J. theor. Biol. (1977) **65**, 735–742

Assessment of Protein Sequence Identity from Amino Acid Composition Data

ATHEL CORNISH-BOWDEN

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

(*Received* 16 *June* 1976, *and in revised form* 27 *September* 1976)

A new index is proposed for assessing the extent of composition divergence between two proteins of equal length. It is defined as half the sum of squares of the differences between the numbers of residues of each type in the two proteins. It is an unbiased estimator of the number of differences between the two sequences, with a coefficient of variation of about 0·4. For unrelated proteins of length N the index is expected to exceed 0·42 N in about 95% of comparisons. The index can also be defined for pairs of proteins of which one is about double the length of the other. Recent data for glucokinase and hexokinase type II, both from rats, are used to illustrate the analysis proposed, and suggest that the two sequences are about 85% identical. Of other indexes currently in use, the one proposed by Marchalonis & Weltman (1971) appears to be the most easily interpretable and is simply related to the one proposed in this paper.

1. Introduction

The amino acid compositions of proteins can be measured far more easily and quickly than their sequences, and consequently there have been many attempts to deduce sequence information from composition data. Several indexes have been proposed for assessing the extent and significance of the differences in composition between pairs of proteins (Metzger, Shapiro, Mosimann & Vinton, 1968; Harris, Kobes, Teller & Rutter, 1969; Marchalonis & Weltman, 1971; Harris & Teller, 1973; Dedman, Gracy & Harris, 1974). These indexes, particularly those of Metzger *et al.* (1968) and of Marchalonis & Weltman (1971), have been determined for many pairs of proteins, and have been used to deduce or exclude possible ancestral relationships between them. But the conclusions have often been expressed in such tentative and unconvincing terms that the purpose of the exercise has not been evident. The reason for this lack of conviction is that almost nothing has been known *a priori* about what relationship, if any, exists between each index and the extent of sequence identity between two

proteins, or about the reliability of any information deduced. Instead, it has been necessary to rely on comparisons between index and sequence for proteins of known sequence. But the scatter of points tends to obscure any relationships that may exist, and in any case such comparisons assume that results for proteins of known sequence apply generally. In this paper I shall show that for two proteins of equal length an index similar to that of Marchalonis & Weltman (1971) provides a direct and unbiased estimate of the number of differences between the two sequences. The precision of this estimate can be calculated *a priori* and so the reliability of deductions about ancestral relationships can be assessed. Straightforward interpretation of the indexes of Harris *et al.* (1969) and of Marchalonis & Weltman (1971) is now possible, because both can readily be converted into the new index with very little calculation.

2. Theory

(A) DEFINITIONS

The index of Marchalonis & Weltman (1971) is denoted by the symbol $S\Delta Q$, and is defined as follows:

$$S\Delta Q = 10^4 \sum_i \left(\frac{n_{iA}}{N_A} - \frac{n_{iB}}{N_B}\right)^2 \qquad (1)$$

where n_{iA} and n_{iB} are the number of residues of the ith type of amino acid in proteins A and B respectively, and

$$N_A = \sum_i n_{iA} \quad \text{and} \quad N_B = \sum_i n_{iB}$$

are the lengths of the two proteins. The summation is ideally carried out over all 20 amino acids. However, protein compositions often do not distinguish between aspartate and asparagine or between glutamate and glutamine, and data for tryptophan and cysteine are often missing; so in practice the summation is usually carried out over 16–18 distinct amino acids.

Unambiguous comparison of two sequences is possible only if they are of the same length, and so I shall initially assume $N_A = N_B = N$. Moreover, as the factor $10^4/N$ in the definition of $S\Delta Q$ complicates the algebra to no purpose, I shall consider a simpler index, $S\Delta n$, defined as follows:

$$S\Delta n = \tfrac{1}{2}\sum (n_{iA} - n_{iB})^2 = \tfrac{1}{2}\sum n_{iA}^2 - \sum n_{iA} n_{iB} + \tfrac{1}{2}\sum n_{iB}^2. \qquad (2$$

This index is an unbiased estimator of the number of loci at which the two sequences are different, M.

