International Journal for Computational Biology (IJCB)
Vol.3, No.1, April 2014, Page: 27-30
ISSN: 2278-8115 ❐ 27

A Mathematical Model of Central Dogma of Molecular Biology employing a Novel Irrational-Integral-Imaginary (i3) Encoding and Numerical Approximation based on Cellular Automaton

Praharshit Sharma[1]*, Bhupendra Kumar Pathak[2], and Tiratha Raj Singh[3]
[1]Bioclues Organization, Kukatpally, Hyderabad, Andhra Pradesh – 500072, India
[2]Department of Mathematics, [3]Department of Biotechnology and Bioinformatics,
Jaypee University of Information Technology, Solan, Himachal Pradesh – 173234, India

Article Info

Article history:

Received Jan 22nd, 2014
Revised Mar 20th, 2014
Accepted Apr 09th, 2014

Keyword:

Cellular Automaton
Central Dogma
Codon Degeneracy
Sequence Encoding

ABSTRACT

Cellular Automaton (CA) is usually used to model the spatio-temporal evolution of dynamical systems. In this work, a special class of the same known as 'Outer-totalistic' Cellular Automaton is applied to examine if there is a rationale behind the correlation between 64 possible codons and the resulting 20 amino-acids. Also, an attempt is made to mathematically model the central dogma of molecular biology in an intelligible format, including transcription and translation. Results suggest that our irrational-integral-imaginary (i3) encoding approach forms not only a satisfactory basis for a mathematical model of translation of mRNA to protein but also that of transcription from ssDNA (single stranded DNA) to mRNA (messenger RNA).

Corresponding Author:

Praharshit Sharma,
Bioclues Organization, Kukatpally,
Hyderabad, Andhra Pradesh –
500072, India
Email: praharshit@bioclues.org

How to Cite:

Sharma P. et. al. A mathematical model of central dogma of molecular biology employing a novel irrational-integral-imaginary (i3) encoding and numerical approximation based on cellular automaton. IJCB. 2014; Volume 3 (Issue 1): 27-30.

1. INTRODUCTION

Computer scientists, engineers, and biologists apply combined efforts to generate biologically meaningful information associated with DNA. The genetic code forms the basis for protein synthesis machinery in an organism through the translation of DNA. DNA (from genes) is modeled as a 4-state, one-dimensional Cellular Automaton (CA) with radius 'r' and proteins as a 20-state, one-dimensional CA with radius 'R'. CA is a spatially explicit form of finite state automata (FSA). Finite automata are a family of mathematical constructs that are defined by a finite set of states, an output alphabet, and the rules that take the automaton from its current state to next state [1]. There are several evidences where principle of finite automaton has been utilized for various biological applications such as systems biology simulation [2], recognition of restriction endonucleases for a given DNA sequence [3], exact string matching [4] and protein structure classification [5]. Based upon the theory of coherent electronic states in nucleotide chains proposed by Achimowicz et al. [6], we arrive at an optimum codon length well in agreement (as a numerical approximation to) with the same. Conventionally, three nucleotides per codon have been assumed. However, Achimowicz et al. [6] have published the concept of coherent electronic states as applied to the process of transcription which yields an optimum value of the number of nucleotide bases per codon to be the fundamental constant *e*, where $e=\sum_{n=0}^{\infty}\frac{1}{n!}$ = *2.718.*

It has also been remarkably showed using the application of one-dimensional cellular automaton theory that one can achieve a highly reasonable approximation very well in agreement with the above breakthrough result[7, 8]. This indeed happens to be a breakthrough result because it correlates well, to a highly reasonable level of numerical approximation, the respective outcomes of calculus (continuous mathematics approach [6]) and cellular automaton theory (discrete mathematics approach [7, 8]).

Let *a*, *c* denote the first, and third nucleotide state variables (A → 0, C → 1, G → 2 and T → 3) respectively [10] of the 64 codons (using 0-based numbering scheme – cf. http://rosalind.info/glossary/0-based-numbering/) and *b* denote the column state variable (1, 2, 3, 4) of any of the four nucleotide bases occupying the second (central) position of the codon (using 1-based numbering scheme – cf. http://rosalind.info/glossary/1-based-numbering/), and *d* denote the computed state variable of the respective translated amino-acid, obtaining which is detailed below.

At time 't_0'outer totalistic cellular automaton is given by *d* such as

$$d : f(b, a + c)\ldots(1),$$

where *f* is a function of *b* and (*a*+*c*), and (*a*+*c*) is the 'outer-total'.

