The amino acid compositions of proteins are correlated with their molecular sizes

Athel CORNISH-BOWDEN
Department of Biochemistry, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, U.K.

(*Received 5 April 1983/Accepted 26 April 1983*)

Natural peptides and small proteins in general have amino acid compositions that diverge much more from the average composition of all proteins than do those of proteins. The effect is large and consistent enough to provide a rough check on the measured molecular mass of a protein and to indicate whether it is likely to have a significantly repetitive structure. For example, the α-chain of tropomyosin, a highly repetitive protein, has no amino acid composition that would be characteristic of a much smaller protein. The observation provides support for the suggestion [Taylor, Britton & van Heyningen (1983) *Biochem. J.* **209**, 897–899] that tetanus toxin resembles a trimer of the light chain produced by proteolysis.

At first sight, the idea that the amino acid composition of a protein might provide a guide to its molecular size is so obviously absurd that it is unworthy of serious consideration. Nonetheless, the divergence of the composition of any protein from the average composition of a large set of diverse proteins correlates remarkably well with the reciprocal of the number of residues in it. The correlation is not good enough to offer a useful way of measuring molecular mass as such, but it can indicate whether a protein may possess a previously unrecognized oligomeric structure or a highly repetitive sequence.

The amino acid compositions of proteins are often compared by means of the index $S\Delta Q$ proposed by Marchalonis & Weltman (1971), which is defined as the sum of the squares of the differences between two proteins of the mole percentages of the 18 kinds of amino acid commonly distinguished in composition measurements. Although Marchalonis & Weltman (1971) suggested that $S\Delta Q$ values less than 100 might constitute evidence of relatedness, it is now clear, both from theoretical considerations (Cornish-Bowden, 1977, 1979) and from observations with proteins of known sequence (Cornish-Bowden, 1979, 1980), that small $S\Delta Q$ values are nearly always observed in comparisons between large proteins, regardless of whether they are related or not, but almost never in comparisons between unrelated small proteins. Thus in practice $S\Delta Q$, in common with most of the other composition indexes that are in widespread use, tells us more about the sizes of the proteins compared than it does about whether they are related. This raises the question, therefore, of whether $S\Delta Q$ can be used as a measure of molecular size.

To avoid the need to consider proteins in pairs, the composition of each protein can be compared with the average composition of a large set of diverse proteins. Thus a new index $S\Delta Q_0$ can be defined as follows:

$$S\Delta Q_0 = \sum (X_i - X_{i0})^2$$

where X_i is the mole percentage of the ith kind of amino acid in the protein considered and X_{i0} is the corresponding average value from Table 1. For calculating the averages given in this Table, I used the compositions of 118 proteins from different 'superfamilies' (Dayhoff, 1978), i.e. 118 proteins for which no sequence relatedness has been detected.

For each protein, the value of $S\Delta Q_0$ is plotted in Fig. 1 against $1/N$, the reciprocal of the number of residues. It is evident from inspection that there is a remarkably good proportionality, which is confirmed by the high value of +0.79 for Spearman's rank correlation coefficient (Kendall, 1970). Although any protein must be more than just a random sequence of amino acid residues (Holmquist & Moise, 1975; Cornish-Bowden & Marson, 1977), it is noteworthy that a simple model edicts that $S\Delta Q_0$ should be proportional to $1/N$. If the amino acids are distributed at random, and X_{i0} is the probability (in percent) that the ith type of amino acid will occur at any locus, then the properties of the multinomial distribution (Cornish-Bowden, 1977, 1979; Cornish-Bowden & Marson, 1977) lead to the conclusion that the expected value of $S\Delta Q_0$, i.e. the mean of its distribution, is $(10^4 - \sum X_{i0}^2)/N$, or $9300/N$. The median value of $N \cdot S\Delta Q_0$ for the 118 proteins is actually 18900, more than double 9300, but it is nonetheless remarkable that a model that is obviously too simple to be correct can account for as much of the variability of $S\Delta Q_0$ as it does.

The proteins with the highest and lowest values of $N{\cdot}S\Delta Q_0$ are identified in Fig. 1. There is no obvious mechanism for generating very low values, and so those are probably nothing more than the expected tail of the distribution. Very high values, however, may be attributed to special characteristics of the proteins concerned. Protamine and galline, for example, are proteins that interact strongly with DNA by virtue of very high positive charges per molecule, the result of an arginine content greater than 50%; metallothionein is a protein with a large capacity for binding Zn^{2+} and Cd^{2+} ions, achieved in a similar way by the presence of a high proportion of cysteine residues.

