Old Herborn University Seminar Monograph 7:
Immune system and microflora.
Editors: Peter J. Heidt, Volker Rusch, and Dirk van der Waaij.
Herborn Litterae, Herborn-Dill, Germany: 133-148 (2003).

CHARACTERISATION OF BACTERIAL SPECIES BY IMAGE ANALYSIS

B.C. MEIJER and G.J. KOOTSTRA

Laboratory for Medical Microbiology, University of Groningen,
Groningen, the Netherlands

SUMMARY

Microscopic images of pure cultures of eight different bacterial species were measured using digital image analysis. The morphometrical characteristics of cells from the same culture turned out to be varied, but consistent differences were shown to exist between species. Those differences can be used for rapid automatic detection of variations in the composition of gut microflora.

INTRODUCTION

The resident flora of the human gut is very complex: It consists of around 400 species, most of which are anaerobic (*Moore* and *Holdeman*, 1974). Culturing and identifying all of them is extremely time-consuming and therefore impractical and too expensive for routine use. Other methods for obtaining information about gut flora are needed.

Digital image analysis is a relatively new research method, which has not yet been much used in microbiology (e.g. *Sager* et al., 1988, *Reid* and *Hawthorn*, 1988, *Fernandes* et al., 1988). Similar work in industrial mycology using a Coulter counter instead of an image analyser has been reported (*van Huyns* et al., 1988). In our laboratory we intend to apply digital image analysis to the analysis of faecal flora. In very general terms our strategy is as follows: A large number (around 500) of bacterial cells are measured, yielding measurement values (e.g. its length, projected surface area, perimeter) for each cell. Every bacterium will thus be represented by a vector in the space of possible measurements. We shall call the distribution of those vectors for a population of bacteria its morphometrical distribution. From the morphometrical distribution we shall try to obtain information about the flora.

Before dealing with the complex gut flora we first study pure cultures, which are much simpler and better characterised. This approach may provide us with realistic expectations of the kind of information we may expect from image analysis when we study gut flora.

Even between cells in a single pure culture of a common species such as *E. coli* a great morphological diversity exists. *E. coli* cells are not all "Gram-negative rods": Some are coccoid, some are filamentous. Nevertheless a Gram-stained culture of *E. coli* has a characteristic microscopic appearance, to a large extent due to the morphologic diversity itself. We infer that it may be possible to characterise and distinguish pure cultures from different species by the morphometrical distributions of their bacterial cells. It is probably difficult, if not impossible, to deduce the species of one bacterium from its morphology, or

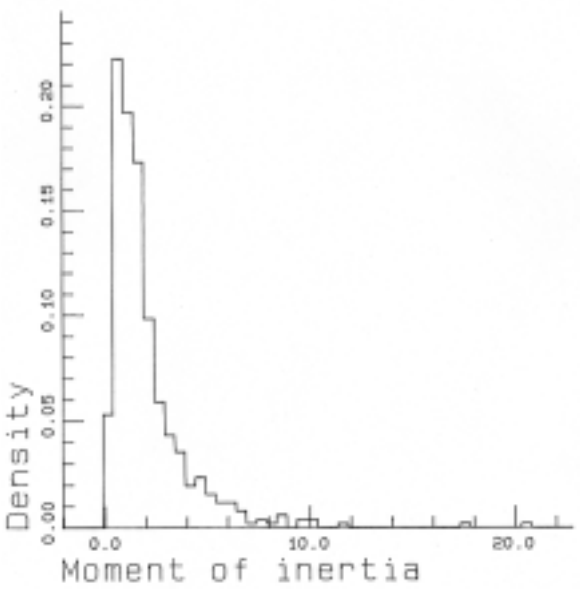

Figure 1: Distribution of the moment of inertia, I, for *K. pneumoniae*.

to disentangle the flora of the gut into it's many constituent species by morphometry.

Our present study is an attempt to characterise bacterial pure cultures of various species and culture ages. We chose aerobic species for two reasons. Firstly, they are easier to culture than anaerobes. Secondly, the questions we try to resolve by this experiment are very general: What do the morphometrical distributions for the various species look like; are they different between one species and another; do they depend on culture age; and how much information about a mixed culture can be obtained from its morphometrical distribution? Results and conclusions obtained for aerobes will guide further studies on the complex anaerobic faecal flora.

