APPLIED MICROBIOLOGY, Oct. 1974, p. 691-695

Vol. 28, No. 4
Printed in U.S.A.

Mathematical Analysis of the API Enteric 20 Profile Register Using a Computer Diagnostic Model

E. ARTHUR ROBERTSON AND JAMES D. MACLOWRY

Clinical Pathology Department, Clinical Center, National Institutes of Health, Bethesda, Maryland 20014

Received for publication 31 May 1974

Results for 21 biochemical tests using the API-20 Enteric kit were obtained from the manufacturer's files for 27,820 bacterial isolates. These isolates were identified by the API Profile Register and also by a computer diagnostic model which estimates the relative likelihoods of various identifications. The computer confirmed the identification in the API Profile Register for 99.36% of the isolates. The manufacturer has reviewed areas of the API Profile Register questioned by the computer analysis; a number of resulting modifications to the API Profile Register have been incorporated in an update letter. This computer model provides a convenient and powerful way to interpret a large number of test results for bacterial identification. This study also demonstrates the use of a large collection of isolates to refine the data matrix used by the diagnostic model.

The computer identification of bacteria using mathematical models has been shown to be possible and accurate (7-9). A powerful and practical diagnostic application of the computer model involves the interpretation of large numbers of biochemical reactions. Even with detailed percentage charts of expected biochemical results for a variety of organisms, it is often difficult to decide which of the tests on an unknown isolate that deviate from the general description of an organism should be considered or ignored. For the microbiologist not working in a specialized laboratory, it may be difficult to make these judgments correctly and consistently. A commercially available set of biochemical reactions, API 20 Enteric (Analytab Products Inc., Carle Place, N.Y.), is a specific case in point. This diagnostic kit utilizes 21 different biochemical tests for the diagnosis of *Enterobacteriaceae*. The early versions of this product provided only a "plus or minus" chart for each test reaction for each organism. It was quickly obvious that this large number of tests required a more sophisticated approach for diagnostic interpretation. As a result, the manufacturer provided a technique which reduces the 21 test results to a seven-digit profile number, which is interpreted by searching a directory of profile numbers, the API Profile Register (Analytab Products Inc., Carle Place, N.Y.).

The question asked in this study was, how would a proven diagnostic computer model identify the test patterns included in the Profile Register? This paper describes a comparison between the Profile Register and computer identification for each of the profile numbers. In addition, the effect of deleting profiles and including additional diagnostic categories is discussed.

MATERIALS AND METHODS

The API Profile Register provides a scheme for identifying *Enterobacteriaceae* on the basis of the 21 biochemical tests performed with the API 20 Enteric kit. The biochemical reactions beta-galactosidase, arginine dihydrolase, lysine decarboxylase, ornithine decarboxylase, citrate [Simmons], hydrogen sulfide, urease, tryptophane deaminase, indole, Voges-Proskauer, gelatin, glucose, mannitol, inositol, sorbitol, rhamnose, sucrose, melibiose, amygdalin, arabinose, and oxidase are read as positive or negative at 18 to 24 h. Using the plastic API Coder (Analytab Products Inc., Carle Place, N.Y.), these 21 test results are reduced to a unique seven-digit profile number. The user looks up the profile number in the Profile Register which lists the genus and usually the species corresponding to the observed pattern of test results. Along with the Profile Register the manufacturer supplies a percent chart showing the expected frequency of positive results for each biochemical test for each of 31 diagnostic categories.

This study used the original (January 1973) edition of the percent chart and Profile Register (original profiles) and the manufacturer's first update letter (November 1973), in accordance with which some profiles were added, some were deleted and the identification of still others changed (updated profiles). In addition, an intermediate profile list (revised profiles) was prepared, consisting of the original profiles and the deletions and changes specified by the update letter, but not containing the new profiles of the update letter.

The data used in this study were obtained from the manufacturer's files of test results on bacterial isolates studied in a number of laboratories using the API-20 Enteric kit. None of the isolates was available to us for microbiological study.

The 1,148 original profiles encompassed 24,058 isolates from the manufacturers files, the 1,107 revised profiles encompassed 24,040 isolates, and the 1,260 updated profiles, 27,847 isolates. The manufacturer made available additional information required in the study: the number of isolates used in establishing the percent chart; actual percentages where the percent chart was incomplete; the number of isolates actually found for each profile number; and the percent positive reactions for subgroups of certain species.

