*J. theor. Biol.* (1978) **74**, 155–161

# Interpretation of the Difference Index as a Guide to Protein Sequence Identity

ATHEL CORNISH-BOWDEN

*Department of Biochemistry, University of Birmingham, P.O. Box* 363, *Birmingham B*15 2*TT, England*

(*Received* 12 *September* 1977)

The expected behaviour of the "difference index" (Metzger, Shapiro, Mosimann & Vinton, 1968) has been examined under simple assumptions. Provided that it is only used to compare the compositions of proteins of equal or near-equal length, it can be interpreted to provide an estimate of the number of differences between the two sequences. Results with several pairs of proteins are in good agreement with estimates obtained by an alternative approach described previously (Cornish-Bowden, 1977).

## 1. Introduction

In a previous paper (Cornish-Bowden, 1977), I described a simple way of using amino acid compositions to estimate the extent of sequence identity between two proteins of equal length. The analysis permits a simple interpretation of the two indexes of compositional difference proposed by Harris, Kobes, Teller & Rutter (1969) and by Marchalonis & Weltman (1971); but it does not apply to the widely used "difference index" of Metzger, Shapiro, Mosimann & Vinton (1968). I have now examined the theoretical properties of the difference index, in the hope of making the abundant data that have been published with it more accessible to interpretation.

## 2. Theory

For the same reasons as before (Cornish-Bowden, 1977), I shall consider two sequences A and B with the same number $N$ of amino acid residues in each. With this restriction the difference index $DI$ of Metzger *et al.* (1968) is defined as follows:

$$DI = \frac{50}{N}\sum_i |n_{iA} - n_{iB}|, \qquad (1)$$

where $n_{iA}$ and $n_{iB}$ are the numbers of residues of the $i$th type of amino acid in A and B respectively. The summation is carried out over as many types of



0022-5193/78/180155+07 $02.00/0 

amino acid as are distinguished by the data, usually 18 as glutamate is not usually distinguished from glutamine or aspartate from asparagine.

Just as the simple properties of the index of Marchalonis & Weltman (1971) are obscured by the inclusion of the factor $10^4/N^2$ in its definition (Cornish-Bowden, 1977), so also is the behaviour of $DI$ complicated by the factor $50/N$ in equation (1). So I shall instead consider the "difference total", $DT$, defined as follows:

$$DT = \tfrac{1}{2}\sum_{i=1}^{18}|n_{i\mathrm{A}} - n_{i\mathrm{B}}| \qquad (2)$$

This quantity has a minimum value of zero, which occurs if identical compositions are compared, and a maximum value of $N$, which occurs if (improbably) the two compositions have no type of amino acid in common. More realistically, and more usefully, $DT$ cannot exceed the number $M$ of loci at which the two sequences are different; so it provides an exact lower limit for this number.

To determine the expected value of $DT$, I shall first consider two independent sequences A and B of the same length $L$, with the same probability $p_i$ of finding the $i$th type of amino acid at any locus in either sequence. Then all of the $n_{i\mathrm{A}}$ and $n_{i\mathrm{B}}$ are binomially distributed and the expected value of $DT$ is as follows:

$$E(DT) = \tfrac{1}{2}\sum_{i=1}^{18}\sum_{n_{i\mathrm{A}}=0}^{L}\left[\binom{L}{n_{i\mathrm{A}}}p_i^{n_{i\mathrm{A}}}(1-p_i)^{L-n_{i\mathrm{A}}} \times \right.$$
$$\left. \times \sum_{n_{i\mathrm{B}}=0}^{L}\binom{L}{n_{i\mathrm{B}}}p_i^{n_{i\mathrm{B}}}(1-p_i)^{L-n_{i\mathrm{B}}}|n_{i\mathrm{A}} - n_{i\mathrm{B}}|\right].$$

This expression becomes amenable to analysis if the absolute differences are removed by redefining the limits so that $n_{i\mathrm{B}} \geqslant n_{i\mathrm{A}}$ and using the fact that for every combination with $n_{i\mathrm{B}} < n_{i\mathrm{A}}$ there is another equally probable combination with $n_{i\mathrm{B}} > n_{i\mathrm{A}}$:

$$E(DT) = \sum_{i=1}^{18}\sum_{n_{i\mathrm{A}}=0}^{L-1}\left[\binom{L}{n_{i\mathrm{A}}}p_i^{n_{i\mathrm{A}}}(1-p_i)^{L-n_{i\mathrm{A}}} \times \right.$$
$$\left. \times \sum_{n_{i\mathrm{B}}=n_{i\mathrm{A}}+1}^{L}\binom{L}{n_{i\mathrm{B}}}p_i^{n_{i\mathrm{B}}}(1-p_i)^{L-n_{i\mathrm{B}}}(n_{i\mathrm{B}} - n_{i\mathrm{A}})\right].$$