In other words, the updated value of the central cell '*b*' (at time 't_1') depends upon the old value of the central cell (at time 't_0') and the 'outer-total' (*a*+*c*), both of which are logically independent [9].

2. RESEARCH METHOD

In the genetic code table, we observed that with the two exceptions of 'serine' and 'stop codons', each with a codon degeneracy of *2+4* and *1+2* respectively, the middle nucleotide '*b*' is always retained across all the other possible codon degeneracies, for the rest of the 19 of 20 amino-acids.

We now define *d* as:

$d = ((a + c) - (b \% 1)^{1/4}$, where % stands for the modulo operator.

$\mathrm{d}^4 = (\mathrm{a} + \mathrm{c}) - (\mathrm{b} \% 1)$, then

$\mathrm{d}^4 - (\mathrm{a} + \mathrm{c}) + (\mathrm{b} \% 1) = 0$ ... (2),

It may be noted that (b % 1) = 0 in all the four possible cases of the column state variables since 1 being the "Multiplicative Identity" leaves a remainder of 0 irrespective of whatever natural number it divides; hence the dual exceptions of 'serine' and 'stop codons' being distributed across two different columns (2,4) and (3,4) respectively; counted from left-to-right in the standard genetic code system table – cf. http://en.wikipedia.org/wiki/DNA_codon_table) is handled evidently.

This is a bi-quadratic equation so the total possible values of '*d*' will be four for a single value of '*a*', '*c*' and '*b*'. Here '*a'* and '*c'* are the input states and each input state in genetic code table takes the values (A|C|G|T). '*b*' as already mentioned is the column state variable denoting which particular column (1|2|3|4) the central nucleotide belongs to in the genetic code table , irrespective of whichever of the four possible nucleotides is actually present in the central position as only one of the four nucleotide character is present in each column in central position. This essentially proves that the central nucleotide is effectively nil-contributory to (locked in position with respect to second/ middle column of each of the 64 codons). This prompts us to pose the question, whether the effective codon length is just two, characterized by the first and third nucleotides alone – for a given set of synonymous codons, ranging from 1 to 6.

3. RESULTS AND ANALYSIS

We adopt the alphanumeric convention A → 0, C → 1, G → 2 and T → 3 in the ssDNA prior to transcription [10]. Hence the input states '*a*' and '*c*' can assume values from 0 to 3, for the employed 28-state Outer-Totalistic CA. Performing operation $d^4 - (\mathrm{a} + \mathrm{c}) + (\mathrm{b} \% 1) = 0$ from equation (2) on each *distinct* Outer-Total (since (*b % 1*) = *0* and does not affect the value of *d*), we obtained the following 7x4 matrix of possible values of '*d*' corresponding to the unique *7* of the *16* possible Outer-Totals of '*a*' and '*c*' as illustrated below:

Table 1. Value of outer parameters a, c, and their outer totals.
*Distinct OTs are in the range of 0 to 6, and are 7 in number.

a	***c***	**Distinct OT**	***d^4 = (a+c) – (b % 1)***			
0	0	0	+0	-0	+0*i*	-0*i*
1	0	1	+1	-1	+1*i*	-1*i*
2	0	2	$(+2)^{1/4}$	$-(2)^{1/4}$	$(+2)^{1/4}i$	$-(2)^{1/4}i$
3	0	3	$(+3)^{1/4}$	$-(3)^{1/4}$	$(+3)^{1/4}i$	$-(3)^{1/4}i$
3	1	4	$(+4)^{1/4}$	$-(4)^{1/4}$	$(+4)^{1/4}i$	$-(4)^{1/4}i$
3	2	5	$(+5)^{1/4}$	$-(5)^{1/4}$	$(+5)^{1/4}i$	$-(5)^{1/4}i$
3	3	6	$(+6)^{1/4}$	$-(6)^{1/4}$	$(+6)^{1/4}i$	$-(6)^{1/4}i$

Where *i* stands for the imaginary number (square root of *(-1)* by mathematical definition).

- The obtained values for 'd' in table 2 follow a certain pattern, partitioned as below:
- Integral values (-0, +0, +1, -1) (ssDNA→mRNA transcribed U, A, C, G bases).
- Purely Imaginary values (+/- 0*i*, +/- 1*i*) → (start (f-Met), stop (Opal), stop (Ochre) and stop (Amber) codons respectively).
- Real irrational values (10 Essential Amino-acids) [11] corresponding to rows 3-7 columns 4-5 (table 2).
- Imaginary irrational values (10 Non-essential Amino-acids) [12] corresponding to rows 3-7 columns 6-7 (table 2).

