups of compounds are used for identification. Among the most often used are certain enzymes, DNA, membrane lipopolysaccharides, carbohydrates, proteins, and bacterial fatty acids (11). The usefulness of the various compounds in this respect depends both on the taxonomic effect of differences in their occurrence and on the speed and simplicity of the analytical methods. In this respect, fatty acid analysis seems to be one of the most promising methods (11, 12).

Gas-liquid chromatography (GLC) of bacterial cellular fatty acids (5, 12) is a versatile method, allowing the identification of the vast majority of isolates. However, chromatograms involve the problem of comparison of multivariate recordings. To be useful in a clinical microbiological laboratory, data processing should be automatic. Although computerized taxonomy has been studied extensively (3, 16), the optimal way of applying a computerized comparative analysis of cellular fatty acid data to bacterial identification has not yet been determined. This is especially true of the identification of clinical isolates.

In the present study we collected isolates of clinically important strains and analyzed the occurrence and relative abundance of cellular fatty acids by GLC. We developed a new analytical GLC method which accounts for the variations in the data and thus permits optimal distinction between species. The new exponential function method was compared with several previously known chromatographic comparative methods.

MATERIALS AND METHODS

Bacterial strains. The composition of bacterial fatty acids and its reproducibility within various species were analyzed by using strains of 12 species. The species tested were *Bacteroides fragilis* (66 strains), *B. vulgatus* (15 strains), *B. thetaiotaomicron* (17 strains), *Staphylococcus aureus* (17 strains), *S. epidermidis* (20 strains), *S. hominis* (13 strains), *Clostridium perfringens* (19 strains), *C. difficile* (56 strains), *Listeria monocytogenes* (10 strains), *Escherichia coli* (13 strains), *Yersinia enterocolitica* serotype O3 (12 strains), and *Pseudomonas aeruginosa* (15 strains).

The strains were clinical isolates that were identified by standard clinical microbiological practices. These included growth and colony morphology on different media (menadione-cysteine-enriched brucella blood agar with and without neomycin, bacteroides bile esculin agar, cefoxitin cycloserine fructose agar, and blood agar); Gram staining (in some cases combined with the KOH test [6]); antibiotic susceptibilities; API 20E, API 20NE, API Staph, and API 20A tests (Analytab Products, Plainview, N.Y.); and RapID-ANA test (Innovative Diagnostics Systems, Inc., Atlanta, Ga.). The reverse CAMP test (7) was used to confirm the identification of *C. perfringens*. Autofluorescence was used to confirm the identification of *C. difficile*.

The following type and reference strains were included in the analyzed bacteria: *B. fragilis* (ATCC 23745, ATCC 25285, and ATCC 25280), *B. thetaiotaomicron* (ATCC 29148), *C. difficile* (ATCC 17858, ATCC 17857, and ATCC 9689), *S. aureus* (ATCC 25923), *S. epidermidis* (ATCC 155, ATCC 35984, ATCC 31432, and ATCC 35983), *S. hominis* (ATCC 35981 and ATCC 35982), *P. aeruginosa* (ATCC 9721), and *E. coli* (ATCC 25922).

* Corresponding author.

GLC analysis. Bacterial cellular fatty acid GLC analysis was performed as described previously (L. Miller, Hewlett-Packard gas chromatography application note 228-41; Hewlett-Packard Co., Palo Alto, Calif.). For GLC, aerobic bacteria were cultured on Trypticase soy agar (BBL Microbiology Systems, Cockeysville, Md.) and obligate anaerobes were cultured on Schaedler agar (BBL). The bacteria were collected from the plates, and the material was saponified, methylated, and analyzed as described previously (Miller, Hewlett-Packard note). In brief, the collected material was incubated for 30 min at 100°C in 15% (wt/vol) NaOH in 50% aqueous methanol and then acidified to pH 2 with 6 N aqueous HCl in CH_3OH. The methylated fatty acids were further extracted with ethyl ether and hexane. The GLC analysis was done as described previously (Miller, Hewlett-Packard note) with an HP5890A gas chromatograph (Hewlett-Packard) and an Ultra 2, 004-11-09B fused-silica capillary column (0.2 mm by 25 m; cross-linked 5% phenylmethyl silicone; Hewlett-Packard). Ultra-high-purity helium was used as the carrier gas. The GLC settings were as follows: injection port temperature, 250°C; detector temperature, 300°C; initial column temperature, 170°C, increasing at 5°C/min up to 270°C at 20 min; total analysis time, 25 min; sample volume, 1 μl. The peak retention time and peak area values were recorded by an HP3392A integrator (Hewlett-Packard).

We identified individual fatty acids by comparing their retention times with those of a bacterial fatty acid standard (bacterial acid methyl ester mix CP, 4-7080; Supelco, Inc., Bellefonte, Pa.). Individual fatty acids were identified by using a 0.5% retention time window when they were compared with the fatty acid standard. The analytical reproducibility of the method was assessed by repeated GLC analyses of the fatty acid standard.

Data analysis. GLC data were transferred from the integrator to the computer and stored as data files by using an Olivetti M24 microcomputer with a 650-kilobyte memory and a 20-megabyte Winchester disk. Data transfer and all analysis programs were done locally (supplied by Scientific Expert Systems Ltd., Helsinki, Finland). The transferred data consisted of retention time and peak area values for all peaks recorded by the integrator. Identified peaks corresponding to the fatty acid standard and three frequently appearing unidentified peaks were used in the subsequent analyses.