This expression is very laborious to evaluate, especially if $L$ is large, because of the triple summation. This can be simplified to a double summation by multiplying out the expression in brackets and recognizing it as a power series in $[p_i/(1-p_i)]$ with coefficients that are alternately of the forms

$$\binom{L-1}{n_{i\mathrm{A}}}^2 \text{ and } \binom{L-1}{n_{i\mathrm{A}}}\binom{L-1}{n_{i\mathrm{A}}-1}.$$

Then the expected value of $DT$ can be expressed as follows:

$$E(DT) = L \sum_{i=1}^{18} \left( (1-p_i)^{2L} \left\{ G_i + \right. \right.$$
$$\left. \left. + \sum_{n_{iA}=1}^{L-1} \binom{L-1}{n_{iA}} \left[ \binom{L-1}{n_{iA}-1} + \binom{L-1}{n_{iA}} G_i \right] G_i^{2n_{iA}} \right\} \right), \quad (3)$$

where $G_i$ represents $[p_i/(1-p_i)]$. Now the expected value of $DT$ can readily be computed for any set of $p_i$ values. [Numerical difficulties arise with large values of $L$ when the binomial coefficients exceed the limits permitted by the computer (about $10^{77}$ with the ICL 1906A used in this work), but with fairly simple precautions these can be avoided, at least for $L$ up to about 700.] The result is weakly dependent on the set of $p_i$ values assumed, and all of the results given in this paper were obtained by putting the $p_i$ equal to the average frequencies observed in proteins (from Dayhoff & Hunt, 1972, combining aspartate with asparagine and glutamate with glutamine). For convenience I shall refer to this as the "natural" set of $p_i$ values. The dependence of $E(DT)$ on the set of $p_i$ is negligible in comparison with the inherent uncertainty in deducing sequence information from composition data: for example, values of $E(DT)$ calculated with $p_i = 0{\cdot}0556$ for all $i$ differ by about 3·5% from those calculated with the "natural" set of $p_i$, which are plotted in Fig. 1.

In practice one is unlikely to be concerned with the value of $DT$ for a pair of unrelated proteins; one is more likely to want to know what proportion of loci in the two sequences contain different types of residue. However, the addition of $(N-L)$ loci with identical residues in both sequences affects neither $DT$ nor $L$, the number of loci at which the sequences are independent. Consequently $L$ can still be estimated from $DT$ by means of Fig. 1. A further complication is that there is a small likelihood $\sum p_i^2$ that the same type of amino acid exists at any locus in both sequences even though they are independent. Therefore $L$ is somewhat greater than the number $M$ of loci at which the two sequences are different. To convert the estimate of $L$ obtained from Fig. 1 into an estimate of $M$ one must multiply it by $(1-\sum p_i^2)$, i.e. by 0·93 if "natural" $p_i$ values are used.

In practice it is very laborious to solve equation (3) for $L$ after replacing the expected value of $DT$ with the observed value. I have therefore derived an expression that allows approximate $L$ values to be calculated directly:

$$L \simeq \frac{DT + 0.114DT^2 + 0.01359DT^3}{1 + 0.0531DT + 0.0001694DT^2}. \quad (4)$$

For the "natural" set of $p_i$ values, and with $DT$ in the range 0–55 (corresponding to an $L$ range of 0–600), equation (4) provides solutions to equation (3) correct to within $\pm 2$.

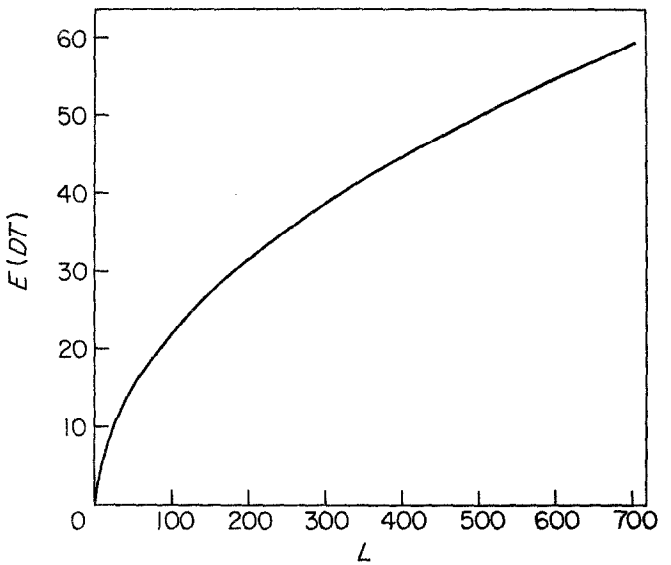


FIG. 1. Expected value of the difference total. Values of E($DT$) are calculated from equation (3) assuming the point probabilities $p_i$ to be equal to the average frequencies for the amino acids observed in proteins (Dayhoff & Hunt, 1972).

