FORMAL LANGUAGE THEORETICAL APPROACH TO INTRACELLULAR BEHAVIOR

VAHE BEDIAN and GABOR T. HERMAN
Departments of Biophysics and Computer Science
The State University of New York at Buffalo

The fact that a gene in the DNA is a string over the alphabet of 64 codons, which describes a string over the alphabet of 20 amino acids is a great scientific discovery with an obvious relationship to formal language theory. We have begun the process of translating into formal language theoretical terminology the concepts associated with the fact that an organism carries a description of its developmental rules in the DNA of every cell. It is our hope that we shall thus be able to formulate precisely (and possibly answer) some problems related to the origin of life: i.e., how the unique, arbitrary genetic code utilized in organisms today could have first arisen.

For example, one may start with the following definitions.

A molecular soup is a 4-tuple $M = \langle A, C, E, G \rangle$, where

A is a finite nonempty set of constructor units (amino acids),

C is a finite nonempty set of descriptor units (codons),

$E \subset A^+$ is the set of constructors (enzymes),

$G \subset C^+$ is the set of descriptors (genes),

such that $\#A \leq \#C \leq \#E \leq \#G$.

A <u>production scheme</u> is a 7-tuple $S = \langle A, C, E, G, F, f_1, f_2\rangle$, where

$\langle A, C, E, G\rangle$ is a molecular soup, the <u>soup</u> of S,

$F \subset E$, $\#F = \#C$,

f_1: $F \to C$,

f_2: $C \to A$.

f_2 can be extended to an f_3 so that f_3: $G \to A^+$, in the usual way.

A <u>production scheme</u> $S = \langle A, C, E, G, F, f_1, f_2\rangle$ is said to be <u>self-coding</u> if and only if

(i) f_1 is one-to-one onto,

(ii) f_2 is onto A,

(iii) f_3 is onto F.

<u>Example</u>. Consider the following molecular soup. $M = \langle A, C, E, G\rangle$, where $A = \{a, b\}$, $C = \{0, 1, 2\}$, $E = \{ab, bab, aa, b\}$, $G = \{02, 212, 10, 0, 11\}$. Let F, f_1, f_2, $\bar{F}$, $\bar{f}_1$ and $\bar{f}_2$ be defined as follows.

$F = \{ab, bab, b\}$,

$f_1(ab) = 0$, $f_1(bab) = 1$, $f_1(b) = 2$,

$f_2(0) = b$, $f_2(1) = a$, $f_2(2) = b$,

$\bar{F} = \{ab, bab, aa\}$,

$\bar{f}_1(ab) = 0$, $\bar{f}_1(bab) = 1$, $\bar{f}_1(aa) = 2$,

$\bar{f}_2(0) = a$, $\bar{f}_2(1) = a$, $\bar{f}_2(2) = b$.

It is easy to show that both $\langle A, C, E, G, F, f_1, f_2\rangle$ and $\langle A, C, E, G, \overline{F}, \overline{f}_1, \overline{f}_2\rangle$ are self-coding production schemes whose soup is M.

Using such definitions we can now state precisely problems like the following: "Characterize those molecular soups for which there is one and only one self-coding production scheme of which it is the soup." (This question is related to the uniqueness of the genetic code in living organisms. However, this might not be the only possible way in which a unique code may arise. A more general question is: "How do the formal, time-independent relationships proposed in our definition constrain the real-time, dynamical behavior of a molecular soup, to result in a unique code?")

Because of the finiteness of the domain and range of f_1 and f_2, it is of course decidable for any molecular soup whether or not it is the soup of a unique self coding production scheme.

The systems we have considered so far are quite simple, but they indicate the type of approach we have in mind. We plan to continue to work to improve these definitions so that one can ask really meaningful questions about the origin of the genetic code.

Acknowledgements. The research for this paper has been supported by NSF grant GJ 998. The authors are grateful to Professor Howard Pattee and Silvano Colombano for illuminating discussions.