Analysis of Changes Occurring in Codon Positions due to Mutations Through the Cellular Automata Transition Rules

Antara Sengupta[1], Sreeya Ghosh[2], Pabitra Pal Choudhury[3]

Department of Computer Science and Engineering, University of Calcutta, West Bengal, India[1]

Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, West Bengal, India[2]

Applied Statistics Unit, Indian Statistical Institute, Kolkata, West Bengal, India[3]

Corresponding author: Antara Sengupta, Email: antara.sngpt@gmail.com

Variation in the nucleotides of a codon may cause variations in the evolutionary patterns of a DNA or amino acid sequence. To address the capability of each position of a codon to have non-synonymous mutations, the concept of degree of mutation has been introduced. The degree of mutation of a particular position of codon defines the number of non-synonymous mutations occurring for the substitution of nucleotides at each position of a codon, when other two positions of that codon remain unaltered. A Cellular Automaton (CA), is used as a tool to model the mutations of any one of the four DNA bases A, C, T and G at a time where the DNA bases correspond to the states of the CA cells. Point mutation (substitution type) of a codon which characterizes changes in the amino acids, have been associated with local transition rules of a CA. Though there can be 4^{4^3} transitions of a 4-state CA with 3-neighbourhood cells, here it has been possible to represent all possible point mutations of a codon in terms of combinations of 16 local transition functions of the CA. Further these rules are divided into 4 classes of equivalence. Also, according to the nature of mutations, the 16 local CA rules of substitutions are classified into 3 sets namely, 'No Mutation', 'Transition' and 'Transversion'. The experiment has been carried out with three sets of single nucleotide variations(SNVs) of three

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 1–21. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-1

different viruses but the symptoms of the diseases caused by them are to some extent similar to each other. They are SARS-CoV-1, SARS-CoV-2 and H1N1 Type A viruses. The aim is to understand the impact of nucleotide substitutions at different positions of a codon with respect to a particular disease phenotype.

Keywords: Codon, Mutation, Cellular Automata, SARS-CoV-1, SARS-CoV-2, H1N1 type A.

1 Introduction

Genetic code defines some rules to translate genetic information encoded in nucleotide triplets or codons into amino acids. It also defines the order of amino acid to be added next during protein synthesis. $4^3 = 64$ codons are there in genetic code table, which encodes 20 standard amino acids and 3 stop codons. Hence, there arises a context of degeneracy. Multiplet structure of DNA sequence [1] specifies that instead of one-to-one mapping a single amino acid can be coded by one, two, three, four or six codons. The codon usage is an important determinant of gene expression and surprisingly transcriptions rather than translations play a key role here [2] [3]. It has been reported that instead of codons or amino acids, codon and amino acid usage is consistent with the forces acting on four DNA bases [4]. Analysis of codon usage gives insight about the evolution of any organism [5]. Selection of codon to code for an amino acid is a natural selection and amino acid composition in protein aims to minimize the the impact of mutations on protein structure [6]. A codon can have mutations at the first, second or third positions. Mutations at the third position of the codon are more likely to be synonymous than mutations that occur at the first or second positions [7]. Hence, probability of substitution of amino acid with a new one due to mutations at third position of a codon is less than that of its first and second positions. The second position of codon is the most conserved position, as every nucleotide change in this position leads to substitution of another amino acid [8]. Hence, change of nucleotides at a particular position of a codon due to substitutions have positional impact on the change in amino acids.

Researchers throughout the globe are trying to figure out the pattern of mutations responsible for a particular genetic disease. Numerous mathematical model based approaches are already introduced to make quantitative understanding of a disease and to apply classification rules to segregate that disease from others. Now-a-days when some Asian countries are witnessing 2nd wave of Corona and 3rd wave has already arrived in other continents like Europe, researchers are trying in every way to understand the pattern of mutations taking place in SARS-

CoV-2. A Plethora of mathematical models are already introduced to reach to that goal [9] [10]. Scientists are working over differentiating coronavirus from influenza virus as both the disease COVID-19 and flu have some similar type of symptoms [11]. Mutations may lead to occur biodiversity. Biodiversity is characterized by the continual replacement of branches in the tree of life, that is clade [12]. People are trying to reach the origin of the tree of life to get some ways of prevention [13].

