JOURNAL OF CLINICAL MICROBIOLOGY, Nov. 1975, p. 391-396
Vol. 2, No. 5

Discriminant Analysis of Antibiotic Susceptibility as a Means of Bacterial Identification

GARY DARLAND

Enterobacteriology Branch, Center for Disease Control, Atlanta, Georgia 30333

Received for publication 7 July 1975

This study shows that antibiotic susceptibility data can be used effectively in the presumptive identification of bacteria. Using 12 antibiotics and determining the zone sizes for each, 82% of the isolates considered were correctly identified without any other information. If the inability to distinguish between *Escherichia coli* and *Shigella* is disregarded, the percentage of correct identification is 92%. The method involves determining a set of discriminant functions and defining each taxon by a unique function. An unknown isolate is identified by evaluating each discriminant function and assigning the isolate to the taxon whose discriminant function has the largest value. A total of 468 isolates were examined. After eliminating the multiply resistant isolates, the remaining 369 isolates were used to determine the discriminant functions for the eight taxa considered.

Clinical microbiologists often utilize antibiotic susceptibility data in an intuitive fashion to achieve a presumptive identification of an unknown isolate. Use of such data has the advantages of being rapidly obtained and at the same time providing the physician with valuable information regarding therapy. The purpose of this paper is to indicate that the presumptive identification of bacteria by using antibiotic susceptibility data is both a valid and accurate procedure.

Several models have been presented for the computer identification of bacteria. For the most part these models have been of two general types. A Bayesian model has been used recently by Friedman and co-workers (9, 10). The necessary estimates of prior probabilities were obtained from the relative frequency with which the various species were isolated.

The second model is a maximum likelihood model. This model is similar to the first, but no estimates of prior probability are used (7, 12). Some normalization method is used in all models of the second type to avoid extremely small numbers (7).

Both types of models require statistically independent variables. It is not at all certain whether this assumption is satisfied in diagnostic tests. Friedman and MacLowry (9) used a Bayesian model to identify clinical isolates on the basis of their antibiotic susceptibility patterns. As they indicate in a discussion of their results, this critical assumption is probably not valid for antibiotic susceptibility data.

Another approach to the computer identification of bacteria is possible. It utilizes the models of multivariate statistical analysis. In these models no a priori assumption regarding the independence of variables is required. Each individual observation is considered an *m*-component vector, and the relationships among these components are examined.

Gyllenberg (11) used one of these models, namely, principal component analysis, for the identification of bacteria. After determining the first three to five principal component scores on representatives of the species being considered, each species is located in hyperspace by determining a centroid. An unknown isolate is then assigned to the taxon to which it is closest in terms of Euclidean distance.

Another of these models, one in which discriminant functions are used (1), is illustrated in this paper. The discriminant functions are derived from antibiotic sensitivity data and, with the exception of *Escherichia coli* and *Shigella*, are shown to correctly identify 90 to 95% of the taxa represented.

The discriminant functions are obtained by solving the following set of equations: $p_i = \acute{a}_i X + c_i$ (equation 1), where p_i is the score for the discriminant function i and X is the data vector for an unknown isolate; a_i is the transposed vector of coefficients (a_i) obtained from $a_i = S^{-1} \bar{x}_i$ (equation 2), where S is the estimate of the common dispersion matrix and $\bar{x}_i$ is the mean vector for the species i. The constant c_i is evaluated at the centroid $c_i = -1/2\, \acute{a}_i \bar{x}_i$ (equation 3). The value of p_i is a scalar quantity, and an unknown isolate is assigned to the species for which the score of p_i is maximal (1, 6). The mathematical details can be found in the textbook by Anderson (1).

MATERIALS AND METHODS

Bacterial isolates. Four hundred and sixty-eight isolates representing six genera and nine species of the *Enterobacteriaceae* were initially included in the study. Strains of *E. coli, Shigella sonnei, S. flexneri, Salmonella typhi, Enterobacter cloacae, E. agglomerans,* and *Serratia marcescens* were chosen from stock cultures in the Enteric Section, Center for Disease Control (CDC), Atlanta, Ga. *Yersinia pseudotuberculosis* isolates were kindly supplied by E. Thal, National Veterinary Institute, Stockholm. *Y. enterocolitica* strains included 14 isolates identified by the CDC, 44 isolates supplied by L. Lafleur, Hopital Sainte Justine, Montreal, Canada, and 18 isolates obtained from R. Sakasaki, National Institute of Health, Tokyo, Japan.