## 3. Examples

The calculation of $DT$ is illustrated in Table 1 with data of Yates & Planqué (1975) for the iron protein components of nitrogenases from two organisms. The calculation of $S\Delta n$, defined as

$$\tfrac{1}{2}\sum_i (n_{iA}-n_{iB})^2,$$

which is an estimator of $M$ (Cornish-Bowden, 1977), is also included for comparison. Reference to Fig. 1 or equation (4) shows that the value of $DT = 38{\cdot}5$ corresponds to a value of $L \simeq 296$, which may be multiplied by 0·93 to convert it into an estimate $M \simeq 275$ of the number of differences between the two sequences. This is in excellent agreement with the value of $S\Delta n = 280 \simeq M$ given by the alternative analysis, and thus the calculations agree in estimating sequence identity of about 55%.

A more extensive comparison of the results from the two methods is shown in Table 2, which is based on data of Habig, Pabst & Jakoby (1974, 1976) for four isoenzymes of glutamine S-transferase from rat liver. Although the six comparisons encompass a wide range of $M$ estimates there is good agreement between the two methods in all cases. It is clear that isoenzymes A and C are very similar, with estimated sequence identity of about 95%, whereas all

TABLE 1

*Calculation of* DT *and* SΔn *for the iron protein components of nitrogenases from two sources*†

| *i* | Amino acid | $n_{iA}$‡ | $n_{iB}$§ | $\lvert n_{iA}-n_{iB}\rvert$ | $(n_{iA}-n_{iB})^2$ |
|---|---|---|---|---|---|
| 1 | Aspartate + asparagine | 58 | 60 | 2 | 4 |
| 2 | Threonine | 24 | 36 | 12 | 144 |
| 3 | Serine | 22 | 24 | 2 | 4 |
| 4 | Glutamate+glutamine | 82 | 84 | 2 | 4 |
| 5 | Proline | 21 | 18 | 3 | 9 |
| 6 | Glycine | 57 | 60 | 3 | 9 |
| 7 | Alanine | 64 | 60 | 4 | 16 |
| 8 | Valine | 56 | 42 | 14 | 196 |
| 9 | Methionine | 29 | 36 | 7 | 49 |
| 10 | Isoleucine | 42 | 48 | 6 | 36 |
| 11 | Leucine | 47 | 42 | 5 | 25 |
| 12 | Tyrosine | 17 | 18 | 1 | 1 |
| 13 | Phenylalanine | 14 | 12 | 2 | 4 |
| 14 | Histidine | 8 | 6 | 2 | 4 |
| 15 | Lysine | 38 | 36 | 2 | 4 |
| 16 | Arginine | 29 | 24 | 5 | 25 |
| 17 | Cysteine | 13 | 18 | 5 | 25 |
| 18 | Tryptophan | 0 | 0 | 0 | 0 |
| | Totals: | 621‖ | 624‖ | 77†† | 559‡‡ |

† Data of Yates & Planqué (1975).
‡ Iron protein from nitrogenase of *Azotobacter chroococcum*.
§ Iron protein from nitrogenase of *Klebsiella pneumoniae*.
‖ The slight difference in *N* values has been ignored.
†† Hence $DT = 77/2 = 38{\cdot}5$ [in agreement with the value of $DI = 6{\cdot}1$, i.e. $38{\cdot}5 \times 100/621$, given by Yates & Planqué (1975)].
‡‡ Hence $S\Delta n = 559/2 = 279{\cdot}5$.

TABLE 2

*Estimated numbers of sequence differences between isoenzymes of glutathione S-transferase*

| | AA | C | B |
|---|---|---|---|
| A | 437 (451) | 15 (16†) | 275 (293) |
| B | 179 (143†) | 221 (261) | |
| C | 367 (394) | | |

Data for isoenzymes A, B and C were taken from Habig *et al.* (1974), data for isoenzyme AA from Habig *et al.* (1976). Small differences in the lengths of the proteins (385, 392, 381 and 398 respectively) were ignored in the calculations. For each comparison the number of sequence differences estimated from *DT* by means of Fig. 1 is shown on the left, and the number obtained according to Cornish-Bowden (1977) is shown following it in parentheses.