MATERIALS AND METHODS

Eight pure cultures were obtained from clinical material. They were of the following species: *Streptococcus pyogenes*, *Escherichia coli*, *Streptococcus* group D (Enterococcus), *Klebsiella pneumoniae*, *Proteus mirabilis*, *Pseudomonas aeruginosa* and *Staphylococcus aureus*. The identifications were made by standard laboratory methods. We subcultured the strains in BHI broth and grew them for four nights; with 24 h intervals we made nigrosin stained slides by a method modified from Fleming's (*Cruickshank* et al., 1975): We prepared staining fluid by adding 10 g of water-soluble nigrosin (GT Gurr) to 100 ml of a 0.25% solution of Tween-80. Equal volumes of culture broth and staining fluid were then mixed. A drop of the suspension was spread on a clean, fat-free microscopic slide as if making a blood smear. In such smears, bacteria show clear against a dark background. We made four slides for each strain. We recorded the time needed for the various steps in preparing the slides.

For image analysis we used a PC/AT

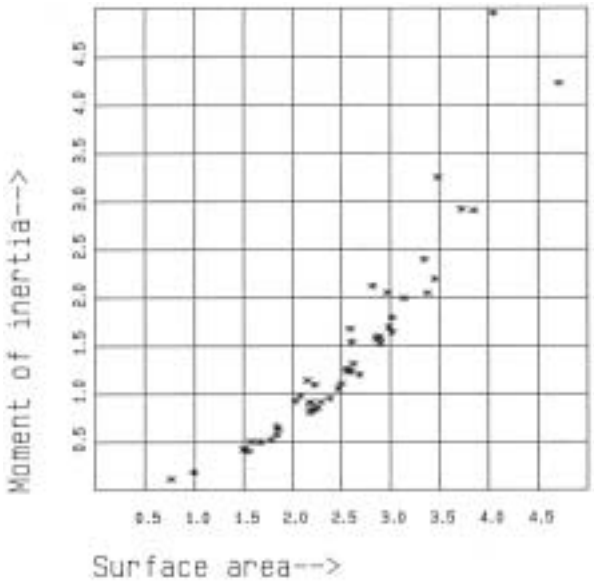

Figure 2: Scatter plot of surface area [$(\mu m)^2$] vs. moment of inertia [$(\mu m)^4$] for *K. pneumoniae*.

compatible computer equipped with a Matrox PIP 1024 video digitiser board (Matrox, Quebec, Canada). Except for the manufacturer's library, we developed the software ourselves.

For about 500 bacteria on each of the slides we measured the surface area (A), the perimeter (P), the moment of inertia (I) and the area of the convex hull (H).

DATA ANALYSIS

Linearisation and elimination of theoretical dependencies

The raw data from the measurements have undesirable properties from a statistical point of view. Firstly, the parameters have skewed distributions, as illustrated in Figure 1 for the moment of inertia of *Klebsiella pneumoniae*. Secondly, all variables depend on the size of the bacterium and so on one another. Finally, those relationships are not linear (Figure 2). New parameters were computed and logarithmic transformation was employed to overcome those problems. The process yielded four new variables: The logarithm of the area (a), the first and second form factors (f_1 and f_2), and the concavity index (c). These were used for subsequent analysis. Figures 3 and 4 show the improvements in distribution characteristics, independence on size and linearity. For a technical discussion we refer to Appendix A.

Mahalanobis transformation

For direct comparisons among species, representation of the measurements in terms of the new variables still carries some drawbacks. In particular, the measurements for a given species will scatter elliptically around the species mean, not spherically. Figure 5 illustrates this for *E. coli*, *K. pneumoniae* and *S. aureus*. We need to transform the whole plane so that the ellipses turn

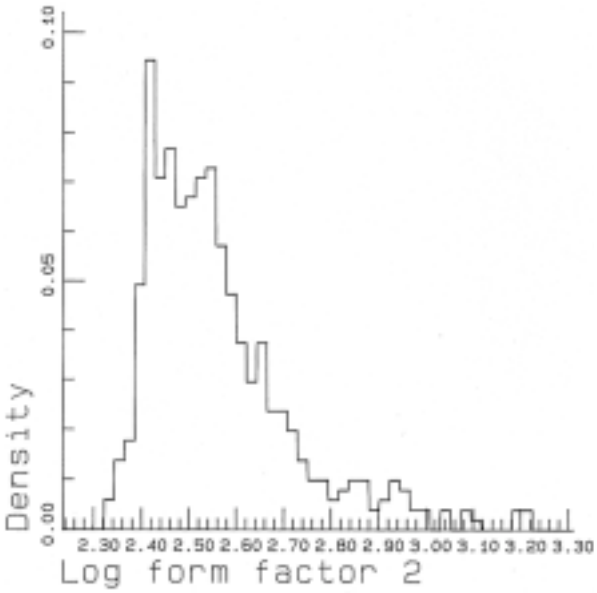

Figure 3: Distribution of the natural logarithm of the second form factor for *K. pneumoniae*.

into circles; the position of their centres will then be such that the distances among them are true measures of dissimilarity (Mahalanobis transformation; Figure 6). In the transformed space, comparison between any pair of species reduces to a comparison of the one-dimensional distributions along the axis connecting the two centres.