Consider two independent sequences A and B, both of the same length L, with the same probability p_i of finding the ith amino acid at any locus in either sequence. Then all of the n_i are binomially distributed and their expected values can be found by standard procedures (Johnson & Kotz, 1969) with the following results:

$$E(n_{iA}) = E(n_{iB}) = Lp_i \tag{3}$$

$$E(n_{iA}^2) = E(n_{iB}^2) = L(L-1)p_i^2 + Lp_i. \tag{4}$$

Since the two sequences are by definition independent, the expected value of $n_{iA}n_{iB}$ is simply

$$E(n_{iA}n_{iB}) = E(n_{iA})E(n_{iB}) = L^2p_i^2. \tag{5}$$

Combining equations (4) and (5) with equation (2) gives

$$E(S\Delta n) = L(1-\sum p_i^2). \tag{6}$$

Consider now the number M of loci at which the amino acid in A is different from that in B. At any locus the likelihood of difference is $(1-\sum p_i^2)$, and the loci are independent, so M is binomially distributed with expected value

$$E(M) = L(1-\sum p_i^2) = E(S\Delta n). \tag{7}$$

Thus $S\Delta n$ is an unbiased estimator of M.

(B) PRECISION OF $S\Delta n$

Although lack of bias is a valuable property in an estimator, it is also useful to have some measure of its precision. In the case of $S\Delta n$ a suitable measure is its variance from M, i.e.

$$\sigma^2(S\Delta n) = E[(S\Delta n - M)^2]. \tag{8}$$

Evaluation of this expression is rather lengthy (see the Appendix), but the result is simple:

$$\sigma^2(S\Delta n) = L^2[2(\sum p_i^2)^2 - 4\sum p_i^3 + 2\sum p_i^2] + L[2\sum p_i^3 - 2(\sum p_i^2)^2]. \tag{9}$$

For any reasonable set of p_i values that might be assumed, the coefficient of L in this equation is negligible compared with that of L^2 [because $\sum p_i^2 \gg (\sum p_i^2)^2 \simeq \sum p_i^3$]. Moreover, equation (7) shows that L can be replaced, to a good approximation, with $M/(1-\sum p_i^2)$, and so the standard deviation of $S\Delta n$ is given by

$$\sigma(S\Delta n) \simeq M[2(\sum p_i^2)^2 - 4\sum p_i^3 + 2\sum p_i^2]^{\frac{1}{2}}/(1-\sum p_i^2). \tag{10}$$

If the p_i are assumed to be equal to the average frequencies observed for the amino acids in proteins (from Dayhoff & Hunt, 1972, combining aspartate with asparagine and glutamate with glutamine, and including tryptophan and cysteine), the numerical value of this expression is as follows:

$$\sigma(S\Delta n) \simeq 0{\cdot}38M \simeq 0{\cdot}35L. \qquad (11)$$

Fortunately this numerical result is only weakly dependent on the p_i values assumed, and so it is of little importance that the values appropriate for a particular comparison are unlikely to be known.

(c) Effect of identities between the two sequences

Thus far I have only considered two independent sequences with no identities apart from those that arise by chance. But for the index to be useful it is necessary to know how the results are affected by introducing an unknown number of loci at which the residues are identical in the two sequences, not by chance but by ancestral relationship. It is obvious that this change has no effect on M, the number of loci at which the two sequences are different. It also has no effect on $S\Delta n$, because any increase in $n_{i\mathrm{A}}$ for any i is matched by an equal increase in $n_{i\mathrm{B}}$, so $(n_{i\mathrm{A}}-n_{i\mathrm{B}})^2$ is unchanged. Thus the results obtained above apply without change, i.e. $S\Delta n$ is still an unbiased estimator of M and its standard deviation is still about $0{\cdot}38M$.