The mean peak area and standard deviation for each fatty acid were calculated for each species. The values were calculated as percentages of the total peak area to eliminate the effect of inoculum size variation. Variations in the peak areas of different fatty acids and their prevalence among different species were analyzed. These values were compared with the mean peak sizes to determine the effect of peak areas on the reproducibility of the results.

Correlation analysis. Fatty acid profiles were stored on the computer to be used for automatic identification of the bacteria. The identification was based on calculating similarity coefficients between individual bacterial strains. This was accomplished by comparing the fatty acid profile of the unknown strain with those of standard strains to find the reference strains that most closely resembled the bacterium being tested.

A selection of individual strains were analyzed by correlation and subsequent cluster analysis to evaluate the ability of these methods to separate the strains into distinguishable species-related clusters. Several methods have been published for calculating the similarity coefficients of GLC data (4). These methods are based on comparing either the ranks of peak areas or absolute peak area values. The methods used in the present study included the Pearson product moment test (4), correlation coefficient (2, 9, 10), the similarity index based on angle separation vectors (8), the Stack method (17), the Spearman coefficient of rank correlation (4), and the similarity index based on overlap of fatty acid profiles (1). In addition, derivatives of the existing methods were produced to study how different factors affect the results. These methods included a variation of the Stack method that compared peak areas with weighted peak sizes, and a variation of the similarity index based only on the presence of the peaks, with or without weighting peak areas. The methods used and their formulas are described in the appendix, which also contains a new exponential function method developed on the basis of the results of the present study. Details of the method are described in Results.

The efficacy of various correlation methods was tested by using groups of 12 species with 6 strains in each (total, 72 strains). The species were selected to include groups which either showed clear qualitative and quantitative differences in their fatty acid profiles or differed only in the amounts of their fatty acids.

The results of the correlation analyses were stored on the computer as similarity index matrices. These were further analyzed by using the weighted pair-group cluster analysis of the arithmetic averages method (14, 15) and are presented as dendrograms for comparison of the clustering efficiencies of the various methods. The following parameters (see Table 1) were calculated for each method to compare their efficacies: the average clustering level within species (ACL), which is the average value of the lowest clustering levels of each species; the intrafamilial clustering index (CI) for the *Bacteroides*, *Staphylococcus*, *Clostridium*, and enterobacterial species, which is equivalent to the ratio of the ACL of the corresponding species and the lowest level of clustering between the same species (this showed the ability of the methods to separate between the corresponding species); the number of strains placed outside their proper clusters; and the number of strains allocated into the wrong clusters.

RESULTS AND DISCUSSION

Bacterial fatty acid composition. Figure 1 shows the average cellular fatty acid compositions and standard deviations within species for the tested strains. In general, both the occurrence of the various fatty acids and their amounts seemed constant within each species. Thus, the reproducibility of the fatty acid data was high enough for bacterial identification. The observed variation was due to differences between different strains. If the same strain was cultured and analyzed repeatedly, the results were practically identical.

Figures 2 and 3 show the prevalence of individual fatty acids and their variation in peak area. The peak areas are presented as percentages of the total fatty acid peak areas. The prevalence of the fatty acids and their coefficients of variation [CVs, calculated as (standard deviation/mean) × 100] are plotted against the mean peak area of each individual fatty acid in each species. No difference was found between different species or different fatty acids in their prevalence or CVs (data not shown). Thus, all individual fatty acids of all 12 tested species are plotted together. If the mean fatty acid peak area was ≥5% of the total area, the peak was present in almost 100% of the strains of the species. Below that level the prevalence began to decrease, so that if the peak area was <1%, a prevalence of 100% was never seen within a species for any fatty acid tested (Fig. 2).