Secondly, we cannot be sure that keeping four variables to describe the data is ideal in terms of the ratio of sig-

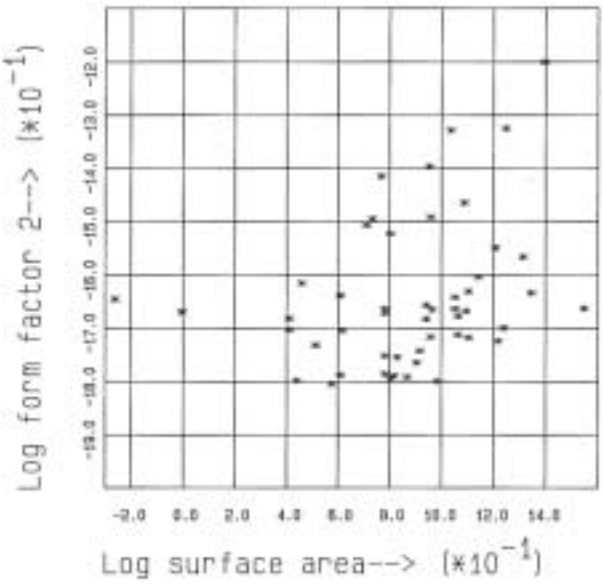

Figure 4: Scatter plot: Log surface area vs. log second form factor for *K. pneumoniae*.

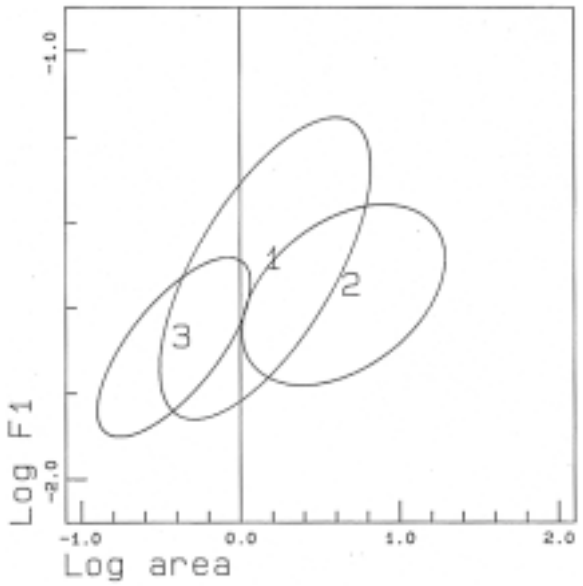

Figure 5: Scattering ellipses for *E. coli* (1), *K. pneumoniae* (2) and *S. aureus* (3) in the plane determined by area and form factor 2.

nal to noise. Besides, the analysis of mixtures, described below, requires economy in the use of variables. In Appendix B a technical method is described that combines Mahalanobis transformation and dimension reduc tion. Application of this method to our measurements yielded three new vari-

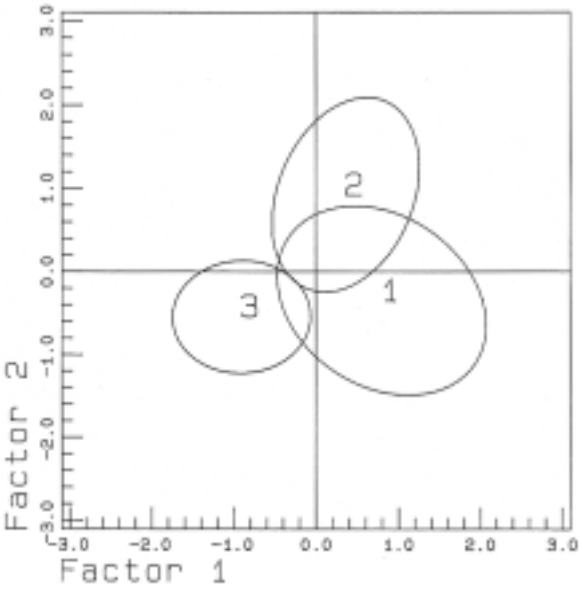

Figure 6: Scattering ellipses for *E. coli* (1), *K. pneumoniae* (2) and *S. aureus* (3) for the principal components derived from area and form factor 2.

