Computer-Assisted Identification of Anaerobic Bacteria

ROGER W. KELLEY* AND SCOTT T. KELLOGG

Department of Microbiology, University of Hawaii, Honolulu, Hawaii 96822

Received for publication 12 September 1977

A computer program was developed to identify anaerobic bacteria by using simultaneous pattern recognition via a Bayesian probabilistic model. The system is intended for use as a rapid, precise, and reproducible aid in the identification of unknown isolates. The program operates on a data base of 28 genera comprising 238 species of anaerobic bacteria that can be separated by the program. Input to the program consists of biochemical and gas chromatographic test results in binary format. The system is flexible and yields outputs of: (i) most probable species, (ii) significant test results conflicting with established data, and (iii) differential tests of significance for missing test results.

A major problem in the field of anaerobic bacteriology is the difficulty in accurately and precisely identifying unknown isolates. Although several computer-assisted identification programs have been reported for the Enterobacteriaceae (6, 7) and for the gram-negative, nonfermentative rods (13), no comparable system is available for the anaerobes with the exception of a brief report (E. W. Rypka and V. R. Dowell, Abstr. Annu. Meet. Am. Soc. Microbiol. 1975, C16, p. 29). Consequently, a computer program was written to assist in the identification of anaerobic bacteria with the purpose of providing for a system that was rapid, accurate, simple to use, inexpensive, reproducible, and consistent with the most detailed reports on anaerobic bacteria. The system was based upon the extensive data found in the Virginia Polytechnic Institute (VPI) Anaerobe Laboratory Manual (10). New species and emendations of existing species from the most recent literature also were included to make the program as current as possible. A short communication was previously published on this system (R. W. Kelley and S. T. Kellogg, Abstr. Annu. Meet. Am. Soc. Microbiol. 1977, C179 p. 65).

MATERIALS AND METHODS

Taxa and tests. The program, called ANROBE, currently accesses a data base of 28 genera with 238 species of anaerobic bacteria that are grouped into files by their morphology, Gram reaction, and ability to form spores. The data base was derived from the VPI Anaerobe Laboratory Manual (10) and also from several more recent literature sources (2-5, 8, 9, 11, 14-20). The genera within the data base include the following: Acidaminococcus, Actinomyces, Arachnia, Bacteroides, Bifidobacterium, Borrelia, Butyrivibrio, Clostridium, Coprococcus, Desulfovomonas, Eubacterium, Fusobacterium, "Gaffky", Lachnospira, Lactobacillus, Leptotrichia, Megasphaera, Peptococcus, Peptostreptococcus, Propionibacterium, Ruminococcus, Sarcina, Selenomonas, Streptococcus, Succinimonas, Succinivibrio, Treponema, and Veillonella. The test batteries used for species identification are those biochemical tests and gas chromatographic (GC) results specified by the VPI manual as necessary for adequate species separation. Many tests, such as glucose fermentation or GC products, were divided into two characters on the basis of final pH or end product concentration (10). The size of the test batteries varied depending on the file accessed, ranging from 29 characters for the spirochetes to 70 characters for the gram-positive nonsporeforming rods.

Program. The program was first written in BASIC and developed on a Hewlett-Packard 2000 Access time-sharing computer (Hewlett-Packard, Palo Alto, Cal.). The program required 4,400 words (16 bit) of memory in addition to 33,000 words of data base storage on disk. Additional versions of ANROBE were later written in FORTRAN for both TSO (Time Sharing Option) and batch submittal on an IBM System 370/158 computer at the University of Hawaii Computing Center.

The method of identification employed in the program is an algorithm based upon a Bayesian probabilistic model for simultaneous pattern recognition (6). This algorithm assumes that the likelihood of occurrence of a particular taxon is equal to that of other taxa. The probability values for each species were generated from ranged values based upon the tables in the VPI manual and sources cited above. An approximation method, similar to one used by Lapage (12), was used because exact frequencies for most species were not available. For example, a +" was interpreted as 95% positive, and a -" was interpreted as 5% positive. This interpretation came from the definition of the VPI manual of a + as 90 to 100% positive and a - as 0 to 10% positive.

