Computer Identification of Bacteria on the Basis of Their Antibiotic Susceptibility Patterns

RICHARD FRIEDMAN AND JAMES MacLOWRY

Research, Development, and Laboratory Automation Section, Clinical Pathology Department, National Institutes of Health, Bethesda, Maryland 20014

A computer program utilizing a Baysean mathematical model was developed to identify bacteria solely on the basis of their antibiotic sensitivities. The model contains probability data on the antibiotic sensitivity patterns for 31 species of bacteria, which account for over 99% of all isolates submitted to our laboratory for testing. During a 4-month test period, antibiotic sensitivity data on 1,000 clinical isolates were processed by the program. The identification achieved by using the model was the same as that of the laboratory for over 86% of the isolates.

Computer based programs for the identification of microorganisms have focused primarily on the gram-negative rods (1, 2, 5, 7, 10). The techniques for identifying these organisms are well documented, and considerable data is available on the results of these procedures (3). Data on the results of tests used to identify other microorganisms are not as readily available; however, considerable quantitative information has been collected on the antibiotic sensitivities of many of these organisms. Although no previous reports have suggested that microorganisms can be identified solely on the basis of their antibiotic sensitivities, Petralli et al. (8) demonstrated sufficient correlation to permit construction of an automated system for quality control in antibiotic sensitivity. Their system relies on the relationship between antibiotic disk sensitivity and classic taxonomy. Given the additional data on antibiotic sensitivity provided by serial dilution techniques (6), it was felt that a program could be developed for accurate classification of microorganisms.

MATERIALS AND METHODS

At the National Institutes of Health, the antibiotic susceptabilities of all clinically significant isolates are tested by serial dilution techniques (6) by using 8 concentrations of 11 of the following antibiotics: ampicillin, carbenicillin, cephalothin, chloramphenicol, colistin, erthromycin, gentamicin, kanamycin, nalidixic acid, nitrofurantoin, oxacillin, penicillin, streptomycin, and tetracycline. Thus, the growth characteristics of each organism are tested in 88 different media (8 dilutions times 11 antibiotics). Technologists have often noted that specific sensitiv-

ity patterns recur with specific organisms. However, the fact that there are a vast number of possible sensitivity patterns (over one billion [8^{11}]) makes it impossible for humans to assimilate this information meaningfully. For this reason, a high-speed digital computer was utilized to identify bacteria solely on the basis of their quantitative antibiotic sensitivity patterns.

Computer program. A Fortran program was written to classify bacteria on the basis of their mean inhibitory antibiotic concentrations (MIC). A score for each possible identification in our data base was determined by the following Baysean formula:

$$Pn = \frac{PP_n \displaystyle\prod_{\substack{all \\ antibiotics \\ tested}} p_{n_{ij}}}{PP_n \displaystyle\prod_{\substack{all \\ antibiotics \\ tested}} p_{n_{ij}} + (1 - PP_n) \displaystyle\prod_{\substack{all \\ antibiotics \\ tested}} q_{n_{ij}}}$$

where Pn equals the relative probability of the unknown organism being organism n; PP_n equals the prior probability of the organisms n (i.e., given a random population, the probability of an unknown being organism n); $p_{n_{ij}}$ equals the probability that, if the unknown were organism n, it would have an MIC j for antibiotic i; and $q_{n_{ij}}$ equals the probability that, if the unknown were not organism n, it would have an MIC j for antibiotic i. (For a detailed discussion of Bayes' theorem and the derivation of this formula, see Fisher [4] and Pratt et al. [9].)

Data base. To calculate the necessary probabilities for the formula (PP_n, $p_{n_{ij}}$, $q_{n_{ij}}$), the results of all antibiotic sensitivity testing performed in our laboratory by the serial-dilution technique over the past 3 years were recorded on punch cards. This provided data on the identification and the antibiotic sensitivity testing results for over 13,000 clinical isolates.

When testing for antibiotic susceptibility, our laboratory divides all isolates into urine source and nonurine source and then tests them against a different range of antibiotic dilutions. To maintain continuity, we divided the 13,000 isolates in a similar manner and developed two distinct data bases.

To make sure sufficient data were available to calculate accurately the necessary probabilities, only those species with at least 15 prior isolates in the data base were included in the identification scheme. Thirty-one species of isolates from nonurine sources had sufficient entries in the data base, and these accounted for over 99% of all nonurine isolates. Eighteen species of isolates from a urine source had sufficient entries, and these accounted for over 99% of all urine isolates (Table 1). Once the necessary probability tables were calculated from the data base, they were stored on high-speed disks for rapid retrieval.

Subgrouping. Preliminary analysis of the data base revealed that within many species there existed two or more distinct subgroupings which could be identified by common antibiotic sensitivity patterns. For example, *Escherichia coli* could be divided into two groups, one sensitive to antibiotic therapy and the other resistant to it. To incorporate this subgrouping into our identification scheme, we submitted all 8,795 isolates in the nonurine data base as unknowns to the program. All isolates that received a relative probability score (Pn) of less than 0.5 for the correct identification were placed in a special subgroup of the correct species classification (i.e., *E. coli* [2]). For six species sufficient isolates (over 20) were assigned to this subgroup to permit adequate group definition. For these species an official subgroup was established and all probabilities were recalculated to include these six additional identifications.

