FUNCTIONAL RELEVANCE OF TRIPLET POSITIONS

Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences, Moscow, Russia

Immauel Kant Baltic Federal University, Kaliningrad, Russia

SUREN ZOLYAN

Institute of Philosophy, Sociology and Law, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia

Abstract

We consider the functional significance of triplet positions of the genetic code. We assume that the genetic code can be described as a language consisting of grammar and vocabulary. Nucleotides perform functions of lexical entities, and positions act as grammatical categories. Each of the three positions has specific structural and functional characteristics, and each nucleotide has its own semantic and syntactic profile.

.

Introduction

I. The combination of the principles of contextual dependence and arbitrariness of sign leads to the conclusion that the primary elements of the genetic code (nucleotides) can be considered not as biochemical constants, but as semiotic, or, more precisely, grammatical variables. Nucleotides perform different functions depending on their position within a triplet. Researchers have already noted this inequality of first, second and third positions and designated it linguistic terms: as a prefix, root, ending, respectively (Rumer 1966; Ratner 1993); or subject, predicate, complement (Lopez-Garcia 2005). The triplet's positions perform a function of a grammatical category: regardless of what this position is filled with, it performs the function assigned to it. Therefore any nucleotide N in the genetic code appears in three forms: N_1; N_2; N_3.

II. The functional distinction between units of the dictionary (nucleotides) and categories of grammar (positions within triplet) allows to identify the formation rules for the meaningful units of the genetic code (doublets and triplets) and explicate their compositional semantics. Instead of linear context-free linguistic models (cf: Ratner 1993; Searls 1999, 2002; 2010; Ji 1997, 1999; Gimona 2006), we suggest to use a three-level categorial grammar, where items are considered context-dependent variables and, simultaneously, context-forming operators.

Acknowledgements

This research is supported by the grant from the Russian Science Foundation (RSF), project no. 17-18-01536 “Knowledge transfer and convergence of methodological practices: Cases of interdisciplinary integration of political, biological, and linguistic research” at the Institute of Scientific Information on Social Sciences of the Russian Academy of Sciences.

Methodology

The grammar of the genetic code determines the formation of triplets (codons). – see Appendix 1. Thus, the crucial distinction is drawn between units of the vocabulary (nuclei acids) and the categories of grammar: the empty positions within triplets (first, second, third), each of them is endowed by its codon-forming functions regardless of which nucleotide it is filled with. From this point, the functional relevance of a nucleotide, or its category, is not its substantial characteristic but is determined by its position within a codon. As one can see, it is not a nucleotide by itself that is important, but the position it occupies: uracil in the first, second and third positions can perform entirely different functions - in the second position it selects a specific class of 5 amino acids and start-codon (Met, Lie, Val; Phe, Leu), in the first position it selects from a particular class given by the second position a specific amino acid:

U_1 - if CA_2 – then Tyr or stop-codon;

if C_2 - then serine;

if U_2 - then Phe or Leu;

if G_2 – then Cys, or Trip, or stop-codon.

In the third position U_3 may have not any distinctive capacity (U_1C_2 N_3), or, in case of either purine R (A or G) or pyrimidine Y (C or U) provide distinction between two amino-acids purines R and pyrimidines (U_2U_2), or between amino-acids and stop - or start-codons. $U_1U_2R_3$ -> Leu, $U_1U_2Y_3$ -> Phe; $U_1A_2R_3$ -> Stop; $U_1A_2R_3$ -> Tyr.

REFERENCES

Crick F.H, Griffith J.S, Orgel L.E. Codes without commas // Proc. Natl. Acad. Sci. U S A. –1957. – Vol. 43(5). – P. 416–421.

Ji, Sunghul. 1997. Isomorphism between cell and human languages: Molecular biological, bioinformatic and linguistic implications. Biosystems 44(1). 17–39.

