ANALYTICAL BIOCHEMISTRY 105, 233–238 (1980)

Critical Values for Testing the Significance of Amino Acid Composition Indexes

ATHEL CORNISH-BOWDEN

Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, England

Received September 25, 1979

Tables are provided for testing the significance of three indexes that are commonly used for expressing the amount of difference between the amino acid compositions of proteins. Two tests are suggested: a "strong" test, which has a negligible danger of indicating relatedness incorrectly, but may fail to detect genuine relatedness; and a "weak" test, which provides an excellent chance of detecting genuine relatedness but may indicate it incorrectly in about 10% of comparisons between unrelated proteins.

The number of amino acid compositions of proteins that have been published is very large (1,2) and continues to grow much faster than the number of published sequences, despite improvements in sequencing techniques. Various indexes are in common use for concisely expressing the amount of difference between compositions (3–6), but it is rare for any but the most trivial conclusions to be drawn from their values, in part because of the absence of a theoretical framework for assessing them. In previous papers I have developed such a theory (7,8) and tested its performance by applying it to proteins of known sequence (9). These studies suggest that useful answers are now possible to the following questions:

(i) Is there a test of composition data that is nearly certain to give a negative result when applied to proteins with unrelated sequences?

(ii) Is there a test of composition data that is nearly certain to give a positive result when applied to proteins with a substantial degree of sequence similarity?

Although past studies have implied that negative answers must be given to both questions, theoretical analysis suggests that this pessimistic conclusion is mistaken and that it has been arrived at by failure to recognize that the significance of the commonly used indexes is strongly dependent on the size of the proteins compared. Allowance for this size dependence permits the formulation of a "strong" test that nearly always gives a negative result when applied to unrelated proteins, and a "weak" test that nearly always gives a positive result when applied to related proteins. The "strong" test has been described previously (9) and the "weak" test is developed in this paper; in addition tables are given that allow corresponding tests to be made of all of the commonly used indexes. Together, the two tests allow a high degree of confidence in classifying any pair of proteins as related, unrelated, or doubtfully related.

DEFINITIONS

Let the number of residues of the ith type of amino acid in protein A be $n_{i\mathrm{A}}$, and let the total number of residues in A be N_{A}. Then the mole fraction of the ith type of amino acid in A is $X_{i\mathrm{A}} = n_{i\mathrm{A}}/N_{\mathrm{A}}$. Let $n_{i\mathrm{B}}$, N_{B}, and $X_{i\mathrm{B}}$ be the corresponding numbers for a second protein B. With these symbols the commonly used indexes of compositional difference are defined as follows:

0003-2697/80/100233-06$02.00/0

$$DI = 50 \sum |X_{iA} - X_{iB}|,$$

$$D = [\sum (X_{iA} - X_{iB})^2]^{1/2},$$

$$S\Delta Q = 10^4 \sum (X_{iA} - X_{iB})^2.$$

In each case the summation is made over as many types of residue as are distinguished by the measurements (usually 18). DI is the "difference index" (3), D is the "compositional divergence" (4,6), and $S\Delta Q$ is the index of Marchalonis and Weltman (5). For theoretical analysis of the properties of these indexes it is convenient to consider the case in which $N_A = N_B = N$, because only then does the extent of sequence difference have an unambiguous meaning. It is also convenient to define a further index $S\Delta n$ (7) for analyzing the properties of D and $S\Delta Q$,

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

and an index DT, the "difference total" (8), for analyzing the properties of DI

$$DT = \tfrac{1}{2} \sum |n_{iA} - n_{iB}|.$$

When the condition $N_A = N_B = N$ holds, these additional indexes are simply and exactly related to the first three:

$$S\Delta n = 0.00005N^2 \cdot S\Delta Q = 0.5N^2D^2,$$

$$DT = 0.01N \cdot DI.$$

It is then a simple matter to calculate critical values for DI, D, and $S\Delta Q$ from those calculated for DT and $S\Delta n$. If N_A and N_B are unequal there are no exact relationships of these kinds, but this does not present a major problem in practice, as discussed below.