The other three proteins in the highest 5% of $N{\cdot}S\Delta Q_0$ values have all been identified (Dayhoff, 1978) as proteins with significantly repetitive sequences. For example, tropomyosin binds seven identical actin molecules in a linear array, and its sequence shows a very pronounced sevenfold periodicity (Longley, 1977). If it were plotted in Fig. 1 according to the size of its approximate repeat unit, therefore, it would appear at a 7-fold higher value of

Table 1. *Average composition of 118 proteins*

The data used for compiling this Table and Fig. 1 were obtained from the *Atlas of Protein Sequence and Structure* (Dayhoff, 1978) and refer to the following 118 proteins: melittin (Ceylon bee); protamine (Caspian sturgeon); thymosin α_1 (cattle); vasoactive intestinal peptide (domestic fowl); calcitonin (rat); proteins of genes J, E, B, D and G (bacteriophage ϕX174); corticotropin (spiny dogfish); crotamine (South American rattlesnake); statherin (human); hevein (rubber tree); monellin, chains A and B (serendipity berry); allergen RA5 (ragweed); purothionin A-I (wheat); 50S-ribosomal proteins L34, L33, L32, L30, L29, L27, L25, L18, L16, L10 and L5 (*Escherichia coli*); coat protein (bacteriophage PF1); toxin II (sea-anemone); relaxin (pig); γ-carboxyglutamic acid-containing protein (cattle); thymopoietin II (cattle); rubredoxin (*Desulfovibrio gigas*); neurotoxin B-IV (heteronemertine worm); ferredoxin (*Clostridium* M-E); pancreatic secretory trypsin inhibitor (human); venom basic proteinase inhibitor II (ringhals); metallothionein (horse); neurotoxin I (North American scorpion); galline (domestic fowl); hirudin (leech); long neurotoxin I (broad-banded blue sea snake); 30S-ribosomal proteins S21, S18, S16, S20, S15, S13, S12, S9 and S6 (*E. coli*); ubiquitin (human); high-potential iron–sulphur protein (*Rhodospirillum gelatinosa*); lipid-binding protein A-II (Rhesus monkey); complement C3A anaphylatoxin (human); murein lipoprotein precursor (*E. coli* B); internal proteins I and II (bacteriophage T4); amyloid protein AA (Pekin duck); apovitellenin I (domestic fowl); haptoglobin α-chain (pig); neurophysin I (cattle); keratin high-sulphur fraction IIB2 (South African Angora goat); feather keratin (silver gull); plastocyanin (broad bean); ribonuclease ST (*Streptococcus erythreus*); enterotoxin β-chain (*Vibrio cholerae*); immunoglobulin κ(b9)-chain constant region (rabbit); testis-specific cytochrome *c* (mouse); immunoglobulin κ-chain V-I (human LAY); thyrotropin β-chain (pig); myohaemerythrin (*Themiste zostericola*); histone H2B (brown trout); ribonuclease (camel); cytochrome c' (*Rhodospirillum rubrum*); lysozyme (chacalaca); glycophorin A (human); prophospholipase A2 (pig); haemoglobin α-chain (tarsier); cytochrome b_5 (cattle); Cu/Zn-superoxide dismutase (cattle); coat protein (tobacco-mosaic virus, cowpea); troponin C (domestic-fowl skeletal muscle); κ-casein (goat); α-crystallin A-chain (kangaroo); troponin I (rabbit skeletal muscle); flavodoxin (*Azotobacter vinelandii*); dihydrofolate reductase (mouse); C-reactive protein (human); adenylate kinase isoenzyme 1 (human); *NO*-diacetylmuramidase (*Charalopsis* sp.); coat protein (alfalfa-mosaic virus); ribitol dehydrogenase (*Klebsiella aerogenes*); triose phosphate isomerase (*Bacillus stearothermophilus*); proteinase (*Streptococcus pyogenes*); troponin T (rabbit skeletal muscle); carbonic anhydrase (rabbit); tropomyosin α-chain (rabbit skeletal muscle); penicillinase (*Bacillus licheniformis*); rhodanese (cattle); carboxypeptidase B (cattle); peroxidase (horseradish); asparaginase (*E. coli*); lactate dehydrogenase M-chain (pig); glyceraldehyde 3-phosphate dehydrogenase (*B. stearothermophilus*); alcohol dehydrogenase (yeast); prochymosin (cattle); assembly protein (bacteriophage MS2); antithrombin III (human); Stuart factor (cattle); NADP-specific glutamate dehydrogenase (*Neurospora crassa*); fibrinogen β-chain (human); RNA replicase β-chain (bacteriophage MS2); serum albumin (human); β-galactosidase (*E. coli*).