Ji, Sunghul. 1999. The linguistics of DNA: Words, sentences, grammar, phonetics, and semantics. Annals of the New York Academy of Science 870. 411–417.

López-García, Á. 2005. The grammar of genes: How the genetic code resembles the linguistic code. Bern: Peter Lang.

Ratner, Vadim. (1993). The genetic languge; grammar, sentences, evolution. Genetika, 29. 709–719. (in Russian).

Ratner V.A. (2000) The Chronicle of the great discovery: ideas and persons]. *Priroda.* 2000. No. 6, pp. 22–30 (in Russian).

Rumer Y B. (1966) About the codon’s systematization in the genetic code. Proc. Acad. Sci. USSR (Doklady) 167, 1393–1394. (in Russian)

Searls, David. 1999. Formal language theory and biological macromolecules. Series in Discrete Mathematics and Theoretical Computer Science 47. 117–140.

Searls, David. 2002. The language of genes. Nature 420(6912). 211–217.

Searls, David. 2010. Molecules, languages, and automata. Lecture Notes in Computer Science 6339. 5–10.

Zolyan, S., (2018) The Genetic code: Grammar, semantics, evolution // METHOD: Moscow Yearbook of Works from Social Sciences / Russian Academy of Sciences, INION. v 8. p. 130 - 184 (In Russian.)

Zolyan, S., Zdanov, R. (2018) Genome as (hyper)text: From metaphor to theory - Semiotica, vol. 2018 – N 6, p. 1 – 18; https://doi.org/10.1515/sem-2016-0214 (with R. Zdanov)

Results

A. FUNCTIONAL CHARACTERISTICS OF POSITIONS

1. The principle of context-sensitivity allows describing cases when biochemically same nucleotides, depending on their location, acquires a different meaning and performs a different function. Regardless of which nucleotide it is filled with, the positions perform the following functions:

1) distinctive: the order of the nucleotides (first, second, and third positions) distinguishes the semantics of one sequence from another. The positions within the triplet are categories with their semantic-syntactic functions: the second position determines some group of amino acids, the first one - identifies the specific amino acid within it.

2) delimitative: the third position marks the end of a three-element sequence of nucleotides, correlated with a particular amino acid. For the half of triplets, the third position plays only a delimitative role, for the other half, both the delimitative and the distinctive;

3) structural - it relates esp. to the third position. In half of the cases, it is redundant from the semantic point of view, but is necessary as a structural unit, since it complements the doublet to the required triplet structure;

4) selective-syntagmatic function: it is performed by all three positions when, in the next stage, a complementary pair ("codon-anticodon") is formed. In the so-called "wobbling" situation, the third position may lose selective characteristics and does not determine which nucleotide of the first anticodon’s position will be attached.

B. OPERATIONAL PROFILES OF NUCLEOTIDES

The distinction between vocabulary (nucleotides) and categories of grammar (empty positions within triplet) allows to identify the formation rules for the significant units of the genetic code (doublets and triplets) and explicate their compositional semantics (correspondence rules between codons and amino acids). The principle of context-sensitivity allows describing cases when biochemically the same sequence of nucleotides, depending on their location, acquires a different meaning, and performs a different function. This explains why sequences of the same nucleotides but ordered differently are associated with different amino acids AUG =/=GUA =/=UAG, as actually they consist of functionally different operators $A_1U_2G_3$ =/=$G_1U_2A_3$ =|/=$U_1A_2G_3$: in this notation, this becomes obvious, that this sequences consisted of completely different elements despite these elements are manifested as the same nucleotides.

As a result, the grammar of the genetic code may be represented as a system of operations or one-side dependencies, regulated by left or right contexts of the nucleotide in the central positions. Thus, any for any nucleotide N_2, there are possible to be in 16 contexts. This allows us to elucidate the individual profile for each of the nucleotides. Any of them is associated with some class of amino acids, and in a few cases with nonsenses. – see: Appendix 2.

