Effect of Atypical Antibiotic Resistance on Microorganism Identification by Pattern Recognition

JAMES C. BOYD†, JOHN W. LEWIS,¹ J. JOSEPH MARR,² ALICE M. HARPER,³ AND BRUCE R. KOWALSKI³

Division of Laboratory Medicine, Departments of Pathology and Medicine, Washington University School of Medicine, St. Louis, Missouri 63110¹; Department of Medicine, St. Louis University School of Medicine, St. Louis, Missouri 63104²; and Laboratory of Chemometrics, University of Washington, Seattle, Washington 98195³

Received for publication 23 August 1978

We classified microorganisms from the clinical laboratory by using information provided by the Gram stain and antibiotic sensitivity profiles obtained with the Bauer-Kirby technique. Approximately 4,000 microorganisms, routinely identified and tested for antibiotic sensitivities in a large hospital microbiology laboratory, were used as a data set for several pattern recognition classification methods: *K*-nearest-neighbor analysis, statistical isolinear multicomponent analysis, Bayesian inference, and linear discriminant analysis. *K*-nearest-neighbor analysis yielded the highest prospective classification accuracy for gram-negative organisms, 90%. When those organisms displaying an atypical antibiotic resistance pattern were excluded from the data, the gram-negative classification accuracy improved to 95%. These results are inferior to currently accepted biochemical identification methods. Microorganisms with atypical antibiotic resistance patterns are likely to be misidentified and are common enough (17% of our isolates) to limit the feasibility of routine identification of microorganisms from their antibiotic sensitivities.

Prior attempts to classify microorganisms on the basis of their antibiotic sensitivities have met with varying success (6, 10, 18, 20). None of these studies has shown a prospective identification accuracy greater than 86%, not high enough to allow routine use of antibiotic sensitivities for organism identification. The question of what effect multiply resistant organisms would have on the classification accuracies of these methods has remained unanswered.

Computer-based pattern recognition methods have been applied successfully to classification problems in a variety of disciplines, including analytical chemistry (14), image processing (19), wave form analysis (11), and optical character recognition (21). These methods appeared promising when applied to the problem of microorganism identification from antibiotic sensitivity data (18); however, this study had the limitation of a small data base which did not include multiply resistant bacterial strains. Other methods used in the identification of microorganisms from their antibiotic sensitivities have required assumptions to be made about the distribution of sensitivity data (10, 20) or about the statistical independence of sensitivity measurements for

different antibiotics (6). Since these assumptions are unlikely to hold for antibiotic sensitivity data, we believed that a study using pattern recognition methods which did not require these assumptions was indicated.

Our study was designed to assess the accuracy of microorganism identification from antibiotic sensitivity data and to evaluate the effect of multiple antibiotic resistance on these identifications.

MATERIALS AND METHODS

Routine antibiotic sensitivities were measured in the Barnes Hospital Microbiology Laboratory by the technique of Bauer et al. (1). Microorganisms were identified in the laboratory by using biochemical methodology. *Enterobacteriaceae* were identified with the Analytab Products Co. API20E system; when necessary, the identification was supplemented with conventional biochemical tests, as recommended by Edwards and Ewing (9). *Pseudomonadaceae* were identified with conventional biochemical tests, as described by Hugh and Gilardi (12). Staphylococcal species were identified with Gram stain, catalase, and coagulase tests. Streptococci were identified by using catalase, growth in bile, hydrolysis of esculin, and growth in 6.5% sodium chloride. Sensitivity zone diameters and organism identifications for 4,213 consecutive, unselected specimens were collected over a 3.5-month period from the on-line computer system serv-

† Present address: University of Virginia Medical Center, Charlottesville, VA 22901.

ing the Barnes Hospital Microbiology Laboratory (17). Included in this data set were 3,448 gram-negative isolates. Organism classes having fewer than 45 isolates were excluded. This yielded a data base having eight classes of gram-negative organisms and three classes of gram-positive organisms (Table 1). These classes encompassed 94% of the isolates in our clinical laboratory. The sensitivity data for infrequently tested antibiotics were not considered in this study. The antibiotics used in identification were as follows: for gram-negative organisms—ampicillin, carbenicillin, cefamandole, cephalothin, colistin, gentamicin, kanamycin, tetracycline, and tobramycin; for gram-positive organisms—ampicillin, cephalothin, chloramphenicol, clindamycin, erythromycin, gentamicin, methicillin, penicillin, tetracycline, tobramycin, and vancomycin.

