SHORT COMMUNICATIONS

A Case for Proportional Similarity in Numerical Taxonomy?

By G. C. WARE and A. J. HEDGES
Department of Bacteriology, The Medical School, Bristol, BS8 1TD

(Received 29 July 1977)

INTRODUCTION

Investigators in numerical taxonomic and related studies have invented a large number of different coefficients to express the degree of similarity between a pair of objects (operational taxonomic units; OTUs) (Sokal & Sneath, 1963; Cheetham & Hazel, 1969). Gower (1971) proposed a general coefficient that permits the mixing of binary, multi-state, and quantitative characters:

\[ S_a = \left( \frac{\sum_{i=1}^{n} w_{ik} \times S_{ijk}}{\sum_{i=1}^{n} w_{ik}} \right) \]

where OTUs \( j \) and \( k \) are compared over \( n \) characters, \( w_{ik} \) is a weight (either 0 or 1) for the \( i \)th character, and \( S_{ijk} \) is a ‘unit’ similarity measure for the comparison of \( j \) and \( k \) with respect to the \( i \)th character. \( S_a \) is, thus, a weighted average of all unit similarity measures for a given pair of OTUs. When the data consist entirely of binary characters and all \( w_i = 1 \), \( S_a \) becomes identical to the much used simple matching coefficient (SBM).

In calculating each unit \( S_{ijk} \), Gower referred the difference observed between the character states in \( j \) and \( k \) to the extreme range observed in that character over the whole set of OTUs, and a similar practice has been followed by others (Carmichael, Julius & Martin, 1965; Anderson, 1971), i.e.

\[ s_{ijk} = 1 - \frac{|X_{ij} - X_{ik}|}{R} \]

where \( X_i \) denotes the state of the \( i \)th character and \( R \) is its range. Most of the similarity coefficients that have so far been applied to the numerical classification of micro-organisms have adopted an equivalent form of ‘scaling’ (Sneath & Sokal, 1973).

RESULTS AND DISCUSSION

As far as we know, no one has yet suggested that each \( S_{ijk} \) be computed as the simple ratio of one character state to the other, i.e.

\[ s_{ijk} = X_{ij}/X_{ik} \quad (X_{ij} \leq X_{ik}) \]

A coefficient, \( S_{aa} \), computed as the weighted or unweighted average of all \( s_{ijk} \) so defined, has the following desirable properties. (a) It has a range from 0 to 1. (b) It can accommodate binary, multi-state, and quantitative characters, provided that 0/0 is set equal to 1. (c) Computer programming is simplified, since all types of data can be dealt with by the same simple algorithm. [A versatile programme (on file as RBA001:BIBLOS, CLSTR1 in the S.W. Universities' Computer Network at Bristol) which utilizes the \( S_{aa} \) coefficient is available from the authors.] (d) For equally weighted binary data, \( S_{aa} = S_{ab} \). (e) OTUs at the extreme range do not show zero similarity. (f) Scaling is unnecessary, since the ratio is independent not
Table 1. Comparison of values of similarity coefficients $S_a$ and $S_{aA}$ for the character, bacterial length

<table>
<thead>
<tr>
<th>OTU</th>
<th>Length</th>
<th>$S_{aA}$</th>
<th>$S_a$</th>
<th>$S_{aA}$</th>
<th>$S_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0.50</td>
<td>0.89</td>
<td>0.50</td>
<td>0.99</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0.33</td>
<td>0.78</td>
<td>0.33</td>
<td>0.97</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>0.25</td>
<td>0.67</td>
<td>0.25</td>
<td>0.97</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>0.20</td>
<td>0.56</td>
<td>0.20</td>
<td>0.96</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>0.10</td>
<td>0.00</td>
<td>0.10</td>
<td>0.91</td>
</tr>
<tr>
<td>G</td>
<td>100</td>
<td></td>
<td></td>
<td>0.01</td>
<td>0.00</td>
</tr>
</tbody>
</table>

* All compared with OTU ‘A’.

only of the units of measurement but also of the range observed – or likely to be observed. This last property would seem to be especially desirable, since it allows the concept of a true parameter of similarity (attainable if all permissible characters were compared for a given pair of OTUs).

Carmichael et al. (1965) discussed what seems to be the main reason for the non-appearance of an $S_{aa}$ type of coefficient. This is a conceptual problem, most easily demonstrated by considering a number of, say, bacteria of different lengths, e.g. 1, 2 and 3 units. If we compare the similarity of the first and second, should we get the same result as when comparing the second and third, i.e. are the similarities $1/2, 2/3$ (ratios), or $1/2, 1/2$ (increments/range)?

Single character similarities computed as $S_a$ and $S_{aA}$ are compared in Table 1 for an extended range of lengths (note particularly the values for OTU ‘F’).

It is true that coefficients of the $S_a$ type can deal with ratios after their logarithmic transformation to linear increments, but there remain both the problem of scaling and the need for a subjective decision as to whether a particular set of data should or should not be transformed.

We are by no means convinced that a simple proportional measure of similarity corresponds less well with our intuitive idea of this elusive property than an incremental measure does, whether standardized or not. We are, however, convinced that a comparison between two OTUs should be independent of the existence or discovery of other OTUs; our impressions of similarity may change but the OTUs do not.

REFERENCES