Table 2. 28 values for the 7x4 matrix of possible values of 'd' corresponding to 7 of the 16 unique Outer-Totals of 'a' and 'c' as illustrated in table 1.

a	c	Outer-Total (OT)	a	c	Outer-Total (OT)
0	0	0	0	2	2
1	0	1	1	2	3
2	0	2	2	2	4
3	0	3	3	2	5
0	1	1	0	3	3
1	1	2	1	3	4
2	1	3	2	3	5
3	1	4	3	3	6

A distinction has been made here between f-Met (N-Formylmethionine) which is a proteinogenic amino acid derived from methionine, and Met by itself which belongs to the original class of essential amino-acids.

4. CONCLUSION

Outer-totalistic cellular automaton concept used here is essentially augmented by a multi-valued function which bridges the gap between 4 DNA bases and 28 outputs (4 bases including Uracil for mRNA and 20 standard amino acids plus 1 start and 3 stop codons). Additionally, our mathematical model correlates these 20 amino-acids as outputs of translation into essential and non-essential amino-acids [11,12]. It is believed that the results obtained for the approximation would provide new directions to the modeling and simulation of central dogma of molecular biology that can be extended as a future work to aspects such as “reverse transcription” and also for “post-translational modifications”. We anticipate that our generalized approach for genetic code modeling would help to better understand the structural, functional and evolutionary aspects of genetic material and its various biological forms.

ACKNOWLEDGEMENTS

The corresponding author wishes to thank Dr. Stephen Wolfram for providing inspiration to pen this article.

REFERENCES

[1] Hopcroft JE, Motwani R, and Ullman JD. Introduction to automata theory, languages and computation. 2nd ed. Massachussets: Addison-Wesley; 1979.
[2] Samarrai W, Yeol JW, Barjis I, and Ryu YS. System biology modeling of protein process using deterministic finite automata (DFA). Proceedings of 9th International Workshop on Cellular Neural Networks and Their Applications, pp. 290-95; 2005.
[3] Singh TR. WebFARM: web server for finite automated restriction mapping. Bioinformation 2010 4(8): 341-343.
[4] Hogeweg P. Multilevel Cellular Automata as a Tool for Studying Bioinformatic Processes. In: Hoekstra AG, Kroc J, Sloot PMA, editors. Simulating complex systems by cellular automata. pp.19-28, Springer; 2010. pp. 19-28.
[5] Singh TR and Pardasani KR. A finite automation model for DNA fragment extraction. GAMS J Math. Biosci. 2007; 3(1) 73-80.
[6] Achimowicz J, Kazimierski K, Wojcik K. Possibility of genetic coding of amino acid sequences by coherent electronic states in nucleotide chains. PhysiolChem Phys. 1981; 13(2):171-3.
[7] Praharshit Sharma. Proof of Achimowicz's result in relation to Genetic Code through a Cellular Automaton Approach. Journal of Errology. Available from: http://www.bioflukes.com/Others/bioflukes/4
[8] Praharshit Sharma. Ab-initio reconstruction of genetic code, prediction of Tm , C-value paradox and correlation with ANISOU data based on cellular automaton theory with further applications to protein secondary structure prediction, codon-bias and degeneracy. F1000Posters 2013, 4: 1315. Available from: http://f1000.com/posters/browse/summary/1094586_
[9] Carsten M and Marc TH. Outer-totalistic Cellular Automaton on graphs. Physics Letters: A. 373(5), 2009, 546–549.
[10] SirakoulisGCh, Karafyllidis I, MizasCh, Mardiris V, Thanailakis A, Tsalides P. A cellular
automaton model for the study of DNA sequence evolution. ComputBiol Med. 2003 Sep; 33(5): 439-53.
[11] Robscheit-Robbins FS, Miller LL, Whipple GH. Plasma protein and hemoglobin production : deletion of individual amino acids from growth mixture of ten essential amino acids. Significant changes in urinary nitrogen. J Exp Med. 1947 Feb 28; 85(3):243-65.
[12] George LN. “Essential” and “nonessential” amino acids in the urine of severely burned patients. J Clin Invest. 1954 June; 33(6): 847–8.