Pattern recognition methods (13) were used to predict organism identities from their antibiotic sensitivity measurements. The basic function of the pattern recognition techniques was to reduce multidimensional data (antibiotic sensitivity measurements for multiple antibiotics) to meaningful information (organism identifications). Once the data were collected, pattern recognition was carried out in several steps: initial data examination, feature selection, classifier training, and classifier testing.

Initial data examination. In this phase, the raw sensitivity measurements were plotted to allow examination of the distribution of antibiotic sensitivity zone diameters and to gather preliminary impressions of how closely the antibiotic sensitivity measurements clustered within each group of organisms.

Frequency histograms of the zone size measurements were the first method used to plot the data. These plots were relied upon to detect non-normal distribution of the measurements for each antibiotic. If the measurements for individual antibiotics did not follow the normal bell-shaped Gaussian distribution, the necessity for using pattern recognition methods which did not rely on the assumption of normally distributed data would be apparent.

To assess how closely the antibiotic sensitivity measurements clustered within each group of organisms, a means was needed for considering the multi-

ple-antibiotic zone size measurements of each isolate simultaneously. One way of accomplishing this was to represent each organism as a labeled point in an N -dimensional plot (N was equal to the number of antibiotics). Each of the N axes would then correspond to an antibiotic; e.g., the ampicillin zone diameter measurement for each organism would serve to position the labeled point for that organism along the ampicillin axis and the measurements for the other antibiotics correspondingly for the other axes. If only two antibiotics were considered, this would reduce to simple two-dimensional plotting. If three antibiotics were plotted, a perspective display could be used to visualize the three-dimensional plot. However, we were concerned with more than three antibiotics and thus with spaces of higher dimensionality than three. These, of course, cannot be visualized directly, but can be manipulated with mathematical methods that follow direct analogies to analyzing two- or three-dimensional space. The principal-components technique (3), which we used to visualize our data points in multidimensional space, is a mathematical method which allows one to project or map the points in higher-dimensional space onto a two-dimensional space (i.e., creating a two-dimensional plot), with the criterion that the data structure in the higher-dimensional space be preserved to the greatest possible extent. In this way (see Fig. 1), we were able to examine directly the clustering relationships which existed among the antibiotic sensitivity data of our organisms when they were considered as points in a multidimensional space.

Feature selection. After some appreciation for the underlying data structure had been gained through initial data examination, the next problem to be solved was to determine which antibiotics supplied the most useful information for identifying organisms. In the pattern recognition literature, this would be called a problem of feature selection. In approaching this problem, we applied several feature selection methods (15, 16) to reduce, if possible, the number of measurements used in subsequent analyses and to enhance the information representation of the data by selecting only those measurements which had value in the identification of organisms. The overall purpose of our feature selection was to extract the most useful information for organism identification from the fewest measurements.

Classifier training and testing. The next step in the pattern recognition process was the development of optimal schemes for organism identification from antibiotic sensitivity measurements. Subsets of 446 and 563 organisms selected at random from the gram-positive and gram-negative groups, respectively, were used in developing organism classification rules. These organisms were called the "training set." By using those antibiotic sensitivity measurements ranked highest by the feature selection methods, organism classification rules were developed from the examples in the training set by application of four different learning methods: K -nearest-neighbor analysis (KNN) (5), statistical isolinear multicomponent analysis (SIMCA) (22), linear discriminant analysis (LDA) (2), and a classifier utilizing the classification rule of Bayes (BAYES) (8). These methods were representative of the major pattern recognition approaches

