JOURNAL OF CLINICAL MICROBIOLOGY, Jan. 1996, p. 179–181 Vol. 34, No. 1
0095-1137/96/$04.00+0

Evaluation of New Computer-Enhanced Identification Program for Microorganisms: Adaptation of BioBASE for Identification of Members of the Family *Enterobacteriaceae*

J. MICHAEL MILLER[1]* AND PETER ALACHI[2]†

Hospital Infections Program, Centers for Disease Control and Prevention, Atlanta, Georgia 30333,[1] and Department of Biology, Northeastern University, Boston, Massachusetts 02115[2]

Received 25 August 1995/Returned for modification 18 September 1995/Accepted 3 October 1995

We report the use of BioBASE, a computer-enhanced numerical identification software package, as a valuable aid for the rapid identification of unknown enteric bacilli when using conventional biochemicals. We compared BioBASE identification results with those of the Centers for Disease Control and Prevention's mainframe computer to determine the former's accuracy in identifying both common and rare unknown isolates of the family *Enterobacteriaceae* by using the same compiled data matrix. Of 293 enteric strains tested by BioBASE, 278 (94.9%) were correctly identified to the species level; 13 (4.4%) were assigned unacceptable or low discrimination profiles, but 8 of these (2.7%) were listed as the correct organisms as the first choice; and 2 (0.7%) were not identified correctly because of their highly unusual biochemical profiles. The software is user friendly, rapid, and accurate and would be of value to any laboratory that uses conventional biochemicals.

Numerical taxonomic methods for the classification and identification of microorganisms, especially bacteria, were introduced in the 1950s (13, 14) and were strengthened in the following decades (1, 2, 4, 8–10, 12, 15–18). Their practicality was not realized until the advent and widespread use of modern computers. Although numerical techniques are extensively used today to identify many microbes (2, 3, 7, 11, 12), there remains a need for a "broad-spectrum" program that can be applied to any microbial group without the high cost of an accompanying identification instrument.

BioBASE, a computer-enhanced identification program for the International Business Machines (IBM) compatible personal computer (PC), provides a comprehensive environment for creating, updating, and manipulating unlimited numbers of microbial databases. An unknown organism belonging to a database can be rapidly identified by using simplified input modes. Furthermore, the software facilitates the process of comparing among different sets of matrix data by using an interface for cluster analysis and single or comparative character profile diagrams.

In the study described here we evaluated BioBASE for its ability to accurately identify members of the family *Enterobacteriaceae* compared with that of the standard identification scheme of the Centers for Disease Control and Prevention's (CDC's) mainframe computer.

Computer system. BioBASE was constructed by using an IBM PC-compatible computer (Leading Edge), with an 80486SX microprocessor running at 25 MHz. The computer was equipped with a high-resolution super VGA color monitor. MS-DOS versions 6.0 and 6.2 were used to run most versions of BioBASE.

System requirements. System requirements are as follows: (i) an IBM PC or IBM-compatible computer with a hard drive (a 386 processor or higher is recommended for speed); (ii) MS-DOS version 2.2 or higher; (iii) at least 2 megabytes of free disk space, and (iv) a graphics card, which is recommended.

Programming thesis. BioBASE was constructed on the basis of the principles of numerical techniques in taxonomy originally introduced by Sneath (13–15) and Sneath and Sokal (16, 17). These techniques determine similarity by comparing an unknown microbe's biochemical profile with the profiles of known taxa previously compiled into a database.

BioBASE uses database files to store physiologic and biochemical profile frequency data available for a particular group of microorganisms. Since numerous strains of a species may be used to describe an expected test reaction, the probability data for tests may range from 0 to 100% positive for each reaction.

To reach an identification verdict, BioBASE calculates identification scores, modal scores, and similarity indices of each taxon and compares them with the unknown's input data (1, 2, 4, 9, 10). With all tests weighted equally, the program performs successive comparisons of the unknown's data profile for overall similarity to the data compiled in the taxonomic matrix. The top-scoring microorganisms are then analyzed and weighted for a decision on identification ranging from unacceptable to excellent.