Cellular Automaton (pl. cellular automata, abbrev. CA) [14] is a discrete model introduced by J.von Neumann and S.Ulam in 1940s for designing self replicating systems. It consists of a finite/countably infinite number of finite-state semi-automata known as 'cells' arranged in an ordered n-dimensional grid. Each cell receives input from the neighbouring cells and changes according to a transition function. Application of cellular automata in bioinformatics is a well-known approach [15]. There has been studies on the evolution of DNA sequence using automata [16], and CA transition rules [17] [18]. CA based models are used to unfold different facts in genomics, proteomics [19] [20]and even for the representation of protein translation using CA rules [21]. However representing all possible point mutations of a codon in terms of CA rules have not been addressed earlier. Variation in mutations and codon selection may cause differences in evolutionary patterns across a DNA or amino acid sequence [7]. In this present study, we have been able to represent all possible changes in amino acids due to point mutations of codons in terms of combinations of 16 local transition rules of a CA. These could further be divided into 4 classes of equivalence. Depending upon the capability of producing a new amino acid, degree of mutations of codons at 3 different positions have been derived. Also, according to the nature of mutations, the 16 local CA rules of substitutions are classified here into 3 sets namely, 'No Mutation', 'Transition' and 'Transversion'.

Recently, attempts has been made to model COVID-19 spread within the framework of Probabilistic CA [22] and Fuzzy CA [23]. Pokkuluri et.al. [24] have constructed CA based classifiers to predict the trend of SARS-CoV-2. Few papers are reported, where dynamics of the influenza infection is described using Beauchemin's CA model [25] [26] [27]. In our work, we have considered SNVs of three different viruses manifesting similar symptoms, namely SARS-CoV-1, SARS-CoV-2 and H1N1 Type A viruses. Our objective is to get a pattern of mutations occurring in these diseases, in the light of degrees of mutations and CA transition functions.

2 Methods and Materials

2.1 Derive Degree of Mutation of Nucleotides at Different Positions of Codon

According to the genetic code table 61 codons code for 20 amino acids and there are three stop codons [28]. The standard classic model of genetic code table consists four rows and four columns. The four rows represents the first base of each codon, the four columns represent the second base and the right side indicates the third base of them. Codon contains combinations of 4 bases A,T,C,G at its 3 positions and as a whole codes for a particular amino acid. Since there are 20 different amino acids and 64 possible codons, more than one codon may code for a single amino acid. Hence, any changes in nucleotides at any positions of codon due to mutation either may change the produced amino acid or can code for the same amino acid and there is a talk about non-synonymous and synonymous mutations respectively. Here in this section it is tried to get a clear view of mapping between codon and amino acid when mutations occur at first, second and third positions of a codon.

[Degree of Mutation of a particular position of codon] Given a codon C with constituent nucleotides say, (N_1, N_2, N_3), where $N_i \in N$ is a particular position of a codon. Now, consider S_i as any one nucleotide among the set of nucleotides S={T,C,A,G} at a particular position N_i in codon C when nucleotides at other two positions are constant. The degree of mutation ($\delta(M)$) at a particular position of codon defines the number of non-synonymous mutations occurred to substitution of nucleotides at that position of a codon, when other two positions of that codon are unaltered.