(d) Implications for the index of Marchalonis & Weltman (1971)

For proteins of equal length the index of Marchalonis & Weltman is proportional to $S\Delta n$, i.e. $S\Delta Q = 2\times 10^4 S\Delta n/N^2$. So the interpretation of $S\Delta Q$ follows simply from the above analysis: it is an unbiased estimator of $2\times 10^4 M/N^2$, with a standard deviation of about $7{\cdot}6\times 10^3 M/N^2$.

(e) Effect of using incomplete data

The effect of using data that do not distinguish between aspartate and asparagine or between glutamate and glutamine is to modify the definition of sequence identity: whatever criterion of identity is used for calculating $S\Delta n$ must also be applied when interpreting it as an estimate of M. This property can be used intentionally if an estimate of sequence similarity is required rather than of sequence identity; in this case it is advisable to recalculate the coefficients in equation (11) with the values of p_i appropriate to the definition of similarity used.

The effect of using data that do not include values for tryptophan is likely to be slight enough to be ignored, because most proteins contain very little tryptophan and the resulting error is likely to be small compared with the statistical error inherent in estimating M from $S\Delta n$. Omission of cysteine is only slightly more serious, because cysteine is also one of the least abundant amino acids.

(F) APPLICATION TO PROTEINS OF UNEQUAL LENGTH

For proteins of equal length, $S\Delta n$ is a more convenient index than $S\Delta Q$ because it provides a direct estimate of the number of sequence differences without any arithmetic manipulation. But it has the disadvantage that it is undefined for proteins of unequal length. It is natural to inquire, therefore, whether an index exists for proteins of unequal length that has a correspondingly simple meaning. First, one must recognize that the difficulty is not simply in comparing the compositions, but would also exist if the sequences were known, because if N_A is not equal to N_B there is no unambiguous way of lining up the sequences for comparison. But if N_A and N_B differ by little enough for such a comparison to be meaningful, one can define $S\Delta n$ more generally as follows:

$$S\Delta n = \tfrac{1}{2}N_A^2 \sum \left(\frac{n_{iA}}{N_A} - \frac{n_{iB}}{N_B} \right)^2 \tag{12}$$

where A is taken (by definition) to be the shorter of the two sequences, i.e. $N_A < N_B$. This has the same meaning as before, i.e. an unbiased estimator of M, but with the added uncertainty that M is no longer uniquely defined.

If N_A and N_B are not approximately equal, but instead one is approximately double the other, say $N_B \simeq 2N_A$, one may wish to investigate the hypothesis that the two sequences have a common ancestor but that gene duplication has occurred in the evolution of the longer sequence. In this case $S\Delta n$ can still be defined by equation (12) and is now an estimator of the M value for comparing sequence A with half of sequence B.

(G) TESTING THE NULL HYPOTHESIS

The null hypothesis is that sequences A and B contain no identities apart from those few that arise by chance, i.e. $L = N \simeq M/(1-\sum p_i^2)$. This can be tested by determining whether $S\Delta n$ is less than $0.42N$, i.e. less than $N - 1{\cdot}65\sigma(S\Delta n)$, from equation (11). This would be a one-tailed test at the 95% confidence level if $S\Delta n$ were strictly normally distributed. In fact it is a

somewhat more stringent test because the distribution of $S\Delta n$ is short-tailed and terminates at zero. A one-tailed test is appropriate because very large values of $S\Delta n$ could not reasonably be interpreted as evidence of ancestral relationship.

3. Results and Discussion

(A) GLUCOKINASE AND HEXOKINASE TYPE II

Data for glucokinase (Holroyde *et al.*, 1976) and hexokinase type II (Holroyde & Trayer, 1976) from rats provide a convenient example for illustrating the method of analysis proposed in this paper. The two compositions were obtained in the same laboratory under similar conditions, and the estimated molecular weight of glucokinase (48,000) is half that of hexokinase type II (96,000). It is reasonable to consider the hypothesis that the two enzymes have arisen from a common ancestor, with gene duplication in the case of hexokinase type II. The calculation of $S\Delta n$ is illustrated in Table 1, and gives a value of 69·2, or about 15% of the length of glucokinase.