General features of the program. BioBASE consists of a database development and management system for entering physiologic and biochemical frequency data for known taxa. For example, the probability matrix for the members of the family *Enterobacteriaceae* consists of two components: species names and their biochemical profiles, expressed as the proportion of positive frequencies ranging from 0 to 100% for each test used. A rapid interface for entering and deciphering the identity of an unknown organism works in either of two ways: (i) by manually entering positive test results or (ii) by entering a profile code number for the unknown organism similar to the octal codes used by rapid identification systems. The program has an interface for comparing the unknown's biochemical profile with that of any taxon in the database. Additionally, the program includes two routines for cluster analysis and profile diagrams.

* Corresponding author. Mailing address: Centers for Disease Control and Prevention, Mailstop C16, Atlanta, GA 30333. Phone: (404) 639-3029. Fax: (404) 639-3241. Electronic mail address: jmm8@cidhip1.em.cdc.gov.

† Corresponding author for software information. Mailing address: Department of Biology, Northeastern University, 414 Mugar Building, Boston, MA 02115. Phone: (617) 373-2118. Fax: (617) 373-3724. Electronic mail address: palachi@lynx.neu.edu.

TABLE 1. Accuracy of BioBASE for identifying members of the family *Enterobacteriaceae*

Organism tested	No. identified correctly[a]/ no. tested (%)
Budvicia aquatica	2/2
Citrobacter davisae	10/10
Citrobacter lapagei	5/5
Citrobacter neteri	1/1
Citrobacter sp. strain 3	1/1
Citrobacter sp. strain 5	1/1
Citrobacter sp. strain 10	1/1
Citrobacter amalonaticus	9/10[b]
Citrobacter braakii	13/13
Citrobacter diversus	9/9
Citrobacter farmeri	10/10
Citrobacter freundii	4/4
Citrobacter sedlakii	3/3
Citrobacter werkmanii	5/5
Citrobacter youngae	2/3[b]
Enterobacter aerogenes	5/5
Enterobacter agglomerans	11/11
Enterobacter amnigenus 1	1/1
Enterobacter amnigenus 2	1/1
Enterobacter asburiae	10/10
Enterobacter cancerogenus (*Enterobacter taylorae*)	10/10
Enterobacter cloacae	9/9
Enterobacter gergoviae	9/10[b]
Enterobacter hormaechei	5/5
Enterobacter intermedium	1/1
Enterobacter sakazakii	7/7
Escherichia coli	3/3
Escherichia fergusonii	3/3
Escherichia hermanii	2/2
Ewingella americana	11/11
Hafnia alvei	10/10
Hafnia alvei biogroup 1	1/1
Klebsiella oxytoca	5/5
Klebsiella ozaenae	5/5
Klebsiella ornithinolytica	5/5
Klebsiella planticola	2/3[b]
Klebsiella pneumoniae	3/5[c]
Klebsiella rhinoscleromatis	3/3
Klebsiella terrigena	0/1[b]
Leclercia adecarboxylata	3/4[b]
Leminorella grimontii	2/2
Leminorella richardii	2/2
Morganella morganii	6/7[b]
Morganella morganii biogroup 1	1/1
Pragia fontium	2/2
Proteus mirabilis	5/5
Proteus penneri	5/5
Proteus vulgaris	5/5
Providencia alcalifaciens	3/3
Providencia heimbachae	2/2
Providencia rettgeri	4/4
Providencia rustigiani	2/2
Providencia stuartii	3/5[d]
Rhanella aquatilis	1/1
Serratia entemophilia	1/1
Serratia ficaria	2/2
Serratia fonticola	3/3
Serratia liquefaciens group	4/4
Serratia marcescens	8/10[e]
Serratia marcescens biogroup 1	2/2
Serratia odorifera biogroup 1	1/1
Serratia odorifera biogroup 2	2/2
Serratia plymuthica	1/1
Serratia rubidea	2/2
Salmonella group 1	3/3
Salmonella typhi	1/1
Salmonella group 2	1/1
Salmonella paratyphi A	1/1
Salmonella sp. (lactose positive)	0/1[f]
Shigella sonnei	2/3[b]
Total	278/293 (94.9)
Total in which first choice was correct	286/293 (97.6)

[a] Correct to genus and species levels according to CDC mainframe computer data.
[b] One strain was identified only to the genus level.
[c] Two strains were identified only to the genus level or were rare biotypes.
[d] Results for two strains with rare biotypes were errors.
[e] Two strains were identified only to the genus level.
[f] One strain was correctly identified to the genus level, but *Escherichia coli* was also listed as a possible choice.

