Vigour and Pattern as Applied to Multistate Quantitative Characters in Taxonomy

By M. J. SACKIN

Department of Microbiology, University of Leicester, Leicester LEI 7RH

(Received 27 August 1980)

This paper presents two extensions of vigour and pattern difference coefficients suitable for quantitative characters and gives some properties of these coefficients, with illustrative examples. The first generalization has been proposed (Sneath, 1979) for use in identification; the second, mentioned by Sneath & Sokal (1973), may be of considerable taxonomic use. Both generalizations when used on binary characters reduce to the original vigour and pattern coefficients, as proposed by Sneath (1968).

INTRODUCTION

With some types of taxonomic data the experimental observations may depend greatly on some influence (such as environment) that affects the characters broadly in the same way. For example, in microbiology if two subcultures of the same strain are tested after short and long incubation times, the former may show many fewer positive test results than the latter: the subcultures will at first sight appear very dissimilar. The method of Sneath (1968) is designed to compensate for such differences and extract any underlying differences that may remain in the pattern of the test results. The overall total difference between the two subcultures may thus be partitioned into a 'pattern difference' and a 'vigour difference'. In the above example, the subculture tested after the shorter incubation time would show the lower vigour, most of the total difference between the two subcultures would be vigour difference, and the pattern difference would be very low.

In general, vigour is the degree of metabolic achievement under defined conditions, and in microbiology, for example, this may be represented by the tendency of a strain to give positive test results. The measures of Sneath (1968) apply to tests (or other characters) that are binary or two-state only, i.e. those coded, for example, as 0 = negative result, 1 = positive result. This paper discusses methods which extend the type of characters to multistate quantitative ones which may take on any value \( x \) lying between 0 and 1 inclusive and which represent the degree or strength of positiveness, from 0 = negative to 1 = strongest positive.

For many numerical classification and identification computer programs the characters must be coded as non-negative integers, for example, 0 = negative, 1 = weak positive, 2 = strong positive, or maybe with more gradations, say 0, 1, 2, 3, 4, 5. Throughout this paper, however, it is assumed for convenience that each character has been linearly transformed into the range 0 to 1 inclusive, by subtracting the maximum value and dividing by the range. The maximum value and the range may correspond either to the most extreme theoretically possible states (in some sense), or else they may correspond to the most extreme states occurring in the particular set of strains under investigation (see Gower, 1978, for discussion). The formulae given below apply to both cases.

It is important to emphasize that all characters must be scorable into a range with 0 representing the most negative state and 1 the most positive, and not with any arbitrary assignment of meaning to the range (for example, 0 = positive, 1 = negative). Thus, the
methods do not allow characters that are qualitative and have three or more states, for example, 0 = red, 1 = blue, 2 = yellow, 3 = brown, where the states are not ordered in any direction of metabolic achievement, because the contribution of such qualitative characters toward vigour would be undefined. However, such characters can usually be broken down into several two-state characters (see Sneath & Sokal, 1973, pp. 149–152), and, if appropriate, they could then be used for vigour and pattern analysis.

Quantitative multistate characters may arise when pairs of taxa are being compared, rather than pairs of strains. The character values for the taxa will normally derive in some way from observations on individual strains. For example, suppose a group of strains is considered to comprise a taxon, and the observations on the strains are two-state tests. Then, typically the character values for the taxon might be the proportions of the strains which show a positive state. Thus, two-state observations on strains may yield multistate characters in taxa. In general, the objects (for example, strains, taxa) are termed OTUs (operational taxonomic units) as in Sneath & Sokal (1973).

Even if all the characters are scored in a range from 0 representing the most negative state to 1 representing the most positive, there is still clearly a danger that removal of vigour difference from the total difference to give pattern difference may remove substantial elements of contrast which are not explained in terms of the metabolic concept of vigour difference given earlier. For example, a strain may give a negative result on a test because it is genetically incapable of reacting positively, however long the incubation period. The applicability of the pattern difference approach must remain a matter for the investigator’s judgement. So far the method has been of value in several studies (Priest et al., 1973; Goodfellow et al., 1976, 1979).