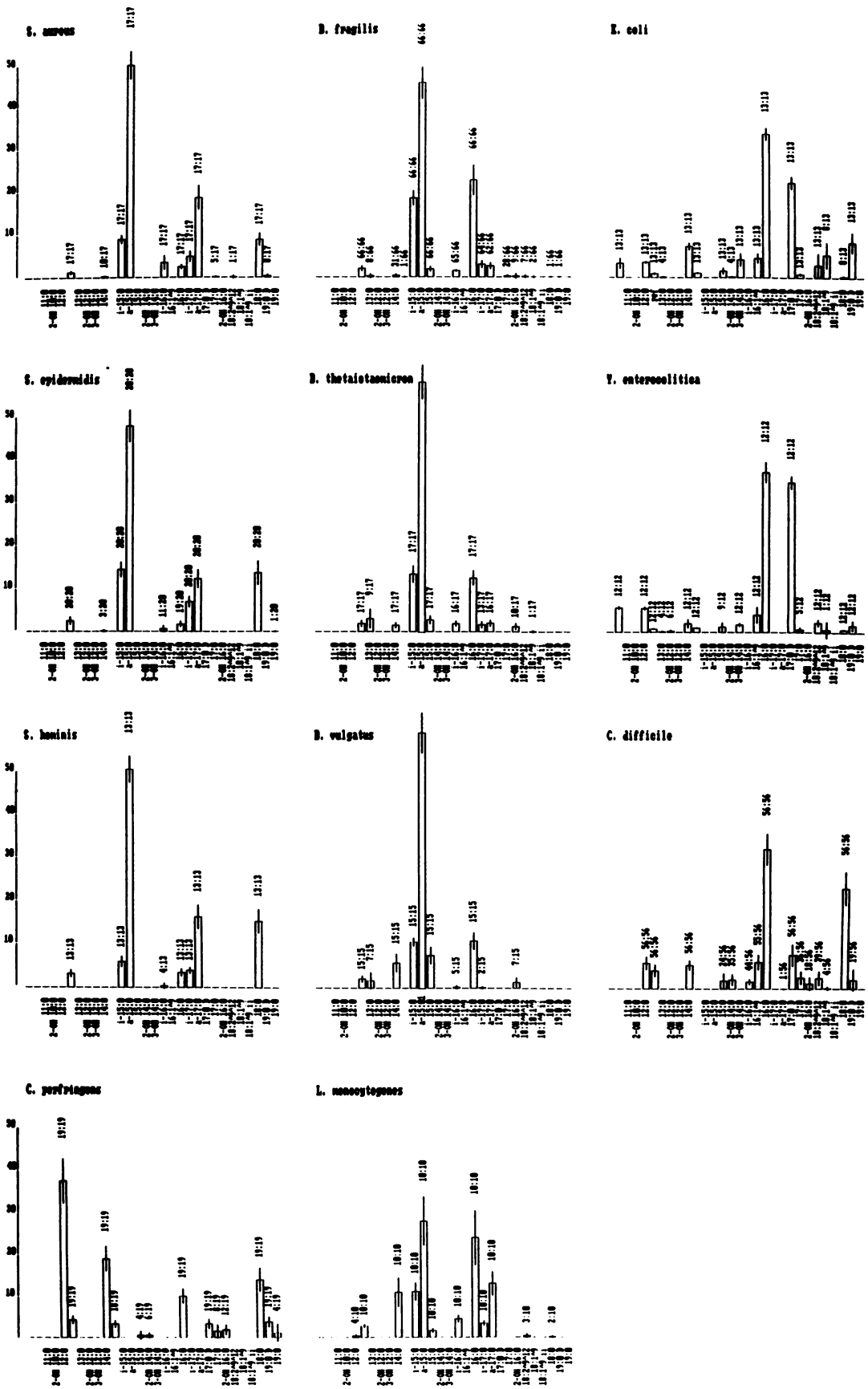


FIG. 1. Peak areas of various fatty acids within analyzed species. The bars show the mean peak areas of fatty acids and standard deviation in all strains (strains lacking the particular fatty acid are also included). The number of strains containing each fatty acid is given above each bar.

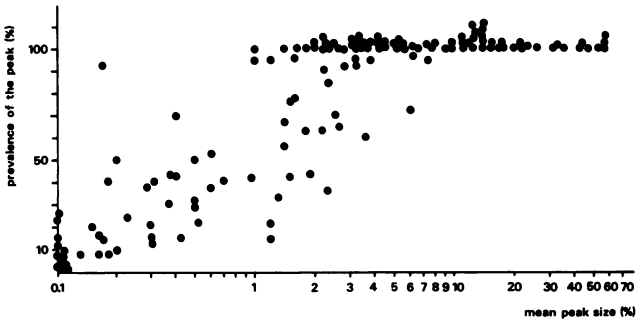


FIG. 2. Effect of the amount of fatty acids on their prevalence among analyzed strains. The prevalence of each fatty acid is presented in relation to the mean peak area of that fatty acid in each species. Fatty acid peak areas are presented as a percentage of the total peak area.

Peak area CVs were inversely correlated with mean peak areas. If the mean peak area was above 1%, CVs were about 30%. When the mean peak area exceeded 10%, CVs decreased gradually, approaching values of 5 to 10%. At mean peak sizes below 1%, CVs were almost constantly above 100% (Fig. 3). These findings formed the basis for the required characteristics of the developed comparison function.

Development of the new comparison function. A crucial point for effective comparison of fatty acid profiles was found to be the possibility of adjusting the method according to variations in the prevalence and amount of the acids. The results showed that for strains of the same species, the reproducibility of peak prevalence and area was very good (prevalence, 100%; CV, <25%) with peaks greater than 5% of the total area and poor (prevalence, <50%, CV, >100%) with peaks less than 1% of the total area (Fig. 1 to 3).

The following initial procedure was used in the analysis. A similarity value was first calculated for each peak pair in the two strains to be compared. We had to decide whether to weight the comparison by peak areas. Weighted or unweighted peak areas could be used for comparison. If the peak area was not weighted, the similarity index was calculated directly as an average of individual peak similarity values (see the equation below). If the peak area was used as an impact factor for the result, the similarity index was calculated as a weighted average of peak similarity values by using the relative areas of the peaks.

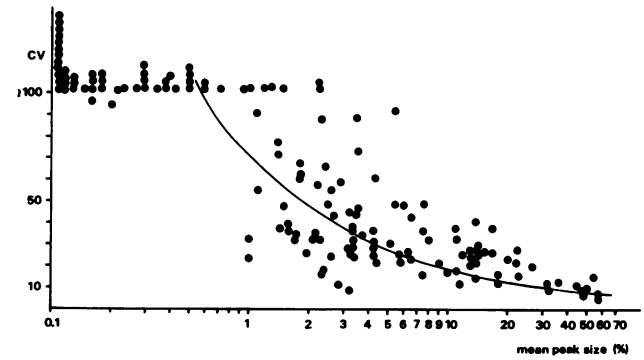


FIG. 3. Effect of the peak areas of fatty acids on the variation of peak sizes. Percent CVs for each fatty acid within each species are presented in relation to the mean peak area of that particular fatty acid within each species. All fatty acid peak areas are presented as a percentage of the total peak area.