TABLE 1. Organism classes studied

Class	Organism	No. of isolates
Gram negative		
1	<i>Enterobacter</i> sp.	233
2	<i>Klebsiella pneumoniae</i>	442
3	<i>E. coli</i>	1,452
4	<i>Pseudomonas aeruginosa</i>	525
5	<i>Proteus mirabilis</i>	396
6	<i>S. marcescens</i>	98
7	<i>C. diversus</i>	46
8	<i>P. morganii</i>	52
Gram positive		
1	<i>Staphylococcus aureus</i>	433
2	<i>Enterococcus</i>	149
3	<i>Staphylococcus epidermidis</i>	141

which have been applied to similar problems. One of our chief objectives in using several different methods was to minimize the likelihood that our results would be biased by the particular method chosen. Comparison of the results obtained by different methods was also expected to be valuable.

Those organisms which were not part of the training set formed a "test set," which was used for prospective evaluation of the performance of the various classification rules developed above in classifier training. This was an important evaluation step since it tested the accuracy of the classifiers which had been developed in the preceding steps, using data other than those used to train the classifiers.

Atypically resistant organisms. To evaluate the effect of multiply and/or atypically resistant species, each gram-negative organism group was split into subgroups by noting the antibiotics to which each isolate was highly resistant. An individual organism was defined to be highly resistant to a particular antibiotic if it grew to the edge of the disc impregnated with that antibiotic. Thus, *Escherichia coli* highly resistant to ampicillin formed one subgroup, whereas *E. coli* highly resistant to tetracycline and cephalothin formed another subgroup.

There is ample evidence in the pattern recognition literature (4, 7) that to evolve a reliable classification rule, one needs at least three to five times as many members in each group of the training set as there are dimensions (antibiotics). Thus, since we had nine antibiotic measurements for each organism, each subgroup had to contain at least 27 members (three times nine antibiotics) to develop a reliable classification rule for that subgroup. Sparsely populated subgroups ($N < 27$) were excluded from study on this basis. This resulted in 14 subgroups, encompassing seven of the original eight organism classes and 83% of the original gram-negative isolates. *Serratia marcescens* had no subgroups with greater than 21 members and thus was totally excluded from this phase of the study. Except for *S. marcescens*, the excluded subgroups represented only a small fraction of the isolates in their respective parent groups and thus were considered to be the organisms with atypical antibiotic resistance patterns. A training process similar to that described above was carried out on the remaining subgroups for three of the four classification methods. Since the performance of LDA had been the worst of the four algorithms employed for the previous phases (*Citrobacter diversus* was completely misclassified by this method as *E. coli*), it was eliminated from further consideration.

RESULTS

Zone size distributions. Histograms of the distribution of zone diameters for individual antibiotics showed a characteristic double-peaked (bimodal) shape for many antibiotics. One of the peaks occurred at a zone size of 6 mm, indicating bacterial species which had grown to the edge of the antibiotic-impregnated disc. The other peak occurred at various zone diameters, depending on the antibiotic sensitivity of the organism. The fact that the bimodality of zone size distributions for individual antibiotics persisted even within

a single class of organisms showed that the antibiotic sensitivity data were not distributed normally. This supported our choice of pattern recognition classification techniques which did not require normally distributed data.

Gram-positive organisms. Only 8 of the 11 available antibiotic sensitivity measurements for each gram-positive isolate were necessary to achieve optimal identification accuracy for the gram-positive organisms; the addition of chloramphenicol, erythromycin, or tobramycin did not improve the identification accuracy. The results of SIMCA classification of a randomly selected subset of gram-positive organisms, using these eight antibiotic sensitivity measurements for each organism, are shown in Table 2. Although the overall classification accuracy of 96% indicated that these species could be reliably identified from their antibiotic sensitivities, the routine identification of these organisms by inspection of colonies and use of the catalase and coagulase tests is both reliable and rapid. Further analysis was confined, therefore, to the more important problem of identifying gram-negative organisms.