METHODS

In the case where all the characters are binary (Sneath, 1968), the vigour of a strain (or, more generally, of an OTU) is the number of characters that have state 1. The vigour difference $D_v$ between the two OTUs is the difference between their vigours, averaged over all the characters. The pattern difference is a measure of the difference which remains when the component due to vigour difference has been removed. For example, all the difference between the two strains with character values $0 \ 0 \ 0 \ 1 \ 1$ (strain 1) $0 \ 1 \ 1 \ 1 \ 1$ (strain 2) is vigour difference. The second strain being the more vigorous in terms of the five characters by which the strains have been described. A pattern difference occurs only if there are characters with negative state in the first strain and positive state in the second strain and characters with positive state in the first strain and negative state in the second strain. For example:

$$
\begin{array}{cccc}
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 1 \\
\end{array}
$$

is vigour difference. Thus, in the usual $2 \times 2$ frequency table for comparison of two OTUs consisting of binary characters:

<table>
<thead>
<tr>
<th>OTU 2</th>
<th>positive</th>
<th>negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTU 1</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

a pattern difference occurs only if both $b$ and $c$ are greater than 0. The total number of positive results, i.e. the vigour. for OTU 1 is $a + c$, and that for OTU 2 is $a + b$. Thus, the vigour difference $D_v$ between OTU 1 and OTU 2 is $(a + c) - (a + b)/(c - b)/n$, where $n$
is the total number of characters examined, ignoring any that have ‘missing’ values in one or both strains, i.e. \( n = a + b + c + d \), as in Sneath (1968). The vigour difference between OTU 2 and OTU 1, incidentally, has the same absolute value as that between OTU 1 and OTU 2 but with opposite sign. Often, however, only the absolute value is of interest, and this is often what is meant by the vigour difference, even though it is not always explicitly stated.

The total difference \( D_T \) is \( (b + c)/n \), and this is \( 1 - S_{SM} \), where \( S_{SM} \) is the well-known simple matching coefficient. Then, the pattern difference \( D_P \) is given by the following partition of \( D_T \):

\[
D_T^2 = D_V^2 + D_p^2
\]
i.e.

\[
D_p^2 = D_T^2 - D_V^2
\]
which gives

\[
D_p = 2\sqrt{bc/n}
\]
As desired, \( D_p \) is zero if \( b \) or \( c \) is zero and is large if both \( b \) and \( c \) are large. It is convenient to use squares in this partitioning since it yields a simple formula for \( D_p \) which has all the desired properties and which lies between 0 and 1.

Note that it is possible (Jones & Sackin, 1980) to base \( D_T, D_V \) and \( D_p \) on Jaccard’s coefficient \( S_j \) rather than on \( S_{SM} \). Jaccard’s coefficient is \( a/(a + b + c) \), i.e. it ignores matching negatives, and the corresponding formulae are obtained simply by taking \( n \) as \( a + b + c \) instead of \( a + b + c + d \).

For the extension of \( D_T, D_V \) and \( D_p \) to multistate characters the following properties seem desirable:

1. If all the characters in the OTUs being compared happen to be two-state then the formulae should reduce to the two-state case above.
2. The pattern difference should be zero if all the differences between the two OTUs are in the same direction, i.e. either \( x_i \geq y_i \) (all \( i \)) or \( x_i \leq y_i \) (all \( i \)), where \( x_i, y_i \) are the character values in the two OTUs for each character \( i \) (\( i = 1, 2, \ldots, n \)). If the pattern difference is zero then the vigour difference should equal the total difference.
3. \( D_T, D_V \) and \( D_p \) should each lie between 0 (no difference) and 1 (maximum possible difference).

One measure which does not satisfy property 2 is the Penrose shape coefficient (Penrose, 1954) called \( C_p \) by Sneath & Sokal (1973). \( C_p \) is the standard deviation of the differences \( x_i - y_i \) in the character values. Thus, even in the (binary) case

\[
\begin{bmatrix}
0 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 1
\end{bmatrix}
\]
\( C_p \) will not be zero and, incidentally, property 1 is not satisfied either.

