Spectrophotometric Titration with DTNB.[41] A typical titration is presented.

1. Equilibrate 3.00-ml aliquots of stock buffer (or of 6.40 *M* buffered guanidinium chloride) at 25.0° in the thermostatted sample and reference compartments of a spectrophotometer. Adjust A_{412} to zero (A_{buffer}, Fig. 1).
2. Add a 100-μl aliquot of stock buffer to the reference cuvette and stir the solution with a flat-ended glass stirring rod.
3. Add a 100-μl aliquot of DTNB solution to the sample cuvette with stirring. Record A_{412} of the DTNB (A_{DTNB}, Fig. 1).
4. Add a 100-μl aliquot of the final diffusate of the protein dialysis to the reference cuvette with stirring.
5. Finally, add a 100-μl aliquot of protein solution to the sample cuvette.[42] After mixing, record A_{412} until the reaction is complete (A_{final}, Fig. 1).

Calculations. The concentration of thiols originally present in the cuvette may be evaluated from Eq. (9) using ΔA_{412}:

$$\Delta A_{412} = A_{\text{final}} - (3.1/3.2)(A_{\text{DTNB}} - A_{\text{buffer}}) \quad (16)$$

When the titration is carried out in stock buffer, ϵ_{412} for TNB^{2-} is 14,150 M^{-1} cm^{-1}, and the titration yields the content of thiols in the native protein that are normally accessible to DTNB. When the titration is carried out in the presence of 6 *M* guanidinium chloride as in Fig. 1, ϵ_{412} for TNB^{2-} is 13,700 M^{-1} cm^{-1} and the total thiol content of the unfolded protein is determined.

[41] We have not found it necessary to exclude oxygen from manipulations in the cuvette or in the spectrophotometer, although it is possible continuously to flush the sample compartment with nitrogen. A cover or stopper is used on the cuvette to minimize evaporation and diffusion of oxygen during slow titrations.

[42] A suitable control experiment must be carried out if a protein contains a chromophore that absorbs significantly at 412 nm under the conditions of the titration.

[9] Relating Proteins by Amino Acid Composition

By ATHEL CORNISH-BOWDEN

The ease of determining amino acid compositions of proteins and the large body of published data[1] make them attractive for assessing relatedness between proteins. Although compositions cannot lead to as precise

[1] See D. M. Kirschenbaum, *Int. J. Biochem.* **13**, 637 (1981) and earlier compilations referenced therein.

ISBN 0-12-181991-4

conclusions as sequences, they are by no means empty of information; interpreted with care they can provide a reliable indication of relatedness.

Indexes for comparing compositions have been available for some years,[2-6] but their usefulness has been severely curtailed by the lack of a theoretical basis for interpreting their values. In practice, therefore, comparisons between compositions have rarely resulted in worthwhile conclusions. More recently, theoretical analysis has led to new composition indexes[7-9] that are more easily interpretable than the older ones and has placed the latter on a sounder theoretical foundation.[10]

Definitions

Let A and B be two proteins with N_A and N_B residues, respectively, of which n_{iA} in A and n_{iB} in B are of the ith type of amino acid, with $i = 1-18$ for the 18 types of amino acid residue commonly distinguished in composition determinations.[11] With these symbols the *difference index*[2] DI is defined as follows:

$$\mathrm{DI} = 50 \sum \left| \frac{n_{iA}}{N_A} - \frac{n_{iB}}{N_B} \right| \tag{1}$$

The *composition divergence*[3,5] D is defined as

$$D = \left[\sum \left(\frac{n_{iA}}{N_A} - \frac{n_{iB}}{N_B} \right)^2 \right]^{1/2} \tag{2}$$

and the index of Marchalonis and Weltman,[4] $S\Delta Q$, is defined as

$$S\Delta Q = 10^4 \sum \left(\frac{n_{iA}}{N_A} - \frac{n_{iB}}{N_B} \right)^2 \tag{3}$$

[2] H. Metzger, M. B. Shapiro, J. E. Mosimann, and J. E. Vinton, *Nature (London)* **219**, 1166 (1968).

[3] C. E. Harris, R. D. Kobes, D. C. Teller, and W. J. Rutter, *Biochemistry* **8**, 2442 (1969).

[4] J. J. Marchalonis and J. K. Weltman, *Comp. Biochem. Physiol. B* **38**, 609 (1971).

[5] C. E. Harris and D. C. Teller, *J. Theor. Biol.* **38**, 347 (1973).

[6] J. R. Dedman, R. W. Gracy, and B. G. Harris, *Comp. Biochem. Physiol. B* **49**, 715 (1974).