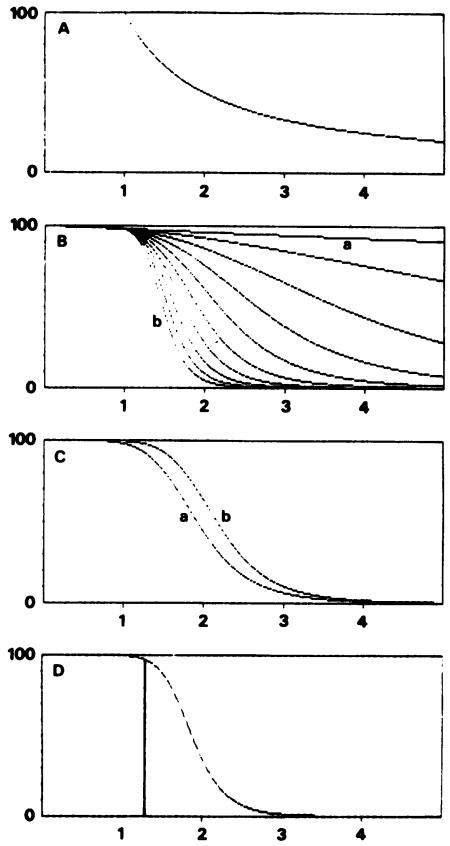


FIG. 4. Shapes of various functions showing the effects of peak area ratios on similarity indices. The horizontal axis shows the peak area (always inverted to be ≥1). The vertical axis shows the similarity index. (A) Similarity index calculated directly as the ratio of peak areas. (B) Similarity index calculated by using the developed function, showing the effect of the constant k_1 on the shape of the function (k_1 has values from 1 to 10, respectively, from line a to line b). (C) Similarity index calculated by using the developed function, showing the effect of the constant k_2 on the slope of the function (k_2 has values of 0 and 0.25 in lines a and b, respectively). (D) Relation between similarity index and peak area ratio when using the developed function. The shape of the function is as used in the analyses. Values of k_1, k_2, and k_3 are given in the text. The vertical line shows the location of peak size ratio 1.3 in the function.

Figure 4A shows a function reflecting the peak similarity value when it is calculated plainly as the ratio of peak areas (x_{ik}/x_{jk}). In Fig. 4A the ratio is inverted so that the value is always greater than or equal to 1. A characteristic of the curve is that the drop of the similarity index is highest at peak ratio values just higher than 1. The approximate CV range, 0 to 30%, reflecting the observed variation of fatty acid peak areas within a species, is in the area of the curve at ratio values between 1 and 1.3. (Two peaks with a 30% difference in their sizes give a ratio of 1.3.) It is just this area on the x_{ik}/x_{jk} curve that shows the highest drop of the function. Thus, the use of the plain peak area ratio is too heavily influenced by variations in fatty acid peak areas.

To eliminate this effect, we developed an exponential function which would show a minimal change in the above area and then drop rapidly, since variations greater than 30%, especially in large peaks (Fig. 3), always signify a lack of identity between the species. Thus, the function was developed so that it allowed a certain amount of peak area variation with a minor impact on the similarity index.

Further weighting of individual peaks by their areas was used in the comparison.

Similarity indices were calculated between individual fatty acid profiles by using the equation

$$C_f = \sum_{k=1}^{n} 0.5(x_{lk} + x_{jk}) f(x_{lk}/x_{jk})$$

where C_f is the similarity index between samples x_l and x_k, k is the peak number, x_{lk} and x_{jk} are the relative areas of the kth peaks of strains x_l and x_j, x_{lk}/x_{jk} is the peak size ratio (x_{lk}/x_{jk} was always inverted to be ≥ 1), $f(x_{lk}/x_{jk})$ is the peak similarity value determining function, which is given by $100k_1/[[(x_{lk}/x_{jk}) - k_2]^{k_3} - 1 + k_1]$, where k_1, k_2, and k_3 are coefficients determining the effect of peak size variation on the similarity coefficient, and $(x_{lk}/x_{jk}) - k_2$ is determined such that if $[(x_{lk}/x_{jk}) - k_2] < 0$, then $[(x_{lk}/x_{jk}) - k_2] = 0$.

When two samples, x_l and x_j, were compared, each peak in x_l was compared with the corresponding peak in x_j. If the same peak was found in both samples, the relative amounts were compared by using the function $f(x_{lk}/x_{jk})$. If the values were equal, the value of $f(x_{lk}/x_{jk})$ was 100. If the peak was absent in curve x_j, the value was 0. If differences occurred in the amounts of the peaks, the value was between 100 and 0. The constants k_1, k_2, and k_3 controlled the effect of peak area difference on the similarity coefficient. Figures 4B and C show the effects of coefficients k_2 and k_3 on the peak similarity value. The constant k_2 moved the curve in a left-right direction, determining the length of the lag period before the increase in the peak area ratio began to decrease the similarity index. The constant k_3 affected the slope of the curve, determining how rapidly the function decreased after the lag period. With k_1 values from 0 to 10, the slope of the decreasing part of the function changed from an almost horizontal position (Fig. 4B, curve a) to an almost vertical one (Fig. 4B, curve b). Analyses of about 4,500 strains of normal clinical isolates gave the following optimal values, which were obtained empirically: $k_1 = 50$, $k_2 = 0.25$, $k_3 = 6.0$. Figure 4D shows the shape of the function based on these selected values.