A computer model involving the calculation of relative likelihoods was used. Test patterns were identified in the following way. The program stored a data matrix listing the expected frequency of positive results for each test for each diagnostic category. When a pattern of test results was presented, the likelihood that it could have been produced by an organism belonging to the first diagnostic class was estimated by calculating the probability of the observed result for each test and multiplying these probabilities together. The likelihood that the observed pattern would be produced by an organism belonging to each of the other categories was calculated in turn. The program then identified the organism as belonging to the diagnostic class yielding the highest likelihood. An earlier version of this program using relative likelihoods in the context of Bayes' theorem has been described in some detail by Friedman (7).

A basic data matrix was constructed using the percent chart supplemented by the additional data from the manufacturer. Using this basic data matrix, the computer assigned an identification to each pattern of the original profile list and the revised profile list. For each pattern the identification in the register was compared with that proposed by the computer and the agreements and discrepancies tabulated.

When the Profile Register listed two possible diagnoses, the pattern was processed under both. The percentage of agreements was computed on the basis of both the number of different profiles and the number of actual bacterial isolates. Agreement was defined as an isolate being placed in the same genus or species by both the Profile Register and the computer program.

An attempt was made to find biochemical subgroups of the original diagnostic categories which were sufficiently unique to justify their being treated as separate categories in the data matrix. To this end, the data matrix was expanded from 32 diagnostic categories to include data for 46 proposed subgroups (1 to 6). Using this expanded data matrix, the program identified the revised profiles. For the purpose of evaluating the usefulness of a particular diagnostic category to the computer model, it was assumed that the Profile Register identifications were correct. Those diagnostic categories which decreased the overall percentage of agreements were removed from the data matrix, whereas those which increased agreements or made no change were retained. This process was repeated several times. When most of the undesirable categories had been deleted, those which made no net change in overall accuracy were deleted as well in order to decrease the time and storage used by the program. This resulted in an increase from 32 to a total of 40 diagnostic categories in the final expanded data matrix.

The diagnostic program was written in Fortran and runs in batch mode on a Control Data Corporation 3200 computer with 32,000 24-bit words of core memory. The supporting programs for data analysis are also Fortran programs but run on an IBM 360/370 system in time-sharing mode.

RESULTS

The results of the comparison between the Profile Register and computer identifications are summarized in Tables 1 and 2. Using the basic data matrix and the original 1,148 profiles, 98.85% of the isolates and 90.24% of the profiles received the same identification by both methods. The expanded data matrix and original profile list produced agreement on 99.33% of the isolates and 93.55% of the profiles. When the fully updated version of the Profile Register was evaluated using the expanded data matrix, there was agreement on 99.36% of the 27,847 isolates and 95.16% of the profiles. In Table 2 isolates are grouped according to the Profile Register identification. For each group the computer identifications are listed, along with the number of profiles and isolates involved. Examination of the table shows that for 16 profiles representing 60 isolates the register and the computer agreed on the genus while differing on the species.

Disagreements are further analyzed in Table 3. The Profile Register and computer identifications disagreed for 61 profiles representing 178 isolates. For 16 of these profiles representing 72 isolates, the register either listed as an alternate diagnosis the one chosen by the computer or indicated that the diagnosis must be confirmed by additional studies. These are termed "disagreements with annotation" in Table 3. Of the remaining disagreements, 16 profiles represent-

TABLE 1. *Comparison of Profile Register and computer identifications*

Data matrix	Profile list	No. of profiles	Agreement (%)	No. of isolates	Agreement (%)
Basic	Original	1,148	90.24	24,058	98.85
Basic	Revised	1,107	91.33	24,040	98.92
Expanded	Original	1,148	93.55	24,058	99.33
Expanded	Revised	1,107	95.03	24,040	99.42
Expanded	Updated	1,260	95.16	27,847	99.36