Gram-negative organisms. Feature selection did not reduce the number of antibiotics needed for optimal gram-negative identification. *C. diversus* and *Proteus morganii* had too few isolates to form both a (retrospective) training set and a (prospective) test set; hence, the prospective identification study included only six organism classes. The accuracy of retrospective and prospective identification, using the data from the nine antibiotics, is shown in Table 3. The best results were obtained with KNN, 90% in the training and test sets. BAYES analysis did well on the training set, 91% overall, but was less accurate (87%) in the prospective identification of the test set organisms. LDA and SIMCA performed more poorly in identifying gram-negative organisms.

Effect of atypical resistance patterns. In an attempt to understand why some isolates were misidentified, those misclassified by KNN were examined for high resistance (as defined above) to each of the antibiotics. The majority of the isolates in the *E. coli* group, for example, were not highly resistant to any of the antibiot-

TABLE 2. Percentages of gram-positive organisms correctly identified

Organism	N	% Correctly identified ^a
<i>S. aureus</i>	273	99
Enterococcus	94	96
<i>S. epidermidis</i>	79	89

^a Classification performed by SIMCA algorithm, using four principal components.

TABLE 3. Percentages of gram-negative organisms correctly identified

Data set	% Correctly identified by:			
	KNN	SIMCA	BAYES	LDA
Training, all classes	90	82	91	77
Training, six classes ^a	90	82	91	78
Test, six classes	90	85	87	82

^a *C. diversus* and *P. morganii* were excluded from the six-class data sets.

ics. However, most misclassified *E. coli* were highly resistant to one or more antibiotics. Inspection of principal-component projections of the *E. coli* sensitivity data indicated that there were several distinct clusters within this class (Fig. 1). Each cluster was found to correspond to a subgroup of *E. coli* which was highly resistant to the same antibiotic(s). The other organism classes exhibited a similar phenomenon of subgrouping according to the pattern of high resistance to an antibiotic or a group of antibiotics. This observation formed the basis for our subgrouping scheme (described above) and provided a mechanism for identifying the atypically resistant organisms. The results of KNN, SIMCA, and BAYES classification after exclusion of the atypically resistant subgroups of organisms (Table 4) indicate that the classification accuracy for all three methods was improved. KNN now correctly classified 95% of the organisms. The remaining misclassifications were due largely to organisms of different species which were highly resistant to the same antibiotic or group of antibiotics.

DISCUSSION

We evaluated several pattern recognition techniques for organism identification based on antibiotic sensitivity patterns. Our data base encompassed the organisms most commonly isolated in a large hospital microbiology laboratory and included all multiply resistant organisms encountered.

Our overall prospective identification accuracy for gram-negative organisms (90% correct) was comparable to the 90% accuracy obtained by MacDonald (18), who also used pattern recognition methods and Bauer-Kirby data. The prospective identification accuracies obtained by other methodologies have generally been lower, 82% for LDA (6) and 86% for BAYES (10). An accuracy of 97% was reported for quadratic discriminant analysis (20), but this method was not tested prospectively to ensure that it identified unknown isolates as accurately as it identified

FIG. 1. Clustering of *E. coli* sensitivity data. Shown is a computer plot of *E. coli* sensitivity zone size measurements which have been projected by using principal-component analysis from nine-dimensional (i.e., nine antibiotics) space onto the two-dimensional plane generated by the second and third principal components. This projection offers the viewer an optimal glimpse at the spatial relationships which exist among sensitivity data of the *E. coli* isolates. Only the five most populous subgroups of *E. coli* (defined by patterns of high antibiotic resistance) are plotted here. Each subgroup is depicted by a different plotting symbol. Three distinct clusters of organisms can be viewed from this projection representing: (lower left) nonresistant organisms (diamond) or organisms resistant to tetracycline (octagon); (middle) organisms resistant to ampicillin and carbenicillin (star) or to kanamycin and tetracycline (square); (upper right) organisms resistant to ampicillin, carbenicillin, and tetracycline (triangle).

