Comput. Biol. Med. Vol. 23, No. 1, pp. 21–27, 1993
Printed in Great Britain

0010-4825/93 $6.00 + .00

A COMPUTER PROGRAM IN QUICKBASIC FOR THE SELECTION OF TESTS FOR THE IDENTIFICATION OF *HELICOBACTER PYLORI*

T. S. LIM* and B. HO †‡

* Hewlett Packard (Singapore) Pte Ltd, 150 Beach Road #29–00, Gateway West, Singapore 0718; and † Department of Microbiology, National University of Singapore, Lower Kent Ridge Road, Singapore 0511

(*Received* 22 *January* 1992; *in revised form* 15 *September* 1992; *received for publication* 15 *October* 1992)

Abstract—Reliable microbiological tests are needed for the identification of bacteria. A program has been written in QuickBasic to identify such tests by using a formula that is based on Gyllenberg's Sum of $C(i)$ and Gyllenberg's Rank $R(i)$. A total of 139 papers on a newly isolated bacterium, *Helicobacter pylori*, was used as data source for the coding of test results into an input file. The program outputs a list that aids in the determination of suitable tests for the identification of *H. pylori*. These tests chosen by the formula were found to be correctly identified as supported by later publications on the bacterium.

Basic Tests Selection Identification Bacteria *Helicobacter pylori*

INTRODUCTION

Helicobacter pylori is a Gram-negative bacterium that is found to be closely associated with active chronic gastritis and duodenal disease [1]. A number of explanations have been offered as to how the bacterium is able to survive in the acidic conditions of the stomach. It has been found that the bacterium adheres to the surface of the gastric epithelial cells and produces a large quantity of urease to neutralize the immediate acidic surrounding [2]. *Helicobacter pylori* has been successfully eliminated with multiple antibiotic therapy. However, it has been found to be resistant to non-antimicrobials such as cimetidine, sucralfate, famotidine and ranitidine [3].

Researchers have used different physiological characteristics, antibiotic sensitivity, serological tests, biochemical assays and histological staining to identify the bacterium [4]. Rapid identification of *H. pylori* relies mainly on the presence of urease and the source of the clinical samples [1, 4]. This paper attempts to identify the most promising tests that can be used in routine diagnosis of *H. pylori*. The data are based on 139 research papers documenting work on the microorganism, carried out worldwide during the period 1984–1989.

MATHEMATICAL ASPECTS OF THE APPROACH

In the classification of bacteria, characters that are useful in differentiating a particular taxon from another can be determined by applying Gyllenberg's Sum of $C(i)$ and Gyllenberg's Rank $R(i)$ [5].

Assuming there are q groups of taxa, Gyllenberg's Sum of $C(i)$ is defined as:

$$\sum_{j=1}^{q} (0.5 + |0.5 - P_{ij}|),$$

‡ Author to whom correspondence should be addressed.

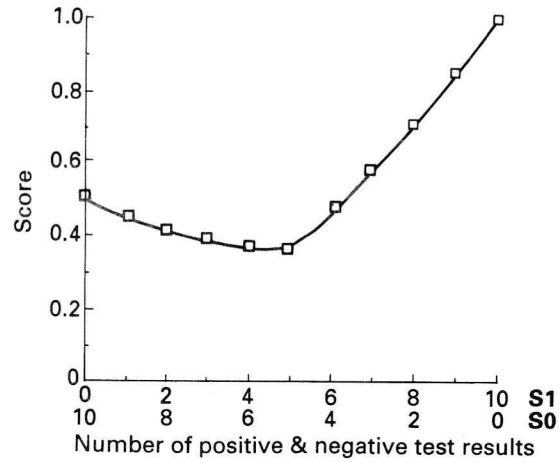

Fig. 1. Graph showing how $S1$ and $S0$ affect the scoring (the score for each set of $S1$ and $S0$ is calculated with $T_a = 10$). $S1$: the number of positive test results; $S0$: the number of negative test results.

where P_{ij} is the proportion of positives for the ith character of the jth taxon.

If F is a cut-off level for a character of a taxon to be considered positive or negative, then Gyllenberg's Rank $R(i)$ is defined as

$$q_0 q_1 \,(\text{Sum } C(i)),$$

where q_1 is the number of taxa with $P_{ij} >= F$; q_0 is the number of taxa with $P_{ij} <= F$; and F ranges between 0.5 and 1.0.

Gyllenberg's Rank measures how successful a character is in differentiating the taxa. When the various taxa do not express a character consistently, q_0 and q_1 become low and this also results in a lower score for $R(i)$. Therefore, a character with low value for $R(i)$ is useful in differentiating the taxa but not for identifying an individual taxon.

In view of expressing the suitability of tests for identifying a taxon, Gyllenberg's Rank is applied such that tests with higher rankings are favoured. Tests with higher rankings will have consistent results and are therefore suitable for identifying the taxon.