Antibiotic susceptibility tests. High potency disks (BBL) of the following antibiotics were used: colistin, nalidixic acid, sulfadiazine, gentamycin, streptomycin, kanamycin, tetracycline, chloramphenicol, penicillin, ampicillin, carbenicillin, and caphalothin. The procedures of Bauer et al. (2) were followed. The diameters of the zones of inhibition were measured to the nearest millimeter. Each isolate could then be represented by a 12-component vector, where each element of the vector was the diameter associated with a single antibiotic.

Analytical procedures. Preliminary data analysis was performed by using the method of prinicpal component analysis (1). This method was chosen to determine whether the isolates under study could be legitimately considered as representing different populations (5).

The principal component analysis was performed by using a slightly modified version of FACTO (IBM, System/360, Scientific Subroutine Package, Version III). The first three principal components were used to determine whether the isolates represented a homogeneous population.

A discussion of discriminant functions can be found by Anderson (1). In essence, the method requires an a priori definition of groups from which are derived a set of discriminant functions. An observation vector representing an unknown isolate is used in these functions, and a single scalar quantity is determined; one scalar is determined for each discriminant function. The unknown can then be assigned to the group for which this number is maximum.

For our purposes the predefined groups represented the taxa listed above, with the exception that two *Shigella* species were combined into a single taxon. The discriminant functions were derived by using BMD07M (BMD, Biomedical Computer Programs, University of California Press, Berkeley, 1970). In the examples below we assumed that all taxa would be encountered with equal probability.

RESULTS

Antibiotic susceptibility of the isolates. Before initiating the discriminant analysis of antibiotic susceptibility, we recognized that the widespread occurrence of resistance transfer factor(s) within the *Enterobacteriaceae* would complicate the analysis. For this reason we attempted to eliminate those isolates suspected of containing episomal elements determining multiple antibiotic resistance. The following criteria had to be met for an isolate to be considered a member of this group. First, an isolate had to lack any detectable zone of inhibition around three or more of the following antibiotics: sulfadiazine, streptomycin, kanamycin, tetracycline, chloramphenicol, and ampicillin. Second, if over one-third of the isolates belonging to a given taxon showed no zone of inhibition around one of these antibiotics, that particular antibiotic was not considered in the definition of "multiply resistant strains." This definition resulted in the elimination of 99 isolates (27 *E. coli*, 7 *Shigella*, 41 *Salmonella typhi*, 1 *E. cloacae*, 2 *E. agglomerans*, and 21 *S. marcescens*) from discriminant analysis. The discriminant functions were determined on the remaining 369 strains.

After the "multiply resistant isolates" were excluded, the average diameter of the radial diffusion zone for each species was determined. The results are summarized in Table 1. Casual inspection of the data indicates that differences between the groups do exist. For example, the average diameter of the radial diffusion zone (i.e., zone size) for cephalothin ranges from over 33 mm for *Y. pseudotuberculosis* to 6.1 mm for *S. marcescens*. However, it would be very difficult to use the data in Table 1 to identify an unknown isolate. With the exception of occasional resistant mutants, the distribution of zone sizes for any antibiotic within any one species is essentially normal.

Principal component analysis. Before the necessary discriminant functions could be defined, it was necessary to determine whether the isolates represented different populations. Principal component analysis was used for this purpose (5). If the cultures from each species represented separate random samples from a single population, the statistics calculated for each species, that is, the mean and variance of principal component scores, should be approximately equal and should approximate the corresponding population parameters.

Table 2 presents the group means and variances for the first three principal components. The inequality of the group means and variances indicates that the species do, in fact, represent several populations.

Calculation of the discriminant functions. To determine the discriminant functions, an a priori definition of groups is required. For the

TABLE 1. *Mean diameter of radial diffusion zones observed for 8 bacteria and 12 antibiotics*