TABLE 2. *Summary of diagnoses using expanded data matrix and updated profiles*

Profile Register identification[a] (computer identification)[b]	No. of profiles	No. of isolates	Profile Register identification (computer identification)	No. of profiles	No. of isolates
Escherichia coli	324	12,422			
E. coli	305	12,345	*Enterobacter hafniae*	1	2
Shigella species	1	6	*Citrobacter freundii*	4	4
Edwardsiella	1	0	*Klebsiella pneumonia*	1	3
Salmonella enteritidis	4	40	*Enterobacter aerogenes*	8	484
Salmonella typhi	1	1	*Enterobacter aerogenes*	8	484
Citrobacter freundii	5	15	*Enterobacter hafniae*	36	155
Citrobacter diversus	1	0	*Enterobacter hafniae*	36	155
Klebsiella ozaenae	2	3	*Enterobacter agglomerans*	38	89
Enterobacter hafniae	3	4	*Enterobacter agglomerans*	33	62
Enterobacter agglomerans	1	8	*Escherichia coli*	2	0
Shigella species	47	132	*Klebsiella pneumoniae*	2	26
Shigella species	44	128	*Citrobacter freundii*	1	1
Escherichia coli	1	3	*Pectobacterium*	2	0
Salmonella typhi	1	1	*Enterobacter agglomerans*	2	0
Salmonella enteritidis	1	0	*Serratia marcescens*	58	424
Edwardsiella	3	20	*Serratia marcescens*	58	424
Edwardsiella	3	20	*Serratia liquefaciens*	62	352
Salmonella typhi	4	12	*Serratia liquefaciens*	59	342
Salmonella typhi	4	12	*Serratia marcescens*	1	8
Salmonella cholerae-suis	12	0	*Enterobacter cloacae*	1	1
Salmonella cholerae-suis	11	0	*Enterobacter agglomerans*	1	1
Enterobacter hafniae	1	0	*Serratia rubidae*	2	2
Salmonella enteritidis	54	231	*Serratia rubidae*	2	2
Salmonella enteritidis	53	213	*Proteus vulgaris*	57	306
Enterobacter hafniae	1	18	*Proteus vulgaris*	53	300
Arizona	18	59	*Proteus morganii*	1	5
Arizona	17	58	*Proteus rettgeri*	1	1
Enterobacter hafniae	1	1	*Providence alcalifaciens*	1	0
Citrobacter species	36	135	*Shigella* species	1	0
Citrobacter freundii	6	17	*Proteus mirabilis*	101	3,636
Citrobacter diversus	30	118	*Proteus mirabilis*	95	3,621
Citrobacter freundii	120	577	*Proteus vulgaris*	1	0
Citrobacter freundii	113	552	*Proteus morganii*	3	14
Citrobacter diversus	4	21	*Yersinia enterocolitica*	2	1
Salmonella enteritidis	2	4	*Proteus morganii*	20	1,102
Arizona	1	0	*Proteus morganii*	20	1,102
Klebsiella species	4	17	*Proteus rettgeri*	50	181
Enterobacter agglomerans	4	17	*Proteus rettgeri*	48	179
Klebsiella pneumoniae	366	5,088	*Providence alcalifaciens*	1	1
Klebsiella pneumoniae	364	5,075	*Shigella* species	1	1
Enterobacter agglomerans		10	*Providence alcalifaciens*	9	61
Serratia rubidae	1	3	*Providence alcalifaciens*	9	61
Klebsiella ozaenae	10	18	*Providence stuartii*	20	492
Klebsiella ozaenae	9	17	*Providence stuartii*	19	487
Klebsiella rhinoscleromatis	1	1	*Shigella* species	1	5
Klebsiella rhinoscleromatis	2	2	*Yersinia enterocolitica*	26	50
Klebsiella rhinoscleromatis	2	2	*Yersinia enterocolitica*	20	50
Enterobacter cloacae	71	1,800	*Salmonella enteritidis*	2	0
Enterobacter cloacae	62	1,783	*Citrobacter freundii*	2	0
Enterobacter aerogenes	3	8	*Citrobacter diversus*	2	0

[a] The organism and values given are for Profile Register identification.
[b] Indented entries refer to the computer identification.

ing 32 isolates have already been scheduled for deletion from future editions of the Profile Register by the manufacturer. The remaining 29 profiles representing 74 isolates are called "unqualified disagreements" in Table 3: they comprise 2.30% of the profiles and 0.27% of the isolates studied. The percentage of agreements between the register and the program was considerably higher when tabulated on the basis of numbers of isolates rather than on numbers of test patterns. Patterns correctly identified represented an average of 23.1 isolates per

TABLE 3. *Analysis of disagreements using expanded data matrix and updated profiles*

Tabulation	Total no.	Total % disagreements	% Disagreements with annotation	% Scheduled for deletion	% Unqualified disagreements
Profiles	1,260	4.84	1.27	1.27	2.30
Isolates	27,847	0.64	0.26	0.11	0.27

pattern, while those misidentified represented an average of 2.9 isolates per pattern. Thus the disagreements tended to involve patterns actually observed only infrequently.