If the frequency of tests is T, the frequencies of tests with positive and negative results are $S1$ and $S0$, respectively, and the total number of different tests surveyed is T_a, then the formula used to select a suitable character is stated as:

$$\frac{(0.5 + |0.5 - S0/T|)(0.5 + |0.5 - S1/T|)(\log T) + (S1/T).}{(\log T_a + 1)}$$

Since $(0.5 + |0.5 - S0/T|)$ will give the same value as $(0.5 + |0.5 - S1/T|)$, the formula is now simplified to:

$$\frac{(0.5 + |0.5 - S1/T|)^2(\log T) + (S1/T)}{(\log T_a + 1)},$$

where $(0.5 + |0.5 - S1/T|)^2$ determines whether a test is reliable. A test is reliable if it has been reported to give either mostly positive or mostly negative results, with $S1/T$ tending towards 1 or 0, respectively. Therefore, the expression yields a high value which implies that the test is more reliable. $(\log T)$ indicates how well accepted the test is. It measures the frequency of tests that have been mentioned in the papers surveyed. $(S1/T)$ gives higher scores to tests with positive results. Tests with positive results are considered to be more diagnostic and they are evidence for the presence of a taxon. $(\log T_a + 1)$ normalizes the value obtained to the range of 0 to 1.

It is observed from Fig. 1 that the formula gives the highest score to the test that always gives positive results (10 $S1$ and 0 $S0$). The lowest score is given to a test that has a highly unreliable result (5 $S1$ and 5 $S0$). Tests that show negative results (0 $S1$ and 10 $S0$) are

also given much lower scores than those with positive results. A high value obtained for the simplified formula thus indicates that the test is reliable, diagnostic and well accepted.

DESCRIPTION OF PROGRAM

A random sample of 139 papers on *H. pylori* was collected (from 1984 to 1989). Tests that are mentioned in the papers are coded into an input file as a matrix. Results of 211 different tests ranging from biochemical to therapeutic agents are coded as follows: 1, negative result; 2, doubtful result (+/−); 3, weakly positive result; 4, positive result; blank, tests not mentioned in paper.

The first row of the matrix stores the references, while the second row stores the year when the papers were published. Subsequent rows store the results of the tests. Individual papers are coded as the columns of the matrix.

The program first reads and displays the title of papers and the year of publication. It then ranks the tests by applying the formula onto the test results. Finally, bubble sort is carried out to arrange those tests that have the highest scores at the top of an output file.

The program is written in QuickBasic and is run on Philips P3204 that has an 80 MB harddisk. A sample of screen output is shown below.

RESULTS AND DISCUSSION

The program first lists the various papers that have been collected. The papers have been given a code each. The program outputs the year of publication as the next row. A sample of the screen output is shown below:

165 29 116 33 34 . . .

1984 1985 1985 1985 1985 . . .

Subsequently, the program prints the results collected for each test. It lists the test number, name and number of negative and positive test results mentioned in the papers:

9. Test:..... Acid Schiff

   Negative Results: 0 Positive Results: 2

10. Test:..... Acridine orange

    Negative Results: 0 Positive Results: 10

11. Test:..... Adipate
    Negative Results: 1 Positive Results: 0

The program then reports that bubble sorting is being performed on the tests so that ranking of tests will be arranged by the score in descending order as shown in Table 1. Finally, the program lists the sorted tests. The output list contains the names of the tests, the number of positive ($S1$) and negative ($S0$) test results mentioned, the total number of test results for each test performed and the scores of the respective tests. Table 1 depicts the best tests scored by the program.

Table 1. The best tests scored by the modified formula

Test name	$S0$	$S1$	Total	Score*
Gram stain	0	50	50	0.8118955
Catalase	0	45	45	0.7981309
Oxidase	0	39	39	0.7794358
Urease	2	50	52	0.7665052

* The score for each test is calculated with $T_a = 211$, where T_a is based on the total number of tests analysed from 139 papers.

The results obtained concur with the conventional tests used, except for the urease test that was reported to be negative in two of the papers [6, 7]. In a later paper [8], the test was reported positive by one of these authors. Thus, analysis of a larger pool of papers will further minimize such an error. Nevertheless, the inconsistency with test results from literature survey is observed to be well handled by the present formula since the urease test is still being ranked relatively high. The approach described is thus especially useful for a "newly" discovered microorganism that has initiated much research interest and yet has uncertainty over the most suitable test for its identification. *Helicobacter pylori* that was discovered in 1984 is one good example [6].

The selection of *H. pylori* takes advantage of the knowledge gained by the various researchers. It effectively reduced labour, cost and time as the program identifies a list of important tests from which investigators can easily select to suit their needs. In this instance, such a selection can also possibly lead to accelerated development of commercial test kits for *H. pylori*. Besides, the rankings can be performed for other organisms, drugs and even chemicals. This is easily achieved by recoding of the input file with results obtained from the survey of literature on other subjects.