Note that if the desirable properties are satisfied then it is possible for two OTUs \( A \) and \( B \) to have no pattern difference, for \( B \) and \( C \) also to have no pattern difference, but for the pattern difference between \( A \) and \( C \) to be greater than zero. This can happen both for binary and quantitative data. For example, if the three OTUs \( A, B \) and \( C \) have character values as follows (just two binary characters):

\[
A: \ 0 \ 1 \\
B: \ 0 \ 0 \\
C: \ 1 \ 0
\]
then \( D_p \) is zero between \( B \) and each of the other strains but \( D_p \) between \( A \) and \( C \) is 1. It is impossible to represent \( A, B \) and \( C \) as points forming a triangle in ordinary (two-dimensional Euclidean) space with the distances of the sides proportional to the corresponding pattern differences: the triangle law is not satisfied. In this example, \( B \) is an OTU with zero vigour, and it will show no pattern difference with any other OTU, all the difference being in vigour; yet clearly the other OTUs may exhibit pattern differences among themselves.
More generally, pattern difference cannot be a metric. A measure $\phi$ between pairs of OTUs is a metric if the following four axioms (Sneath & Sokal, 1973, p. 120) are satisfied for all $A$, $B$, $C$:

1. $\phi(A, B) \geq 0$ and $\phi(A, A) = \phi(B, B) = 0$.
2. $\phi(A, B) = \phi(B, A)$.
3. $\phi(A, C) \leq \phi(A, B) + \phi(B, C)$.
4. If $A \neq B$ then $\phi(A, B) > 0$.

Pattern difference satisfies axioms 1 and 2 but not 3 and 4, axiom 3 being the triangle inequality. In practice, if there are not enormous differences in the vigours of the OTUs under examination, then pattern difference is often a semimetric, i.e. axiom 3 would in practice usually be satisfied but not axiom 4.

The presence of OTUs with very low or very high vigour will cause difficulties with clustering methods. A single strain with zero vigour or one with unit vigour will cause single linkage clustering to collapse, but with average linkage in practice usually only small-scale local collapsing occurs and much structure remains (see the applications cited at the end of the Introduction).

Two extensions of vigour and pattern to allow quantitative characters will now be presented. Both satisfy the three desirable properties, and both seem useful, though for slightly different purposes.

The first generalization, in which total vigour and pattern difference will be denoted by $D_{T1}$, $D_{V1}$ and $D_{P2}$, respectively, has been described by Sneath (1979). Let $x_i, y_i$, as earlier, be the $i$th character states in the two QTUs, let $\Delta_i = x_i - y_i$ be their difference, and let $n$ be the number of characters considered, so that $i$ runs from 1 to $n$. As in the binary case, characters with matching zeros ($x_i = y_i = 0$) may be ignored, but it would probably be much more usual to count them here. Then:

$$D_{T1} = \frac{(\sum \Delta_i^2)}{n}$$
$$D_{V1} = \frac{(\sum |\Delta_i|)}{n}$$
$$D_{P2}^2 = D_{T1}^2 - D_{V1}^2$$

where the summation is over $i = 1, 2, \ldots, n$. $D_{P1}$ works out as $(2/n) \sqrt{\sum \Delta_i^2 \Delta_j^2}$, where the summation $\sum'$ is over all pairs $i, j$ $(i < j)$ for which $\Delta_i, \Delta_j$ are of different sign. [Note: in Sneath, 1979, a summation sign was inadvertently omitted from the start of the formula for $D_p$ in paragraph (e) on p. 197.]

$D_{T1}$ is the squared taxonomic distance $d^2$ (notation as in Sneath & Sokal, 1973, and elsewhere).

In the second generalization, in which the coefficients will be denoted by $D_{T2}$, $D_{V2}$ and $D_{P2}$, $D_{T2}$ is the mean character difference (m.c.d.) which is the same as $1 - S_G$, where $S_G$ is Gower’s similarity coefficient (Gower, 1971), provided each character has a range which runs from 0 to 1. These are precisely the types of character allowed for here (see above), and for the identity to hold, the characters may be any combination of (a) multistate quantitative, converted into range 0 to 1, (b) two-state, counting matching negatives, and (c) two-state, ignoring matching negatives.