Selection criteria. Early experience with the model indicated that those cases in which the model incorrectly identified an isolate could be divided into three groups: (i) those isolates for which no identification received a score above 0.5; (ii) those isolates for which two or more identifications received a score above 0.9; and (iii) those isolates for which two or more identifications were within 0.1 of the highest score. This indicated that the score pattern of the unknown isolate could be used to predict the accuracy of the program's identification. If, for example, a score pattern for an unknown fell into one of the above categories, then the identification was correct in less than 25% of the cases. However, if an isolate's score pattern was such that one identification received a score above 0.95 and no other identification received a score above 0.5, the model's classification was correct over 96% of the time. If one utilized selection criteria to reject isolates with score patterns previously determined to be frequently incorrect, the model's overall accuracy increased. As these criteria became more "selective," an increasing number of isolates were rejected because their score patterns failed to fall within the predetermined limits. By carefully setting these criteria, one could preselect a subpopulation of clinical isolates for which the model's identification was extremely accurate. To increase the accuracy of

TABLE 1. *Distribution of organisms tested for antibiotic sensitivity, 1968 to 1971*

Organism	Distribution %
Nonurine	
Staphylococcus aureus	22.7
Pseudomonas aeruginosa	14.3
Klebsiella sp.	11.5
Escherichia coli	10.2
Proteus mirabilis	5.5
Staphylococcus epidermidis	4.5
Streptococcus faecalis	3.2
Enterobacter aerogenes	3.1
Enterobacter cloacae	2.8
Enterobacter sp.	2.2
Pseudomonas aeruginosa mucoid	1.9
Haemophilus parahaemolyticus	1.6
Serratia sp.	1.6
Pseudomonas sp.	1.5
Citrobacter sp.	1.5
Bacteroides fragilis	1.4
Pseudomonas sp. *mucoid*	1.2
Proteus morganii	1.0
Streptococcus viridans	1.0
Haemophilus parainfluenzae	1.0
Herellea vaginicola	1.0
Enterobacter cloacae, atypical	0.9
Haemophilus influenzae	0.8
Pseudomonas maltophilia	0.8
Streptococcus beta hemolytic	0.6
Haemophilus sp.	0.4
Enterobacter hafniae	0.4
Bacillus sp.	0.3
Salmonella sp.	0.3
Proteus vulgaris	0.3
Bacteroides sp.	0.3
Other	0.2
Urines	
E. coli	40.4
P. mirabilis	16.5
Klebsiella sp.	12.1
S. faecalis	9.1
P. aeruginosa	8.1
Escherichia sp.	1.8
S. epidermidis	1.5
S. beta hemolytic	1.5
E. aerogenes	1.5
S. aureus	1.1
P. morganni	1.1
Serratia sp.	1.1
Providence sp.	1.0
P. vulgaris	0.9
Citrobacter sp.	0.7
Pseudomonas sp.	0.7
E. cloacae	0.4
E. cloacea, typical	0.4
Other	0.1

the model while still making an identification in a majority of the cases, the following selection criteria

TABLE 2. *Sample output, organism 8027, source wound*[a]

Identification	Relative probability
Bacillus sp.	0.001
Bacteroides sp.	0.000
Bacteroides fragilis	0.000
Citrobacter	0.051
Enterobacter sp.	0.340
Enterobacter cloacae	0.289
Enterobacter cloacae, atypical	0.056
Enterobacter aerogenes	0.287
Enterobacter hafniae	0.004
Escherichia coli	0.001
Haemophilus sp.	0.000
Haemophilus influenzae	0.000
Haemophilus parahaemolyticus	0.000
Haemophilus parainfluenzae	0.000
Herellea vaginicola	0.006
Klebsiella sp.	0.993
Proteus vulgaris	0.005
Proteus mirabilis	0.000
Proteus morganii	0.000
Pseudomonas sp.	0.000
Pseudomonas aeruginosa	0.000
Pseudomonas sp. mucoid	0.000
Pseudomonas aeruginosa mucoid	0.000
Pseudomonas maltophilia	0.000
Salmonella sp.	0.005
Serratia sp.	0.000
Staphylococcus aureus	0.000
Staphylococcus epidermidis	0.000
Streptococcus beta hemolytic	0.000
Streptococcus viridans	0.000
Streptococcus faecalis	0.000

[a] Most Likely Identification, *Klebsiella* sp. (identification acceptable by "selection criteria").

TABLE 3. *Effect of subgrouping*

Species	No. of isolates in non-urine data base	No. of isolates in subgroup	Correctly identified by computer[a] (%)	
			Without sub-grouping	With sub-grouping
Enterobacter cloacae	246	26	35	42
Escherichia coli	862	152	65	73
Klebsiella sp.	943	82	67	78
Pseudomonas aeruginosa	1,398	87	91	95
Staphylococcus aureus	2,004	162	82	85
Staphylococcus epidermidis	415	45	48	61

[a] Most Likely Identification (same as laboratory identification).

were incorporated into the model. They required that the scoring pattern be such that at least one identifi-

cation have a score above 0.9, with no more than one identification having a score above 0.975.