RESULTS AND DISCUSSION

Theory predicts that $S\Delta n$ should exceed $0.42N$ in 95% of comparisons between pairs of proteins with N residues each that have the same statistical properties but are otherwise unrelated (7). In practice the frequency with which this test, which for convenience I shall refer to as the "strong" test, gives an incorrect indication of relatedness is much less than the calculated 5% (9). This is primarily because unrelated proteins cannot reasonably be assumed to have the same statistical properties and if this assumption is not made then $S\Delta n$ is expected to overestimate the amount of sequence difference (9). This suggests that one may with considerable safety use a less demanding test: for example, one might take $S\Delta n$ as significant if it predicted *any* sequence similarity in excess of the 7% of identities expected from chance alone. In this test, which I shall refer to as the "weak" test, $S\Delta n$ would be taken as significant if it were less than $0.93N$.

To examine the validity of the "weak" test I have reexamined the 163 comparisons that were listed in Table 2 of Ref. (9), which were used before for checking the validity of the "strong" test. The 163 pairs of proteins were selected in the following way. Eighty-three proteins of known sequence were taken from the Atlas of Protein Sequence and Structure (10), with number of residues ranging from 24 to 374, such that no two proteins were listed in the Atlas under the same name. The 83 proteins were then arranged in order of increasing length and each was compared with its nearest and next-nearest neighbors in the order, so that there were 163 comparisons altogether. As the sequences were known in all cases it was possible to compare the results given by the "strong" test of the compositions with those expected from the sequences. In fact, 19 of the 163 pairs showed significant sequence similarity in a 99.9% test of the sequences, and of these 19 there were 7 pairs with $S\Delta n$ less than $0.42N$, but 11 of the other 12 gave $S\Delta n$ less than $0.93N$. Thus although the "strong" test identified only 7 of the 19 indisputably related pairs, the "weak" test identified 18 out of 19. In addition, however, there were 20 pairs for which similarity had not been detected in the sequences but which gave $S\Delta n$ less than $0.93N$ (though larger than $0.42N$). These 20 questionable results are listed in Table 1,

TABLE 1

PAIRS OF PROTEINS WITH SIGNIFICANTLY SIMILAR COMPOSITIONS[a]

Protein A	Protein B	N_A	N_B	$S\Delta n$[b]	M[c]
Bombinin	Melittin	24	26	21.9	21
Melittin	Secretin	26	27	24.0	25
Glucagon	Calcitonin	29	32	21.8	29
Corticotropin	Gastric inhibitory polypeptide	39	43	34.6	39
Rubredoxin	Ferredoxin	52	55	40.8	47
Rubredoxin	Basic trypsin inhibitor	52	58	44.0	53
Basic trypsin inhibitor	Neurotoxin β	58	61	51.8	54
Erabutoxin A	α-Bungarotoxin	62	74	44.4	64[d]
Cytochrome c_{551}	Parathyroid hormone	82	84	70.9	78
Thioredoxin	Cytochrome c_2	108	112	85.6	96[e]
Thioredoxin	Hemerythrin	108	113	88.3	102
Adrenodoxin	Lactalbumin	118	123	77.3	113
Nerve growth factor	Ribonuclease	118	124	90.0	109
Azurin	Avidin	128	128	116.0	116
Histone IIB1	Histone III	129	135	89.0	119
Hemoglobin γ-chain	Aspartate transcarbamylase, R-chain	146	152	136.0	138
Nuclease	Myoglobin	149	153	119.6	140
Aspartate transcarbamylase R-chain	Coat protein	152	158	137.0	144
Lactogen	Prolactin	190	198	86.0	174[f]
Elastase	Chymotrypsinogen A	240	245	193.3	220[g]

[a] The table shows results from a study of 163 comparisons between 83 proteins. A full list of these proteins, with species and references, is given in Ref. (9). The present table includes only comparisons for which the sequences appeared unrelated in a 99.9% significance test whereas the compositions indicated less than 93% but more than 42% sequence difference, i.e., the compositions failed the "strong" test suggested previously (7) but satisfied the "weak" test suggested in the present paper.