[7] A. Cornish-Bowden, *J. Theor. Biol.* **65**, 735 (1977).

[8] A. Cornish-Bowden, *J. Theor. Biol.* **74**, 155 (1978).

[9] A. Cornish-Bowden, *J. Theor. Biol.* **76**, 369 (1979).

[10] A. Cornish-Bowden, *Anal. Biochem.* **105**, 233 (1980).

[11] If asparagine can be distinguished from aspartate, and glutamate from glutamine, *for both proteins*, they should be counted separately, but this is not usually possible. If values for other kinds of residue, e.g., cysteine or tryptophan, are missing, they must simply be omitted from the calculations; although this adds to the uncertainty of any predictions made, it is unlikely to invalidate them except in the rare case of proteins that differ by large amounts in the numbers of residues of these kinds that they contain.

Each of the summations in these expressions, together with all the other summations that appear in this chapter, is carried out over as many kinds of amino acid residue as are distinguished by the data.

All these indexes are defined in terms of mole fractions rather than numbers of residues of each type. They can consequently be calculated for proteins of very different size, i.e., N_A quite different from N_B, without producing numbers that are much larger than those typically found in comparisons between proteins of similar size. This property is superficially attractive, but it is not clear that it is a real advantage, especially if it encourages comparisons between proteins for which no relationship can reasonably be postulated.

If $N_A = N_B = N$, it is convenient to use $S\Delta n$, a modified form of $S\Delta Q$ defined as follows[7]:

$$S\Delta n = \tfrac{1}{2}\sum (n_{iA} - n_{iB})^2 \tag{4}$$

This is exactly interconvertible with $S\Delta Q$ and D by the following relationship:

$$S\Delta n = 0.00005N^2 S\Delta Q = 0.5N^2 D^2 \tag{5}$$

and has the theoretical advantage that under certain statistical assumptions (see below) it directly predicts the number of differences between the sequences of A and B.

The difference index DI cannot be exactly related to $S\Delta Q$, D, or $S\Delta n$, but its properties can be analyzed, with the same restriction that $N_A = N_B = N$, in terms of the *difference total*[8] DT, defined as follows:

$$\mathrm{DT} = \tfrac{1}{2}\sum |n_{iA} - n_{iB}| = 0.01N\mathrm{DI} \tag{6}$$

This index, which was referred to but not considered in detail by Prager and Wilson,[12] has the useful property that it is an exact lower limit for the number of differences between the sequences of A and B, but in general it is less convenient than $S\Delta n$ because it does not give a direct prediction of the number of sequence differences without further manipulation.

Theoretical Properties of $S\Delta n$ and DT

The theoretical background for this section is given elsewhere.[7-9] Here I shall simply state the statistical assumptions that need to be made and the conclusions that follow from them. The basic assumptions for analyzing the behavior of both $S\Delta n$ and DT are as follows:

1. A and B are independent random sequences of amino acid residues.

[12] E. M. Prager and A. C. Wilson, *J. Biol. Chem.* **246**, 5978 (1971).

2. The probability p_i of finding a residue of the ith type at any locus is the same for all loci and the same in both A and B.

To these we must add a third restriction, which is not strictly an assumption because it will normally be known whether it is true or not, at least approximately.

3. The total number of residues in A is the same as the total number in B; i.e., $N_A = N_B = N$.

These restrictions may seem to be so severe that $S\Delta n$ and DT can have little predictive value, but in practice they can be relaxed sufficiently to make them acceptable. The two statistical assumptions are required in full only for testing whether A and B are related: if one can conclude that a relationship does exist, subsequent analysis requires the assumption of randomness only at the loci at which the sequences are different. The requirement for equal numbers of residues in the two proteins is also only approximate: provided that N_A and N_B do not differ by more than 18, the properties of $S\Delta n$ and DT are essentially the same as for the case where N_A and N_B are truly equal.[9,13]

With the above assumptions, the relationship between the predicted number of sequence differences and $S\Delta n$ is very simple:

$$S\Delta n = M(1 \pm 0.38) \tag{7}$$

where M is the actual number of sequence differences. Thus $S\Delta n$ is itself an unbiased estimator of the number of sequence differences, with a coefficient of variation of 38%. The corresponding relationship for DT is extremely complicated,[8] but may be approximated (for DT $\leq$ 55) by the following expression[14]:

$$\text{Predicted value of } M = \frac{0.93\text{DT} + 0.106\text{DT}^2 + 0.01264\text{DT}^3}{1 + 0.0531\text{DT} + 0.0001694\text{DT}^2} \tag{8}$$

The statistical uncertainty associated with this prediction is not known, but it is probably similar to the 38% calculated for $S\Delta n$, as there seems to be no strong reason for expecting one method of analysis to be substantially more precise than the other.

