BIOCHEMICAL JOURNAL LETTERS

The prediction of repetitive protein sequences from amino acid compositions

Cornish-Bowden (1983) demonstrated recently that the amino acid compositions of proteins correlate with their molecular sizes. He devised a new index $S\Delta Q_0$ to assess the compositions of proteins relative to an average composition (of 118 proteins), and showed that for most proteins $S\Delta Q_0$ was proportional to the reciprocal of protein size. Of the five proteins with anomalously high values for $S\Delta Q_0$, three had highly repetitive sequences and Cornish-Bowden went on to suggest that anomalously high values could be taken to indicate a repetitive sequence.

I have calculated $S\Delta Q_0$ for two quite different proteins with repetitive sequences, the fibrous protein collagen and the lens protein γ-crystallin. Collagen chains have a repeating triplet structure, $(Gly\text{-}Xaa\text{-}Yaa)_n$, but the collagen genes have retained a 54 base-pair exon size in the triple helical region leading to the concept of an ancestral 54 base-pair sequence, and thus 18 amino acid sequence, for this protein (Yamada, 1980) although some variability had been noted in later work (Wozney *et al.*, 1981). γ-Crystallin has the most symmetrical structure known for a globular protein, with a total of four motifs in two domains and evidence of a double gene duplication (Blundell *et al.*, 1981).

The values of $S\Delta Q_0$ for the α1 and α2 polypeptides of type I collagen are much higher than expected for polypeptides with about 1040 amino acid residues (Fig. 1). When plotted as if they had only 18 residues apiece they fall remarkably close to an extrapolation of the line that Cornish-Bowden (1983) drew through the points for 118 proteins (Fig. 1). This seems powerful support for Cornish-Bowden's suggestion and indicates that the data was available to predict the repetition in the collagen structure years before it was deduced from the exon pattern of the gene.

The value of $S\Delta Q_0$ for γ-crystallin is a little high for its size, but is much lower than would have been predicted for a protein with a four-fold repeat (Fig. 1). Another lens polypeptide, the βBp polypeptide of β-crystallin, shows sequence homology to γ-crystallin (Driessen *et al.*, 1981), and has been predicted to have an almost identical three-dimensional structure (Wistow *et al.*, 1981), and yet its value of $S\Delta Q_0$ falls exactly on Cornish-Bowden's line so that the four-fold repeat seen in the three-dimensional structure, in the amino acid sequence and, for another β-crystallin, in the gene (Inana *et al.*, 1983) would not have been predicted by the Cornish-Bowden approach. An anomalous value for $S\Delta Q_0$ might only be expected where homology of the repetitive sequences is strong, and

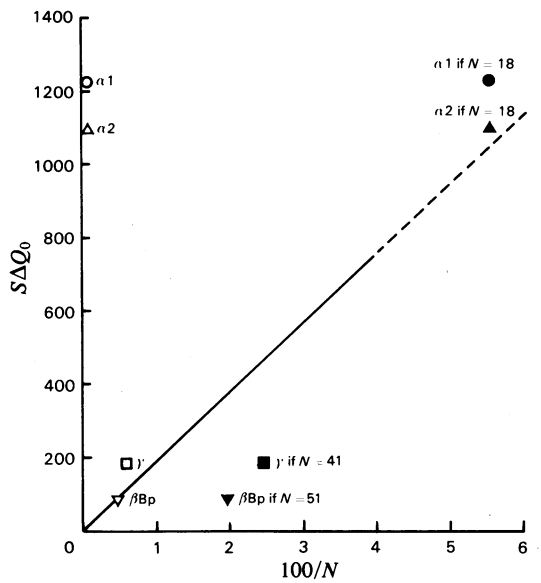

Fig. 1. *Correlation between amino acid composition and reciprocal number of residues*