Antibiotic	Mean diam (mm)								Grand mean
	E. coli (73)[a]	*Shigella* sp. (22)	*S. typhi* (39)	*E. cloacae* (19)	*E. agglomerans* (18)	*W. marcescens* (102)	*Y. enterocolitica* (76)	*Y. pseudotuberculosis* (20)	
Colistin	10.3	12.7	13.8	10.2	12.7	8.9	15.2	10.6	11.6
Nalidixic acid	22.0	22.7	23.2	19.7	24.4	27.6	27.2	30.8	25.3
Sulfadiazine	19.3	11.4	28.1	21.9	25.4	22.7	28.2	25.7	23.3
Gentamycin	18.0	20.5	23.9	20.1	22.4	23.3	22.0	25.0	21.8
Streptomycin	14.4	12.9	14.2	16.4	19.5	17.8	17.4	21.4	16.6
Kanamycin	19.2	21.3	23.2	20.9	24.3	24.6	23.1	28.3	22.9
Tetracycline	17.6	16.8	23.5	16.2	21.7	13.4	21.9	21.4	18.2
Chloramphenicol	22.0	21.1	26.8	20.4	26.7	25.3	26.8	29.9	25.0
Penicillin	6.6	6.8	13.2	6.0	9.3	6.0	6.5	24.7	8.2
Ampicillin	18.5	18.3	24.6	8.6	16.7	11.9	11.2	33.1	16.0
Carbenicillin	23.1	22.7	25.0	21.8	18.9	25.0	12.3	36.3	22.0
Cephalothin	16.4	17.6	25.0	7.3.	22.6	6.1	10.9	33.4	14.2

[a] Number in parentheses represents the number of individual isolates.

TABLE 2. *Summary of principal component analysis*

Group	N[a]	Principal component					
		1		2		3	
		Mean	S^2[b]	Mean	S^2	Mean	S^2
E. coli	73	−0.624	0.46	0.364	0.69	0.229	0.19
Shigella sp.	22	−0.381	1.00	0.787	2.76	0.287	0.49
S. typhi	39	2.142	0.95	0.602	0.37	1.325	0.18
E. cloacae	19	−1.010	0.47	−0.555	0.25	−0.484	0.10
E. agglomerans	18	1.305	1.89	−0.564	0.91	0.835	0.45
S. marcescens	102	0.226	1.65	−1.140	0.52	−1.616	0.58
Y. enterocolitica	76	0.411	1.04	−1.744	0.74	0.974	0.37
Y. pseudotuberculosis	20	4.885	0.57	1.456	0.14	0.116	0.60

[a] N, Number of isolates per group.
[b] S^2, Within-group variance.

purposes of the present discussion, the groups will represent the taxa listed in Materials and Methods. Thirty-three isolates were removed at random from the 369 isolates remaining after the multiply resistant strains were eliminated. These isolates were to be used to test the ability of the discriminant functions to identify unknown isolates. Thus, the discriminant functions were determined on a total of 336 isolates representing eight different groups. The discriminant functions are given in Table 3.

Once the discriminant functions have been determined, an unknown isolate is assigned to the taxon in which the value of $\acute{a}_i X + c_i$ is the largest. In this equation, $\acute{a}_i$ represents the transpose of the discriminant function, X is the observation vector for the unknown isolate in which each element of the vector is the zone size in millimeters on the appropriate antibiotic, and c_i is a constant.

Classification of unknown. To test the ability of the discriminant functions to identify unknown isolates, the functions were tested by using the observation vectors of the 33 isolates that had been omitted from the determination of the discriminant functions. The results of the classification are summarized in Table 4.

The ability of the functions to properly identify unknown isolates is acceptable. With the exception of the confusion presented by the *E. coli/Shigella* group, the identification agreed with conventional methods (8) 92% of the time. The inability to distinguish *E. coli* from *Shigella* was not unexpected, since deoxyribonucleic acid homology studies between the two genera failed to demonstrate any significant differences (3).

In addition to assigning an unknown to the appropriate class, the posterior probability that an unknown organism belongs to the assigned

TABLE 3. *Coefficients for the discriminant functions*

Antibiotic	Discriminant function[a]							
	1	2	3	4	5	6	7	8
Colistin	0.91	1.21	1.22	0.62	1.15	0.12	1.75	0.50
Nalidixic acid	0.68	0.66	−0.05	0.11	0.31	1.10	0.89	1.19
Sulfadiazine	0.16	−0.19	0.42	0.31	0.24	0.22	0.32	0.30
Gentamycin	2.25	2.92	4.38	3.13	2.85	3.46	2.72	2.86
Streptomycin	−0.53	−0.87	−1.01	−0.58	−0.54	−0.85	−0.68	−0.64
Kanamycin	0.07	0.24	−0.38	0.36	0.50	0.37	0.34	0.54
Tetracycline	0.24	0.23	0.52	0.19	0.31	−0.21	0.32	0.06
Chloramphenicol	0.90	1.14	1.17	0.90	1.13	1.17	0.97	1.04
Penicillin	0.16	0.12	1.09	1.01	0.62	1.05	0.92	3.01
Ampicillin	−0.03	−0.20	−0.06	−0.74	−0.46	−0.71	−0.16	−0.59
Carbenicillin	0.32	0.27	0.18	0.56	0.19	0.63	−0.06	0.46
Cephalothin	0.97	1.14	1.36	0.15	1.57	−0.08	0.31	1.66
Constant[b]	−57.02	−69.96	−103.98	−57.53	−86.07	−77.76	−79.94	−143.36