TABLE 1.

Calculation of SΔn *for glucokinase and hexokinase type II*

Amino acid	A: Glucokinase†		B: Hexokinase II‡		$10^4\left(\frac{n_{iA}}{N_A}-\frac{n_{iB}}{N_B}\right)^2$
	n_{iA}	$100n_{iA}/N_A$	n_{iB}	$100n_{iB}/N_B$	
Aspartate + asparagine	46	10·41	92	10·37	0·0016
Threonine	24	5·43	54	6·09	0·4356
Serine	31	7·01	52	5·86	1·3225
Glutamate + glutamine	59	13·35	104	11·72	2·6569
Proline	14	3·17	26	2·93	0·0576
Glycine	40	9·05	86	9·70	0·4225
Alanine	30	6·79	66	7·44	0·4225
Valine	32	7·24	66	7·44	0·0016
Methionine	18	4·07	30	3·38	0·4761
Isoleucine	17	3·85	36	4·06	0·0441
Leucine	41	9·28	92	10·37	1·1881
Tyrosine	8	1·81	16	1·80	0·0001
Phenylalanine	18	4·07	36	4·06	0·0001
Histidine	11	2·49	23	2·59	0·0100
Lysine	25	5·66	52	5·86	0·0400
Arginine	28	6·33	56	6·31	0·0004
Totals:	442	100·01	887	99·98	7·0797§

†From rat liver (Holroyde *et al.*, 1976).
‡From rat skeletal muscle (Holroyde & Trayer, 1976).
§This total gives $S\Delta Q = 7{\cdot}10$, from which $S\Delta n = 69{\cdot}2$ is obtained by multiplying by $N_A^2/20{,}000$.

This is much less than $0{\cdot}42N$, i.e. 186, and allows the null hypothesis to be rejected, i.e. it indicates significant identity between the two sequences.

This example illustrates both the strength and the weakness of $S\Delta n$ as a measure of composition difference. On the one hand the interpretation is simple and direct, as it leads immediately to an estimate of the number of differences between the two sequences. But it is an imprecise estimate, as the observed value of 69 would still be within $1{\cdot}65$ standard deviations even if the true value of M were as high as 185. This imprecision is inherent in the data and, as Fitch (1973) has remarked, it would be unreasonable to expect any method based on composition data alone to yield precise information about sequence. Certainly it would be rash to attempt to construct a phylogenetic tree from composition data, whatever index is used.

(B) OTHER COMPOSITION INDEXES

The results given in this paper have implications about the other indexes that have been proposed. For example, the standard deviation of $S\Delta Q$, the index of Marchalonis & Weltman (1971), is proportional to M/N^2 and so any test of the null hypothesis must take account of the dependence on the lengths of the proteins. It is misleading to regard values of $S\Delta Q$ less than 100 as "significant" simply because 98% of the 820 pairs of unrelated proteins considered by Marchalonis & Weltman (1971) gave values greater than 100. The lengths of the proteins considered were uniform neither between pairs nor within pairs and so the distribution curve cannot readily be interpreted. Similar difficulties must apply to the index of Harris *et al.* (1969), which is equivalent to $0{\cdot}01(S\Delta Q)^{\frac{1}{2}}$. They almost certainly apply also to the index of Metzger *et al.* (1968), which is, however, much more difficult to analyse *a priori* because it is defined in terms of absolute differences rather than squared differences.

REFERENCES

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.

DEDMAN, J. R., GRACY, R. W. & HARRIS, B. G. (1974). *Comp. Biochem. Physiol.* **49B,** 715.

FITCH, W. M. (1973). *A. Rev. Genet.* **7,** 343.

HARRIS, C. E., KOBES, R. D., TELLER, D. C. & RUTTER, W. J. (1969). *Biochemistry* **8,** 2442.