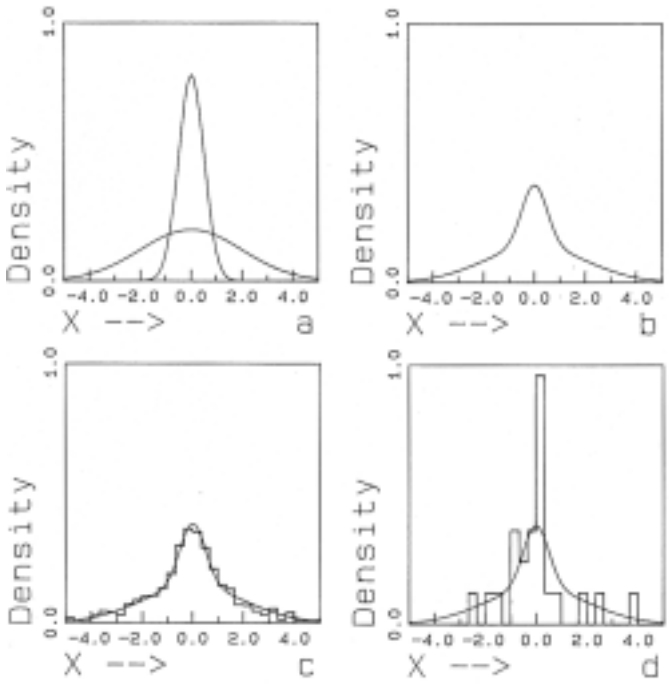

Figure 7: Simulated decomposition experiment with two easily distinguishable parent distributions. a) The two parent distributions; b) The composite distribution; c) Sample density for a sample of 1000; d) Sample density for a sample of 25.

ables: Factors 1, 2 and 3. Factor scores were computed for all the bacteria; the rest of the analysis was performed using only these factor scores.

Multidimensional scaling

At this stage, the dissimilarities among species were given as distances in three-dimensional space. For a more visual representation than a table of numbers can provide, a two-dimensional drawing was constructed by multidimensional scaling. In such drawings the distance relationships are preserved as closely as possible. For technical details we refer to *Mardia* et al., 1979.

Multivariate analysis of variance

A problem not yet addressed is the identification and quantitative description of the sources of variation in the data. How much variation can be ascribed to difference among species, to culture age and to the combination of culture age and species? Are the effects from those sources statistically significant and quantitatively important? Multivariate analysis of variance (MANOVA) was used to answer these questions. This technique is described in detail by *Mardia* et al (1979).

Decomposition of mixtures

A mixture containing many bacteria of more than one species will give rise to a morphometrical distribution, which is a weighted sum of the distributions of the constituent species. Mixtures with fewer bacteria in the same proportions will yield approximations to this sum. Figure 7 illustrates this for a simulated hypothetical pair of species; in this case it would be rather easy to determine the proportions from the larger sample. The

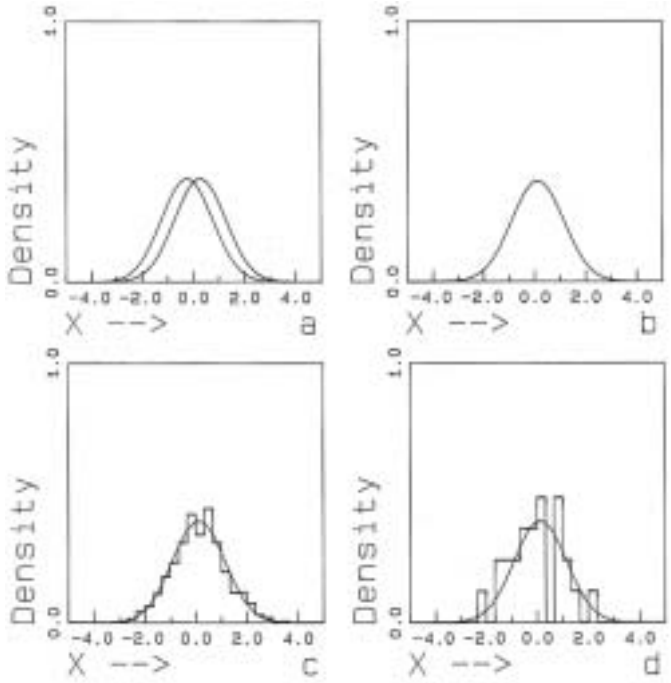

Figure 8: Simulated decomposition experiment with two similar parent distributions.
a: The two parent distributions; b) The composite distribution; c) Sample density for a sample of 1000; d) Sample density for a sample of 25.

smaller sample does not permit accurate reconstruction. Note that both distributions have the same mean: This does not prevent us from decomposing the mixture. In Figure 8 the differences between the species distributions are subtler, and decomposition is more difficult. The results of both simulation experiments are given in Table 1.