The $f(x_{lk}/x_{jk})$ value was further multiplied by the relative peak areas, x_{lk} and x_{jk}, to weight the effects of metabolites according to their peak areas. Thus, the similarity coefficient, C_f, was 100 if the two chromatograms were identical, irrespective of the numbers of peaks and their absolute areas. If the formula was used without weighting the peaks according to their sizes, it took the following form:

$$C_f = \sum_{k=1}^{n} f(x_{lk}/x_{jk})/n$$

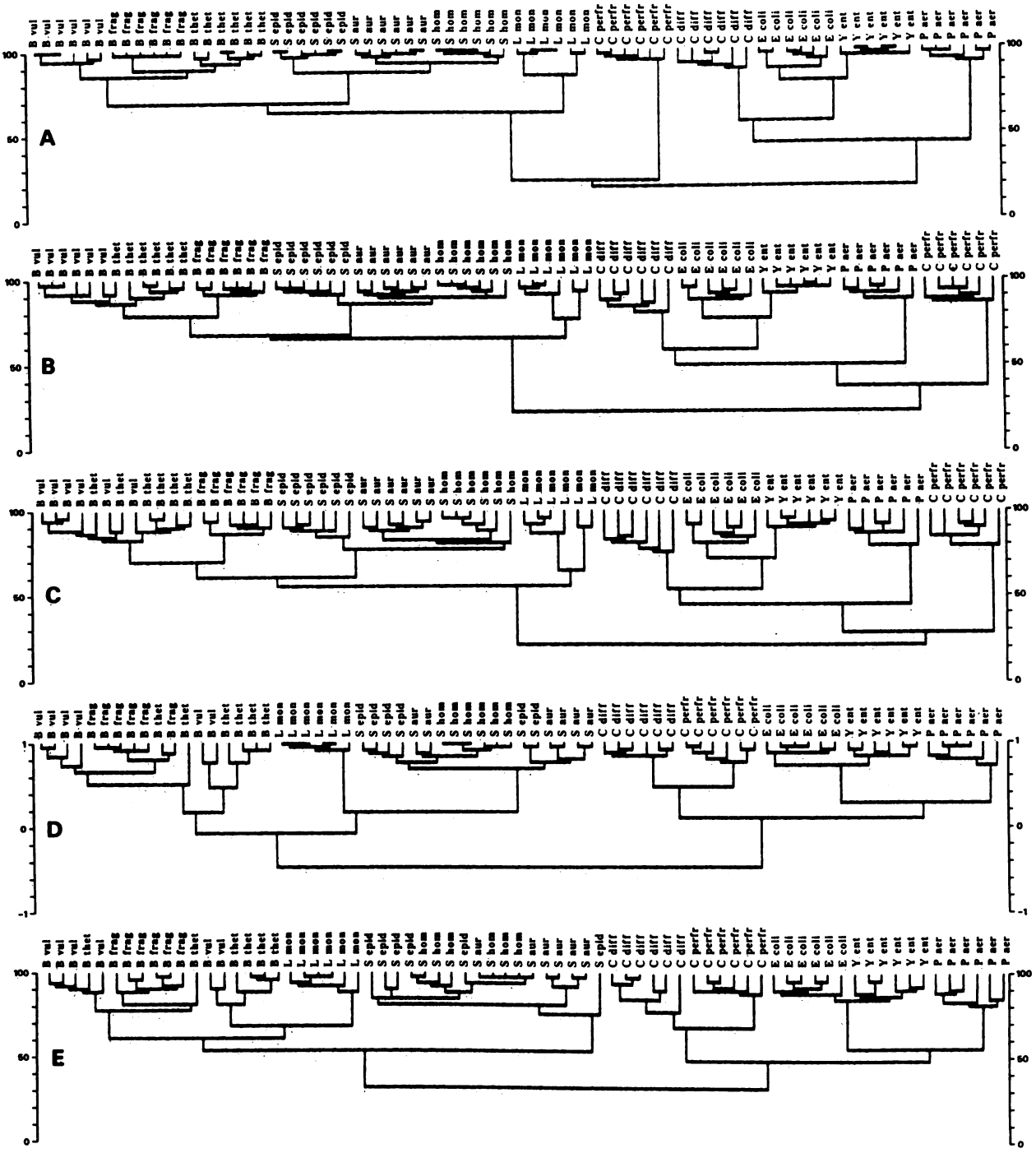
Comparison of the different data processing methods used for species identification. Figure 5 shows tree graphs that present the results of the correlation and subsequent cluster analysis of the 72 strains of the 12 species by the 10 methods. Table 1 shows the calculated index values for each method.

The ACL, calculated as the mean of the lowest clustering levels of each species, reflected how small the variation was within species, optimally 100. There was no marked difference in ACL among the methods. However, the Stack method was poorest in this respect. The species-specific CI values showed the ratio between the ACL (calculated separately for the three *Bacteroides* species, the three *Staphylococcus* species, the two enterobacterial species *Y. enterocolitica* and *E. coli*, and the two *Clostridium* species) and the mean clustering level at which these species were separated from each other. Thus, values close to 1.0 showed a complete inability to separate species. The further the values were below 1.0, the better was the ability to separate species. These intrafamilial index values must be compared with the numbers of species placed outside their own clusters, which tended to increase as the species-specific CI values improved. On the other hand, the number of species misallocated increased as the CI values approached 1.0.

None of the analyzed strains was allocated to a wrong cluster by the developed comparison function (method A), which was capable of separating the three *Bacteroides* species, the three *Staphylococcus* species, and the two enterobacterial species into separate clusters (with CI values within the range of 0.89 to 0.93). Almost equally good results were obtained with the overlapping method (method B). It misallocated only one strain, and the species separation was slightly lower than with the developed function (the CI ranged between 0.84 and 0.95). The ACL values were, however, not as good (87.5 versus 92.4), showing higher variations within clusters. Another method with relatively good results was the Stack method, if it was modified by weighting the peaks according to peak areas (method C). It misallocated six strains to wrong clusters. The other methods were clearly poorer in their ability to separate species. The correlation coefficient method (method G), the angle separation vector method (method H), and the method determining the peak presence (method F) gave similar results; they all had a marked tendency to combine several clusters. The Spearman rank correlation method (method D) and the methods which compared the presence of peaks (method E), peak size ratios directly (the Stack method [method I]), or peak sizes with the developed function (method J) but without weighting the peak areas all showed similar results, with several strains placed outside their own clusters. They also showed low ACL values.

