Lower Limit to the Size of the Primeval Amino Acid Alphabet

Ariel Fernández

Institute for Biophysical Dynamics, Department of Computer Science, Ryerson Hall,
The University of Chicago, Chicago, Illinois 60637.
Permanent address: Indiana University School of Informatics, Center for Computational
Biology and Bioinformatics, 714 N. Senate Ave., Indianapolis, IN 46202. Fax: 3172789217.
E-mail: ariel@uchicago.edu

Z. Naturforsch. **59c**, 151–152 (2004); received June 16/July 17, 2003

Here I systematically examine the information complexity of all primary sequences of natural proteins deposited in the Swiss-Prot database. The sequence complexity is assessed by determining the frequency of occurrence of each amino acid type on sequence windows of fixed length, calculating the Shannon entropy of the window and then averaging over all windows covering the sequence. The minimum value in information content obtained from the present-day record imposes a lower limit in the number of letters that a primeval amino acid alphabet must have had.

Key words: Genetic Code, Amino Acid Alphabet, Translation

Introduction

Studies on the origin of the genetic code led scientists to propose that early protein synthesis carried out by a primeval translation machinery must have involved a simplified or shorter amino acid alphabet (Crick, 1968; Osawa and Jukes, 1989). There are also burgeoning efforts to generate functional folds topologically equivalent to those of known present-day structures using a simplified or reduced set of amino acid types (Akanuma *et al.*, 2002). Thus, I may address the question: Is there evidence in the present record of natural protein sequences revealing how small the primeval amino acid alphabet may have been? In essence I am asking what is the minimal alphabet that can reproduce the complexity of even the simplest present-day sequences.

This basic question will only be answered in part here by determining a lower limit on the size of the primeval alphabet. This lower limit may be estimated by adopting a measure of primary sequence complexity that takes into account the occurrence of each amino acid type and then examining how complexity is distributed on a vast – ideally exhaustive – database of natural sequences. My quest is limited in scope in that I cannot make statements on the actual evolution of the code but rather assess how simple the precursor amino acid alphabet may have been.

Methods

For the purposes of this study I adopt the Swiss-Prot database (Bairoch and Apweiler, 1999), which has the most significant contribution (~ 6%) from *Archaea.* A measure of sequence complexity is provided by the Shannon information content (Romero *et al.*, 2001). This measure was first introduced in information theory of communication to assess the complexity of a message given by a sequence of letters constructed from a fixed alphabet (Shannon, 1946). Thus, if the message consists of a single repeated letter (trivial), we would expect zero complexity, while the maximum complexity is to be achieved when all letters in the alphabet appear in the message with equal frequency $1/W$, where W is the size of the alphabet. Notice that the maximum complexity corresponds to the maximum uncertainty – or minimum *a priori* probability – in regards to which letter will appear at any given position on the sequence. On the other hand, a zero complexity implies that at any given place a fixed amino acid is found with absolute certainty, implying a logarithmic relation between information complexity and probability.

Given these premises, the form of the complexity measure σ is uniquely defined as minus the expected value of the logarithm of the probability p of finding a particular amino acid at a particular position: $\sigma = -<\log p>$, as shown in Feinstein (1958). Thus, the information content or complexity, σ_L, of arbitrary windows of length L along a given amino acid sequence (W = 20 if we assume only natural amino acids) is given as:

$$\sigma_L = - \sum_{i = 1,2,\ldots,20} (n_i/L)\log_2(n_i/L),$$

where the index i labels each of the 20 amino acid types and n_i is the number of amino acids of type i which occur within the window of length L. Thus, n_i/L indicates the frequency of occurrence of amino acid i in the L-window. The complexity σ of an entire sequence is then determined by averaging over all the L-subsequences obtained by sliding the L-window iteratively along the sequence by one amino acid at a time.

Results and Discussion

Irrespective of window length, the maximum information complexity possible is $\sigma = \sigma_L = \log_2 20 \sim 4.32$, corresponding to a randomly generated sequence (1/20 probability of finding any one of the 20 amino acids at any given site). This level of complexity is never realized within the present-day database, independently of the window adopted for sequence interrogation.