A statistical approach to this problem makes use of the maximum likelihood principle. The proportions from the various species assumed to be present in the mixture are adjusted so that the likelihood for the measurements to be as they are is maximal. For a technical description we refer to Appendix C.

Table 1: Results of one-dimensional simulation

Experiment	Mean	Variance	True proportions	Estimated proportions (n=1000)	Estimated proportions (n=25)
1	0.0	0.25	0.7	0.690 (0.03)	0.47 (0.15)
	0.0	4.0	0.3	0.310 (0.03)	0.53 (0.15)
2	-0.25	1.0	0.4	0.345 (0.04)	0.6 (0.4)
	0.25	1.0	0.6	0.655 (0.04)	0.4 (0.4)

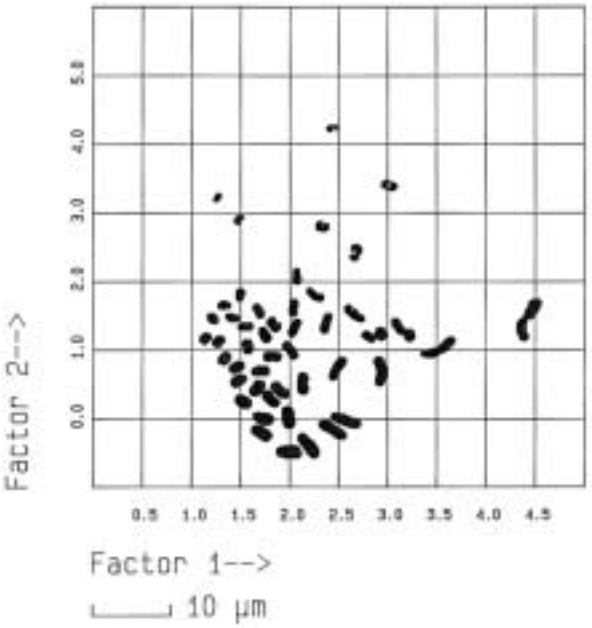

Figure 9: Modified scatter plot of *Klebsiella pneumoniae*: The bacteria have been drawn on the spots corresponding to their factor scores.

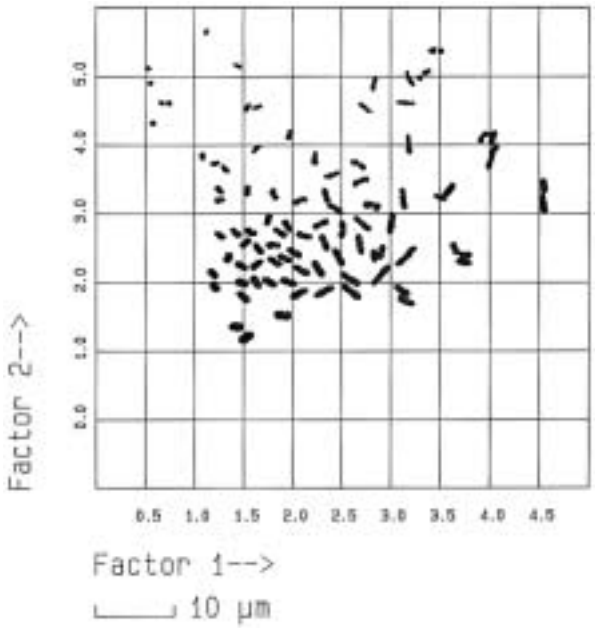

Figure 10: Modified scatter plot of *Pseudomonas aeruginosa*: The bacteria have been drawn on the spots corresponding to their factor scores.