It follows from the theoretical analysis of the properties of $S\Delta n$ that

[13] Note that this condition refers to the absolute, not the relative, difference between N_A and N_B; i.e., the maximum difference is 18, not 18%.

[14] This is obtained by multiplying the expression given in Eq. (4) of Cornish-Bowden[8] by 0.93. The multiplication is needed because the expression given previously was for the predicted number of loci at which the sequences are independent, which is larger than the number at which they are different because about 7% of independent loci contain chance identities.

95% of comparisons between pairs of unrelated proteins satisfying the specified restrictions should give $S\Delta n$ greater than $0.42N$. Consequently, comparison of $S\Delta n$ with $0.42N$ can serve as the basis for a 95% confidence test of the hypothesis that the proteins are unrelated. It is, however, a very conservative test, as will be discussed below.

Effect of Experimental Error

The theoretical analysis outlined assumes that the compositions are known exactly, so that the statistical uncertainty in the resulting prediction derives wholly from the inherent statistical properties of the composition index. In reality, however, experimental error in the measured compositions must add to this uncertainty. Its effect may be judged from the following expression[9] for the observed $S\Delta n$:

$$\text{Observed } S\Delta n = \text{true } S\Delta n + \sum(1 - \rho_i)\sigma_i^2 \tag{9}$$

in which σ_i^2 is the variance of $n_{i\text{A}}$ and also that of $n_{i\text{B}}$, i.e., the variance associated with the determination of the ith kind of amino acid, the ρ_i is the correlation coefficient between the errors in $n_{i\text{A}}$ and $n_{i\text{B}}$. The values of ρ_i are likely to be close to zero if the compositions of A and B are determined by different operators using different equipment. Compositions measured under consistent conditions, however, may show certain kinds of amino acid consistently overestimated and others consistently underestimated: this would give rise to positive ρ_i values and so decrease the measurement error in $S\Delta n$. Negative ρ_i values are not in general very likely, but could conceivably occur if the analyses were done in different laboratories, one of which tended to overestimate the amino acids underestimated by the other, and vice versa.

Two major conclusions follow from these considerations:

1. The measurement error in $S\Delta n$ is minimized by having A and B analyzed under as nearly identical conditions as possible.
2. The maximum bias in $S\Delta n$ due to experimental error is likely to be of the order of $\sum\sigma_i^2$. The magnitude of this quantity depends on the amount of care used, but it is experimentally possible[9] to ensure that it is small compared with the 38% uncertainty that is independent of experimental error.

If replicate analyses are carried out the value of $\sum\sigma_i^2$ can in principle be estimated. This is useful as a rough guide to the likely bias in $S\Delta n$, but it is probably unwise to use it as a numerical correction to $S\Delta n$, because too many unknowns are involved; in particular, no proper correction can be made without knowledge of the ρ_i.

In the remainder of this chapter I shall not explicitly consider the effect of experimental error on the validity of composition comparisons.

Testing Two Proteins of Unknown Sequence for Possible Relatedness

Case 1. Proteins of Equal Size, $N_A = N_B = N$

This is the ideal case. The simplest procedure is to calculate $S\Delta n$ according to Eq. (4) and compare it with $0.42N$ and with $0.93N$.

1. If $S\Delta n < 0.42N$, there is a strong indication, amounting almost to certainty, that the proteins are related.
2. If $0.42N < S\Delta n < 0.93N$, there is a weak indication that the proteins are related.
3. If $S\Delta n > 0.93N$, however, the compositions do not provide grounds for believing a relationship to exist.

The rationale for this test is as follows: 95% of comparisons between unrelated proteins *that satisfy the statistical assumptions specified in the theoretical properties section* give $S\Delta n < 0.42N$, but in reality nearly 100% of comparisons between unrelated proteins fail this "strong test"[9] of relatedness. This discrepancy can be accounted for without jettisoning the basic theory by realizing that genuinely unrelated proteins are unlikely to have exactly the same statistical properties; i.e., they are likely to violate assumption 2 given in the theoretical properties section. If this assumption is not made, the same kind of analysis predicts that $S\Delta n$ should overestimate the actual number of sequence differences.[9] As unrelated proteins typically differ at about 93% of loci (the other 7% being the result of chance identities), one ought to expect most comparisons between unrelated proteins to give $S\Delta n$ greater than $0.93N$. This provides the basis for the alternative "weak test,"[10] in which $S\Delta n < 0.93N$ is taken as an indication of relatedness. The reliability of this test may be judged from the fact that out of 140 comparisons between probably unrelated proteins the "weak test" gave a spurious indication of relatedness in only 16 cases, whereas it correctly indicated relatedness in 22 out of 23 comparisons between related proteins studied at the same time.[10]