† Values less than $0{\cdot}42N$: these indicate compositional similarity significantly greater than expected for unrelated sequences.

other pairs show much less similarity, though isoenzymes B and AA have compositions that are (just) significantly more similar than expected for unrelated proteins.

Several other comparisons confirmed the agreement between the two methods, whether they were applied to small and closely similar proteins, such as the ferredoxins from *Leucaena glauca* (Benson & Yasunobu, 1969) and *Sambucus racemosa* L. (Altosaar, Bohm & Taylor, 1977), or to large and compositionally dissimilar proteins, such as the molybdenum–iron protein components of the nitrogenases from *Azotobacter chroococcum* and *Clostridium pasteurianum* (Yates & Planqué, 1975).

## 4. Discussion

The excellent agreement between the results obtained with $DT$ and $S\Delta n$ suggests that neither index is overwhelmingly better than the other. By extension, it follows that the index of Metzger *et al.* (1968) can be interpreted to provide essentially the same information as those of Harris *et al.* (1969) and Marchalonis & Weltman (1971). It is therefore largely a matter of individual preference which is used, though there are significant advantages in both approaches. The main advantage of $S\Delta n$ is that it yields an unbiased estimate of the number of sequence differences directly, without further calculation and without dependence on the set of $p_i$ used. If $DT$ is used it must be processed further to convert it into an estimate of the number of sequence differences, but it does directly provide an exact lower limit for this number, information that is not simply available from $S\Delta n$. It is obvious that $DT$ cannot be greater than the number of sequence differences, whereas $S\Delta n$ can be much greater—indeed, it can exceed $N$, as it does, for example, in the comparison between isoenzymes A and AA in Table 2.

Unfortunately the absolute differences used in the definition of $DT$ make it impracticable to derive an expression for its variance. Nonetheless, the close agreement in practice between the results with $DT$ and $S\Delta n$ suggests that the approximate significance of $DT$ may be tested by adapting the significance test for $S\Delta n$ (Cornish-Bowden, 1977). If it is true that the two indexes provide essentially the same information, then $DT$ can be taken to be significantly smaller than the value expected for unrelated sequences if it leads to an estimate of fewer than $0{\cdot}42N$ sequence differences. This test can be extended in an obvious way to the difference index $DI$ of Metzger *et al.* (1968), provided that it refers to a comparison between proteins of equal or near-equal length.

In my earlier paper (Cornish-Bowden, 1977) I commented that it would be rash to attempt to construct a phylogenetic tree from composition data.

Shortly after this was published, however, Black & Harkins (1977) presented results that suggested that I had been unduly pessimistic. Their success in relating the amino acid compositions of cytochromes *c* to their phylogenetic relationships, and in constructing a phylogenetic tree from measurements of the index of Harris *et al.* (1969) for pyruvate kinases, argues that the limit in extracting information from composition data has by no means been reached. The results of Black & Harkins (1977) are also of interest in that they provide empirical support for the view that the various composition indexes are of equal value in predicting significant sequence similarities between proteins.

## REFERENCES

ALTOSAAR, I., BOHM, B. A. & TAYLOR, I. E. P. (1977). *Canad. J. Biochem.* **55,** 159.
BENSON, A. M. & YASUNOBU, K. T. (1969). *J. biol. Chem.* **244,** 955.
BLACK, J. A. & HARKINS, R. N. (1977). *J. theor. Biol.* **66,** 281.
CORNISH-BOWDEN, A. (1977). *J. theor. Biol.* **65,** 735.
DAYHOFF, M. O. & HUNT, L. T. (1972). In *Atlas of Protein Sequence and Structure* (M. O. Dayhoff, ed.), vol. 5, p. D-355. Silver Spring: National Biomedical Research Foundation.
HABIG, W. H., PABST, M. J. & JAKOBY, W. B. (1974). *J. biol. Chem.* **249,** 7130.
HABIG, W. H., PABST, M. J. & JAKOBY, W. B. (1976). *Arch. Biochem. Biophys.* **175,** 710.
HARRIS, C. E., KOBES, R. D., TELLER, D. C. & RUTTER, W. J. (1969). *Biochemistry,* **8,** 2442.
MARCHALONIS, J. J. & WELTMAN, J. K. (1971). *Comp. Biochem. Physiol.* **38B,** 609.
METZGER, H., SHAPIRO, M. B., MOSIMANN, J. E. & VINTON, J. E. (1968). *Nature, Lond.* **219,** 1166.
YATES, M. G. & PLANQUÉ, K. (1975). *Eur. J. Biochem.* **60,** 467.