The developed comparison function with weighted peak areas (method A) and the overlapping method (method B) were clearly superior to the others. These were further tested for their ability to separate numerous *Staphylococcus* and *Bacteroides* strains. Of 75 *S. aureus*, *S. epidermidis*, and *S. hominis* strains, 60 and 54 were allocated correctly by the exponential function method and the overlapping method, respectively. Of 113 *B. fragilis*, *B. thetaiotaomicron*, and *B. vulgatus* strains, 97 were allocated correctly by the exponential function method. The overlapping method allocated 86 strains correctly. This difference was due to the inability of the overlapping method to distinguish between *B. thetaiotaomicron* and *B. vulgatus* strains, which were distributed among several closely related subclusters (data not presented).

Clearly, if GLC is used for identification, a computer program must be available for comparison of the fatty acid profiles. Several methods have been published for the correlation analysis of chromatographic data (1). However, to our knowledge, no studies have been published that compare the efficacy of various methods in identifying clinical isolates. The results of the present study show that comparison of GLC data, without an optimal comparison procedure, may not lead to adequate identification of species. The new exponential comparison function was developed to adjust to the variation of chromatographic data. In the present study no false clustering and a high ACL were seen. The good results in comparison with other methods evidently depended on selecting a shape of the discriminant function which filtered out the noise included in the data but not the parts relevant to correct identification.



The correlation analysis function method separated the analyzed strains into species-specific clusters, with small variation within clusters. It was the only method which also separated the three *Bacteroides* spp. and the three *Staphylococcus* spp. into separate clusters. Separation between these species depended mainly on differences in the amounts of the fatty acids and not in their presence. Previously published correlation analysis methods have been compared for their ability to analyze bacterial fatty acid data (1). Hypothetical data and fatty acid profiles from coryneform bacteria were analyzed by using the correlation coefficient method (method G), the angle separation vector method (method H), and overlapping of chromatograms (method B). The overlapping method gave the best results. In the present work it was the only method with results comparable to those of the developed-function method. However, when used for the analysis of a large number of strains, it was found to be poorer at analyzing profiles with quantitative variations only.

All other methods clearly gave poorer results. This was due to their inability to adjust to variation of data and inconsistent prevalence of small peaks. Their results cannot