If one wishes to test a value given in the literature or elsewhere for one of the original indexes, DI, D, or $S\Delta Q$, it is not necessary to reexamine the data to calculate $S\Delta n$, because the critical values of these indexes based on the behavior of $S\Delta n$ are given in Table I. Although it is a little less convenient to use these critical values than to use the values of 100 for $S\Delta Q$[4] and 0.07 for D[5] suggested by the original authors, it is a great deal

TABLE I
CRITICAL VALUES OF DI, $S\Delta Q$, AND D[a]

		"Strong" test[b]			"Weak" test[c]		
N[d]	M_r[e]	DI	$S\Delta Q$	D	DI	$S\Delta Q$	D
10	1100	32.5	840	0.290	57.9	1860	0.431
20	2200	27.1	420	0.205	44.7	930	0.305
30	3300	23.5	280	0.167	37.7	620	0.249
40	4300	21.0	210	0.145	33.0	465	0.216
50	5400	19.2	168	0.130	29.8	372	0.193
60	6500	17.7	140	0.118	27.3	310	0.176
70	7600	16.6	120	0.110	25.4	266	0.163
80	8700	15.6	105	0.102	23.8	233	0.152
90	9700	14.8	93.3	0.0966	22.4	207	0.144
100	10800	14.1	84.0	0.0917	21.4	186	0.136
120	13000	12.9	79.0	0.0837	19.6	155	0.124
140	15100	12.0	60.0	0.0775	18.1	133	0.115
160	17300	11.3	52.5	0.0725	17.0	116	0.108
180	19500	10.7	46.7	0.0683	16.0	103	0.102
200	21600	10.1	42.0	0.0648	15.2	93.0	0.0964
250	27000	9.09	33.6	0.0580	13.6	74.4	0.0863
300	32400	8.32	28.0	0.0529	12.5	62.0	0.0787
350	37800	7.71	24.0	0.0490	11.5	53.1	0.0729
400	43200	7.22	21.0	0.0458	10.8	46.5	0.0682
450	48600	6.82	18.7	0.0432	10.2	41.3	0.0643
500	54000	6.47	16.8	0.0410	9.66	37.2	0.0610
550	59400	6.17	15.3	0.0391	9.22	33.8	0.0582
600	64800	5.91	14.0	0.0374	8.82	31.0	0.0557
650	70200	5.68	12.9	0.0359	8.48	28.6	0.0535
700	75600	5.48	12.0	0.0346	8.17	26.6	0.0515

[a] Reproduced, with permission, from Cornish-Bowden.[10]

[b] Observed values less than those tabulated provide strong evidence that the proteins compared are related.

[c] Observed values less than those tabulated provide weak evidence that the proteins compared are related.

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

[e] Approximate value of the relative molecular mass corresponding to each value of N, calculated from $M_r \simeq 108N$. If the proteins compared have different M_r values, the larger should be used.

more reliable because it takes account of the very strong dependence of the significance of each index on the total number of residues. With the tests suggested originally one would expect to have almost no chance of detecting a genuine relationship between small proteins and an excessive

TABLE II
CORRECTING $S\Delta n$ FOR DIFFERENCES IN SIZE[a]

$\lvert N_A - N_B \rvert$	Correction	$\lvert N_A - N_B \rvert$	Correction
0	0.0	10	+1.9
1	+0.5	11	+1.7
2	+0.9	12	+1.4
3	+1.3	13	+1.0
4	+1.6	14	+0.6
5	+1.8	15	+0.2
6	+2.0	16	−0.4
7	+2.0	17	−1.0
8	+2.0	18	−1.7
9	+2.0	19	−2.5

[a] If $S\Delta n$ is calculated as $\frac{1}{2}\sum(n_{iA} - n_{iB})^2$, but N_A is different from N_B, the corrected value of $S\Delta n$ may be obtained by adding the appropriate correction from this table.

likelihood of making incorrect predictions of relatedness between large proteins. This is indeed what one observes.[15]

Case 2. Proteins of Approximately Equal Size, $\lvert N_A - N_B \rvert \leq 18$

This case can be analyzed in essentially the same way as case 1. The value of $S\Delta n$ can be corrected if desired by redefining it as follows[9]:

$$S\Delta n = \tfrac{1}{2}\sum(n_{iA} - n_{iB})^2 - 0.035(N_A - N_B)^2 + 0.535|N_A - N_B| \quad (10)$$

or (equivalently) calculating it in the ordinary way by means of Eq. (4) and adding the corrections shown in Table II. It will be seen, however, that the corrections are small, and in practice the error caused by ignoring them is likely to be trivial by comparison with other sources of error. It follows, therefore, that there will be little error associated with using Table I to test DI, D, or $S\Delta Q$ when the proteins compared differ in length by 18 or fewer residues.If this is done, it is cautious to take N for insertion into Table I as the larger of N_A and N_B, as this will decrease the likelihood of getting a significant result.