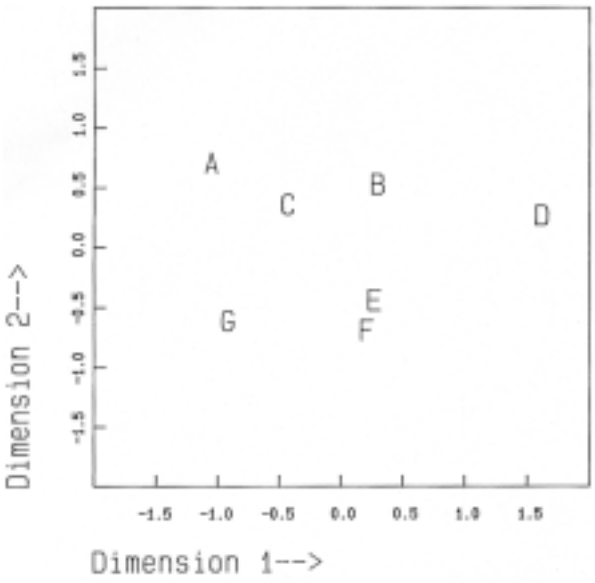

Figure 11: Distances between species centroids, drawn in two dimensions by multidimensional scaling. Meaning of the letters: A: *S. pyoqenes*. B: *E. coli*. C: *Enterococcus* sp. D: *K. pneumoniae*. E: *P. mirabilis*. F: *P. aeruginosa*. G: *S. aureus*.

To test the method, we prepared two random samples from the data files of the seven species, containing known proportions from each, and then computed maximum likelihood estimates for the proportions. The first sample contained 140 bacteria from each of the seven species. In the second only *Streptococcus pyogenes*, *E. coli* and *S. aureus* were present: 330 of each.

RESULTS

Twenty slides could be made in 6 h. Recording and measuring the images from those twenty slides took 26 additional hours, 11 of which required the presence of a human operator.

In Figures 9 and 10 we give two examples of the distribution of forms from pure cultures. Both figures are modified scatter plots in which the bacteria are drawn on the spots corresponding to their first two principal component scores.

Basic statistics of our measurements are given in Table 2. The centroids in principal component space for the species were all significantly different from one another ($p<0.1\%$). Table 3 records the standardised distances between species centroids: The unit of distance on each component is the pooled within-species standard deviation for that component. We derived Figure 11 from the distance table by multidimensional scaling.

The results of variance analysis are summarised in Table 4. We see that the effects of species, culture age and the combination of both are all significant, but the species effect dominates the others by a factor of at least 10.

Among all species distribution moments of up to third order were compared. Of the 399 comparisons made, only 39 yielded no significant difference at the 5% level (Student's t test).

Table 2 : basic statistics for the slides

Species	Day	N	Area (μm^2)	First form factor (-)	Second form factor (-)	Concavity factor (-)
S. pyogenes	1	631	1.50 (0.04)	16.7 (0.30)	0.25 (0.004)	0.15 (0.009)
	2	719	1.41 (0.04)	16.7 (0.30)	0.26 (0.005)	0.14 (0.007)
	3	615	1.47 (0.05)	16.8 (0.30)	0.26 (0.006)	0.15 (0.010)
	4	680	1.91 (0.06)	17.9 (0.30)	0.27 (0.005)	0.17 (0.009)
E. coli	1	380	2.25 (0.05)	16.4 (0.20)	0.26 (0.004)	0.08 (0.004)
	2	473	1.94 (0.06)	17.2 (0.30)	0.29 (0.006)	0.09 (0.005)
	3	493	2.07 (0.06)	18.6 (0.30)	0.31 (0.007)	0.14 (0.007)
	4	572	2.06 (0.05)	17.3 (0.20)	0.28 (0.005)	0.10 (0.005)
Enterococcus	1	418	1.97 (0.04)	18.1 (0.30)	0.30 (0.006)	0.17 (0.010)
	2	549	1.64 (0.04)	16.7 (0.30)	0.26 (0.004)	0.13 (0.007)
	3	498	1.61 (0.04)	16.4 (0.30)	0.26 (0.005)	0.13 (0.007)
	4	692	1.39 (0.04)	14.8 (0.20)	0.22 (0.003)	0.09 (0.005)
K. pneumoniae	1	508	2.93 (0.04)	13.1 (0.10)	0.20 (0.002)	0.04 (0.003)
	2	673	2.04 (0.04)	14.6 (0.20)	0.22 (0.002)	0.08 (0.003)
	3	811	2.05 (0.04)	14.1 (0.10)	0.21 (0.002)	0.06 (0.003)
	4	293	2.18 (0.07)	13.6 (0.20)	0.21 (0.003)	0.05 (0.004)
P. mirabilis	1	463	1.50 (0.03)	14.7 (0.20)	0.23 (0.003)	0.06 (0.003)
	2	634	1.50 (0.03)	14.9 (0.20)	0.23 (0.002)	0.07 (0.0031
	3	800	1.21 (0.02)	15.0 (0.20)	0.24 (0.003)	0.07 (0.003)
	4	1000	1.41 (0.03)	15.6 (0.20)	0.24 (0.003)	0.08 (0.004)
P. aeruginosa	1	721	1.41 (0.02)	15.5 (0.10)	0.25 (0.003)	0.07 (0.003)
	2	970	1.17 (0.02)	14.7 (0.10)	0.23 (0.002)	0.05 (0.003)
	3	1442	1.07 (0.01)	14.3 (0.07)	0.22 (0.001)	0.05 (0.002)
	4	1732	0.94 (0.01)	13.3 (0.05)	0.20 (0.001)	0.04 (0.001)
S. aureus	1	626	0.44 (0.01)	12.5 (0.10)	0.18 (0.002)	0.03 (0.003)
	2	865	0.88 (0.01)	13.6 (0.10)	0.20 (0.002)	0.07 (0.003)
	3	565	0.77 (0.02)	13.1 (0.10)	0.19 (0.002)	0.05 (0.004)
	4	542	0.85 (0.02)	13.1 (0.10)	0.19 (0.002)	0.05 (0.004)