Then, with the same notation as in the first generalization:

$$D_{T2} = (\sum |\Delta_i|)/n$$
$$D_{V2} = (\sum \Delta_i)/n$$

and

$$D_{P2}^2 = D_{T2}^2 - D_{V2}^2$$

which yields

$$D_{P2} = (2/n)\sqrt{-\frac{\sum \Delta_i \Delta_j}{n}}$$
Table 1. Values of total, vigour and pattern difference coefficients between OTU L and each of the OTUs M₁ to M₁₂.

Each OTU is described in terms of six characters (1 to 6). For details of the coefficients, see text.

<table>
<thead>
<tr>
<th>OTU</th>
<th>Character values</th>
<th>Coefficients for OTU L compared with each OTU M in turn</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>D₁₁</td>
<td>D₁₂</td>
</tr>
<tr>
<td>L</td>
<td>0 0 0 0 0 0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M₁</td>
<td>0 0 0 0 0 1</td>
<td>0.5167</td>
<td>0.5167</td>
</tr>
<tr>
<td>M₂</td>
<td>0 0 0 0 0 1</td>
<td>0.3500</td>
<td>0.3500</td>
</tr>
<tr>
<td>M₃</td>
<td>0 0 0 0 0 1</td>
<td>0.3833</td>
<td>0.3800</td>
</tr>
<tr>
<td>M₄</td>
<td>0 0 1 0 0 0</td>
<td>0.5167</td>
<td>0.4333</td>
</tr>
<tr>
<td>M₅</td>
<td>0 1 0 0 0 0</td>
<td>0.6167</td>
<td>0.4033</td>
</tr>
<tr>
<td>M₆</td>
<td>1 0 0 0 0 0</td>
<td>0.6833</td>
<td>0.3500</td>
</tr>
<tr>
<td>M₇</td>
<td>0.5 0.5 0.5 0.5 0.5 0.5</td>
<td>0.1667</td>
<td>0.0533</td>
</tr>
<tr>
<td>M₈</td>
<td>1 1 1 0 0 0</td>
<td>0.7833</td>
<td>0.1533</td>
</tr>
<tr>
<td>M₉</td>
<td>1 1 0 0 0 0</td>
<td>0.7833</td>
<td>0.2367</td>
</tr>
<tr>
<td>M₁₀</td>
<td>0.5 0.7 1 0 0.1 0.1</td>
<td>0.5300</td>
<td>0.2800</td>
</tr>
<tr>
<td>M₁₁</td>
<td>0.1 0.3 0.6 0.8 0.9 0.9</td>
<td>0.0100</td>
<td>0</td>
</tr>
</tbody>
</table>

Vigour and pattern on quantitative characters
DISCUSSION

The two generalizations seem appropriate for slightly different purposes. The first has been described and used (Sneth, 1979) as part of the taxon radius model for (microbial) identification. The second, mentioned by Sneth & Sokal (1973, pp. 171–172), would seem appropriate in a numerical taxonomic study in which $S_G$ has been used on quantitative characters whose states are ordered in the direction of increasing vigour. This is the situation in many microbiological studies: hence the partitioning of $1 - S_G$ into $D_{p2}$ and $D_{v2}$ may frequently be of taxonomic use. The method will partition the complement of $S_G$ into vigour and pattern components just as existing methods partition the complement of $S_{SM}$ or $S_{P}$.

The examples in Table 1 will help to highlight the properties of the coefficients. In the table, OTU $L$ is being compared with each of OTUs $M_1, M_2, \ldots, M_{12}$ in turn, in terms of six characters all of whose states lie between 0 and 1. The table shows the total, vigour and pattern differences for the two methods as well as $d_p$ which is $\sqrt{D_{p1}}$. The coefficient $d_p$ has been used (Sneth, 1979) because $D_{p1}$ is still a squared distance and also because $d_p$ seems in practice to be more nearly normally distributed than $D_{p1}$ for the members of a cluster of OTUs when compared with the centroid of the cluster. This allows one, albeit with due caution, to use normal probabilities when testing an unknown for membership to a group by the criterion of $d_p$.

