Notes brèves

A BETTER ASSESSMENT OF VARIABILITY OF QUALITATIVE CHARACTERS FOR THE COMPUTER IDENTIFICATION PROGRAM *NEMAID*

Renaud Fortuner*

The computer program NEMAID helps with the identification of nematodes by providing an estimate of the similarity between a sample to be identified and all the species in a genus. This estimate is obtained by matching each character as it appears in the sample and in each successive species, and scoring the match from zero (characters dissimilar) to one (characters identical). Scores of all the characters are averaged for the final coefficient of general similarity.

Measurements are considered to be identical (score $S = 1$) as long as the sample mean $\bar{x}_m$ differs from the species mean $\bar{x}_s$ by less than a value c, which is the intraspecific variability of the measurement in the genus. If R is the range of specific values of the measurement in the genus, the score S is equal to :

$$S = 1 - \frac{|\bar{x}_m - \bar{x}_s| - c}{R - c}$$

The intraspecific variability of the qualitative characters is taken into consideration by coding and scoring independently each state of these characters. Intermediate state scores are averaged to obtain the final score of the character (Fortuner, 1983; Fortuner & Wong, 1984).

For example, some populations of *Helicotylenchus pseudorobustus* (Steiner) Golden have all individuals with a Y-shaped fusion of inner lateral field lines on tail, while in other populations, some specimens have a Y-shaped fusion and others have a U-shaped fusion (Fortuner, Maggenti & Whittaker, 1984). The composite description of this species, including variability observed in all its populations, is coded 1 (present) for state Y and also 1 for state U, because both states are present.

Let's compare a non variable population of *H. pseudorobustus* (i.e., with all specimens Y-shaped) to the composite description of this species. The current program NEMAID scores positive matches (1-1) as 1 and scores mismatches (1-0 or 0-1) as 0. The negative matches (0-0) are neutralized. These intermediate scores are averaged to obtain the final score for the character.

The final score calculated by NEMAID (0.50) is rather low in spite of the fact that the sample does belong to *H. pseudorobustus.* Generally speaking, the method above gives an incorrect result when a sample is compared to a species with variable qualitative characters.

Professor Sneath has suggested *(in litt.)* the use of the method of Lapage *et al.* (1973) recently reviewed by Willcox, Lapage and Holmes (1980), developed for computer assisted bacterial identification. With this method, each character state is represented for each species by the percentage of populations of the species where this state has been observed.

Individuals with a Y-fusion have been observed in eleven samples (92 %), and with a U-fusion in seven samples (58 %), out of twelve large samples of *H. pseudorobustus* studied by Fortuner, Maggenti and Whittaker (1984). In another species, *H. dihystera* (Cobb) Sher, the inner line fusion pattern is not variable and all samples have 100 % individuals with a Y-fusion (Fortuner, Merny & Roux, 1981). In Table 2, for Willcox's method, the 100 % and 0 % values are represented respectively by 0.99 and 0.01 " because bacterial strains are susceptible to variation and one can not be sure that every strain in a particular taxon will always be positive or negative for a specific test " (Stevens, 1980).

Table 1

NEMAID identification score.

Character states	*Sample codes*	H. pseudo-robustus *codes*	*Intermediate score*	*Final score*
Y-junction	1	1	1	$\frac{(1 + 0)}{2} = 0.50$
U-junction	0	1	0	

Table 2

Willcox's identification matrix.

Character states	H. pseudorobustus	H. dihystera
Y-junction	0.92	0.99
U-junction	0.58	0.01

* Associate in the Division of Nematology, University of California, Davis; and California Department of Food and Agriculture, Analysis & Identification, Room 340, 1220 N Street, Sacramento, CA 95814, USA.