TABLE 4. Percentages of gram-negative organisms correctly identified after exclusion of atypically resistant organisms

Classification method	% Correctly identified	
	Training set	Test set
KNN	96	95
SIMCA	95	93
BAYES	98	94

those in the training set. The total number and types of antibiotics to which bacterial sensitivities were measured and the methods for determining sensitivities were not the same in all studies. Our investigation of gram-negative organisms employed 9 antibiotics, whereas Darland (6) and MacDonald (18) used 12. Friedman and MacLowry (10) used eight dilutions of 11 antibiotics in a tube dilution system, whereas

Sielaff et al. (20) utilized the light-scattering indexes of 18 antibiotics measured by an Auto-bac I. Further differences existed in the number and the types of bacterial species used in each of these studies. The effects of these other variables are not known; therefore, comparison among these studies is difficult.

Many investigators have suspected that the occurrence of multiple- or atypical antibiotic resistance patterns would adversely affect the accuracy of organism identification. Some have excluded multiply resistant organisms from their studies (6, 18), whereas others have reported no apparent effect of multiple resistance on their identification accuracy (20). Since KNN classification had correctly identified the greatest percentage of unknown organisms in our study, we reasoned that an examination of the KNN-misclassified organisms would be helpful in determining the factors responsible for our low overall organism identification rate. Our investigation of the antibiotic resistance patterns of KNN-misclassified organisms established that atypical antibiotic resistance was a major factor reducing the accuracy of organism identifications obtained from antibiotic sensitivity profiles, regardless of what computer classification technique was used. Subgrouping our organisms according to patterns of high antibiotic resistance showed that organisms having common antibiotic resistance patterns were readily identified, whereas those having atypical patterns (including multiply resistant species) were quite often misidentified. Organisms falling into the subgroups that we characterized as having atypical antibiotic resistance patterns comprised about 17% of our isolates and accounted for over one-half of the incorrectly identified organisms. Excluding the atypically resistant groups of organisms from our data base improved the overall KNN organism identification accuracy from 90 to 95% (Table 4).

The high prevalence of atypically resistant organisms and the fact that these organisms account for the majority of organism misidentifications help explain the disappointing organism identification rates noted in most studies (including this one). Several possible approaches to obviate this problem merit further exploration. First, a means might be devised to separate organisms with atypical resistance patterns. All other organisms could then be given provisional identifications with an acceptable error rate. A second possibility would be to use antimicrobial agents other than clinically employed antibiotics for organism sensitivity testing and identification. In this way, the atypical resistance patterns resulting from resistance transfer plasmids could be minimized. However, with this method the

advantage of organism identification from routinely obtained antibiotic sensitivities would be lost. A third possibility would be to identify organisms by using data from routine antibiotic sensitivities combined with a subset of the more traditionally employed biochemical tests. In this manner, the additional information available from antibiotic sensitivity data might allow a reduction in the number of biochemical tests needed for organism identification.

Given the 90% correct overall identification rate noted in our study, it does not appear practical to identify microorganisms exclusively from their antibiotic sensitivity profiles as a routine practice. An additional concern is that, although sensitivity patterns are approximately the same across the country, there is enough interhospital variation to cast doubt on the general application of this method. This variation and the intrahospital variation with time would probably require that each hospital develop its own data base and periodically update its classifiers. If means can be developed to overcome the detrimental effects of atypically resistant organisms on the accuracy of identification, then microorganism identification using antibiotic sensitivity measurements might be a useful addition to the armamentarium of the clinical microbiology laboratory.

ACKNOWLEDGMENTS

We acknowledge the support of National Institutes of Health National Research Service Award fellowship 5 F32 6M05009-02 (J.C.B.), National Science Foundation grant BMS-71-01597A01 (J.J.M.), and Public Health Service grant RO1-MB-00184-02 (A.M.H. and B.R.K.).