An "identification score" was computed by the procedure of Willcox and Lapage (21). Specifically, if $P(t|R)$ is the probability that an organism giving a set of character results $R$ is a member of taxon $t$, then $P(t|R) = \frac{\prod P(R_i|t)}{\prod P(R_i)}$, where $P(R_i) = P(r_1|t)P(r_2|t) \ldots P(r_n|t)$. $P(R_i)$ is the probability

507
that a member of taxon \( t_i \) will give results \( R \), with \( n \) characters, and \( r_1 \ldots r_n \) individual character results.

Test results against the computer identification were determined by computing if the absolute differences between the data base and the user inputs were greater than 60%. Missing test results, which could be used to differentiate possible species, were examined in a similar manner by comparing the data base results for the appropriate pairs of organisms.

RESULTS AND DISCUSSION

The program ANROBE was found to be capable of separating by probabilistic methods all 238 species in the data base. The Hewlett-Packard 2000 minicomputer was able to identify unknowns with speeds varying from 6 to 7 s for the gram-negative cocci and spirochetes (about equivalent size data matrices) to about 4 min for the gram-positive nonsporing rods (largest data matrix). With the IBM 370, under TSO, identification took less than 1 s of execution time. All identifications cost about 10¢ per identification on either the Hewlett-Packard 2000 minicomputer or the IBM 370.

Figure 1 illustrates two example identifications of Eubacterium contortum. The first example (accession no. 1) shows the effect of GC result deletions and incorrect test entry. Two characters, specifically amygdalin final pH >5.5 and mannitol final pH >5.5, were deliberately entered incorrectly with respect to expected E. contortum behavior to test whether the program could still correctly identify E. contortum. Although E. contortum was identified, the identification score was not at a certainty level of \( \geq 0.99 \). The two incorrectly entered characters were found to be in contradiction with respect to the E. contortum in the reference data base. However, when the GC results were entered and the two corrected aberrant characters were reentered (see Fig. 1, accession no. 2), an excellent score of \( \geq 0.999 \) was obtained.

In Fig. 2, data for an unknown gram-positive coccus were also entered without GC results. This yielded a poor identification score of 0.523. GC results may not always be required to correctly identify some isolates. For example, when data for E. contortum with no GC results and no aberrant tests were entered, an identification score of 0.992 was obtained. However, in Fig. 2, species identification could not be achieved with missing GC results. Therefore, we recommend that the program be used with as complete a set of data as possible. The feature of having only those missing tests listed that will aid in further discrimination between most probable species can help prevent the performance of unnecessary tests.

Although programs using probabilistic identification matrices will not correctly identify an unknown isolate if its reference species is not present in the data base, they will specify the species that are most similar to the unknown. In Fig. 3 an isolate (Coprococcus eutactus) was tested on an older data base that did not contain the following species of gram-positive cocci: C. comes, C. eutactus, C. catus, Ruminococcus torques, R. callidus, R. gnavus, R. lactaris, R. obeum, Streptococcus hansenii, and Peptococcus niger. It can be seen in the first example (accession no. 4) that the program identified Sarcina ventriculi as being most similar to the unknown isolate (C. eutactus), although the identification score of 0.58 was very poor and there were five test disagreements. When the same results were entered (accession no. 5) and computed with the expanded file containing the above ten species, an excellent identification for C. eutactus was obtained. This example illustrates that increased diversity in the data base yields a greater likelihood for a correct identification.

The theorem of Bayes (1) is based upon the assumption that the characters used in computing a probability are mutually independent. However, many bacteriological tests are correlated with one another to varying degrees. In an attempt to measure the degree of deviation from the concept of mutual independence, a correlation coefficient analysis was performed. It was found, for example, that amongst the gram-positive nonsporing rods (80 species by 70 characters), the apparently linked characters (pH’s and end product concentrations) correlated (\( P < 0.01 \)) in only 0.46% of the total possible intercorrelations, and no correlations of 1.00 were ever found. Thus, no attempt was made to alter the VPI manual character batteries used throughout this study.