We can now compare the sample used for the first example (Y-junction positive; U-junction negative) to the two species *H. pseudorobustus* and *H. dihystera*. The Willcox's method utilizes the score in the identification matrix if the unknown is positive for a particular state (here for Y-junction). If the unknown is negative (here the sample is negative for U-junction), 1.00 minus the score in the matrix is used. Willcox's probabilities are calculated as follows :

— with *H. pseudorobustus* :

$$0.92 \times (1.00 - 0.58) = 0.3864$$

— with *H. dihystera* : $0.99 \times (1.00 - 0.01) = 0.9801$

These scores indicate that the unknown is closer to *H. dihystera* than to *H. pseudorobustus*. Actually, because *H. pseudorobustus* is variable for the inner line fusion pattern, some populations of this species are 100 % Y-positive. The unknown, which is also 100 % Y-positive, should have received as high a score with *H. pseudorobustus* as with *H. dihystera*.

Willcox's method cannot take into proper account the variability of some qualitative characters in nematodes.

The percentages of specimens positive for each character state could be averaged accross the populations observed for each species. The mean *m* and standard deviation s.d. of the distribution of these percentages could be calculated. The intraspecific variability of the characters could be taken into account by accepting a sample as similar to the species only when the percentage of positive specimens for the sample falls within an interval +/— 2 s.d. centered on the mean percentage *m*.

Such a method would assume that the percentages of positive specimens in various populations of a species are normally distributed. Normality of such distributions have been tested (Tab. 3) in *H. pseudorobustus* for the fusion of inner lateral field lines, and in *H. dihystera* for tail shape (type-3 tails). It is evident from Table 3 that the percentages are not normally distributed.

With the identification program NEMAID, a sample should be scored as similar to a species for a character state if the percentage of positive specimens for this state in the sample is within the range of percentages observed in various populations that belong to this species. In the example used above, a sample with 100 % Y-positive specimens should be scored as similar *(S = 1)* to *H. pseudorobustus* for this character, because some at least of the populations of *H. pseudorobustus* are 100 % Y-positive. In this case, because mean = 59.5 % and standard deviation = 47.4 % among various populations of *H. pseudorobustus* (Tab. 3), use of a weighted average (mean +/— 2 st. dev.) would provide the desired result. However in other instances, weighted averages would be erroneous. Type-3 shape in *H. dihystera* has mean = 38.3 % and standard deviation = 30.7 % (Tab. 3). Samples would be considered as similar to *H. dihystera* for percentage of positive type-3 tail specimens falling within the interval 0-99.7 %. The true range of observed percentages is only 0-90 % (Tab. 3).

Better practical results are achieved when two parameters *R* and *c* are calculated for each species. *R* is the mid-range of the various percentages observed in several populations of the species, and *c* is equal to half the percentage variation. A sample with U % positive specimens is considered similar to the species for the character when U falls within the range *R* +/— *c*.

The algorithms for comparison of qualitative characters can be described as follows :

1. For each species, the percentage of positive specimens for each character state is recorded in several populations, and the minimum and maximum values of these percentages are noted (for example in *H. pseudorobustus*, the percentage of positive Y-junction specimens varies from 0 to 100 % in different populations).
2. Two parameters *R* (mid percentage range) and *c* (half percentage variation) are calculated for each species character state with :

$$R = \frac{\text{(Max. percentage + Min. percentage)}}{2}$$

$$c = \frac{\text{(Max. percentage — Min. percentage)}}{2}$$

   Table 4 shows the calculation of *R* and *c* for the inner lines junction in the two species from the previous examples.
3. To compare an unknown sample with the species, the percentage of positive individuals in the sample

Table 3

Distribution of percentages of positive specimens for two qualitative characters in several *(n)* populations of a species.

Character	*n*	*mean*	*median*	*mode*	*st. dev.*	*range*	*skew.*	*kurt.*
Y-fusion inner lines	10	59.50	86.50	100.00	47.423	0-100	— 0.31	— 2.00
Type-3 tail shape	11	38.27	30.00	not unique	30.723	0-90	0.49	— 1.25

(U %) is calculated for each character state. Then an intermediate score *S* is calculated with :

$$S = 1 - (1\ U\text{-}R\,| - c)$$

Table 4

Improved identification matrix.