LITERATURE CITED

1. Bauer, A. W., W. M. Kirby, J. C. Sherris, and M. Turck. 1966. Antibiotic susceptibility testing by a standardized single disc method. *Am. J. Clin. Pathol.* **45**:493-496.
2. Bender, C. F., and B. R. Kowalski. 1974. Multiclass linear classifier for spectral interpretation (pattern recognition). *Anal. Chem.* **46**:294-296.
3. Chien, Y. T., and K. S. Fu. 1967. On the generalized Karhunen-Loeve expansion. *IEEE Trans. Inf. Theory* **13**:518-520.
4. Cover, T. M. 1965. Geometrical and statistical properties of systems of linear inequalities with applications to pattern recognition. *IEEE Trans. Electron. Comput.* **14**:326-334.
5. Cover, T. M., and P. E. Hart. 1967. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**:21-27.
6. Darland, G. 1975. Discriminant analysis of antibiotic susceptibility as a means of bacterial identification. *J. Clin. Microbiol.* **2**:391-396.
7. Demartini, J., and A. Vincent-Carrefour. 1977. Topics in pattern recognition, p. 107-126. In A. Remond (ed.), *EEG informatics. A didactic review of methods and applications of EEG data processing*. Elsevier Scientific Publishing Co., New York.
8. Duda, R. O., and P. E. Hart. 1973. Bayes decision theory, p. 10-43. In *Pattern classification and scene analysis*.

- John Wiley & Sons, Inc., New York.
9. **Edwards, P. R., and W. H. Ewing.** 1972. Identification of enterobacteriaceae. Burgess Publishing Co., Minneapolis.
 10. **Friedman, R., and J. MacLowry.** 1973. Computer identification of bacteria on the basis of their antibiotic susceptibility patterns. *Appl. Microbiol.* **26**:314-317.
 11. **Fu, K. S.** 1976. Pattern recognition in remote sensing of the earth's resources. *IEEE Trans. Geosci. Electron.* **14**:10-18.
 12. **Hugh, R., and G. L. Gilardi.** 1974. *Pseudomonas*, p. 250-269. In E. H. Lennette, E. H. Spaulding, and J. P. Truant (ed.), *Manual of clinical microbiology*. American Society for Microbiology, Washington, D.C.
 13. **Koskinen, J. R., and B. R. Kowalski.** 1975. Interactive pattern recognition in the chemical laboratory. *J. Chem. Inf. Comput. Sci.* **15**:119-123.
 14. **Kowalski, B. R.** 1975. Measurements analysis. *Anal. Chem.* **47**:1152A-1162A.
 15. **Kowalski, B. R., and C. F. Bender.** 1975. Pattern recognition. A powerful approach to interpreting chemical data. *J. Am. Chem. Soc.* **94**:5632-5639.
 16. **Kowalski, B. R., and C. F. Bender.** 1976. An orthogonal feature selection method. *Pattern Recognition* **8**:1-4.
 17. **Lewis, J. W., and J. J. Marr.** 1975. A low-cost clinical microbiology computer system. p. 97-100. In J. Zimmerman (ed.), *Proceedings of the 1975 Mumps User's Group Meeting*. Mumps User's Group, St. Louis.
 18. **MacDonald, J. C.** 1978. Pattern recognition in microbiology. *Am. Lab.* **10**:78-85.
 19. **Rosenfeld, A.** 1976. *Digital picture analysis*. Springer Publishing Co., New York.
 20. **Sielaff, B. H., E. A. Johnson, and J. M. Matsen.** 1976. Computer-assisted bacterial identification utilizing antimicrobial susceptibility profiles generated by Autobac 1. *J. Clin. Microbiol.* **3**:105-109.
 21. **Ullmann, J. R.** 1973. *Pattern recognition techniques*. Crane, Russak and Co., New York.
 22. **Wold, S.** 1976. Pattern recognition by means of disjoint principal components models. *Pattern Recognition* **8**: 127-139.