A few observations may be made from Table 1. First, although each of the OTUs $M_3, M_4, M_5, M_6$, and $M_7$ seem equally vigorous (their character values each consist of five 0s and one 1), their vigour differences with $L$ are the same (0.4333) according to the $D_{v2}$ formula only, not by the $D_{v1}$ formula. This is because large character state differences have more than a proportionate effect than small ones in the $D_{v1}$ formula (and $D_{p1}$ too) because of the squared nature of these formulae. Thus, for example, in the ($L,M_3$) and ($L,M_4$) comparisons, the large (positive) effect of character 6 on $D_{v1}$ in the ($L,M_3$) comparison more than outweighs the difference in the contribution for character 3, giving $D_{v1}$ higher for ($L,M_3$) than for ($L,M_4$).

The second observation is that $D_{p2}$ always lies in between $D_{p1}$ and $d_p$. In fact it can be proved that this will always happen, or, more precisely that

$$D_{p1} < D_{p2} < d_p$$

(These relations are termed inequalities, even though equality can occur as special cases.)

Given, as usual, that all character values lie in the range 0 to 1, inclusive, the left-hand inequality follows immediately from the formulae for $D_{p1}$ and $D_{p2}$. Equality occurs either (a) when all terms in the summation $\Sigma'$ are unity, i.e. when the characters are two-state, and the resulting coefficient will indeed be the same as $D_p$ as defined by Sneth (1968) for binary characters, or (b) if every $\Delta_j$ is of the same sign (or zero), in which case both $D_{p1}$ and $D_{p2}$ will be zero.

To prove the right-hand inequality, we require to prove that

$$(2/n)\sqrt{-\Sigma' \Delta_i \Delta_j} \leq \sqrt{(2/n)\Sigma' \Delta_i^2 \Delta_j}$$

i.e. that

$$(2/n)(-\Sigma' \Delta_i \Delta_j) \leq \Sigma' \Delta_i^2 \Delta_j$$

i.e. that

$$(4/n^2)(\Sigma' \Delta_i \Delta_j)^2 \leq \Sigma' \Delta_i^2 \Delta_j$$

i.e. that

$$(4/n^2)(\Sigma \Delta_j)^2 (\Sigma \Delta_i)^2 \leq \Sigma \Delta_i^2 \Sigma \Delta_j^2$$

where the summation of the $i$-subscripted terms is over the $p$, say, positive $\Delta$s and the summation of the $j$-subscripted terms is over the $q$, say, negative $\Delta$s, the remaining $n - p - q$ \Delta$s, if any, being zero (i.e. $p + q \leq n$).

Thus, it is necessary to prove that

$$\frac{4}{n^2} \leq \frac{\Sigma \Delta_i^2}{(\Sigma \Delta_j)^2} \frac{\Sigma \Delta_j^2}{(\Sigma \Delta_i)^2}$$

(2)
The conditions for equality in (2) are the same as those for the right-hand inequality of (1). Note, however, that (2) is determinate only if both $p$ and $q$ are non-zero. If $p$ or $q$ is zero then so are $D_{p^2}$ and $d_p$, so that in this case the right-hand inequality of (1) is satisfied. The remainder of the proof is for $p \neq 0, q \neq 0$.

Now,

$$\left( \sum \Delta_i \right)^2 \leq p \sum \Delta_i^2$$  \hspace{1cm} (3)

equality occurring when all the $\Delta_i$s are equal. The proof of (3) follows that for Cauchy's inequality:

$$\left( \sum a_i b_i \right)^2 \leq \sum a_i^2 \sum b_i^2$$

(Hardy et al., 1952), taking the special case $a_i = \Delta_i, b_i = 1$.

For,

$$p \sum \Delta_i^2 - (\sum \Delta_i)^2 = \frac{1}{2} \sum_{i<k} [\Delta_i^2 + \Delta_k^2 - \Delta_i \Delta_k]$$

where the summations are over all $p^2$ pairs of terms for which $\Delta_i$ and $\Delta_k$ are both positive.

$$= \sum \sum (\Delta_i^2 + \Delta_k^2 - 2\Delta_i \Delta_k)$$

$$= \sum \sum (\Delta_i - \Delta_k)^2$$

$$= 0 \text{ if all } p \text{ values of } \Delta_i \text{ are equal}$$

$$> 0 \text{ otherwise}$$

Similarly

$$\left( \sum \Delta_i \right)^2 \leq q \sum \Delta_i^2$$

Hence

$$\frac{1}{pq} \leq \frac{\sum \Delta_i^2}{(\sum \Delta_i)^2} \frac{\sum \Delta_i^2}{(\sum \Delta_i)^2}$$  \hspace{1cm} (4)

equality in (4) occurring when all the $\Delta_i$s are equal and all the $\Delta_j$s are equal.