[b] Corrected for differences between N_A and N_B as described in Ref. (9). In all cases the corrections required were trivial (less than two).

[c] Number of residues in the longer sequence unmatched by identical residues (apart from possible differences in amide assignments) in the shorter sequence when the two were aligned in the best alignment that allowed no gaps in the longer sequence and no internal gaps in the shorter sequence.

[d] Decreased to 51 if aligned with internal gaps as on pp. 150–151 of Ref. (11).

[e] Significant in a 95% test.

[f] Decreased to 144 if aligned with internal gaps as on p. D-374 of Ref. (10).

[g] Decreased to 150 if aligned with internal gaps as on p. D-372 of Ref. (10).

from which it is apparent that at least four comparisons were between proteins with related sequences: three of these were not identified by the original sequence test because satisfactory alignment required internal gaps, which were not permitted; the fourth failed to show significant sequence similarity in the 99.9% test that was used, but would have been significant in a 95% test. Some of the remaining 16 comparisons may also have been between proteins more closely related than the sequence test suggested: for example, bombinin and melittin, which have only 24 and 26 residues, respectively, share the same *N*-terminal tetrapeptide but otherwise show little sequence similarity. Thus, out of about 140 comparisons between probably unrelated proteins the "weak" composition test indicated no similarity in 124 cases and incorrectly indicated similarity in only 16, whereas it detected the genuine similarity in 22 out of 23 cases (the exception

TABLE 2

CRITICAL VALUES OF DI, $S\Delta Q$, AND D

	"Strong" test[b]			"Weak" test[c]		
N[a]	DI	$S\Delta Q$	D	DI	$S\Delta Q$	D
10	32.5	840	0.290	57.9	1860	0.431
20	27.1	420	0.205	44.7	930	0.305
30	23.5	280	0.167	37.7	620	0.249
40	21.0	210	0.145	33.0	465	0.216
50	19.2	168	0.130	29.8	372	0.193
60	17.7	140	0.118	27.3	310	0.176
70	16.6	120	0.110	25.4	266	0.163
80	15.6	105	0.102	23.8	233	0.152
90	14.8	93.3	0.0966	22.4	207	0.144
100	14.1	84.0	0.0917	21.4	186	0.136
120	12.9	79.0	0.0837	19.6	155	0.124
140	12.0	60.0	0.0775	18.1	133	0.115
160	11.3	52.5	0.0725	17.0	116	0.108
180	10.7	46.7	0.0683	16.0	103	0.102
200	10.1	42.0	0.0648	15.2	93.0	0.0964
250	9.09	33.6	0.0580	13.6	74.4	0.0863
300	8.32	28.0	0.0529	12.5	62.0	0.0787
350	7.71	24.0	0.0490	11.5	53.1	0.0729
400	7.22	21.0	0.0458	10.8	46.5	0.0682
450	6.82	18.7	0.0432	10.2	41.3	0.0643
500	6.47	16.8	0.0410	9.66	37.2	0.0610
550	6.17	15.3	0.0391	9.22	33.8	0.0582
600	5.91	14.0	0.0374	8.82	31.0	0.0557
650	5.68	12.9	0.0359	8.48	28.6	0.0535
700	5.48	12.0	0.0346	8.17	26.6	0.0515

[a] N is defined as the larger of N_A and N_B, the numbers of residues in the two proteins compared.

[b] Observed values less than those tabulated indicate near certainty of relationship.