The continuous line represents the median value of $N \cdot S\Delta Q_0$ and is taken from Cornish-Bowden (1983) and extrapolated (broken line). Values for $S\Delta Q_0$ for collagen α1 and α2 chains are plotted with the full residue numbers (α1, ○; α2, △); and assuming only 18 amino acid residues (α1, ●; α2, ▲). Values for $S\Delta Q_0$ for γII-crystallin and the βBp polypeptide of β-crystallin are plotted using the full residue numbers (γII, □; βBp, ▽) and assuming a four-fold repeat (γII, ■, βBp, ▼). Amino acid compositions are taken from Epstein *et al.* (1971) for the collagen chains, Croft (1972) for γII-crystallin and Driessen *et al.* (1981) for βBp polypeptide.

so the values for γ-crystallin and the βBp polypeptide are not altogether surprising because the sequence homology is weak (Driessen *et al.*, 1981). However, the sequence repeat in the collagen chains is not striking.

In spite of the weak result from the lens proteins the striking result with collagen should encourage those who find that their favourite protein has a $S\Delta Q_0$ index double the expected value, or greater, to look for other evidence of a repetitive structure.

John J. HARDING
Nuffield Laboratory of Ophthalmology, University of Oxford, Walton Street, Oxford OX2 6AW, U.K.

(*Received 13 July 1983*)

Blundell, T., Lindley, P., Miller, D., Moss, D., Slingsby, C., Tickle, I., Turnell, B. & Wistow, G. (1981) *Nature* (*London*) **289**, 771–777
Cornish-Bowden, A. (1983) *Biochem. J.* **213**, 271–274
Croft, L. R. (1972) *Biochem. J.* **128**, 961–970
Driessen, H. P. C., Herbrink, P., Bloemendal, H. & de Jong, W. W. (1981) *Eur. J. Biochem.* **121**, 83–91
Epstein, E. H., Scott, R. D., Miller, E. J. & Piez, K. A. (1971) *J. Biol. Chem.* **246**, 1718–1724
Inana, G., Piatigorsky, J., Norman, B., Slingsby, C. & Blundell, T. (1983) *Nature* (*London*) **302**, 310–315
Wistow, G., Slingsby, C., Blundell, T., Driessen, H., de Jong, W. & Bloemendal, H. (1981) *FEBS Lett.* **133**, 9–16
Wozney, J., Hanahan, D., Tate, V., Boedtker, H. & Doty, P. (1981) *Nature* (*London*) **294**, 129–135
Yamada, Y. (1980) *Cell* **22**, 887–892

The prediction of repetitive protein sequences from amino acid compositions: a comment

The very large value of $S\Delta Q_0$ for the collagen polypeptides noted by Harding (1984) is very striking, and, as he remarks, provides powerful support for the ideas I proposed earlier (Cornish-Bowden, 1983). I did not consider collagen chains previously because attention was restricted to the 118 'superfamilies' for which complete sequences were available from the most recent supplement of the *Atlas of Protein Sequence and Structure* (Dayhoff, 1978). However, if collagen had been included it would have displayed a more divergent value of $S\Delta Q_0$ than any of those that I did consider.

A crystallin chain (the A-chain of α-crystallin from kangaroo) was included in my study, but as noted by Harding (1984) it does not give a high enough value of $S\Delta Q_0$ to suggest the repetitive structure that it possesses. This is not in itself surprising, because even if large $S\Delta Q_0$ values provide strong evidence of a repetitive structure, as I have argued, it does not follow that all repetitive proteins will have large $S\Delta Q_0$ values; without a mechanism to generate very low $S\Delta Q_0$ values it would probably be rash to attach much significance to them. Nonetheless, the value for the βBp polypeptide of β-crystallin calculated by Harding (1984) is so low if one treats it as a 51-residue polypeptide that a plausible explanation would be welcome.

My paper (Cornish-Bowden, 1983) unfortunately contained several typographical errors, some of which seriously affected the sense. These were corrected in *Biochem. J.* (1983) **213**, following p. 770.

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

(*Received 14 October 1983*)

Cornish-Bowden, A. (1983) *Biochem. J.* **213**, 271–274
Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, National Biomedical Research Foundation, Silver Spring, MD
Harding, J. J. (1984) *Biochem. J.* **217**, 339–340