It has been noted that when any two positions of a codon are constant, it is possible to make change in 3 times with maximum three nucleotides, as the position is initially being occupied by any one of four nucleotides. Due to the changes occurring in 1^{st} or 2^{nd} positions of a codon total changes (non-synonymous) occurring in corresponding amino acids vary from 2 to 3. In this way due to the changes occurring in 1^{st} or 2^{nd} positions of a codon total changes (non-synonymous) occurring in corresponding amino acids vary from 0 to 1 (shown in Table 1). In some cases changes in codon does not change the amino acid produced. Thus total number of times codons get changes in its each position is 48 as 16 possible di-nucleotides can exists in other two positions as a whole. As an example, when T is constant at both 2^{nd} and 3^{nd} positions, due to change in nucleotides (A/T/C/G) at first position the total numbers of amino acids can be changed is 3 according to genetic code table and the amino acids are F, L, I and V. Hence, the

Table 1: Degree of mutations ($\delta(M)$) of all 64 codons

1st Position	2nd Position	3rd Position	$\delta(M)$ at 1st position	Name of AA
T/C/A/G	T	T	3	F/L/I/V
T/C/A/G	T	C	3	F/L/I/V
T/C/A/G	T	A	2	L/M/V
T/C/A/G	T	G	2	L/M/V
T/C/A/G	C	T	3	S/P/T/A
T/C/A/G	C	C	3	S/P/T/A
T/C/A/G	C	A	3	S/P/T/A
T/C/A/G	C	G	3	S/P/T/A
T/C/A/G	A	T	3	Y/H/N/D
T/C/A/G	A	C	3	Y/H/N/D
T/C/A/G	A	A	3	STOP CODON/Q/K/E
T/C/A/G	A	G	3	STOP CODON/Q/K/E
T/C/A/G	G	T	3	C/R/S/G
T/C/A/G	G	C	3	C/R/S/G
T/C/A/G	G	A	2	STOP CODON/R/R/G
T/C/A/G	G	G	2	W/R/R/G
1st Position	**2nd Position**	**3rd Position**	**$\delta(M)$ at 2nd position**	**Name of AA**
T	T/C/A/G	T	3	F/S/Y/C
T	T/C/A/G	C	3	F/S/Y/C
T	T/C/A/G	A	2	L/S/STOP CODON/STOP CODON
T	T/C/A/G	G	3	L/S/STOP CODON/W
C	T/C/A/G	T	3	L/P/H/R
C	T/C/A/G	C	3	L/P/H/R
C	T/C/A/G	A	3	L/P/Q/R
C	T/C/A/G	G	3	L/P/Q/R
A	T/C/A/G	T	3	I/T/N/S
A	T/C/A/G	C	3	I/T/N/S
A	T/C/A/G	A	3	I/T/K/R
A	T/C/A/G	G	3	I/T/K/R
G	T/C/A/G	T	3	V/A/D/G
G	T/C/A/G	C	3	V/A/D/G
G	T/C/A/G	A	3	V/A/E/G
G	T/C/A/G	G	3	V/A/E/G
1st Position	**2nd Position**	**3rd Position**	**$\delta(M)$ at 3rd position**	**Name of AA**
T	T	T/C/A/G	1	F/L
T	C	T/C/A/G	0	S
T	A	T/C/A/G	1	Y/STOP CODON
T	G	T/C/A/G	2	C/STOP CODON/W
C	T	T/C/A/G	0	L
C	C	T/C/A/G	0	P
C	A	T/C/A/G	1	H/Q
C	G	T/C/A/G	0	R
A	T	T/C/A/G	1	I/M
A	C	T/C/A/G	0	T
A	A	T/C/A/G	1	N/K
A	G	T/C/A/G	1	S/R
G	T	T/C/A/G	0	V
G	C	T/C/A/G	0	A
G	A	T/C/A/G	1	D/E
G	G	T/C/A/G	0	G

degree of mutation $\delta(M)$ here is 3.

