J Clin Pathol 1989;**42**:763–766

Efficient selection of tests for bacteriological typing schemes

M A GASTON, P R HUNTER *Central Public Health Laboratory, Colindale, London*

SUMMARY To simplify the selection of tests for bacteriological typing methods, such as bacteriophage, bacteriocin, and biotyping, a computerised method was assessed. This uses a numerical index of discrimination (*D*) to facilitate the selection of an efficient typing set. The computer programs take the most discriminatory test as the initial test in the partial typing set, and then select the next test by combining each of the remaining candidates with the partial set and choosing the test which maximises *D*. This cycle is repeated until the remaining candidates do not increase the discriminatory power of the typing set. Options are provided for the investigator to pre-select certain tests for inclusion or exclusion from the typing set.

It is concluded that the numerical index *D* is a simple means of test selection, but it must be emphasised that it is important to combine its use with data on the incidence of reaction in each test, on reproducibility, and on the similarity among tests.

One of the difficulties associated with the development of epidemiological identification (typing) schemes for bacteria is the selection of the most efficient set of tests. This problem is particularly acute for non-specialist clinical and medical microbiologists setting up methods to study novel organisms or to investigate local outbreaks.

The efficiency of typing methods are measured primarily on the basis of typability, reproducibility, and discrimination. The first two factors are relatively easy to quantify and there are definitive examples to guide investigators.[1,2] Discrimination is the most difficult factor to quantify but it is arguably the most important at the initial stages of selecting the typing set. There is little point in assessing the reproducibility and typability of a method that fails to discriminate between strains of the target organism.

The problem of test selection is particularly acute in the areas of biochemical, bacteriophage, and bacteriocin typing, where investigators may have a large number of candidate tests which must be reduced to a more practical number for day to day use. In our experience, each of the candidate tests will have been evaluated on a large collection of representative strains, and there is a sizeable matrix of N strains by T tests, which forms the core data for the selection process. Even for bacteriologists with a background in epidemiological studies, it is difficult to interpret these data without recourse to numerical analysis.

Accepted for publication 2 February 1989

Bergan evaluated the use of similarity coefficients in selecting a bacteriophage typing set for *Pseudomonas aeruginosa*.[3] Such coefficients are valuable aids to the identification of identical or very similar tests that would be redundant if included in the typing set. These techniques are therefore most useful as negative selectors for rejecting a proportion of the candidate tests.[4,5]

Quantitative methods that could facilitate the positive selection of typing tests would be useful aids to the selection process and several indices have been proposed to quantify the ability of individual tests to separate strains or biological groups.[6-8] Tests with high values of discrimination are not necessarily useful when combined in typing sets as they may provide redundant information. The selection of combinations of tests is more complex.[8-10] Ideally, the discriminatory ability of every possible combination of tests should be determined, but this approach is not practicable, as even for a limited set of 19 candidate tests there are over 500 000 possible combinations. An approximation is to select tests sequentially, such that discrimination is maximised at each stage.

Recently, we described a numerical index which quantifies the ability of typing schemes to discriminate between strains.[11] This permits easy comparison of different schemes. A second and powerful use for this index is the quantitation of the discrimination of partial sets of tests. Thus combinations of tests can be evaluated and the data used to build up efficient typing schemes. We developed several computer programs using this approach which facilitate the selection and

analysis of sets of typing reagents. Although the central algorithm is based on simple probability theory, no mathematical skill is required to operate the programs or to interpret the results.

Methods

Test results are entered as positive (1), negative (0), or variable (V). The most discriminatory tests are selected from the results by two programs SEL and CHOISEL. The primary algorithm in these similar programs uses a numerical index D to quantify the ability of combinations of tests to subdivide the strains in the data file. D is derived from elementary probability theory and is given by equation I:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{s} n_j (n_j - 1) \qquad \text{I}$$

where N is the total number of strains in the population, s is the total number of types described, and n_j is the number of strains belonging to the jth type. Alternatively, D can be defined by equation II:

$$D = 1 - \frac{1}{N(N-1)} \sum_{j=1}^{N} a_j \qquad \text{II}$$

where a_j is the number of strains that are indistinguishable from the jth strain. This is the more flexible of the two definitions and allows the selection of tests where more than one reaction difference is required to separate strains or where equivocal reactions are present.