User evaluation of the computer diagnosis should take into account the following: (i) the identification score, with a score \( \geq 0.99 \) reflecting a good identification; (ii) how well the most probable species is separated from the next most likely one; and (iii) whether or not any of the entered characters contradict the computer diagnosis. Contradictory tests should always be repeated before one concludes that the isolate is not in the data base. ANROBE was designed to assist in identification of unknown isolates and is not intended to replace other means of identification. Subjective evaluation of test characteristics not in the data base, such as morphology and cell arrangement, may be useful in confirming identification.

In summary, this system of computer-assisted identification should prove to be a powerful aid...
COMPUTER-ASSISTED IDENTIFICATION OF BACTERIA

EXE-ANROBE

**ANROBE**

**IDENTIFICATION OF ANAEROBIC BACTERIA**

*INSTRUCTIONS IN THE USE OF THIS SYSTEM* ?NO

*ENTER 1 (G+ NONSPORING ROD), 2 (G+ COCCUS), 3 (G+ SPORING ROD), 4 (G- ROD), 5 (G- COCCUS), 6 (G- SPIROCHAETE)* ?!

*WOULD YOU LIKE A LIST OF ALL BIOCHEMICAL TESTS* ?NO

*WOULD YOU LIKE AN EVALUATION OF YOUR MISSING TESTS* ?YES

*WOULD YOU LIKE A LIST OF G+ NONSP RODS IN DATA BASE* ?NO

*DO YOU WISH TO ENTER TEST RESULTS HORIZONTALLY* ?YES

*WHAT IS YOUR ACCESSION NUMBER* ?

**COMPUTER DIAGNOSIS**

**IDENTIFICATION SCORE**

EU. CONTORTUM 0.961464
LACT. PLANTARUM 0.022957
BIF. ADOLESCENTIS VAR. C 0.010784

*ISOLATE TEST RESULTS AGAINST EU. CONTORTUM*

AMYG > .05 +
MANL > .05 +

*SIGNIFICANT TESTS NOT DONE TO DIFFERENTIATE EU. CONTORTUM FROM LACT. PLANTARUM*

ACET > .94
LACT > -.94

*SIGNIFICANT TESTS NOT DONE TO DIFFERENTIATE EU. CONTORTUM FROM BIF. ADOLESCENTIS VAR. C*

LACT > -.94
SUC < -.70

*COMPUTER ANALYSIS COMPLETED FOR ACCESSION # 1

*ANOTHER ISOLATE IDENTIFICATION* ?YES

*WOULD YOU LIKE DIFFERENT OUTPUT* ?NO

*WHAT IS YOUR ACCESSION NUMBER* ?

**COMPUTER DIAGNOSIS**

**IDENTIFICATION SCORE**

EU. CONTORTUM 0.999978

*ISOLATE TEST RESULTS AGAINST EU. CONTORTUM*

NONE

*COMPUTER ANALYSIS COMPLETED FOR ACCESSION # 2

*ANOTHER ISOLATE IDENTIFICATION* ?NO

DONE

**Fig. 1.** Computer identification printout showing effect of GC test deletions and aberrant test results. Bottom threshold for identification scores was set at 0.01. Abbreviated tests are printed vertically, e.g., first test is amygdalin pH < 5.5. An input for < and > means a final pH above 6.0; an - + input means a final pH 5.5 to 6.0; and an input + - means a final pH below 5.5.

to both research and clinical anaerobic bacteriologists. It offers the advantage of a rapid, precise, and reproducible means of identification of unknowns, and a data base that is easily updated as new species or emendations of existing species are published.

Copies of the system and an accompanying user guide are available from the authors.
ACKNOWLEDGMENTS

We are grateful to W. E. C. Moore for discussion concerning this project and to H. A. Frank for reviewing the manuscript.

LITERATURE CITED


of determinative bacteriology, 8th ed. The Williams & Wilkins Co., Baltimore.