Case 3. One Protein Approximately Twice as Large as the Other, $N_B \simeq 2N_A$

In this case the hypothesis to be tested is that B resembles a dimer of A. Provided that B can plausibly be assumed to be dimer-like, i.e., to have a sequence in which the second half is the same or approximately the same as the first half, the appropriate procedure is to halve all the values

[15] A. Cornish-Bowden, unpublished work.

of n_{iB} before calculating $S\Delta n$ and continue as for case 1 or 2 as appropriate.[16] Alternatively, one can use Table I to test the significance of DI, D, or $S\Delta Q$, taking N as the larger of N_A and $N_B/2$.

In principle, the same sort of approach could be used in the more general case where N_B/N_A is approximately equal to a small integer greater than 2. Although this is not fundamentally wrong, any conclusions drawn from such an analysis should be tempered with the knowledge that they depend on the plausibility of assuming that B consists of the same sequence repeated more than twice.

Case 4. Proteins of Unrelated Size

My own belief is that there is little point in comparing the compositions of proteins of unrelated size. The literature teems with such comparisons, however, so presumably my opinion is not generally shared. If Table I is used for testing such pairs of proteins little harm is likely to be done, as studies of proteins of known sequence[10] have shown that there is negligible danger of concluding that a relationship exists when it does not; there is also very little chance of detecting one that does exist.

Comparing Related Proteins

When a group of proteins are known to be related, their amino acid compositions can be used to quantify the degree of relatedness and to generate a phylogenetic tree. The assumptions needed for this purpose are similar to those needed for testing whether a relationship exists, but are relaxed to such an extent that they become plausible. It is not necessary to assume that the proteins compared are random sequences of residues, but only that a degree of randomness exists at loci where they are different. For these loci only, one must assume that the probability of finding any particular kind of residue is the same for both proteins. It then follows,[7] at least for proteins of equal size, that $S\Delta n$ is an unbiased estimator of the number of sequence differences, with a coefficient of variation of 38%.

If N_A is different from N_B, the "number of sequence differences" needs to be defined: if it is taken to be the number of residues in the longer sequence that are unmatched by indistinguishable (i.e., identical apart from possible differences in presence or the absence of amide groups) residues in the shorter sequence when the two are aligned,[17] then it is the

[16] This differs somewhat from the analysis suggested previously,[7] which was made obsolete by subsequent investigation[9] of the effect of allowing N_A to be different from N_B.

[17] There is an implicit assumption that a unique alignment exists that would be obvious if the sequences were known. This assumption is not unreasonable for proteins with a substantial degree of relatedness, though it would have little meaning if one tried to apply it to unrelated proteins.

quantity estimated by the modified form of $S\Delta n$ defined by Eq. (10). According to this definition, blanks in the shorter sequence are counted as differences, but blanks in the longer sequence (if any are needed to achieve alignment) are not.

Studies with groups of proteins of known sequence (insulin, snake venom toxins, cytochrome c, pancreatic ribonuclease, α chain of hemoglobin, β and δ chains of hemoglobin, myoglobin) showed that the actual behavior of $S\Delta n$ is very similar to what was predicted[9]: there was a slight tendency to exaggerate the amount of difference, i.e., there was a positive bias of around 10%, but the coefficients of variation were clustered around the predicted value of 38%.

The statistical uncertainty of 38% may seem to be so large that even if $S\Delta n$ does behave as predicted it is likely to be too unreliable to make compositions a practical alternative to sequences in the construction of phylogenetic trees. In fact, however, several studies,[9,18–20] one of which[9] is illustrated in Fig. 1, have shown that the results are much closer to those given by the corresponding sequences than one might expect. In examining Fig. 1, the main features are obvious, but two points should be noted: first, a serious typographical error in the original publication[9] has been corrected; second, although the composition tree wrongly separates the long neurotoxins into two groups, among themselves the long neurotoxins are clustered by their compositions almost exactly the same as they are clustered by their sequences.