Attempts are presently being made to improve the accuracy of the formula by including, on the one hand, the tests that give weakly positive results and on the other, those tests that give doubtful results. These results could possibly be handled by the present formula.

Further inaccuracy in the formula may also accrue because some investigators publish more of their work than others. Hence, this gives a false impression that the tests conducted are well accepted by the other investigators. This problem is especially acute in the early phases of research of a newly discovered microorganism when it is likely that only a handful of investigators is involved. However, for *H. pylori*, there is a large number of well-balanced publications on tests employed for its identification. Thus, *H. pylori* is a suitable candidate for this program.

The sample of papers used as the source for this study was up to early 1989. The most potential tests identified by the formula were found to be correctly selected by the program and shown to be useful as quoted by works done in later years [1, 3, 4].

SUMMARY

It is crucial during routine bacterial identification that tests performed are diagnostic and reliable to ensure accuracy and ease of detection. The possibility of identifying such tests for *H. pylori* by computer is investigated. Test results from a collection of 139 papers on this bacterium are coded into an input file. Using a modification of the formula based on Gyllenberg Sum of $C(i)$ and Gyllenberg Rank $R(i)$, tests are scored and ranked in an output file. The best tests, in descending order, are found to be Gram stain (0.8118955), Catalase (0.7981309), Oxidase (0.7794358) and Urease (0.7665052). This selection effectively reduces labour, cost and time, and may lead to the possible development of commercial test kits for *H. pylori* and other organisms.

Acknowledgements—This work was supported by grant RP 900366 from the National University of Singapore (NUS). T.S. Lim has been a graduate student of Microbiology at NUS.

REFERENCES

1. G. E. Buck, *Campylobacter pylori* and gastroduodenal disease, *Clin. Microbiol. Rev.* **3**, 1–12 (1990).
2. H. L. T. Mobley, L. T. Hu and P. A. Foxall, *Helicobacter pylori* urease: properties and role in pathogenesis, *Scand. J. Gastroenterol. Suppl.* **187**, 39 (1991).
3. B. J. Marshall, *Campylobacter pylori*: its link to gastritis and peptic ulcer disease, *Rev. infect. Dis.* **12** Suppl. 1, S87 (1990).
4. J. D. Dick, *Helicobacter (Campylobacter) pylori*: a new twist to an old disease, *A. Rev. Microbiol.* **44**, 249 (1990).
5. P. H. A. Sneath, Basic program for character separation indices from an identification matrix of percent positive characters, *Comput. Geosci.* **5** (1979).
6. B. J. Marshall, H. Royce, D. F. Annear, C. S. Goodwin, J. W. Pearman, J. R. Warren and J. A. Armstrong, Original isolation of *Campylobacter pylori* from human gastric mucosa, *Microbiol. Lett.* **25**, 83–88 (1984).

7. G. Kaper and N. Dickgiesser, Isolation from gastric epithelium of campylobacter like bacteria that are distinct from *C. pyloridis, Lancet* **i**, 111 (1985).
8. B. J. Marshall, J. A. Armstrong, D. B. Gechie and R. J. Glancy, Attempt to fulfil Koch's postulates for *pyloric campylobacter, Med. J. Aust.* **142**, 436 (1985).

About the Author—T. S. LIM completed his B.Sc. at the National University of Singapore (NUS) in 1990. He subsequently received his postgraduate diploma in Knowledge Engineering at Institute of System Science (NUS) in 1991. T. S. Lim is presently working on an Artificial Intelligence project at Hewlett Packard Singapore Pte Ltd. His major interests are in expert systems and genetic algorithms.

About the Author—B. HO received his Ph.D. degree in Microbiology from the University of Wales, U.K., in 1977. He joined the National University of Singapore in 1978. His research interest is in the characterization of *Helicobacter pylori* and the computer base bacterial taxonomy.