2.2 Cellular Automata and Mutations of Nucleotides

A **CA**(denoted by $\mathcal{C}_{\tau}^{Q}$)(reported in [29] [14] [30]) is a triplet $(Q, Q^{\mathbb{Z}}, \tau)$, where,

- Q is a finite state set
- $Q^{\mathbb{Z}} = \{C|C : \mathbb{Z} \to Q\}$ is the set of all global configurations C

- $\tau : Q^{\mathbb{Z}} \to Q^{\mathbb{Z}}$ is a global transition function

For $i \in \mathbb{Z}, r \in \mathrm{N}$, let $S_i = \{i-r, ..., i-1, i, i+1, ..., i+r\} \subseteq \mathbb{Z}$. S_i is the neighbourhood of the i^{th} cell having 2r+1 cells. r is the radius of the neighbourhood of a cell.
It follows that $\mathbb{Z} = \bigcup_i S_i$
A restriction from $\mathbb{Z}$ to S_i induces a restriction of C to $\overline{c_i}$ given by $\overline{c_i} : S_i \to Q$; where $\overline{c_i}$ may be called the **local configuration** and S_i the **neighbourhood** of the i^{th} cell.
The mapping $\mu_i : Q^{S_i} \to Q$ is known as a **local transition function** for the i^{th} cell.
Thus $\forall i \in \mathbb{Z}, \mu_i(\overline{c_i}) \in Q$ and it follows that,

$$\tau(C) = \tau(..., c_{i-1}, c_i, c_{i+1}, ...) = ...\mu_{i-1}(\overline{c_{i-1}}).\mu_i(\overline{c_i}).\mu_{i+1}(\overline{c_{i+1}}).....$$

Corollary 2.1 *An m-celled neighbourhood of an i^{th} cell can also be considered to be a subset of $\mathbb{Z}$ having the form $S_{Li} = \{i-(m-1), ..., i-1, i\}$ or $S_{Ri} = \{i, i+1, ..., i+(m-1)\}$ for some $m \in \mathrm{N}$ such $\mathbb{Z} = \bigcup_i S_{Li}$ or $\mathbb{Z} = \bigcup_i S_{Ri}$*

2.3 Representation of Mutations Occurring in Different Codon Positions Using CA Transitions

A codon is composed of 3 nucleotides. In order to associate a codon with a local configuration of Cellular Automata(CA), 3-celled neighbourhoods have been considered here for any i^{th} cell of the CA. Point mutation of a codon can thus be represented by local transitions of such a CA. 16 substitutions are possible with four DNA bases A, C, T and G. They are

$$A \to A,\ \ A \to T,\ \ A \to C,\ \ A \to G,$$

$$T \to T,\ \ T \to C,\ \ T \to G,\ \ T \to A,$$

$$C \to C,\ \ C \to T,\ \ C \to A,\ \ C \to G,$$

$$G \to G,\ \ G \to T,\ \ G \to C,\ \ G \to A$$

They can be represented in terms of combinations of 16 local transition functions of CA.

Let us consider the global configuration of a CA to be composed of local configurations having three cells corresponding to the three nucleotide positions of a codon. The position of the codon at which the point mutation occurs, is denoted

by the i^{th} cell and the other two nucleotides which remain fixed are denoted by x and y where $x, y \in Q = \{A, T, C, G\}$. If point mutation occurs at the third position then the neighbourhood of the i^{th} cell is considered as

$$S_i = (c_{i-2}, c_{i-1}, c_i)$$

The local configuration of the i^{th} cell maybe denoted by (x, y, c_i) such that $c_{i-2} = x$ and $c_{i-1} = y$. The local transition function for i^{th} cell denoted by $\mu_{R(xyi)}$ is,

$$\mu_{R(xyi)}(\overline{c_i}) = \mu_{R(xyi)}(x, y, c_i)$$

where $R(xyi)$ is the rule number for some $R \in \{0, 1, 2, ..., 15\}$. The rules for third position mutation corresponding to first and second position constant nucleotides x, y is computed as follows :