It is not always remembered that amino acid sequences are themselves subject to statistical uncertainties if used to estimate the time since separation of two evolutionary lines; for sequences that differ at very few loci, this uncertainty may be as large as or larger than the 38% uncertainty inherent in the use of $S\Delta n$ as an estimator of the number of sequences. Thus in clustering closely related proteins there may be little or no advantage in using sequences rather than compositions, and the cost in materials, experimental effort, and time will certainly be much greater. This expectation is confirmed by computer simulation[20] of a known evolutionary tree: it is found that unless there are more than about eight differences between the sequences of the most dissimilar proteins considered one cannot expect an appreciable advantage to accrue from knowing the sequences. Moreover, if results from composition determinations on several different kinds of protein are combined, they are capable of yielding more reliable results than can be obtained from sequence determinations of a single kind of protein.[20]

[18] J. A. Black and R. N. Harkins, *J. Theor. Biol.* **66**, 281 (1977).
[19] W. S. Davidson and T. G. Flynn, *J. Mol. Evol.* **14**, 251 (1979).
[20] A. Cornish-Bowden, *Biochem. J.* **191**, 349 (1980).

Trees constructed from both sequences and compositions of snake venom toxins (Fig. 1), α chains of hemoglobin,[20] as well as unpublished results with insulin, myoglobin, cytochrome c, β and δ chains of hemoglobin, and others, bear out these predictions. In all cases the sequences gave better results, but the advantage was usually small and not always in favor of the sequences in matters of detail. Thus, although sequences can cer-

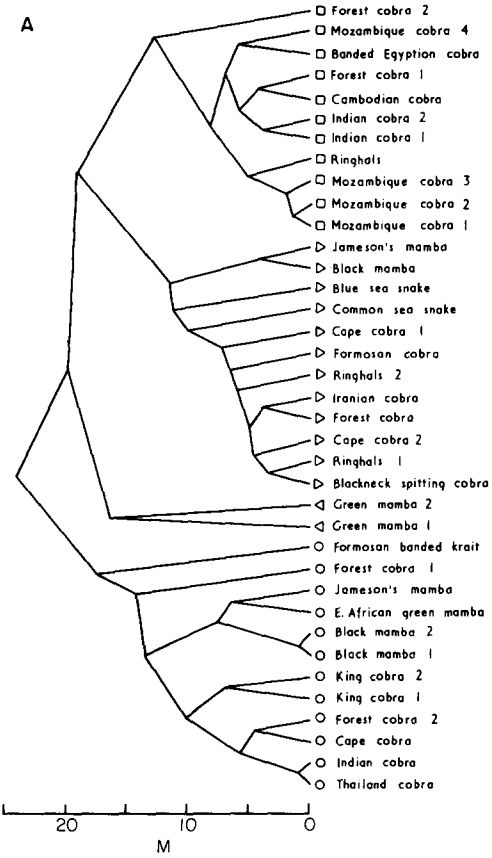

FIG. 1. Phylogenetic trees for 37 snake venom toxins. The trees were calculated by the UPGMA method[23] from (A) values of M, the number of unpaired residues in the longer sequence when the two are aligned as on p. 150 of "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 2; National Biomedical Research Foundation, Silver Spring, Maryland, 1976; (B) values of $S\Delta n$ calculated from the compositions according to Eq. (10). The main classes of toxin are indicated by the following symbols: ○, long neurotoxins; ◁, short toxins; ▷, short neurotoxins; □, cytotoxins. Reproduced, after correcting a serious typographical error affecting 14 of the symbols, from Cornish-Bowden.[9]

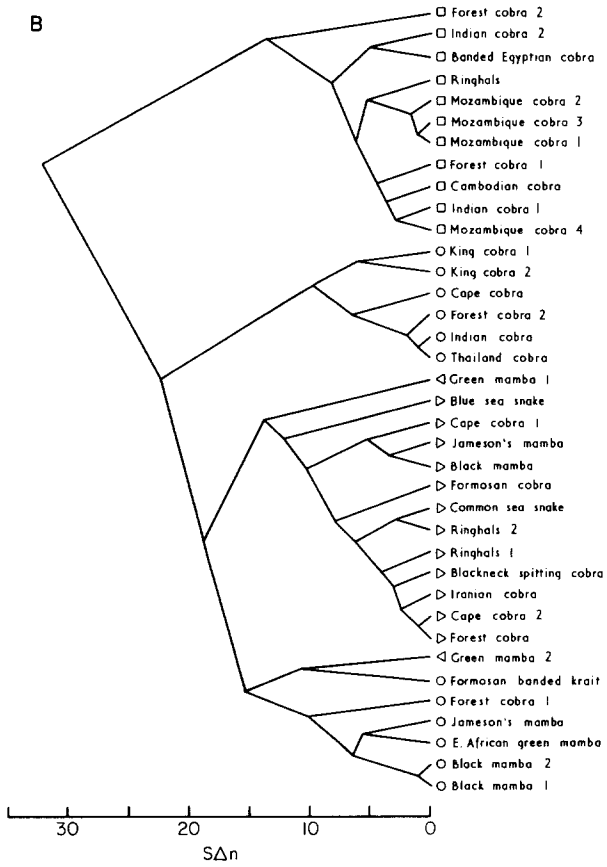

FIG. 1. (*continued*)

tainly be expected to perform better than compositions, the extra effort required for measuring them may well not be cost-effective.