Factors influencing identification. The final identification of an unknown organism is based on several factors: (i) transcription accuracy of frequency data when creating a database, (ii) selection of tests that separate the species, (iii) selection of a reliable data matrix with correct and complete frequency data, (iv) proper interpretation of the unknown's tests results, (v) familiarity with program features and usage, and (vi) independent verification of final identification. Since the analysis is probabilistic in nature, identification may be inconclusive and may necessitate further tests.

Organism database. The data for the "ENTERIC" database were taken from an updated version of previously published biochemical charts (5, 6). Forty-seven biochemical reactions were entered as percent positive for each taxon in the database, as described in the software instructions. After the database was completed, each entry was checked for its accuracy against its biochemical profile in the published chart.

Evaluation method. The biochemical profiles for 293 isolates of the family *Enterobacteriaceae* of human origin previously identified in the enteric laboratory of the Hospital Infections Program, CDC, and analyzed by the CDC mainframe computer program were also analyzed by the BioBASE program. The same biochemical profile data entered into the mainframe computer for identification were entered into the BioBASE program. Since many of these organisms have identical biochemical profiles, we chose not to repeatedly enter identical data for large numbers of strains, hence the seemingly limited number of strains tested.

BioBASE was used to identify 293 previously identified strains of the family *Enterobacteriaceae*. The reference identification provided by the mainframe computer program of the enteric laboratories of CDC was considered correct, and the BioBASE identification result was compared with the reference identification result for each of the strains tested. Table 1 lists the taxa tested and shows the results obtained with BioBASE.

In the present study, 278 of 293 (94.9%) enteric strains tested were correctly identified to the species level by BioBASE. Only two errors (0.7%, incorrect genus) occurred; two *Providencia stuartii* strains with rare profiles were shown by BioBASE to be *Tatumella* spp. Indeed, on its initial identification, the mainframe computer's identification agreed with that of BioBASE but when the resulting data were compared with those for all other strains in the CDC database, these two unusual isolates were more closely related to *P. stuartii* than to *Tatumella* spp. No other such error occurred in the study.

For 13 (4.4%) organisms, BioBASE did not report a genus or a species but, instead, indicated an unacceptable profile or a low discrimination index that prevented a genus or species report and that encouraged the use of additional tests to make a final determination. In every case, a list of the three to five strains most closely fitting the test profile was shown along with a probability and a similarity index for each organism. For 8 of these 13 organisms, the first choice listed was correct. Two strains of highly unusual *Klebsiella pneumoniae* exhibited a first choice of either *Klebsiella ozaenae* or *Klebsiella planticola*, with *Klebsiella pneumoniae* being a second choice, indicating that further work would be necessary to correctly report the isolates. Two strains of *Serratia marcescens* of rare biotypes were identified by BioBASE to be *Serratia liquefaciens* in one case and *Serratia marcescens* biogroup 1 in the other. The *Serratia liquefaciens* report was due to a highly unusual profile for the *Serratia marcescens* isolate. One *Citrobacter youngae* isolate was reported only as *Citrobacter* genus, with a first choice of *Citrobacter sedlakii*. The total number of isolates for which the first choice listed was correct was 286 of 293 (97.6%) isolates tested.

The version of BioBASE that we evaluated does not list the specific additional tests to be performed when a clear decision is not generated. However, with a simple keystroke, reactions atypical of the reported unknown can be evaluated against each biochemical in the database. At this point, interpretive judgment must prevail.

In addition, the program automatically creates an octal code based on the individual biochemicals used at the local laboratory to create the database. The user has the option of entering the raw biochemical results into the program to generate a genus and species response, or the user may have an octal code in hand, generated by the same order of biochemicals, whose numbers can be entered into the program to generate the same organism report.

BioBASE was designed to conform to the routine identification protocols of virtually any laboratory, whereby the user can input the chosen method of identification, regardless of whether that method is a commercial system rendering an octal code after inoculation, incubation, and interpretation or whether conventional biochemicals are used. As a result, BioBASE is available as a computer-assisted adjunct to interpreting the results of the biochemical tests used to identify members of the family *Enterobacteriaceae*. The present study incorporated the conventional, reference biochemicals used at CDC (6).

BioBASE is not an identification "instrument." It is a software package designed to receive laboratory input data and interpret them on the basis of a comparison with its database profiles. We found this product to be extremely fast, accurate, and user friendly. Its accuracy compared with that of our mainframe computer's algorithms was impressive in the present application to members of the family *Enterobacteriaceae*. BioBASE would be of value to any reference laboratory that uses conventional biochemicals to provide reference identifications of this group of organisms.