The programs start by selecting the most discriminatory individual test as the initial test in the partial typing set. This test is then combined with each of the remaining tests in turn and the value for D calculated. The test that maximises D is chosen as the next test and added to the partial typing set. This cycle is repeated until the remaining tests provide no extra discrimination. Where there are more than one equally good tests, the first in numerical order is chosen, although the others are printed for further analysis.

Of the two programs, SEL has the fastest execution time (about 10 times faster) and selects tests with the assumption that a single reaction difference between strains is significant. CHOISEL, based on equation II, allows the investigator to choose the number of test differences that are required before strains are considered to have been separated. This second program is much slower as a table of reaction differences must be calculated for each combination of tests. Both programs allow the operator to select candidate tests to be included or excluded from the final set, the programs then select the remaining tests.

The option of allowing one or two differences in reaction before strains are separated has been provided so that safeguards can be built into the typing set, to circumvent the poor reproducibility of some biological test systems. As CHOISEL works by choosing the combination of tests which produce the maximum value for D it cannot select the initial one or two tests if two or three tests are required to separate strain pairs—that is, in the initial tests, no strains are considered to be separate—and therefore $D = 0$ for all combinations. CHOISEL therefore "cheats" during the selection of the initials tests by assuming any test difference to be significant.

As an example of test selection for epidemiological studies table 1 gives a subset of data showing the patterns of inhibition produced by 19 strains of *Serratia marcescens* examined for bacteriocinogenic activity on 17 strains. Data were selected from a larger matrix, generated in a epidemiological study of *S marcescens* serotype 014 strains from clinical material.

Table 1 *Inhibition of Serratia marcescens serotype 014 strains by bacteriocinogenic strains*

	Inhibition by bacteriocin producer strains:*																		
Strains	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
1	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
3	−	−	−	−	+	+	+	−	−	+	+	−	−	−	−	−	+	+	+
4	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
6	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	+	+
8	−	+	−	−	+	−	+	+	+	−	+	+	−	−	−	−	−	+	+
9	−	−	−	−	+	+	+	−	−	−	+	+	−	+	−	−	+	+	+
13	−	−	−	−	+	−	−	+	+	−	−	−	−	−	−	−	−	−	−
15	−	+	−	+	+	+	+	−	−	−	−	−	−	+	−	−	+	+	+
18	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−
21	−	+	−	−	+	−	−	+	+	−	−	−	−	+	+	+	+	+	−
26	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−
27	−	−	−	+	+	−	+	−	−	+	+	+	−	−	−	−	−	−	−
28	−	−	−	+	+	+	+	+	+	−	+	+	−	+	−	+	+	+	+
31	−	−	−	−	+	−	−	−	−	−	−	+	−	−	−	−	−	−	−
42	+	−	−	+	+	+	+	+	+	−	+	+	−	+	−	−	+	+	+
68	−	−	−	−	+	−	−	−	−	−	−	−	+	−	−	−	−	−	−
71	−	−	+	+	−	−	+	−	−	−	−	+	−	+	−	−	−	−	−

*Data modified by exclusion of weak reactions.

Table 2 *Discriminatory tests selected based on inhibition data given in table 1*

	Automatic selection			*Preselection of test 18*		
Rank	*Test*	*D index (× 100)*	*Equally good tests*	*Test*	*D index (× 100)*	*Equally good tests*
1	7	52·9	18	18	52·9	NA
2	8	74·3	9, 14	12	77·9	—
3	5	86·0	—	14	88·2	—
4	14	92·6	—	8	93·4	9
5	12	96·3	—	5	96·3	—
6	6	97·8	13. 18, 19	7	97·8	10, 11
7	1	98·5	3, 13, 16	1	98·5	3, 13, 16
8	3	99·3	13	3	99·3	13
9	13	100·0	—	13	100·0	—
	Automatic selection requiring two reaction differences			*Automatic selection requiring three reaction differences*		
1	7	0	18	7	0	18
2	11	39·7	—	11	0	—
3	4	57·4	—	12	30·1	19
4	8	64·0	9	4	44·2	6, 14, 17, 18, 19
5	9	77·2	—	19	55·9	—
6	6	80·9	14, 17	8	61·0	9
7	18	86·0	19	9	67·6	17
8	14	87·5	—	14	75·0	17
9	10	91·9	—	5	80·1	—
10				6	83·8	—
11				10	86·7	—
12				3	88·2	13, 18
13				13	90·4	—
14				18	91·2	—
15				2	91·9	15, 16, 17

NA: not applicable with a preselected test.