[a] 1, *E. coli*; 2, *Shigella* sp.; 3, *S. typhi*; 4, *E. cloacae*; 5, *E. agglomerans*; 6, *S. marcescens*; 7, *Y. enterocolitica*; 8, *Y. pseudotuberculosis*. Total number of isolates used to determine function was 336. The numbers in the columns represent the coefficients (a'_i).

[b] $c_i = -{}^1/_2\, a'\bar{x}_i$.

TABLE 4. *Classification of 33 unknown isolates by discriminant analysis*

Group[a]	Discriminant function group							
	E. coli	*Shigella* sp.	*S. typhi*	*E. cloacae*	*E. agglomerans*	*S. marcescens*	*Y. enterocolitica*	*Y. pseudotuberculosis*
E. coli	4	3						
Shigella sp.	1	1						
S. typhi			3					
E. cloacae				1				
E. agglomerans					1			
S. marcescens				1		9		
Y. enterocolitica				1			6	
Y. pseudotuberculosis								2

[a] Groups were established by conventional biochemical tests (8).

group given the value of the discriminant function can be calculated by solving the following equation:

$$P_{ik} = \frac{\exp(p_{ik})}{\sum_{i=1}^{q} \exp(p_{ik})} \qquad (4)$$

where P_{ik} represents the posterior probability of case k coming from group i, p_{ik} is the value of the discriminant function for case k evaluated for the discriminant function i (equation 1), and q represents the total number of discriminant functions available. An example of the type of results obtained are shown in Table 5. Although in all four cases the unknown is assigned to one of the eight groups, in one instance the probability associated with this identification was not high. That is, isolate 31 was assigned to *Y. enterocolitica*, but with a posterior probability of only 0.58. In this case objective consideration of the posterior probabilities indicates that identification of the isolate as *Y. enterocolitica* is tenuous at best.

DISCUSSION

The preceding data illustrate another approach to the computer identification of bacteria. Although the mathematics at first seems cumbersome, several computer programs are available that perform the necessary calculations (6).

The question that must be answered in problems of identification is: Given an individual isolate with certain measured characteristics, from which population does it come? Phrasing the problem in this manner immediately sug-

TABLE 5. *Posterior probabilities of unknown organism isolate belonging to the group to which it is assigned*

Isolate no.	Probability of identification							
	E. coli	*Shigella* sp.	*S. typhi*	*E. cloacae*	*E. agglomerans*	*S. marcescens*	*Y. enterocolitica*	*Y. pseudotuberculosis*
1	0.027	*0.973*[a]	0.000	0.000	0.000	0.000	0.000	0.000
11	0.000	0.000	*0.995*	0.000	0.000	0.000	0.000	0.000
21	0.000	0.000	0.000	0.086	0.000	*0.907*	0.006	0.000
31	0.002	0.415	0.000	0.000	0.003	0.000	*0.580*	0.000

[a] Italicized number indicates taxon to which unknown was assigned.

gests the two major dilemmas that are common to all attempts to devise a computer identification algorithm. First, all the pertinent taxa involved must have been included and adequately defined. An individual cannot be assigned to a taxon that does not "exist." Furthermore, when defining the groups extreme care must be taken that the definitions are based on random samples from the appropriate populations. In the example above, we considered only eight different populations, all of which are defined by conventional biochemical methods (8).

These taxa were then "redefined" in terms of antibiotic susceptibility data by the derivation of the eight discriminant functions (Table 3). The use of discriminant functions has one major advantage over likelihood models (7, 12) and Bayesian models (9, 10); namely, the tests being measured need not be independent. The discriminant functions are derived in such a way that differences between the groups is maximized. Since equation 2 involved an estimate of the common dispersion matrix, a critical assumption is that the group dispersion matrixes be equal. This assumption may not be valid in the above example. However, Cooley and Lohnes (4) suggested that the model is reasonably robust and tolerates deviations from this assumption reasonably well. In view of this fact, no attempt was made to test the equality of the dispersion matrixes.