HARRIS, C. E. & TELLER, D. C. (1973). *J. theor. Biol.* **38,** 347.

HOLROYDE, M. J., ALLEN, M. B., STORER, A. C., WARSY, A. S., CHESHER, J. M. E., TRAYER, I. P., CORNISH-BOWDEN, A. & WALKER, D. G. (1976). *Biochem. J.* **153,** 363.

HOLROYDE, M. J. & TRAYER, I. P. (1976). *FEBS Lett.* **62,** 215.

JOHNSON, N. L. & KOTZ, S. (1969). *Discrete Distributions*, pp. 51, 284. Boston: Houghton Mifflin.

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.

APPENDIX

Outline Derivation of the Variance of $S\Delta n$†

The variance of $S\Delta n$ is defined as follows:

$$\sigma^2(S\Delta n) = E[(S\Delta n - M)^2] = E[(S\Delta n)^2] - 2E(M\ S\Delta n) + E(M^2) \quad \text{(A1)}$$

and may be evaluated by deriving expressions for the three terms on the right-hand side.

$E[(S\Delta n)^2]$ may be evaluated by defining

$$S\Delta n = \sum_i q_i, \qquad (S\Delta n)^2 = \sum_i \sum_j q_i q_j,$$

where $q_i = \frac{1}{2}(n_{iA} - n_{iB})^2$, and expanding the expression $4q_iq_j = (n_{iA} - n_{iB})^2 \times (n_{jA} - n_{jB})^2$. As A and B are assumed to have the same probability distribution, $E(n_{iB}^2 n_{jB}^2) = E(n_{iA}^2 n_{jA}^2)$, etc.; and, as they are assumed to be independent, $E(n_{iA}^2 n_{jA} n_{jB}) = E(n_{iA}^2 n_{jA})E(n_{jB}) = E(n_{iA} n_{jA})E(n_{jA})$, etc. Each of the terms $E(n_{iA}^2 n_{jA}^2)$ etc. can be evaluated by standard procedures (Johnson & Kotz, 1969), deriving results for $j = i$ separately from those for $j \neq i$. Eventually one obtains the following expression:

$$E[(S\Delta n)^2] = L^2[3(\sum p_i^2)^2 - 4\sum p_i^3 + 1] + L[-3(\sum p_i^2)^2 + 4\sum p_i^3 - \sum p_i^2]. \quad \text{(A2)}$$

$E(M\ S\Delta n)$ may be evaluated by defining $M = \sum m_i = \frac{1}{2}\sum (m_{iA} + m_{iB})$, where m_{iA} is the number of loci with the ith amino acid in sequence A and a *different* amino acid in sequence B, and m_{iB} is defined analogously. Then a derivation similar to the one for $E[(S\Delta n)^2]$ leads to the following result:

$$E(M\ S\Delta n) = L^2[(\sum p_i^2)^2 - 2\sum p_i^2 + 1] + L[-(\sum p_i^2)^2 + \sum p_i^3]. \quad \text{(A3)}$$

$E(M^2)$ may be evaluated by consideration of the fact that at any locus the probability is $(1 - \sum p_i^2)$ that there are different amino acids in the two sequences. So M is binomially distributed and M^2 has the following expected value:

$$E(M^2) = L^2[(\sum p_i^2)^2 - 2\sum p_i^2 + 1] + L[-(\sum p_i^2)^2 + \sum p_i^2]. \quad \text{(A4)}$$

Finally, equations (A2)–(A4) may be substituted into equation (A1) to give the following expression for the variance of $S\Delta n$:

$$\sigma^2(S\Delta n) = L^2[2(\sum p_i^2)^2 - 4\sum p_i^3 + 2\sum p_i^2] + L[2\sum p_i^3 - 2(\sum p_i^2)^2] \quad \text{(A5)}$$

which is equation (9) of the main paper.

†A fuller derivation is available from the author on request, and will be included with reprints.