THE AMOUNT OF INFORMATION STORED IN PROTEINS AND OTHER SHORT BIOLOGICAL CODE SEQUENCES

THOMAS A. REICHERT
CARNEGIE-MELLON UNIVERSITY

1. Introduction

These remarks were made to the conference assembly in a context not unlike that of a surprise witness for the defense. This work was so hot off the press that there had been no time to communicate it before the conference itself. I am grateful to Dr. Lila Gatlin for the opportunity to make this presentation. All of the work to be discussed here has been done in collaboration with A. K. C. Wong, also of the Biotechnology Program at Carnegie-Mellon University.

In the last year, we have developed a measure of the amount of information required to perform genetic mutations, together with an algorithm utilizing these measures, for aligning amino acid and RNA code sequences [9], [11]. In the process of this development, we attempted to calculate the amount of information which was stored in a protein's amino acid sequence. We had, at the time, only the tools of the conventional communications form of information theory. Thus, we attempted the calculation using the two expressions:

$$H = -\sum_{i=1}^{a} p(i) \log p(i) \tag{1}$$

or

$$I_{\text{self}} = -\sum_{i=1}^{a} n_i \log p(i). \tag{2}$$

Equation (1) is the expression for the entropy of a discrete information source operating with an alphabet of a letters. This quantity is also interchangeably called the information content of such a source. Since information and entropy are more nearly opposites than synonyms, this equivalence has always been confusing. Indeed, the values of H obtained for a set of cytochrome c sequences, by allowing the amino acid frequencies in each sequence to determine the alphabet character probabilities used in equation (1), displayed the then embarrassing trend of higher information content with lower organism complexity. Since the method used to estimate these probabilities is not generally known, let me describe it here.

It was Laplace, I believe, who first noted that the frequency limit estimate of the probability of an event's occurrence was applicable only in the limit of infi-