[c] Observed values less than those tabulated indicate that relationship is likely.

was the comparison between erabutoxin from sea snake and neurotoxin α from Cape cobra). So by using the "weak" test one can substantially improve the chance of detecting genuine relatedness while keeping spurious indications of relatedness at low frequency.

The two tests of $S\Delta n$ may be converted into tests of $S\Delta Q$ and D that are exactly equivalent to those for $S\Delta n$ if $N_A = N_B$, and approximately equivalent if N_A and N_B differ by less than about 18 (9). These tests are tabulated in Table 2 together with tests of DI calculated from the corresponding tests of DT. The latter are justified by the observation that in practice DT and $S\Delta n$ make closely similar predictions (8). If N_A and N_B are unequal it is prudent to decrease the likelihood of spurious positive results by defining N as the *larger* of N_A and N_B when testing DI, $S\Delta Q$, or D. For $S\Delta n$ it is possible to correct for differences in N_A and N_B (9), but as this correction is trivial if $|N_A - N_B|$ does not exceed about 18 the lack of a similar correction for the other indexes is not a major problem.

The use of Table 2 is illustrated in Table 3 by some data of Cerff and Chambers (13) for two glyceraldehyde-3-phosphate dehydrogenases from white mustard. The right-hand column of the table gives $S\Delta Q = 20.6$ (in agreement with the value of 21 calculated by the original authors), which is exactly equivalent to $D = 0.0454$. The two enzymes were estimated to have 366 and 362 residues, respectively, and so we take $N = 366$ as the larger of these in using Table 2: this shows that the critical values in the "strong" test are about 23 for $S\Delta Q$ and about 0.048 for D, these values being obtained by interpolation between the values tabulated for $N = 350$ and $N = 400$. The observed values just satisfy the "strong" test and so indicate with near certainty that the two unknown protein sequences are related. Table 3 also shows that $DI = 7.75$ for the same data. Reference to Table 2 shows that this value narrowly fails to satisfy the "strong" test, though it satisfies the "weak" test very easily. As DI and $S\Delta Q$ are not precisely interconvertible (unlike $S\Delta Q$ and D), one should not be surprised if, as in this example, they occasionally give results close to but on the opposite sides of the borderline for one or other of the tests.

An important difference between the tests shown in Table 2 and those suggested or implied by previous workers is that the new tests take account of the strong dependence

TABLE 3

COMPOSITIONS OF GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASES FROM WHITE MUSTARD[a]

Amino acid	$100X_A$	$100X_B$	$100\lvert X_A - X_B\rvert$	$10^4(X_A - X_B)^2$
Ala	9.4	8.8	0.6	0.36
Val	9.2	7.2	2.0	4.00
Leu	8.5	6.5	2.0	4.00
Ile	5.0	4.6	0.4	0.16
Pro	4.4	3.6	0.8	0.64
Phe	2.7	4.0	1.3	1.69
Trp	1.7	1.7	0.0	0.00
Met	1.3	2.1	0.8	0.64
Gly	9.2	9.6	0.4	0.16
Ser	7.2	7.4	0.2	0.04
Thr	5.5	6.1	0.6	0.36
Cys	2.4	1.9	0.5	0.25
Tyr	1.7	2.6	0.9	0.81
Asx	13.6	12.8	0.8	0.64
Glx	6.0	7.4	1.4	1.96
Lys	6.8	8.9	2.1	4.41
Arg	4.0	3.3	0.7	0.49
His	1.6	1.6	0.0	0.00
Total	100.2	100.1	15.5[b]	20.61[c]

[a] Data of Cerff and Chambers (13) for NADP-linked isoenzyme I (protein A) and NAD-specific enzyme (protein B).

[b] Hence DI = 15.5/2 = 7.75.