Species	*Character state*	*Percentage of positive specimens* *Min.*	*Max.*	*R*	*c*
H. pseudorobustus	Y-junction	0	100	0.50	0.50
	U-junction	0	100	0.50	0.50
H. dihystera	Y-junction	100	100	1.00	0.00
	U-junction	0	0	0.00	0.00

Intermediate scores are averaged for the final character score.

This method can be tested with the sample used in the previous example. It has *U* = 100 % for Y-junction and *U* = 0 % for U-junction. The score computation is shown on Table 5.

The final score (1.00 in both cases) indicates that the sample is similar to both species for the character.

H. pseudorobustus includes populations with different values for the character inner lines junction pattern. For example, the type population has 20 % specimens Y-positive and 80 % specimens U-positive. The comparison of this population with the composite description of the species gives the following scores :

Y-junction : s = 1 — (|0.20 — 0.50| — 0.50) = 1.20

U-junction : s = 1 — (|0.80 — 0.50|— 0.50) = 1.20

Here the final score (1.20) is arbitrarily set to 1.00 because no similarity score can be higher than one with the program NEMAID. It can be calculated that the type population of *H. pseudorobustus* has a score of 0.20 with *H. dihystera.* Finally one sample of *H. pseudorobustus* from Germany had 100 % specimens U-positive (Fortuner, Maggenti & Whittaker, 1984). According to the improved method its scores would be 1.00 with *H. pseudorobustus,* 0.00 with *H. dihystera.*

Table 5

Identification score calculated with improved method.

Species	*Character state*	*Intermediate score*	*Final score*
H. pseudorobustus	Y-junction	s = 1 — (|1.00 — 0.50| — 0.50) = 1	$\frac{(1 + 1)}{2} = 1$
	U-junction	s = 1 — (|0.00 — 0.50| — 0.50) = 1	
H. dihystera	Y-junction	s = 1 — (|1.00 — 1.00| — 0.00) = 1	$\frac{(1 + 1)}{2} = 1$
	U-junction	s = 1 — (|0.00 — 0.00| — 0.00) = 1	

The improved computation method gives a more accurate score when a sample is compared to a variable species. NEMAID has been rewritten for IBM-PC and IBM-compatible microcomputers. An updated program, NEMAID-3, will use this improved method.

ACKNOWLEDGEMENTS

I thank Professor Sneath, University of Leicester, England for his comments and helpful suggestions.

REFERENCES

FORTUNER, R. (1983). Computer assisted semi-automatic identification of *Helicotylenchus* species. The program NEMAID. *Calif. Pl. Pest Dis. Reptr;* 2 : 45-48.

FORTUNER, R., MAGGENTI, A. R. & WHITTAKER, L. M. (1984). Morphometrical variability in *Helicotylenchus* Steiner, 1945. 4 : Study of field populations of *H. pseudorobustus* and related species. *Revue Nématol.,* 7 : 121-135.

FORTUNER, R., MERNY, G. & ROUX, C. (1981). Morphometrical variability in *Helicotylenchus* Steiner, 1945. 3 : Observations on African populations of *Helicotylenchus dihystera* and considerations on related species. *Revue Nématol.,* 4 : 235-260.

FORTUNER, R. & WONG, Y. (1984). Review of the genus *Helicotylenchus* Steiner, 1945. 1 : A computer program for identification of the species. *Revue Nématol.,* 7 : 385-392.

LAPAGE, S. P., BASCOMB, S., WILLCOX, W. R. & CURTIS, M. A. (1973). Identification of bacteria by computer : general aspects and perspectives. *Microbiol.,* 77 : 273-290.

STEVENS, M. (1980). Computer assisted bacterial identification. *Med. Lab. Sc.,* 37 : 223-228.

WILLCOX, W. R., LAPAGE, S. P. & HOLMES, B. (1980). A review of numerical methods in bacterial identification. *Ant. Leeuwenhock,* 46 : 233-299.

Accepté pour publication le 23 mars 1986.