There have been a number of applications to protein composition data of the methods discussed in this chapter. One of these,[20a] a study of the CuZn superoxide dismutases from the ponyfish and its symbiotic bacterium *Photobacter leiognathi*, deserves special mention as an example of how protein compositions have been able to provide biological information of general interest and importance. The typically fish-like composition of the bacterial enzyme, similar to but clearly different from that of the host fish and grossly different from that of any known bacterial en-

[20a] J. P. Martin and I. Fridovich, *J. Biol. Chem.* **256**, 6080 (1981).

zyme, was taken as strong evidence that genetic information can be transferred from a eukaryote to a prokaryote.

Constructing a Phylogenetic Tree from Composition Data

Many methods have been used for constructing phylogenetic trees from sequence data,[21-23] but the choice for composition data is more restricted: the ancestral-sequence method cannot be used with compositions and one is forced to use a matrix method, i.e., a method that depends on carrying out a series of operations on a table of differences between the proteins. Fortunately, however, results from computer simulation of known trees suggest that matrix methods may well be preferable even when other possibilities exist.[24] Even within the large set of different matrix methods, the choice for compositions is much more limited than it is for sequences. This is because without knowledge of which particular substitutions have occurred there is no way of taking account of the fact that, for example, a proline–tryptophan substitution is much more drastic than an aspartate–glutamate substitution. Thus the only reasonable choice for the entries in the initial table of differences is an unweighted composition index such as $S\Delta n$.

The initial table will typically have the appearance of example a shown in Table III. There are many ways of converting such a table into a tree, but fortunately the method that is easiest to understand and the most widely used, the *unweighted pair-group method with arithmetic averages* (UPGMA), is also a very satisfactory method and is the only one described here. The steps that have to be carried out are as follows:

1. Construct a difference matrix (a) of the kind shown in Table III.
2. Identify the smallest entry in it [in (a) of Table III, the 3 between A and D].
3. Cluster the proteins concerned (A and D), i.e., replace them with a single composite set of entries for AD: differences that do not involve A or D are left unchanged; the others are placed by the average entries for A and D. This gives matrix (b) of Table III as the new matrix.
4. Treat the new matrix in the same way, i.e., cluster AD with B, but

[21] W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967).

[22] M. O. Dayhoff, C. M. Park, and P. J. McLaughlin, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, pp. 7–16. National Biomedical Research Foundation, Silver Spring, Maryland, 1972.

[23] P. H. A. Sneath and R. R. Sokal, "Numerical Taxonomy," pp. 188–308. Freeman, San Francisco, California, 1973.

[24] M. O. Dayhoff, *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **35**, 2132 (1976).

TABLE III
CLUSTERING BY THE UNWEIGHTED PAIR-GROUP METHOD WITH ARITHMETIC AVERAGES (UPGMA)[a]

B	6.0				B	**5.0**			C	9.7			CE	9.6
C	11.0	8.0			C	10.5	8.0		E	9.5	**5.0**			
D	**3.0**	4.0	9.0		E	10.0	9.0	5.0						
E	10.0	9.0	5.0	10.0										
	A	B	C	D		AD	B	C		ADB	C			ADB
		(a)				(b)				(c)				(d)

[a] A matrix showing the differences, as measured for example by $S\Delta n$, between five proteins A, B, C, D, and E is shown above (a). The method of clustering proteins and converting each matrix into a smaller one is described in the text. In each case the difference between the proteins or groups of proteins to be clustered next is shown in boldface type.

when averaging give each entry weight proportional to the number of proteins from the original matrix that gave rise to it.[25] Thus in calculating the difference between ADB and C, we must give twice as much weight to the difference between AD and C as to the difference between B and C, so the result is $(10.5 \times 2 + 8)/3 = 9.7$. The new matrix is (c) in Table III.

5. Continue in the same way until the final matrix consists of a single entry. For Table III this requires only one more step to give matrix (d).
6. Draw a tree showing the clusters that have been found (Fig. 2).