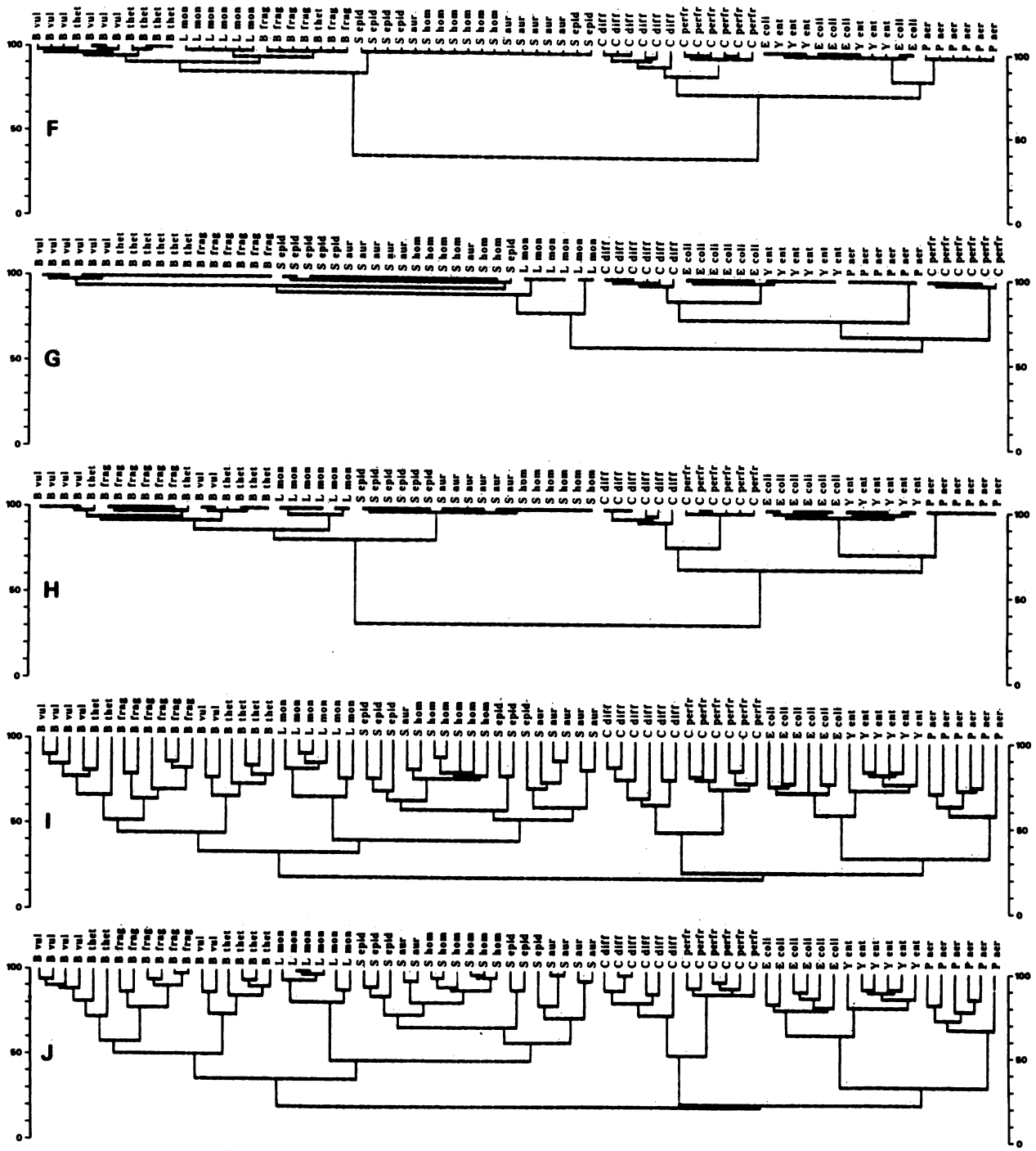


FIG. 5. Dendrograms showing the results of the various analysis methods of the GLC data. The methods, presented as letters from A to J, are the same as in the appendix.

be improved by raising the detection threshold to a level at which the presence of peaks would have been more constant. Wherever the threshold was, there would always be species having peaks with sizes varying about the selected threshold level. These peaks would appear inconsistently in GLC data and cause false clustering. Thus, weighting of peak areas is essential for optimal comparison.

The selected principle in the analysis procedure in identifying the clinical isolates was the correlation and subsequent cluster analysis of the GLC data. This method has clear advantages in clinical microbiology. Since all analyses are done by comparing chromatograms of presently and previously analyzed strains, the user can create reference files. The system can thus have different applications, depending on the type of microorganisms to be analyzed (13). We are currently using the present analysis method in automatic identification of clinical isolates. The fatty acid profile of the analyzed strain is transferred on-line from the chromatograph to the computer and immediately compared with a set of reference strains. The results of the analysis are printed as

TABLE 1. Efficacy of comparison methods in species identification of analyzed strains

Method[a]	ACL	Intrafamilial CI[b] of:				No. of species outside own cluster[c]	No. of species in wrong clusters[c]
		Bacteroides strains	*Staphylococcus* strains	Enterobacterial strains	*Clostridium* strains		
A. Exponential function weighted by peak size	92.4	0.89	0.93	0.86	0.13	0	0
B. Overlap	87.5	0.92	0.95	0.84	0.37	1	1
C. Stack method weighted by peak size	73.8	0.82	0.91	0.81	0.28	1	6
D. Spearman rank[d]	90.8	0.78	0.93	0.92	0.81	8	1
E. Presence of peaks	85.7	0.74	0.93	0.95	0.79	9	5
F. Presence of peaks weighted by peak size	97.2	0.94	1.0	1.0	0.89	0	30
G. Correlation coefficient	97.1	0.99	0.99	0.99	0.64	0	36
H. Angle separation vector	97.5	0.94	0.99	0.98	0.80	3	26
I. Stack method	64.2	0.69	0.82	0.78	0.63	8	3
J. Exponential function not weighted by peak size	78.1	0.57	0.78	0.57	0.78	9	1

[a] Methods and letters are the same as in the appendix.
[b] CI = 1 means complete inability to distinguish between species.
[c] Total number of analyzed strains was 72.
[d] The range −1 to 0.1 is rescaled to 0 to 100 for comparability with other results.

a list of strains with the closest resemblance to the tested strain. The tested strain is also placed in a dendrogram consisting of reference strains. The part of the dendrogram containing the test strain and those with the closest resemblance is printed with the results. The location of the tested strain in the reference dendrogram shows the level of correlation between the test strain and reference strains, as well as the correlation among the references themselves, indicating the confidence level of the identification. Further, if the test strain is not among the references, the dendrogram shows how closely it is related to certain reference species. The system has been tested with about 4,500 strains and has proved useful in rapid identification of clinical isolates.

Depending on the type of microorganisms analyzed and hence on data variation, the optimal comparison function may differ from the present one. The parameters defining the slope of the present discriminating function can be determined for each collection of standard reference isolates individually.

APPENDIX

A list of the correlation analysis methods used in the study is given below. Method A involves an exponential comparison function with weighted peak areas. It is explained in the text. Method B involves a similarity index based on overlap of fatty acid profiles:

$$C_0 = 100 - 0.5 \sum_{k=1}^{n} |x_{lk} - x_{jk}|$$

Method C is the Stack method with weighted peak areas:

$$C_{sw} = \sum_{k=1}^{n} 0.5(x_{lk} + x_{jk})(x_{lk}/x_{jk})/100$$

(x_{lk}/x_{jk}) is always inverted to be ≤ 1. Method D is the Spearman coefficient of rank correlation method:

$$C_r = 1 - 6 \sum_{k=1}^{n} d_{ij}^2 \Big/ (n^3 - n)$$

where d_{ij} is the difference of ranks of the peak pair. Method E is the similarity index method based on the presence of peaks:

$$C_p = \sum_{k=1}^{n} f(x_{lk},x_{jk}) \Big/ n$$

where $f(x_{lk},x_{jk}) = 100$ if x_{lk} and x_{jk} are present in both curves and $f(x_{lk},x_{jk}) = 0$ if x_{lk} or x_{jk} is missing in either curve. Method F is the similarity index method based on the presence of peaks with weighted peak areas:

$$C_{pw} = \sum_{k=1}^{n} 0.5(x_{lk} + x_{jk}) f(x_{lk},x_{jk}) \Big/ 100$$

where $f(x_{lk},x_{jk}) = 100$ if x_{lk} and x_{jk} are present in both curves and $f(x_{lk},x_{jk}) = 0$ if x_{lk} or x_{jk} is missing in either curve. Method G is the correlation coefficient method:

$$C_c = 50(1 + r_{ij}),$$

where

$$r_{ij} = \sum_{k=1}^{n} (x_{lk} - x_l)(x_{jk} - x_j) \Big/ \left[\sum (x_{lk} - x_l)^2\right]^{1/2} \left[\sum (x_{jk} - x_j)^2\right]^{1/2}$$