The natural complexities found after exhaustive interrogation of the Swiss-Prot database lie invariably within the range $2.807 < \sigma < 4.243$, for all lengths L lower or equal to 20 investigated. The relative abundance of windows of different complexities is given in Fig. 1. I collected statistics on distribution of complexity adopting eight windows sizes: L = 20, 30, 40, 45, 50, 55, 60 and 65. The sizes L = 45, 60 (Fig. 1) produce the broadest dispersion in the distribution, actually realizing the empirical limits 2.807 and 4.243 as the minimum and maximum, respectively.

At this point I may address the problem of finding a minimal alphabet that would be needed to produce the lowest possible complexity $\sigma = 2.807$. Since the maximum complexity that may be achieved with an alphabet of M letters is $\log_2$M, and given that the lowest complexity found is $\sigma_{45} = \sigma_{60} = 2.807$, we may conclude that there are no traces in the existing record that a primeval amino acid alphabet could have had less than seven letters. This is so because of the arithmetic inequalities: $2.807 < \log_2 7 < 2.808$ and because $\log_2 7$ is the maximum complexity that a seven-letter alphabet can achieve.

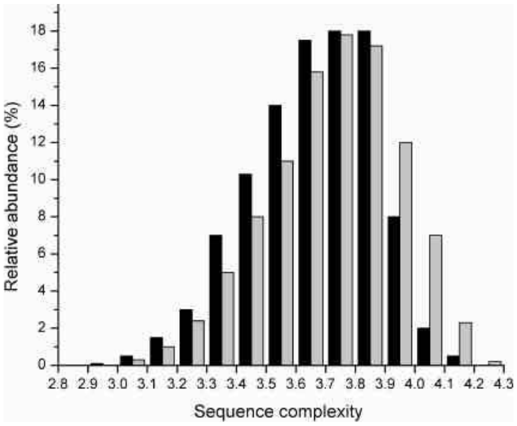

Fig. 1. Relative abundance of sequences grouped according to their complexities. The window lengths are fixed at L = 45 (black bars) and L = 60 (gray bars). The windows are grouped in intervals of σ_L-values of length 0.1, partitioning the range of complexities $2.8 < \sigma_L < 4.3$ that holds for all lengths L less or equal to 20 investigated. Actually, no value higher than 4.243 or lower than 2.807 has been found for any window length. Only 10^{-4}% of sequences lie in the interval $2.8 < \sigma_L < 2.9$. The Swiss-Prot database was exhaustively examined: all 28,740,215 windows with L = 45 and all 25,231,084 windows with L = 60 were interrogated.

It is worth emphasizing that I am not claiming the present-day 20-letter alphabet evolved from a reduced set of seven letters. The original alphabet might have contained 7 or more amino acids. What is rigorously true is that the most rudimentary level of complexity found in the present record requires at least 7 letters to span it, and thus the primeval alphabet should have had at least 7 letters.

Akanuma S., Kigawa T., and Yokoyama S. (2002), Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc. Natl. Acad. Sci. USA **99**, 13549–13553.

Bairoch A. and Apweiler R. (1999), The SWISS-PROT sequence databank and its supplement TrEMBL in 1999. Nucleic Acids Res. **27**, 49–54.

Crick F. H. C. (1968), The origin of the genetic code. J. Mol. Biol. **38**, 367–379.

Feinstein F. (1958), Foundations of Information Theory. McGraw Hill, New York.

Osawa S. and Jukes T. H. (1989), Codon reassignment (codon capture) during evolution. J. Mol. Evol. **28**, 271–278.

Romero P., Obradovic Z., Li X., Garner E. C., Brown C. J., and Dunker A. K. (2001), Sequence complexity of disordered protein. Proteins **42**, 38–48.

Shannon C. E. (1946), A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423.

Satz und Druck: AZ Druck und Datentechnik GmbH, Kempten