The numbers between parentheses are the standard errors for the estimated means.

The results of our analysis of mixed data files are summarised in Table 5.

Determination of the most likely species for the 28 slides yielded correct classifications for all slides.

DISCUSSION

From the modified scatter plots we see that the variables measured reflect differences in shape and size. The morphometrical distribution for *Klebsiella* is, for example, easily distinguished from that for *Pseudomonas*. On the other hand, there is a substantial area in which both distributions overlap.

Table 3 : Distances between centroids for species in principal components

	1	2	3	4	5	6
1 S. pyoqenes	0.000					
2 E. coli	1.382	0.000				
3 Enterococcus sp.	0.642	0.865	0.000			
4 K. pneumoniae	2.145	1.593	1.595	0.000		
5 P. mirabilis	1.563	0.839	0.962	1.382	0.000	
6 P. aeruqinosa	1.696	1.053	1.148	1.614	0.291	0.000
7 S. aureus	1.172	1.607	1.020	2.094	1.145	1.xxx

The simplest mixed data file, containing only three dissimilar species, is dealt with satisfactorily (see Table 5, columns II and III). The estimates for the proportions are correct within their error margins, both when all seven species are used, and when only the three species actually present are employed to explain the moments of the resultant morphometrical distribution. The mixture of all seven species however is not resolved by our method. Note especially the obviously incorrect results for *Proteus* and *Pseudomonas*. When species are morphologically similar, there is not enough information in the mixture to distinguish between them. Obviously this problem depends on the number of species used: The similarity between a given species and the one most resembling rises with the size of the group to choose from.

Normal gut flora may contain up to about 400 different species. Even when many bacteria are measured, we cannot expect accurate estimates for the num-

Table 4: Analysis of variance

Source of variance	Degrees of freedom	Covariance matrix			F	Wilks' lambda
Species	6	372.4			414.54	
		-280.0	926.5		1018.31	
		88.2	-23.0	475.5	520.88	
						0.583
Age	3	10.5			11.72	
		0.2	70.0		76.98	
		15.9	7.1	25.0	27.39	
						0.982
Species and age combined	18	30.1			33.47	
		-21.4	78.4		86.15	
		6.4	12.5	21.0	23.03	
						0.880
Residual	19337	0.898				
		0.001	0.910			
		-0.007	0.012	0.913		

Table 5: Analysis of mixed data files

Species	I*			II			III		
	True	Comp	(SE)**	True	Comp	(SE)	True	Comp	(SE)
S. pyogenes	0.14	0.12	(0.02)	0.00	0.03	(0.03)	0.00	-	
E. coli	0.14	0.29	(0.04)	0.33	0.32	(0.02)	0.33	0.32	(0.02)
Enterococcus sp.	0.14	0.03	(0.04)	0.00	0.01	(0.03)	0.00	-	
K. pneumoniae	0.14	0.21	(0.03)	0.33	0.35	(0.02)	0.33	0.32	(0.02)
P. mirabilis	0.14	-0.37	(0.08)	0.00	0.00	(0.20)	0.00	-	
P. aerigunosa	0.14	0.54	(0.10)	0.00	0.00	(0.20)	0.00	-	
S. aureus	0.14	0.16	(0.02)	0.33	0.30	(0.03)	0.33	0.36	(0.02)

*: I = equal numbers of all species were used for the mixture.
II = species 2, 4 and 7 were present in equal numbers; data from all species were used to resolve the mixture.
III = The same mixture as in II, now resolved using only data from the species present.
**: True: true proportions.
Comp: computed proportions.
SE: estimated standard error for the computed proportions.

bers of all those species present in a specimen. On the other hand, we have shown the species to differ with regard to the means and higher moments of their morphometrical distributions. Those measures will therefore be sensitive to the balance between species composing the microflora.