Method H is the similarity index method based on the angle of separation vectors:

$$C_a = \sum_{k=1}^{n} x_{lk}^{1/2} x_{jk}^{1/2}$$

Method I is the Stack method, not weighted by peak areas:

$$C_s = \sum_{k=1}^{n} (x_{lk}/x_{jk})n$$

where (x_{lk}/x_{jk}) is always inverted to be ≥ 1. Method J is the exponential comparison function, not weighted by peak areas. It is explained in the text. For all of the above methods, x_{lk} and x_{jk} are percentage areas of the kth peaks of organisms l and j, respectively; x_l and x_j are mean percentage peak areas of organisms l and j; and n is the number of different peaks in two compared organisms.

ACKNOWLEDGMENTS

We thank Kirsti Tuomela and Marja-Riitta Teräsjärvi for their excellent technical assistance and Eija Nordlund for help in preparation of the manuscript.

LITERATURE CITED

1. **Bousfield, I. J., G. L. Smith, T. R. Dando, and G. Hobbs.** 1983. Numerical analysis of total fatty acid profiles in the identification of coryneform, nocardioform, and some other bacteria. J. Gen. Microbiol. **129**:375–394.
2. **Drucker, D. B.** 1974. Chemotaxonomic fatty acid fingerprints of some streptococci with subsequent statistical analysis. Can. J. Microbiol. **20**:1723–1728.
3. **Drucker, D. B.** 1981. Analysis of structural components of

microorganisms by gas chromatography, p. 166–291. *In* D. B. Drucker (ed.), Microbiological applications of gas chromatography. Cambridge University Press, Cambridge.

4. **Drucker, D. B.** 1981. Computation of gas chromatographic data, p. 400–423. *In* D. B. Drucker (ed.), Microbiological applications of gas chromatography. Cambridge University Press, Cambridge.
5. **Guerrant, G. O., M. A. Lambert, and C. W. Moss.** 1982. Analysis of short-chain acids from anaerobic bacteria by high-performance liquid chromatography. J. Clin. Microbiol. **16**:355–360.
6. **Halebian, S., B. Harris, S. M. Finegold, and R. D. Rolfe.** 1981. Rapid method that aids in distinguishing gram-positive from gram-negative anaerobic bacteria. J. Clin. Microbiol. **13**:444–448.
7. **Hansen, M. V., and L. P. Elliott.** 1980. New presumptive identification test for *Clostridium perfringens*: reverse CAMP test. J. Clin. Microbiol. **12**:617–619.
8. **Ikemoto, S., H. Kuiraishi, K. Komagata, R. Azuma, T. Suto, and H. Murooka.** 1978. Cellular fatty acid composition in *Pseudomonas* species. J. Gen. Appl. Microbiol. **24**:199–213.
9. **Janzen, E., K. Bryn, T. Bergan, and K. Bovre.** 1974. Gas chromatography of bacterial whole cell methanolysates. V. Fatty acid composition of neisseriae and moraxellae. Acta Pathol. Microbiol. Scand. Sect. B **82**:769–779.
10. **Janzen, E., K. Bryn, T. Bergan, and K. Bovre.** 1974. Gas chromatography of bacterial whole cell methanolysates. VI. Fatty acid composition of strains within Micrococcaceae. Acta Pathol. Microbiol. Scand. Sect. B **82**:785–798.
11. **Janzen, E.** 1984. Analysis of cellular components in bacterial classification and diagnosis, p. 257–302. *In* G. Odhan, L. Larsson, and P.-A. Mårdh (ed.), Gas chromatography mass spectrometry applications in microbiology. Plenum Publishing Corp., New York.
12. **Moss, C. W., and O. L. Nunez-Montiel.** 1982. Analysis of short-chain acids from bacteria by gas-liquid chromatography with a fused-silica capillary column. J. Clin. Microbiol. **15**:308–311.
13. **Rizzo, A. F., H. Korkeala, and I. Mononen.** 1987. Gas-chromatographic analysis of cellular fatty acids and neutral monosaccharides in the identification of lactobacilli. Appl. Environ. Microbiol. **53**:2883–2888.
14. **Sneath, P. H. A., and R. R. Sokal.** 1973. The estimation of taxonomic resemblance, p. 114–187. *In* D. Kennedy and R. D. Park (ed.), Numerical taxonomy. W. H. Freeman and Co., San Francisco.
15. **Sneath, P. H. A., and R. R. Sokal.** 1973. Taxonomic structure, p. 288–308. *In* D. Kennedy and R. D. Park (ed.), Numerical taxonomy. W. H. Freeman and Co., San Francisco.
16. **Sokal, R. R.** 1985. The principles of numerical taxonomy: twenty-five years later, p. 1–20. *In* M. Goodfellow, D. Jones, and F. G. Priest (ed.), Computer assisted bacterial systematics. Academic Press, Inc. (London), Ltd., London.
17. **Stack, M. V., H. D. Donoghue, J. E. Tyler, and M. Marshall.** 1976. Comparison of oral streptococci by pyrolysis GLC, p. 57–68. *In* C. E. R. Jones and C. A. Cramers (ed.), Analytical pyrolysis. Elsevier Biomedical Press, Amsterdam.