	Average composition	
Amino acid	(total no. of residues)	(residues/ 100 residues)
Ala	1530	8.45
Arg	1061	5.86
Asx	1761	9.72
Cys	444	2.45
Glx	1924	10.62
Gly	1332	7.35
His	394	2.17
Ile	839	4.63
Leu	1435	7.92
Lys	1256	6.93
Met	333	1.84
Phe	669	3.69
Pro	825	4.53
Ser	1204	6.65
Thr	1036	5.72
Trp	237	1.31
Tyr	602	3.32
Val	1235	6.81
Total	18115	99.97

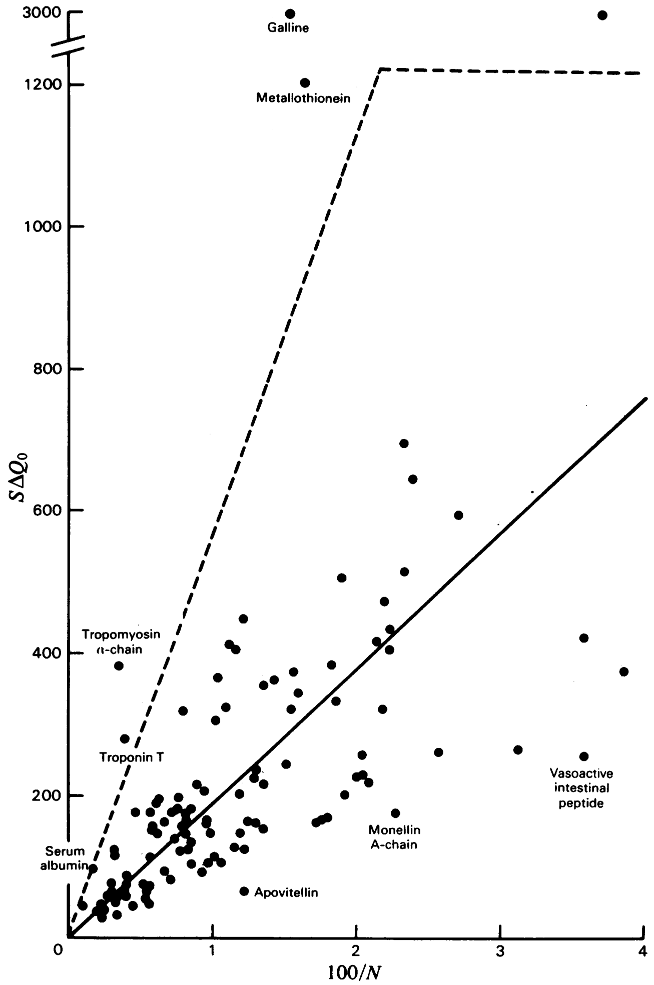

Fig. 1. *Correlation between amino acid composition and reciprocal number of residues*
For each of the 118 proteins listed in Table 1, the divergence of the amino acid composition from the average composition shown in Table 1 is expressed as $S\Delta Q_0$ and plotted against 100/N, where N is the number of residues. The continuous line represents the median value of $N \cdot S\Delta Q_0 = 18900$. The broken line separates values more than 3 times this median, and proteins with larger values of $N \cdot S\Delta Q_0$ are identified, as are three proteins with very low values of $N \cdot S\Delta Q_0$.

100/N, i.e. 2.45, close to the median line. Serum albumin and troponin T are less striking examples of the same sort of behaviour.

It follows from these observations that a value of $S\Delta Q_0$ much larger than expected from the estimated number of residues, for a protein of unknown sequence, constitutes a strong suggestion that this sequence is repetitive, or even that a supposed monomer is actually separable into subunits.

In view of these results, it seemed of interest to examine the value of $S\Delta Q_0$ for tetanus toxin, a potent neurotoxin secreted by *Clostridium tetani* after infection of a wound. The toxin is synthesized as a single chain of 150kDa, which is converted by proteolysis into a 100kDa heavy chain and a 50kDa light chain (Taylor *et al.*, 1983). The amino acid compositions of these two chains are so similar that Taylor *et al.* (1983) have suggested that the whole toxin resembles a trimer of the light chain. The value of $S\Delta Q_0$ for the whole toxin is 58.3, about 4 times higher than expected for a protein with an estimated 1317 residues, but typical for a protein of about 437, the number estimated for the light chain. It is therefore clear that, even if no information existed about the individual chains, the amino acid composition of tetanus toxin would be highly suggestive of a repetitive sequence.

References

Cornish-Bowden, A. (1977) *J. Theor. Biol.* **65**, 735–742

Cornish-Bowden, A. (1979) *J. Theor. Biol.* **76**, 369–386

Cornish-Bowden, A. (1980) *Anal. Biochem.* **105**, 233–238

Cornish-Bowden, A. & Marson, A. (1977) *J. Mol. Evol.* **10**, 231–240

Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, National Biomedical Research Foundation, Silver Spring

Holmquist, R. & Moise, H. (1975) *J. Mol. Evol.* **6**, 1–14

Kendall, M. G. (1970) *Rank Correlation Methods*, 4th edn., Griffin, London

Longley, J. (1977) *Int. J. Pept. Protein Res.* **9**, 49–51

Marchalonis, J. J. & Weltman, J. K. (1971) *Comp. Biochem. Physiol. B* **38**, 609–625

Taylor, C. F., Britton, P. & van Heyningen, S. (1983) *Biochem. J.* **209**, 897–899