REFERENCES

1. **Bascomb, S., S. P. Lapage, M. A. Curtis, and W. R. Willcox.** 1973. Identification of bacteria by computer: identification of reference strains. J. Gen. Microbiol. **77:**291–315.
2. **Boeufgras, J. M., J. L. Blazer, F. Allards, and I. Diaz.** 1987. A new computer program for routine interpretation of API identification systems. *In* 2nd Conference on Taxonomy and Automatic Identification of Bacteria. Scientific Computer Department, API Systems, La Balme-Les Grottes, France.
3. **Cox, R. P., and J. K. Thomsen.** 1990. Computer-aided identification of lactic acid bacteria using the API CHL system. Lett. Appl. Microbiol. **10:**257–259.
4. **Dybowski, W., and D. A. Franklin.** 1968. Conditional probability and the identification of bacteria: a pilot study. J. Gen. Microbiol. **54:**215–229.
5. **Farmer, J. J., III, B. R. Davis, F. W. Hickman-Brenner, A. McWhorter, G. P. Huntley-Carter, M. A. Asbury, C. Riddle, H. G. Wathen, C. Elias, G. R. Fanning, A. B. Steigerwalt, C. M. O'Hara, G. K. Morris, P. B. Smith, and D. J. Brenner.** 1985. Biochemical identification of new species and biogroups of *Enterobacteriaceae* isolated from clinical specimens. J. Clin. Microbiol. **21:**46–76.
6. **Farmer, J. J., III, and M. T. Kelly.** 1991. *Enterobacteriaceae*, p. 360–383. *In* A. Balows, W. J. Hausler, Jr., K. L. Herrmann, H. D. Isenberg, and H. J. Shadomy (ed.), Manual of clinical microbiology, 5th ed. American Society for Microbiology, Washington, D.C.
7. **Freney, J., M. T. Duperron, C. Courtier, W. Hansen, F. Allard, J. M. Boeufgras, D. Monget, and J. Fleurette.** 1991. Evaluation of API Coryne in comparison with conventional methods for identifying coryneform bacteria. J. Clin. Microbiol. **29:**38–41.
8. **Gyllenberg, H. G.** 1965. A model for computer identification of microorganisms. J. Gen. Microbiol. **39:**401–405.
9. **Lapage, S. P., S. Bascomb, W. R. Willcox, and M. A. Curtis.** 1970. Computer identification of bacteria, p. 1–22. *In* A. Baillie and R. J. Gilbert (ed.), Automation, mechanization and data handling in microbiology. Society for Applied Bacteriology technical series no. 4. Academic Press, London.
10. **Lapage, S. P., S. Bascomb, W. R. Willcox, and M. A. Curtis.** 1973. Identification of bacteria by computer: general aspects and perspectives. J. Gen. Microbiol. **77:**273–290.
11. **Rhoden, D. L., G. A. Hancock, and J. M. Miller.** 1993. Numerical approach to reference identification of *Staphylococcus*, *Stomatococcus*, and *Micrococcus* spp. J. Clin. Microbiol. **31:**490–493.
12. **Schnider, J.** 1979. Computer-aided numerical identification of gram negative fermentative rods on a desk-top computer. J. Appl. Bacteriol. **47:**45–51.
13. **Sneath, P. H. A.** 1957. Some thoughts on bacterial classification. J. Gen. Microbiol. **17:**184–200.
14. **Sneath, P. H. A.** 1957. The application of computers to taxonomy. J. Gen. Microbiol. **17:**201–226.
15. **Sneath, P. H. A.** 1984. Bacterial classification II. Numerical taxonomy, p. 5–7. *In* N. R. Krieg and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 1. The Williams & Wilkins Co., Baltimore.
16. **Sneath, P. H. A., and R. R. Sokal.** 1962. Numerical taxonomy. Nature (London) **193:**855–860.
17. **Sneath, P. H. A., and R. R. Sokal.** 1973. Numerical taxonomy. The principles and practice of numerical classification. W. H. Freeman & Co., San Francisco.
18. **Willcox, W. R., S. P. Lapage, S. Bascomb, and M. A. Curtis.** 1973. Identification of bacteria by computer: theory and programming. J. Gen. Microbiol. **77:**317–330.