DISCUSSION

When performing 21 tests (each of which can be positive or negative), there are 2^{21} or 2,097,152 possible result combinations. Interpretation of these data by manual methods becomes unwieldy, e.g., a directory listing the interpretations of all the possible patterns would require 4,000 pages if printed in the format used by metropolitan phone directories. This computer program makes practical the simultaneous interpretation of any number of test results in identifying a bacterium. Also new or hypothetical patterns may be evaluated by the program in the same way as familiar patterns.

Early in this study the manufacturer was advised of over 100 profiles about which the computer analysis raised questions. The manufacturer undertook an extensive review of these profiles. Many of the changes and deletions specified in the November 1973 update letter reflect the results of this review. Many of the problems pointed out by the computer analysis might not otherwise have been discovered until much later since they generally involved test patterns which are observed only infrequently. Before adding new profiles to the register, the manufacturer now submits them for analysis by the computer program.

The observation that the percentage of disagreements dropped from 1.15% to 0.67% of the isolates in the original register when additional diagnostic subcategories were included suggests that many of the disagreements centered on small, atypical subgroups existing within the larger diagnostic categories. As the number of isolates in each pattern increases, the quality of the data matrix will be further enhanced. Since the authors did not have access to the original cultures, it was not possible to definitively arbitrate disagreements between the register and the program.

This program may be used to check the adequacy of the tests chosen to identify an isolate and the appropriateness of the interpretation of the results. If the identifications assigned by the technologist and the computer differ, or if two or more diagnostic categories receive almost equal scores, further study may be needed.

The computer model is a very useful teaching tool for microbiologists and technologists. One can easily observe the change in the likelihood of various identifications when a single test or set of tests is changed. There is often a great deal of mystique involved in which test to "weight" more heavily in traditional taxonomy; this question is readily explored with the model.

The methodology used in this study could be applied to the evaluation of other diagnostic schemes which explicitly or implicity enumerate acceptable test patterns and the corresponding identifications (for example, binary decision tree systems). The computer model will also be used to identify other families of organisms for which a data matrix can be constructed.

Finally, this study illustrates the usefulness of a large collection of well-identified isolates, each of which has been subjected to the same battery of tests. The ability to evaluate diagnostic subgroups by comparing the overall success of the computer program in identifying the 24,000 isolates before and after the new categories were added to the data matrix was illustrated. Another study now in progress uses the same isolates to determine how much accuracy would be lost if only a subset of the 21 tests were performed.

ACKNOWLEDGMENT

We acknowledge the cooperation of Pierre Janin of Analytab Products, Inc. for making available to us unpublished statistical information from API records for use in this study.

LITERATURE CITED

1. Ewing, W. H. 1971. Biochemical characterization of *Citrobacter freundii* and *Citrobacter diversus*. U.S. Dept. of Health, Education and Welfare, Center for Disease Control, Atlanta.
2. Ewing, W. H., M. Ball, F. Bartes, and A. C. McWhorter. 1970. Biochemical reactions of certain species and bioserotypes of Salmonella. J. Infect. Dis. **121**:288-294.
3. Ewing, W. H., M. Ball, F. Bartes, and A. C. McWhorter. 1970. Supplement to biochemical reactions of certain species and bioserotypes of *Salmonella*. U.S. Dept. of Health, Education and Welfare, Center for Disease Control, Atlanta.
4. Ewing, W. H., B. R. Davis, and W. J. Martin. 1972. Biochemical characterization of *Escherichia coli*. U.S. Dept. of Health, Education and Welfare, Center for

Disease Control, Atlanta.
5. Ewing, W. H., and M. A. Fife. 1972. Biochemical characterization of *Enterobacter agglomerans*. U.S. Dept. of Health, Education and Welfare, Center for Disease Control, Atlanta.
6. Ewing, W. H. et al. 1971. Biochemical reactions of Shigella. U.S. Dept. of Health, Education and Welfare, Center for Disease Control, Atlanta.
7. Friedman, R. B., D. Bruce, J. MacLowry, and V. Brenner. 1973. Computer-assisted identification of bacteria. Amer. J. Clin. Pathol. **60**:395-403.
8. Friedman, R., and J. MacLowry. 1973. Computer identification of bacteria on the basis of their antibiotic susceptibility patterns. Appl. Microbiol. **26**:314-317.
9. Lapage, S. P., Shoshana Boscomb, W. R. Wilcox, and M. A. Curtis. 1973. Identification of bacteria by computer: general aspects and perspectives. J. Gen. Microbiol. **77**:273-290.