An advantage of an identification scheme based on antibiotic susceptibility is the speed with which the data can be obtained. The time between obtaining a pure culture and completing its identification is about 24 h. Most clinical laboratories probably determine antibiotic susceptibility for reasons related to therapy. It would be economical to be able to utilize these data in the identification of isolates.

The general procedure for this type of identification scheme would be to determine the discriminant function for each group being considered. One discriminant function is needed for each species. Once the functions have been determined, they can be suitably stored in the computer and recalled as needed. An unknown isolate is then tested on the set of antibiotics being used. All of the antibiotics must be employed. An observation vector is thus determined for the unknown. This vector is used in the discriminant function to achieve an identification. Not only is an identification obtained, information regarding the reliability of the classification is also given. Although only eight taxa were considered in this paper, there is little difficulty extending the approach to as many groups as desired.

Although *E. coli* could not be distinguished from *Shigella* by the use of discriminant functions, 82% of the rather limited number of organisms with which it was confronted were correctly identified. If *E. coli* and *Shigella* are considered as a single group, as suggested by deoxyribonucleic acid homology (3), 94% (31/33) of the isolates were correctly identified. This is a reasonably high degree of accuracy for a presumptive test. The accuracy may be increased in several ways. First, the covariance matrix can be weighted so that the prior probability of isolating a particular species is considered. Several rapid biochemical tests may also be included to increase resolution between such species as *E. coli* and *Shigella*. Finally, although the zone of inhibition is used as the indicator of antibiotic susceptibility in the model, it may be better to consider the minimum inhibitory concentration.

A satisfactory computer model for identification must be capable of identifying atypical isolates as well as "typical" individuals. In this case, the unusual isolates would be represented by the multiply resistant strains. In the above model a multiply resistant isolate may be considered in two ways. First, the use of minimum inhibitory concentrations could help determine differences between "resistant" isolates. Second, if zone sizes are to be used, it may be possible to define a group that contains multiply resistant isolates regardless of the taxa to which they actually belong. Isolates classified in this group could then be analyzed independently. Preliminary studies have tended to support this approach. It does, however, suffer from the disadvantage of requiring two separate analyses.

In conclusion, a systematic analysis of antibiotic susceptibility data, as described, would greatly facilitate presumptive identification of clinical isolates. Since such information is already being used, although often intuitively, some attempt should be made to objectively analyze these data. The advantages of using antibiotic susceptibility data to identify microorganisms are obvious, and with the growing tendency toward centralized data processing the analytical chore is not overwhelming.

LITERATURE CITED

1. Anderson, T. W. 1958. An introduction to multivariate statistics. John Wiley and Sons, New York.
2. Bauer, R. W., W. M. M. Kirby, J. C. Sherris, and M. Turck. 1966. Antibiotic susceptibility testing by standardized single disk method. Am. J. Clin. Pathol. 45:493-496.
3. Brenner, D. J., G. R. Fanning, F. J. Skerman, and S. Falkow. 1972. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. J. Bacteriol. 109:953-965.
4. Cooley, W. W., and P. R. Lohnes. 1971. Multivariate data analysis. John Wiley and Sons, New York.
5. Darland, G. 1975. Principal component analysis of infraspecific variation in bacteria. Applied Microbiol. 30:282-289.
6. Dixon, W. J. 1970. BMD biomedical computer programs, 2nd ed. University of California Press, Berkeley.
7. Dybowski, W., and D. A. Franklin. 1968. Conditional probability and the identification of bacteria: a pilot study. J. Gen. Microbiol. 54:215-229.
8. Edwards, P. R., and W. H. Ewing. 1972. Identification of *Enterobacteriaceae,* 3rd ed. Burgess Publishing Co., Minneapolis.
9. Friedman, R., and J. MacLowry. 1973. Computer identification of bacteria on the basis of their antibiotic susceptibility patterns. Appl. Microbiol. 26:314-317.
10. Friedman, R. B., D. Bruce, J. MacLowry, and V. Brenner. 1973. Computer assisted identification of bacteria. Am. J. Clin. Pathol. 60:395-403.
11. Gyllenberg, H. G. 1965. A model for computer identification of microorganisms. J. Gen. Microbiol. 39:401-405.
12. Lapage, S. P., S. Bascomb, W. R. Willox, and M. A. Curtis. 1970. Computer identification of bacteria, p. 1-22. *In* A. Bailie and R. J. Gilbert (ed.), Automation, mechanization and data handling in microbiology. Academic Press Inc., London.