APPENDIX

```
REM ** THIS PROGRAM SCORES AND OUTPUTS **
REM ** A LIST OF TESTS,  OF WHICH THE **
REM ** MOST SUITABLE TESTS FOR THE **
REM ** IDENTIFICATION OF ORGANISMS ARE **
REM ** RANKED ON TOP OF THE LIST **


REM ** INITIALIZATION **

NOOFPAPERS = 139
NOOFTESTS = 211
 '139 papers on Helicobacter pylori have been surveyed'
 '211 tests are mentioned in these papers'

DIM A(4, NOOFTESTS)
 'array to store P, S0, S1 and S'
 'P is the score calculated for each test'
 'S0 is the number of negative test results'
 'S1 is the number of positive test results'
 'S is the total number of tests mentioned'

DIM PAPERNAME$(NOOFPAPERS)
 'array to store code names of papers'

DIM YEAR(NOOFPAPERS)
DIM YEAR$(NOOFPAPERS)
 'arrays to store dates of papers'

DIM RESULT$(NOOFPAPERS)
 'array to store results of tests'

DIM TESTNAMES$(NOOFTESTS)
 'array to store names of tests'

OPEN "A:INPUT.IN" FOR INPUT AS #2
 'the results of all the tests are stored in a:INPUT.IN'
 'the results are stored as a 139 X 211 matrix'

OPEN "A:OUTPUT.OUT" FOR OUTPUT AS #1
 'A:OUTPUT.OUT contains the list of ranked tests'


REM ** READING THE NAME AND THE YEAR OF THE PAPERS **

CLS
FOR PAP1 = 1 TO NOOFPAPERS
        INPUT #2, PAPERNAME$(PAP1)
        PRINT PAPERNAME$(PAP1);
NEXT PAP1

FOR PAP2 = 1 TO NOOFPAPERS
        INPUT #2, YEAR$(PAP2)
        YEAR(PAP2) = VAL(YEAR$(PAP2))
        PRINT YEAR(PAP2);
```

```
NEXT PAP2
 'this row of data is useful for linear regression studies'
 'for predicting the popularity trend of the tests in the'
 'next few years'

REM ** TO COUNT S0 AND S1 AND TO COMPUTE P **

CLS
FOR ROW = 1 TO NOOFTESTS
 'looping for every test in the input file'
        PRINT ROW;
        INPUT #2, TESTNAMES$(ROW)
        PRINT "TEST:.......", TESTNAMES$(ROW)
        SUBTOT = 0: S = 0:  S0 = 0: S1 = 0
        FOR PAP3 = 1 TO NOOFPAPERS
         'looping for all the papers of each test'
         'and count S0 and S1'
                INPUT #2, RESULT$(PAP3)
                RESULT = VAL(RESULT$(PAP3))
                IF RESULT$(PAP3) = "     " THEN GOTO JUMP1
                 'incrementing of S0 or S1 is skipped if'
                 'there is no result'
                IF RESULT = 1 THEN S0 = S0 + 1
                IF RESULT = 4 THEN S1 = S1 + 1
JUMP1:
        NEXT PAP3
        PRINT "NEGATIVE RESULTS:"; S0; "POSITIVE RESULTS:"; S1
        S = S0 + S1
        IF S1 <> 0 OR S0 <> 0 THEN P = (S1 / S + (.5 + ABS(.5 - (S1 / S))) *
         (.5 + ABS(.5 - (S0 / S))) * (.434294481# * LOG(S))) /
         (1 + (.434294481# * LOG(NOOFTESTS))) ELSE P = 0
         'P is computed for each test'
        A(1, ROW) = P: A(2, ROW) = S0: A(3, ROW) = S1: A(4, ROW) = S
PRINT
NEXT ROW


REM ** BUBBLE SORT **

PRINT "BUBBLE SORTING NOW"
REPEAT:
FINISH = 0
FOR COUNT = 1 TO (NOOFTESTS - 1)
        IF A(1, COUNT) >= A(1, COUNT + 1) THEN GOTO NOSWITCH

        TESTNAMES$ = TESTNAMES$(COUNT)
        TESTNAMES$(COUNT) = TESTNAMES$(COUNT + 1)
        TESTNAMES$(COUNT + 1) = TESTNAMES$

        'sorting P'
        A1 = A(1, COUNT)
        A(1, COUNT) = A(1, COUNT + 1)
        A(1, COUNT + 1) = A1

        'sorting S0'
        A2 = A(2, COUNT)
        A(2, COUNT) = A(2, COUNT + 1)
        A(2, COUNT + 1) = A2

        'sorting S1'
        A3 = A(3, COUNT)
        A(3, COUNT) = A(3, COUNT + 1)
        A(3, COUNT + 1) = A3

        'sorting S'
        A4 = A(4, COUNT)
        A(4, COUNT) = A(4, COUNT + 1)
        A(4, COUNT + 1) = A4

        FINISH = 1
NOSWITCH:
NEXT COUNT
PRINT ".";
IF FINISH = 1 THEN GOTO REPEAT
```

```
REM ** OUTPUTTING THE SORTED TESTS **

CLS
PRINT #1, "NO.                    TESTS                    S0              S1
TOTAL            SCORE"
FOR COUNT = 1 TO NOOFTESTS
        PRINT COUNT, TESTNAMES$(COUNT), A(2, COUNT), A(3, COUNT),
          A(4, COUNT), A(1, COUNT)
        PRINT #1, COUNT, TESTNAMES$(COUNT), A(2, COUNT), A(3, COUNT),
          A(4, COUNT), A(1, COUNT)
NEXT COUNT
```