BIOCHEMICAL LETTERS JOURNAL

Failure of analysis of amino acid composition as a probe for sequence similarity in tetanus toxin

Tetanus toxin is a potent neurotoxin, whose mechanism of action is little understood (for review, see van Heyningen, 1986). It is encoded on a plasmid in *Clostridium tetani* and synthesized as a single polypeptide chain, but this is cleaved *in vivo* into two chains: a heavy chain (M_r 93000) and a light chain (M_r 52000) joined by a single disulphide bond. The complete nucleotide sequence of the toxin gene has recently been determined independently by two groups (Eisel *et al.*, 1986; Fairweather & Lyness, 1986). Results from the two groups were identical.

A few years ago, we (Taylor *et al.*, 1983) presented a hypothesis that the sequences of the heavy and light chains might have marked similarities; in particular, that the heavy chain was an approximate dimer of the light chain, suggesting a gene triplication event sometime in the evolution of the toxin. This suggestion was made, not on the basis of experimentally obtained amino acid sequences (which were then available for only small parts of the molecule), but on our observation that the amino acid compositions of the two chains were remarkably similar. Analysis of these compositions by the method of Cornish-Bowden (1979, 1983) satisfied his most stringent test for sequence similarity, suggesting very strongly that the heavy and light chains were related. (An index, $S\Delta n$, is calculated from the amino acid compositions to estimate the number of differences in sequence between two proteins of about the same molecular mass. If N is the total number of residues in the smaller of the two proteins, then, when $S\Delta n < 0.42N$, there is a strong indication that two proteins are related, when $0.42N < S\Delta n < 0.93N$, the indication is weak, and when $S\Delta n > 0.93N$, it is likely that the proteins are unrelated. $S\Delta n$ is calculated as shown in Table 1.)

Unfortunately, as reported by Eisel *et al.* (1986), there is no similarity between the two chains, or evidence of repetition within the heavy chain. We have made an extensive series of self-comparisons of the uncleaved polypeptide chain. 'Dot plots' were drawn using a wide range of window lengths, and scoring both exact matches and amino-acid similarities. These procedures are described and discussed by Collins & Coulson (1987). No significant repeats were detected in any of these plots.

To try to confirm these negative results, we also compared the two chains with each other, using the 'best local similarity' algorithm of Smith & Waterman (1981) [program GAP of the uwgcg package (Devereux *et al.*, 1984)]. In this case also, a variety of scoring schemes for amino acid compositions and indel penalties was used. Each scoring scheme produced a different pair of 'most similar' sequences, and this, combined with the weakness

Table 1. Comparison of the chains of tetanus toxin

$S\Delta n$ was calculated from the equation of Cornish-Bowden (1979, 1983):

$$S\Delta n = 0.5 (n_{iH} - n_{iL})^2 - 0.035(N_H - N_L)^2 + 0.535(N_H - N_L)$$

where N_H is half the number of residues in the heavy chain = 429 = N , and N_L is the number of residues in the light chain = 456. Thus $S\Delta n = 151.68$, $0.42N = 180.1$ and $0.93N = 398.97$.

	Number of residues from sequence			
	H-chain			
Amino acid	Whole	Half (n_{iH})	L-chain (n_{iL})	Difference squared [($n_{iH} - n_{iL}$) ²]
Ala	37	18.5	15	12.25
Arg	24	12	14	4
Asn	79	39.5	42	6.25
Asp	55	27.5	33	30.25
Cys	5	2.5	5	6.25
Gln	27	13.5	16	6.25
Glu	46	23	24	1
Gly	39	19.5	24	20.25
His	8	4	6	4
Ile	90	45	44	1
Leu	76	38	38	0
Lys	74	37	33	16
Met	12	6	12	36
Phe	33	16.5	23	42.25
Pro	29	14.5	24	90.25
Ser	68	34	31	9
Thr	49	24.5	24	0.25
Trp	12	6	1	25
Tyr	51	21.5	28	6.25
Val	44	22	19	9
Sum				325.5

of the individual alignments, implies that there are no significant similarities between the two chains.

It was possible that this error was not due to a breakdown in the theory, but rather to poor results in the experimental determination of the compositions. Table 1 shows the analysis of Cornish-Bowden worked out again using composition data got from the published sequences. Results with these 'correct' amino-acid compositions still meet the most stringent criterion, although less markedly so. A similar calculation was made to compare the *N*-terminal and *C*-terminal halves of the heavy chain from composition, and this gave a value of $S\Delta n$ of 765. This value is well outside the range within which the Cornish-Bowden criterion would suggest any relation between the two half-sequences.

We are grateful to the Medical Research Council for a grant (to S. van H.).

Simon VAN HEYNINGEN* and
Andrew F. W. COULSON†

*Department of Biochemistry, University of Edinburgh, Hugh Robson Building, George Square, Edinburgh EH8 9XD, and
†Department of Molecular Biology, University of Edinburgh, King's Buildings, Mayfield Road, Edinburgh EH9 3JR, U.K.

Collins, J. F. & Coulson, A. F. W. (1987) in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach* (Bishop, M. & Rawlings, C., eds.), IRL Press, Oxford, in the press
Cornish-Bowden, A. (1979) *J. Theor. Biol.* **76**, 369–386

Cornish-Bowden, A. (1983) *Methods Enzymol.* **91**, 60–75
Devereux, J., Haeberli, P. & Smithies, O. M. (1984) *Nucleic Acids Res.* **12**, 387–395
Eisel, U., Jarausch, W., Goretzki, K., Henschen, A., Engels, J., Weller, U., Hudel, M., Habermann, E. & Niemann, H. (1986) *EMBO J.* **5**, 2495–2502
Fairweather, N. F. & Lyness, V. A. (1986) *Nucleic Acids Res.* **14**, 7809–7812
Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197
Taylor, C. F., Britton, P. & van Heyningen, S. (1983) *Biochem. J.* **209**, 897–899
van Heyningen, S. (1986) in *Pharmacology of Bacterial Toxins* (Dorner, F. & Drews, J., eds.), pp. 549–569, Pergamon Press, Oxford

Received 10 December 1986