Table 2 gives the tests selected by the programs. The first set, automatic selection, produced by either SEL or CHOISEL, is a completely automatic selection of tests based on the premise that any test difference separates pairs of strains. The second set, also produced by SEL or CHOISEL, pre-selected test 18 (the inhibition produced by strain 18) and shows an increase in discrimination given by the partial sets of two, three, and four tests. The third and fourth sets, produced by CHOISEL alone, represent an automatic selection of tests based on the assumption that two and three differences, respectively, are required to separate pairs of strains.

CHOISEL has some similarities to the approach of Rypka *et al*[9] and the sequential method of Willcox and Lapage,[12] but the algorithm presented here is both simple and flexible and offers an advantage over several other methods of test selection in that a quantified index *D* is produced that allows the investigator to gauge the efficacy of the selected tests. For example, in the first and second series in table 2, nine out of a possible 19 tests gave 100% discrimination but over 95% discrimination was achieved with only five tests. So for purely practical reasons the investigator may choose this shortened set of tests and still expect a high level of discrimination.

The programs were written in BASIC for an Acorn Achimedes/BBC Microcomputer system and contain relatively few machine specific instructions. They could therefore be readily implemented on other microcomputer systems. The only potential drawback with the current programming is the amount of memory required by CHOISEL which dimensions integer arrays of N × N. Thus when a large number of strains are used in the analysis a very large amount of computer memory (RAM) is necessary. To circumvent this problem we developed an alternative but slower version (SLOWSEL) that requires much less memory.

Given a sufficiently discriminating set of initial data these programs will produce a minimal typing set—that is, a set of tests with just enough tests to distinguish each strain. Thus any variation in the patterns of reactions of a test may decrease the overall discrimination of the selected tests. The set produced by this approach does not necessarily represent the set with the fewest possible number of tests. Identifying the theoretical minimum set of tests, possibly by comparing every possible combination of tests, may take a prohibitive amount of computer time, although steps may be taken to limit the number of comparisons required.[12]

The ability to select positively combinations of tests greatly simplifies the development of efficient epidemiological typing methods. We believe that the numerical index *D* provides a simple measure for test selection and that it could be a useful tool for many investigators. It is important, however, that the results

of this sort of analysis are put into context. These programs work on one set of data which may not be absolutely reproducible. Indeed, individual tests may be highly irreproducible, therefore positive test selection should be used in combination with data on the incidence of reaction of each test, on reproducibility, and on the apparent similarity between tests (based on similarity coefficients).

Copies of the programs described above and a FORTRAN version of CHOISEL are available from the authors.

References

1 Williams REO, Rippon JE. Bacteriophage typing of Staphylococcus aureus. *J Hyg (Camb)* 1952;**50**:320–53.
2 Anderson ES, Williams REO. Bacteriophage typing of enteric pathogens and staphylococci and its use in epidemiology. *J Clin Pathol* 1956;**9**:94–127.
3 Bergan T. Comparison of numerical procedures for grouping Pseudomonas bacteriophages according to lytic spectra. *Acta Pathol Microbiol Scand* 1972;**80B**:55–70.
4 Gaston MA. Isolation and selection of a bacteriophage-typing set for Enterobacter cloacae. *J Med Microbiol* 1987;**24**:285–90.
5 Gaston MA, Ayling-Smith BA, Pitt TL. New bacteriophage typing scheme for subdivision of the frequent capsular serotypes of Klebsiella spp. *J Clin Microbiol* 1987;**25**:1228–32.
6 Gyllenberg H. A general method for deriving determination schemes for random collections of microbial isolates. *Ann Acad Sci Fenn* 1963;**69**:1–23.
7 Niemela SI, Hopkins JW, Quadling C. Selecting an economical binary test battery for a set of microbial cultures. *Can J Microbiol* 1968;**14**:271–9.
8 Sneath PHA. Basic program for character separation indices from an identification matrix of percent positive characters. *Comp Geosciences* 1979;**5**:349–57.
9 Rypka EW, Clapper WE, Bowen IG, Babb R. A model for the identification of bacteria. *J Gen Microbiol* 1967;**46**:407–24.
10 Rypka EW, Babb R. Automatic construction and use of an identification scheme. *Med Res Eng* 1970;**9**:9–19.
11 Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *J Clin Microbiol* 1988;**26**:2465–6.
12 Willcox WR, Lapage SP. Automatic construction of diagnostic tables. *Comp J* 1972;**15**:263–7.

Requests for reprints to: M A Gaston, Wellcome Research Laboratories, Langley Court, Beckenham, Kent BR3 3BS.