J. theor. Biol. (1981) **91**, 379–380

LETTERS TO THE EDITOR

Overlapping Genes and Information Theory†

The discovery of overlapping genes in the viruses ϕX174 and SV40 has generated a number of interesting questions. Here we are concerned with the contribution of information theory to the analysis of overlapping genes. Several papers have dealt with this connection (Figueroa, *et al.*, 1977; Sander & Schulz, 1979; Yockey, 1979; Smith & Waterman, 1980; Fitch & Siegel, 1980). The purpose of this letter is to comment on Yockey's 1979 application of information theory to these problems.

Yockey states: "It is clear that . . . there is not enough information in a DNA or RNA sequence to code two overlapping protein sequences each of which has the information content or biochemical specificity of an 'average protein'." The idea underlying this sentence is that the information $I(\mathbf{F}_1)$ associated with the protein coded for in the main frame $\mathbf{F}_1$ and the information $I(\mathbf{F}_2)$ associated with the protein coded for in another frame $\mathbf{F}_2$ are additive; that is

$$I(\mathbf{F}_1 \,\&\, \mathbf{F}_2) = I(\mathbf{F}_1) + I(\mathbf{F}_2).$$

The information of the double-coding DNA sequence is denoted by $I(\mathbf{F}_1 \,\&\, \mathbf{F}_2)$. If, however, $I(\mathbf{F}_1)$ and $I(\mathbf{F}_2)$ are both near their maximum possible values, then the value of the sum $I(\mathbf{F}_1) + I(\mathbf{F}_2)$ would exceed the information of the DNA sequence. Thus Yockey writes, "Consequently, in practice, to accommodate the genetic code, it must be expected that the sum of the information content of the transcribed overlapping sequences will be considerably less than the maximum information content of the DNA or RNA sequence". He shows that this is indeed true for cytochrome *c* sequences.

The difficulty with this analysis is that

$$I(\mathbf{F}_1 \,\&\, \mathbf{F}_2) = I(\mathbf{F}_1) + I(\mathbf{F}_2)$$

holds only when $\mathbf{F}_1$ and $\mathbf{F}_2$ are probabilistically independent. Of course, $\mathbf{F}_1$ and $\mathbf{F}_2$ are highly dependent, a fact recognized in Yockey's paper, but not utilized. This oversight underlies much of Yockey's analysis and conclusions.

† This work was performed under the auspices of the U.S. Department of Energy.

0022-5193/81/140379+02 $02.00/0

The correct formula is

$$I(\mathbf{F}_1 \,\&\, \mathbf{F}_2) = I(\mathbf{F}_1) + I(\mathbf{F}_2|\mathbf{F}_1),$$

where $I(\mathbf{F}_2|\mathbf{F}_1)$ is the conditional information of frame 2 given frame 1. Such conditional information is calculated from conditional probabilities of codons in frame 2 given codons in frame 1. An analysis based on this approach is given by Smith & Waterman (1980) and, among other things, can be used to indicate certain reading frames that should be very rare in nature.

Non-independence of sequence elements is a general problem and has long been studied. Within a single protein or nucleic acid sequence, the sequence elements are not order-independent. These first-order dependencies have been used to find the information associated with such sequences (Smith, 1969; Gatlin, 1972). Here, and in Smith & Waterman (1980), we are suggesting that when overlapping reading frames are studied, independence, in particular, cannot be assumed even as a first approximation.

Los Alamos Scientific Laboratory
University of California,
Los Alamos, New Mexico 87545, *U.S.A.*

T. F. SMITH
M. S. WATERMAN

(*Received* 15 *August* 1980, *and in revised form* 17 *December* 1980)

REFERENCES

FIGUEROA, K., SEPULVEDA, A., SOTO, M. A. & TOHA, J. (1977). *Naturforsch.* **32c,** 850.
FITCH, W. & SIEGEL, A. (1980). *Math. Biosci.* **49,** 27.
GATLIN, L. L. (1972). *Information Theory and the Living System.* New York: Columbia University Press.
SANDER, C. & SCHULZ, G. E. (1979). *J. mol. Evol.* **13,** 245.
SMITH, T. F. (1969). *Math. Biosci.* **4,** 179.
SMITH, T. F. & WATERMAN, M. S. (1980). *Math. Biosci.* **49,** 17.
YOCKEY, H. P. (1979). *J. theor. Biol.* **80,** 21.