Crick called the genetic code a code without commas , that is, without delimiters (Crick et al. 1957), but this is not so - with the three-element reading frame and codon, there is no need for a special element to perform this function..

APPENDIX 1

THE CATEGORIAL GRAMMAR OF THE GENETIC CODE

The grammar of the genetic code determines the formation of triplets (codons). – see Appendix 1. It consists of two variables X and Y; they can be interpreted as any nucleotide in the initial position (X and the resulting codon Y. These micro-grammar rules describe the derivation of Y from X based on the "empty" positions. These rules are represented as the X's left and right context – they are operators (or functions) correlating this X with Y. Correspondingly, the initial central position may be mentioned as the basic unit X, the left context (the first position) as an operator which transformed X into a doublet and it is assigned the category "X / (X/У)" (something, which in conjunction with X generates "(X/У)": (X; X / (X/У)) → (X/У). The third position may be considered as an operator (X/У)У, in conjunction with the doublet (X/У) transforming it into the triplet Y: (X/У); (X/ У)У →Y

Thus, the formation of the triplet is identified as three steps compositions:

1) X

2) X; (X/Y)

3) (X; X/Y) Y

The rules of grammar are separated from the is the basic dictionary, it consisted of four units - nucleotides A (adenine), U (uracil), G (glycine), C (cytosine). Besides the initial position, any of them can stand for the other positions and assign the categories $N_{.X.;}$ $N_{(X;\ (X/Y)}$: $N_{(X;\ X/Y)\ Y,}$ or, as the categories coincide with positions within codons correspondingly, $N_{2;}$ $N_{1;}$ N_3

APPENDIX 2

C_2 correlates with 4 amino acids.

Left context (1 position) Right context (3-rd position)

1. A_1 C_2 – => Treonin; N_3 (non- relevant)
2. G_1 C_2– => Alanine; N_3 (non- relevant)
3. U_1 C_2 – => Serine ; N_3 (non- relevant)
4. C_1 C_2 – => Proline N_3 (non- relevant)

A_2 - is associated with 7 amino acids and stop codon:

1. A_1 A_2 R_3 => Lys;
2. A_1 A_2 Y_3 => Asn;
3. G_1 A_2 R_3 => Glu ;
4. A_1 A_2 Y_3 => Asp ;
5. U_1 A_2 R_3 => Stop;
6. U_1 A_2 Y_3 => Tyr ;
7. C_1 A_2 R_3 => Gln ;
8. C_1 A_2 Y_3 => His;

The **G_2** is associated with 5 amino acids and stop codon:

(A_1 G_2 Y_3 - => Ser, A_1 G_2 R_3 => Arg; G_1 G2 N_3- => Gly; C_1 G_2N_3 => Arg); U_1 G_2 Y_3 => Cys, U_1 G_2 R_3 G_3 => Trip, G_2 R_3 A_3 => Stop;

U_2 is associated with 5 amino acids and start codon:

U_1 U_2 Y_3 => Phe ; U_1 U_2 R_3 => Leu; C_1 U_2 N_3 => Leu ; G_1 U_2 N_3 => Val;

A_1 U_2 Y_3=> Ile, A_1 U_2 A_3=> Lie; A_1U_2 G_3 => Met, Start;

1) U_2

2) G_1U_2 => valin vs (non-G $)_1$ U_2.

3) (non-G $)_1$ U_2 => C_1 => Leu vs (non-G $)_1$&(non-G $)_1$ U_2

4) (non-G $)_1$&(non-G $)_1$(~C)U_2 =>$U_1U_2N_3$ vs $A_1U_2N_3$

5) $U_1U_2N_3$=>$U_1U_2Y_3$ => Phe vs $A_1U_2R_3$ =>Leu

6) $A_1U_2N_3$ => $A_1U_2Y_3$& $A_1U_2A_3$ => Lie vs $A_1U_2G_3$

7) $A_1U_2G_3$ => Met vs Start (in the beginning)