[c] This value gives $S\Delta Q$ directly, i.e., $S\Delta Q = 20.6$; $D = 0.01 \cdot S\Delta Q^{1/2} = 0.0454$.

of the expected values of the indexes on N. By contrast, critical values have previously been suggested without regard to N: 100 for $S\Delta Q$ (5) and 0.07 for D (6); although Metzger and co-workers (3) did not explicitly propose a critical value for DI their Fig. 1 implies a value of about 13. Comparison of these values with those in Table 2 shows that they are too small for small proteins and too large for large proteins. With the critical value of 100 for $S\Delta Q$, for example, one would have almost no chance of detecting a genuine relationship between proteins with fewer than about 85 residues each, and one should expect many spuriously positive results with proteins with more than about 180 residues each.

It is highly questionable whether it is useful to compare the compositions of proteins of unknown sequence and appreciably different size, because of the difficulty in defining the extent of sequence difference in an unambiguous way. [For simplicity I shall not here discuss the exceptional but potentially interesting case where N_B is approximately a small multiple of N_A (7,9).] Nonetheless, many such comparisons have been made, and it is appropriate to conclude by inquiring whether there is any appreciable danger in applying the tests given in Table 2 when N_A and N_B are very different. To answer this I calculated $S\Delta Q$ for all 3403 pairs of the 83 proteins referred to above and listed elsewhere (9). In applying the tests I took N as the larger of N_A and N_B. The "strong" test gave a positive result in only 16 cases: even if all of these were spurious this would be a failure rate of only 0.5%; in fact, however, 10 of the 16 showed unmistakeable evidence of relationship when the sequences were plotted against one another by the method of Gibbs and Mc-

Intyre (12). Similarly, the "weak" test gave a positive result in only 119 comparisons, or about 3.5%, and the relationship was genuine in at least 34 of these. Thus although it remains arguable whether there is much point in comparing the compositions of proteins of appreciably different sizes it seems clear that there is little danger of deducing a spurious relationship from such a comparison.

One result from this last study proved especially tantalizing: comparison of bovine posterior pituitary peptide ($N_A = 48$) with bovine neurophysin ($N_B = 97$) gave $S\Delta Q = 47.5$. Not only does this value satisfy the "strong" test very easily, as the critical value is 86.6, but the fact that N_B is almost exactly double N_A suggests that neurophysin may closely resemble a dimer of posterior pituitary peptide. Disappointingly, the sequences show no trace of this supposed relationship. Fortunately, however, this proved to be the only such case in 3403 comparisons and can hardly therefore be regarded as typical.

REFERENCES

1. Kirschenbaum, D. M. (1977) *Anal. Biochem.* **83,** 484–520.
2. Kirschenbaum, D. M. (1977) *Anal. Biochem.* **83,** 521–550.
3. Metzger, H., Shapiro, M. B., Mosimann, J. E., and Vinton, J. E. (1968) *Nature (London)* **219,** 1166–1168.
4. Harris, C. E., Kobes, R. D., Teller, D. C., and Rutter, W. J. (1969) *Biochemistry* **8,** 2442–2454.
5. Marchalonis, J. J., and Weltman, J. K. (1971) *Comp. Biochem. Physiol.* **38B,** 609–625.
6. Harris, C. E., and Teller, D. C. (1973) *J. Theor. Biol.* **38,** 347–362.
7. Cornish-Bowden, A. (1977) *J. Theor. Biol.* **65,** 735–742.
8. Cornish-Bowden, A. (1978) *J. Theor. Biol.* **74,** 155–161.
9. Cornish-Bowden, A. (1979) *J. Theor. Biol.* **76,** 369–386.
10. Dayhoff, M. O. (1972) Atlas of Protein Sequence and Structure, Vol. 5, Natl. Biomed. Res. Found., Silver Spring, Md.
11. Dayhoff, M. O. (1976) Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 2, Natl. Biomed. Res. Found., Silver Spring, Md.
12. Gibbs, A. J., and McIntyre, G. A. (1970) *Eur. J. Biochem.* **16,** 1–11.
13. Cerff, R., and Chambers, S. E. (1979) *J. Biol. Chem.* **254,** 6094–6098.