Convergent evolution of lysozyme sequences?

SIR—As Stewart *et al.*[1] have pointed out, there are theoretical reasons to expect the evolution of protein sequences to be divergent, and there exists abundant experimental evidence from many kinds of protein to suggest that it is. The theoretical argument applies with particular force to the bacteriolytic function of lysozyme, which can be imitated with random copolymers of glutamate and phenylalanine[2,3]. As there are many more than twenty families of proteins for which at least six sequences are available for comparison, one ought to be surprised if one of them did not display characteristics with a likelihood less than 5%, and so to be convinced of the reality of convergence in the evolution of sequences one can hardly be satisfied with a result significant in a 95% confidence test.

The sequences of lysozyme from langur, baboon, human, rat, cattle and horse contain four loci at which the langur and cattle enzymes share identical residues not found in any other vertebrate lysozymes apart from other ruminant or colobine stomach enzymes, whereas other pairs from the six sequences share either one such identity or none[1]. As the seven identities could in principle be distributed at random among 15 pairs of sequences, one can readily calculate the probability that four or more would occur by chance in the same pair of sequences as $15p^4(p^3 + 7p^2q + 21pq^2 + 35q^3) = 0.0088$, where $p = 1/15$ and $q = 14/15$. This is smaller than implied by the reported significance in a 95% test[1], but is too large to exclude the possibility that one is dealing with more than the expected tail of the distribution.

When the lysozyme sequences were analysed on the assumption that the true phylogenetic tree linking them is the ordinary biological tree in which the langur is more closely related to the baboon than to cattle, five amino-acid substitutions were placed on the lineage leading to langur from the ancestor unique to langur and baboon. The probability of observing as many as five on this lineage was also assessed as less than 5%. But the particular placing of substitutions represents an interpretation of the observations, not the observations as such, and the putative substitutions cannot be analysed statistically as if they were actually observed. At locus 50, for example, the occurrence of glutamate in cattle and langur but glutamine in the rat and baboon was interpreted as two

parallel replacements of glutamine by glutamate in the langur and cattle lineages, though the data were also consistent with two parallel replacements of glutamate by glutamine in the rat and baboon lineages. It could, of course, be argued that it is more parsimonious to take glutamine rather than glutamate as ancestral, because then glutamine in the rat can become glutamate in the ancestor of horse and cattle lysozymes, which requires only one base change to give glycine in the horse and none to retain glutamate in cattle. But this imples that the appearance of glutamate in the ancestor of the horse and cattle was not an adaptation to foregut fermentation, and makes it difficult to argue simultaneously that it was adaptive in the langur. In any case, the worse-than-random performance of methods based on the genetic code in cases where they can be checked[4] suggests that it is safer simply to count amino-acid differences.

At first glance, the six sequences appear to contain abundant data that led to the initial observation that an unexpected tree showing the langur and cattle as close relatives required as few substitutions as the biologically reasonable tree. In reality, however, this result was derived from data at very few loci. Of 130 loci altogether, 47 show no variation and are not used by any method of tree construction. A further 63 show variations of the kind where any amino acid other than the main one occurs once only. Such variations contribute to tree-construction methods such as UPGMA[5] that are based on the table of differences, but not to minimum-length ('parsimony') methods if no assumptions are made about the likely evolutionary route from one amino acid to another[6]. Only 20 lead to variations in the lengths of the possible trees linking the 6 sequences, and 7 of these (loci 23, 37, 41, 88, 90, 117 and 119) do not discriminate between the two trees of particular interest. Even considering only those loci that are ignored by the minimum-length approach, UPGMA generates a tree that relates the langur and baboon sequences closely to one another, less closely to the human sequence, and more distantly to the others.

ATHEL CORNISH-BOWDEN
Centre de Biochimie et de Biologie Moléculaire,
CNRS, 31 chemin Joseph-Aiguier, BP 71, 13402 Marseille Cedex 9, France

STEWART *ET AL.* REPLY — We do not agree with Cornish-Bowden's statistical criticism. Our lysozyme sequences[1] were not sampled at random as he assumes. Rather, we chose representatives from two lineages that convergently lost an old function while acquiring the same set of new functions. We also required that the times when the new functions arose be recent enough to avoid the problem of multiple hits at the same amino-acid site, yet old enough to let a significant number of amino-acid replacements accumulate. The lineages leading to langur and cow stomach lysozymes meet these requirements and they were the first functionally convergent pair to be analysed explicitly with the approach we described[1]. Thus, there are no empirical grounds for regarding this set of lysozyme sequences as the tail of a distribution based on numerous other protein sets, each containing a pair of functionally convergent sequences that exhibit no convergence in primary sequence. For these reasons, we continue to consider stomach lysozymes to provide the most plausible case available for convergent evolution of protein primary structures in response to known new selection pressures impinging at known times; nevertheless, we admit that the case is not yet conclusive.

Cornish-Bowden raises two other points that reflect his unwillingness to adopt methods of analysing evolutionary history and his preference of distance methods. This reticence may stem in part from his attempt to analyse variation at position 50 in six mammalian lysozymes; if his analysis had included the other lysozymes of known sequence, as ours did[1], it would have been clear that convergence is the simplest explanation for the uniquely shared residues at this position. His preference for UPGMA is misguided because this method does not generate evolutionary trees; it is a purely phenetic method, which merely summarizes the distances between the sequences[5]. The parsimony method we used represents a genealogical analysis that takes account of the nature and locations of the sequence differences and of variation in evolutionary rates among lineages. It is only through parsimony analysis that parallel and convergent events can be distinguished from divergent ones. If investigators were to follow Cornish-Bowden's recommendation "simply to count amino-acid differences" thus precluding genealogical analysis, their ability to study the evolutionary process would be severely limited.

CARO-BETH STEWART
Hormone Research Institute, University of California at San Francisco, San Francisco, California 94143, USA
JAMES W. SCHILLING
California Biotechnology, Inc., Mountain View, California 94043, USA
ALLAN C. WILSON
University of California at Berkeley, Berkeley, California 94720, USA

1. Stewart, C.-B., Schilling, J.W. & Wilson, A.C. *Nature* **330**, 401–404 (1987).
2. Naithan, V.K. & Dhar, M.M. *Biochem. biophys. Res. Commun.* **29**, 368–372 (1967).
3. Gutte, B., Moser, R., Klauser, S. & Weilenmann, M. in *Dynamics of Biochemical Systems* (eds Ricard, J. & Cornish-Bowden, A.) 259–272 (Plenum, New York, 1984).
4. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, 1983).
5. Sneath, P.H.A. & Sokal, R.R. *Numerical Taxonomy* (Freeman, San Francisco, 1973).
6. Cornish-Bowden, A. *J. theor. Biol.* **101**, 317–319 (1983).
7. Stewart, C.-B. & Wilson, A.C. *Cold Spring Harb. Symp. quant. Biol.* **52**, 891–899 (1987).