Identification of bacteria. To process an unknown isolate, the antibiotic sensitivities of the organism were entered into the computer via a single punch card. The program then selected the necessary probabilities from disk memory and, by using the Baysean formula, calculated a relative probability score (ranging from 0 to 1) for each species in the identification scheme. The computer then selected the identification which received the highest score and designated it the "Most Likely Identification" (MLI) (Table 2).

Test trials. To test the model, the antibiotic susceptibility patterns for the next 1,000 clinical isolates submitted to our laboratory for testing were punched on computer cards. These data were entered into the program, and the computer's MLI was compared with the laboratory's identification.

RESULTS

The program's MLI was the same as the laboratory's identification for 82.3% of the isolates. When subgrouping was utilized (for nonurine isolates) the overall accuracy increased to 85.6%. For each of the six species where sufficient data were available to permit subgrouping, the program's accuracy showed a definite improvement (Table 3).

When "selection criteria" were incorporated into the model, the program rejected 356 (36%) of the test isolates because the scoring patterns failed to fall within the predetermined limits. For the remaining 644 (64%) isolates, the program's identification was correct in over 92% of the cases.

DISCUSSION

These results indicate that identification of microorganisms on the basis of their antibiotic sensitivities is feasible. However, the system fails to meet the requirements for accuracy currently employed in most clinical laboratories. Whether such accuracy can be obtained in a model that ignores the morphology and metabolic characteristics which currently serve as the basis of classification and identification is problematic. However, the accuracy of this system can be made to approximate more closely the acceptable laboratory levels if the following modifications are made.

First, in analyzing the 1,000 test isolates, it was noted that for species with more than 400 entries (seven species) in the data base, the MLI was correct in 90% of the cases. In contrast, for isolates with less than 75 entries (seven species), this figure fell to 60%. Therefore, as the data base is increased, the accuracy of the model should improve.

Second, Bayes' theorem requires that all

factors used in calculating probability should be mutually independent. This is not true in our current model, as many of the antibiotics now used have similar modes of action and therefore overlapping spectra. This lack of independence tends to place excessive weight on certain characteristics and to decrease the "actual" number of factors used in calculating the probabilities. The accuracy of the model could be enhanced either by substituting new antibiotics with totally independent spectra or by calculating the interdependence of the antibiotics now used and modifying the model accordingly.

Third, the model's ability to correctly identify isolates could be enhanced by the increased use of subgrouping.

Fourth, the accuracy of the model could be improved through the use of stricter "selection" criteria.

Fifth, a method must be incorporated for continually updating the data base. This is necessary if the program is to compensate for changes in antibiotic sensitivity resulting from genetic variation.

The model does not currently have the accuracy necessary to replace more conventional methods of laboratory identification. However, it does demonstrate that identification on the basis of antibiotic sensitivity is feasible, and with full implementation of the steps noted above, might well reach acceptable standards.

Presently the program is useful for rapid preliminary identification of isolates and as a means of quality control in the laboratory.

LITERATURE CITED

1. Dominque, G. J., F. Dean, and J. R. Miller. 1969. A diagnostic scheme for identifying Enterobacteraceae and miscellaneous gram negative bacilli. J. Clin. Pathol. 51:62-70.
2. Dybowski, W., and D. A. Franklin. 1968. Conditional probability and the identification of bacteria; a pilot study. J. Gen. Microbiol. 54:215-229.
3. Edwards, P. R., and W. H. Ewing. 1964. Identification of Enterobacteriacae. Burgess Publishing Co., Minneapolis.
4. Fisher, R. A. 1950. Contributions to mathematical statistics. John Wiley & Sons, Inc., New York.
5. Lapage, S. P., S. Bascomb, W. R. Wilcox, et al. 1970. In A. Baillie and R. J. Gilbert (ed.), Automation, mechanization and data handling in microbiology, p. 1-22. Academic Press Inc., London.
6. MacLowry, J. D., M. J. Jaqua, and S. T. Selepak. 1970. Detailed methodology and implementation of a semiautomated serial dilution microtechnique for antimicrobial susceptibility testing. Appl. Microbiol. 21:46-53.
7. Payne, L. C. 1963. Towards medical automation. World Med. Electron. 2:1.
8. Petralli, J., E. Russell, A. Kataoka, and T. C. Merigan. 1970. On-line computer quality control of antibiotic-sensitivity testing. New Engl. J. Med. 283:735-739.
9. Pratt, J. W., H. Raiffa, and R. Schlaifer. 1965. Introduction to statistical decision theory, p. 10.1-10.15. McGraw Hill, New York.
10. Rypka, E. W., W. E. Clapper, I. G. Bowen, and R. Babb. 1967. A model for the identification of bacteria. J. Gen. Microbiol. 46:407-424.