Now

$$\frac{4}{n^2} \leq \frac{4}{(p + q)^2}$$  \hspace{1cm} (5)

because $p + q \leq n$, equality occurring when $p + q = n$, and

$$\frac{4}{(p + q)^2} \leq \frac{1}{pq}$$  \hspace{1cm} (6)

because

$$4pq \leq (p + q)^2$$

i.e.

$$0 \leq (p - q)^2$$

equality occurring when $p = q$.

Combining results (4), (5) and (6),

$$\frac{4}{n^2} \leq \frac{4}{(p + q)^2} \leq \frac{1}{pq} \leq \frac{\sum \Delta_i^2}{(\sum \Delta_i)^2} \frac{\sum \Delta_i^2}{(\sum \Delta_i)^2}$$

which proves (2) and hence the right-hand inequality of (1). Equality in the right-hand inequality of (1) thus occurs when all four conditions, (a) $p + q = n$, (b) $p = q$, (c) all the $\Delta_i$s
are equal, (d) all the $\Delta s$ are equal, are satisfied, for example, in cases $(L,M_{11})$ and $(L,M_{12})$ in Table 1. In the $(L,M_{11})$ comparison, $\Delta_i = 0.5$ (this applies to the first three characters), $\Delta_i = -0.9$ (characters 4, 5, 6), and both $dp$ and $D_{p2}$ work out as $\sqrt{\Delta_1\Delta_2}$, i.e. $\sqrt{0.5 \times 0.9} = \sqrt{0.45} = 0.6708$. Equality also occurs when $p$ or $q$ is zero, i.e. there is no pattern difference.

Thus, $D_{p1} < D_{p2} < dp$ always, and (i) $D_{p1} = D_{p2}$ when (A) the characters are all two-state, or (B) there is no pattern difference, and (ii) $D_{p2} = dp$ when (C) $p = q$ and $p + q = n$ and $\Delta_i$ are constant and $\Delta_i$ are constant, or (D) there is no pattern difference.

The comparisons $(L,M_{11})$ and $(L,M_{12})$ are cases where $D_{p2}$ and $dp$ are identical. In $(L,M_8)$ and especially $(L,M_9)$ they are close. Where there is much variation in the $\Delta$s from one character to another and where the variation over most, but not all, of the characters is in the same direction (i.e. $p$, $q$ are very different), then condition (C) is far from being satisfied and one would expect $D_{p2}$ to be nearer to $D_{p1}$ than to $dp$, for example the case $(L,M_3)$.

Thus, $D_{p1}$ and its square root $dp$ are more sensitive than $D_{p2}$ to variation within the positive and negative character differences and in the numbers of the two kinds (positive and negative) of character difference. They then give more weight than $D_{p2}$ to large differences in character values. Where all the differences are small, as in $(L,M_4)$ or $(L,M_{13})$, $D_{p1}$ in particular will be very small, being the square of an already small number.

In conclusion, the three pattern difference coefficients, $D_{p1}$, $D_{p2}$ and $dp$, all show the same kinds of trend and are hence fairly interchangeable. However, $D_{p2}$ seems the most appropriate for numerical taxonomic studies on quantitative characters in which Gower’s coefficient $S_G$ is being used, because it is a partitioning of the complement of $S_G$, but $dp$ seems more appropriate in the taxon radius model for identification, because of its tendency to be distributed approximately normally. It will be interesting to use these coefficients in large-scale studies.

This work draws on many ideas of P. H. A. Sneath. I thank Professor Sneath and Dr Dorothy Jones for reading the manuscript and Professor R. O. Davies for help in the proof involving Cauchy’s inequality. I also thank the M.R.C. for support (grant G977/613/S).

REFERENCES