A distinguishing and biologically important property of normal gut flora is its morphological diversity (*Baquero* et al., 1988). Loss of diversity will result in a narrower distribution. Digital image analysis may then provide a fast and reliable indication of diversity loss, such as occurs in overgrowth or with the therapeutic use of antibiotics.

The modified scatter plot is useful as an aid to the human viewer. It provides a systematical survey of large numbers of bacteria, grouped automatically according to their size and shape. Moreover, when viewed from a larger distance, it gives an indication of the form of the morphometrical distribution of all the bacteria together. Plots of this kind may be the most useful products of morphometrical analysis of mixed floras.

The discrimination between species by image analysis, though rather successful in our small group of species, is not practically useful. When more species are considered, morphometrical space would quickly get too crowded for accurate identification to be possible. In our view the great effort is needed to build a database and to chart the effects of environmental influences such as antibiotic use on bacterial shapes is not worthwhile.

What counts in practice is a simple and rapid method to determine a significant change in the composition of the intestinal microflora, for instance during antibiotic therapy. Some antibiotics, such as the β-lactams, influence the shape of bacteria directly, and so could disturb the morphometrical distribution of a strain, and its analysis. For the analysis of mixed flora this is problem is not as serious as it seems. If a certain strain in the flora is morphologically changed by the antibiotic, it is to be expected that the strain will not reproduce as effectively as it did before the therapy. It will therefore be overgrown by more resistant strains, whose morphol-

ogy remains unchanged. For this reason, we may disregard the direct effect of antibiotics on morphology.

Pure bacterial cultures display consistent differences in morphometric characteristics, as determined by image analysis. These differences are useful for the automatic detection of variations in gut flora. The method is fast enough for experimental use in a clinical setting. With more efficient image processing machines practical applicability is likely to improve.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. Dr. D. van der Waaij for his continuing involvement with the project and his numerous helpful suggestions at all stages of the work, and to Dr. Ir. G.J. te Neerman and Prof. Dr. W. Schaafsma for their help with the statistical analysis.

LITERATURE

Baquero, F., Fernandez-Jorge, A., Vicente, M.F., Alós, J.I., and Reig, M.: Diversity analysis of the human intestinal flora: A simple method based on bacterial morphotypes. Microb. Ecol. Health Dis. 1, 101-108 (1988).

Cruickshank, R., Duguid, P, Marmion, B.P., and Swain, R.H.A.: Staining methods. In: Medical Microbiology (Eds.: Cruickshank, R., Duguid, P., Marmion, B.P., and Swain, R.H.A.). Twelfth edition. Churchill Livingstone, Edinburgh, 33-34 (1975).

Fernandes, M.A, Jackman, P.J., Clark, S.A., and Gunard, S.R.: Detection and quantification of microorganisms in a heterogeneous foodstuff by image analysis. Comput. Appl. Biosci. 4, 291-295 (1988).

van Huyns, N., Guns, M., and Declaire, M.: Differentiation of *Saccharomyces cerevisiae* strains with a microcomputer-assisted Coulter counter. Microbios 57, 41-48 (1988).

Mardia, K.Y., Kent, T., and Bibby, J.M.: Multivariate Analysis. Academic Press, London (1979).

Moore, W.E.C. and Holdeman, L.V.: Human faecal flora: The normal flora of 20 Japanese Hawaiians. Appl. Environm. Microbiol. 27. 961-979 (1974).

Reid, G. and Hawthorn, L.A.: Use of an image analyzer to study bacterial adherence to polymer and prosthetic surfaces. Annual meeting of the American Society for Microbiology, Miami Beach, Florida, USA (1988).

Sager, B.M., Sekelsky, J.J, Matsumura, P., and Adler, J.: Use of a computer to assay motility in bacteria. Anal. Biochem. 173, 271-277 (1988).

APPENDIX A:
DERIVED VARIABLES; LOGARITHMIC TRANSFORMATION

The following formulae were used to find the transformed variables a, f_1, f_2 and c:

$a = \ln A$, the logarithm of the area ; {A.1}

$f_1 = 2 \ln P - \ln A$, the first form factor; {A.2}

$f_2 = \ln I - 2 \ln A$, the second form factor; {A.3}

$c = \ln H - \ln A$, the concavity factor. {A.4}

The first variable (a), measures size. f_1 indicates the irregularity of the contour; f_2 reflects circularity of the overall form of the bacterium, and c will be large if the bacterium is curved or otherwise concave.