It sometimes happens [e.g., see matrix (b) in Table III] that the smallest entry in the matrix is not unique. In the example shown to illustrate the UPGMA method it makes no difference to the final result if one clusters C and E before AD and B, but this is by no means always true. Full discussion of this point is beyond the scope of this chapter, and the reader should refer to Sneath and Sokal,[23] which contains a very extensive account of clustering methods.

Combining Sequence and Composition Data

It may happen that one wishes to cluster a group of proteins for which some sequences are known but others are not. If $S\Delta n$ is used as composi-

[25] The object is to give equal weight to all the entries in the original matrix. In the alternative *weighted pair-group method with arithmetic averages* (WPGMA),[23] there is no weighting during clustering, and as a result the original numbers become unequally weighted. This confusing nomenclature is unfortunately well established.

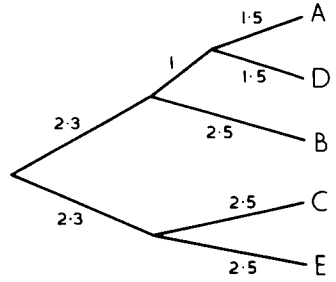

FIG. 2. Tree resulting from the application of the UPGMA method set out in Table III. Addition of the numbers along the path connecting any two proteins shows the distance at which they were clustered; e.g., the distance between B and D is 2.5 + 1 + 1.5 = 5, which corresponds to the 5 shown in boldface type in Table III, matrix (b).

tion index, this should present no problem because it directly estimates the number of sequence differences and so the initial matrix can be constructed with $S\Delta n$ values replaced by numbers of sequence differences wherever these are known. (These should not count aspartate and asparagine or glutamate and glutamine as different unless *all* of the compositions used do so as well.) In the subsequent clustering process a case could be made for giving more weight to numbers derived from sequences than to those derived from compositions, but no study of this has been carried out. If composition indexes other than $S\Delta n$ are used they ought to be transformed into estimates of the numbers of sequence differences by means of Eqs. (5), (6), or (8) as appropriate before mixing with numbers from sequence determinations.

Proteins That Can Be Separated into Subunits

If a protein to be studied can easily be separated into nonidentical subunits or polypeptide chains that correspond from organism to organism, it is better to compare the individual chains rather than the complete proteins. For example, the complete insulin molecules from human and sheep differ at four loci, of which three are in the A chain and one is in the B chain. The value of $S\Delta n$ for the complete molecule is 6, overestimating the actual number of sequence differences by 50%. If the chains are considered separately, however, we find $S\Delta n = 3$ for the A chain and $S\Delta n = 1$ for the B chain, both of which are correct values for the numbers of sequence differences. Although one cannot always expect as large an improvement as this, one can expect some improvement on average because the more one can analyze a protein as fragments, the closer one is to using sequence information rather than composition information. In fact, of 153 comparisons between 18 insulins, 102 were improved by replacing

TABLE IV
PERFORMANCE OF $S\Delta n$ WITH INSULIN CHAINS[a]

Protein	Mean $S\Delta n/M$[b]	Coefficient of variation[c] (%)
A chain only	0.96	27.3
B chain only	1.06	23.3
Whole protein	1.24	35.6
Sum of A and B chains[d]	1.02	17.6

[a] All the 18 insulins shown on p. 128 of "Atlas of Protein Sequence and Structure" (M. O. Dayhoff, ed.), Vol. 5, Suppl. 2; National Biomedical Research Foundation, Silver Spring, Maryland, 1976, were used, so that there were 153 pairwise combinations.

[b] The theoretical mean would be 1.00 if $S\Delta n$ were a true unbiased estimator of M, the number of sequence differences.

[c] Calculated with 17 degrees of freedom, i.e., one less than the number of proteins in the sample (see Cornish-Bowden[9]).

[d] $S\Delta n$ and M were obtained by adding together the values for the separate chains.

$S\Delta n$ calculated for the whole protein by the sum of the $S\Delta n$ values for the A and B chains, 29 were unchanged, and only 22 were made worse. A statistical summary of the results of this study is shown in Table IV, from which it may be seen that not only was the precision greatly improved, but in addition the appreciable bias in the results for the whole protein was largely eliminated by considering the two chains separately.

Note Added in Proof

The original analysis[7] of the theoretical properties of $S\Delta n$ contained an error, which was unfortunately carried over to later articles[8–10] and the present chapter. For 95% of comparisons between proteins satisfying the statistical restrictions given in the Theoretical Properties section, $S\Delta n$ exceeds a value 1.65 standard deviations below its expected value, i.e., below $0.93N$, rather than below N, as stated. The effect of this is that the nominal confidence level of the "strong" test is about 92.5% rather than the intended 95%. Although this correction is necessary for establishing the theoretical properties of $S\Delta n$ it has little practical consequence.