$$\mu_{0(xyi)} : (x, y, T) \to T \quad Rule\ 0(xyi)$$

$$\mu_{1(xyi)} : (x, y, T) \to C \quad Rule\ 1(xyi)$$

$$\mu_{2(xyi)} : (x, y, T) \to A \quad Rule\ 2(xyi)$$

$$\mu_{3(xyi)} : (x, y, T) \to G \quad Rule\ 3(xyi)$$

$$\mu_{4(xyi)} : (x, y, C) \to T \quad Rule\ 4(xyi)$$

$$\mu_{5(xyi)} : (x, y, C) \to C \quad Rule\ 5(xyi)$$

$$\mu_{6(xyi)} : (x, y, C) \to A \quad Rule\ 6(xyi)$$

$$\mu_{7(xyi)} : (x, y, C) \to G \quad Rule\ 7(xyi)$$

$$\mu_{8(xyi)} : (x, y, A) \to T \quad Rule\ 8(xyi)$$

$$\mu_{9(xyi)} : (x, y, A) \to C \quad Rule\ 9(xyi)$$

$$\mu_{10(xyi)} : (x, y, A) \to A \quad Rule\ 10(xyi)$$

$$\mu_{11(xyi)} : (x, y, A) \to G \quad Rule\ 11(xyi)$$

$$\mu_{12(xyi)} : (x, y, G) \to T \quad Rule\ 12(xyi)$$

$$\mu_{13(xyi)} : (x, y, G) \to C \quad Rule\ 13(xyi)$$

$$\mu_{14(xyi)} : (x, y, G) \to A \quad Rule\ 14(xyi)$$

$$\mu_{15(xyi)} : (x, y, G) \to G \quad Rule\ 15(xyi)$$

Example 2.1 *For constant first and second necleotides AA, Rule* $6(AAi)$ *represented by* $\mu_{6(AAi)} : (A, A, C) \rightarrow A$ *changes nucleotide* C *in the third position to nucleotide* A *corresponding to the mutation of codon* AAC *to* AAA *for amino acid Asn to Lys.*

Substitutions at second and first positions can be computed similar to that of the third position point mutation by changing the neighbourhood of the i^{th} cell as follows.

Corollary 2.2 *If point mutation occurs at the second position then the neighbourhood of the* i^{th} *cell is considered as*

$$S_i = (c_{i-1}, c_i, c_{i+1})$$

The local configuration of the i^{th} *cell maybe denoted by* (x, c_i, y) *such that* $c_{i-1} = x$ *and* $c_{i+1} = y$. *The local transition function for* i^{th} *cell denoted by* $\mu_{R(xiy)}$ *is,*

$$\mu_{R(xiy)}(\overline{c_i}) = \mu_{R(xiy)}(x, c_i, y)$$

The rules $R(xiy)$ *for second position mutation are as follows :*

$$\begin{aligned}
\mu_{0(xiy)} &: (x, T, y) \rightarrow T \quad \textit{Rule } 0(xiy) \\
\mu_{1(xiy)} &: (x, T, y) \rightarrow C \quad \textit{Rule } 1(xiy) \\
\mu_{2(xiy)} &: (x, T, y) \rightarrow A \quad \textit{Rule } 2(xiy) \\
&\vdots \qquad \cdots \quad \cdots \qquad \vdots \\
\mu_{15(xiy)} &: (x, G, y) \rightarrow G \quad \textit{Rule } 15(xiy)
\end{aligned}$$

Corollary 2.3 *If point mutation occurs at the first position then the neighbourhood of the* i^{th} *cell is considered as*

$$S_i = (c_i, c_{i+1}, c_{i+2})$$

The local configuration of the i^{th} *cell maybe denoted by* (c_i, x, y) *such that* $c_{i+1} = x$ *and* $c_{i+2} = y$. *The local transition function for* i^{th} *cell denoted by* $\mu_{R(xyi)}$ *is,*

$$\mu_{R(ixy)}(\overline{c_i}) = \mu_{R(ixy)}(c_i, x, y)$$

The rules for first position mutation are as follows :

$$\begin{aligned}
\mu_{0(ixy)} &: (T,x,y) \to T \quad Rule\ 0(ixy) \\
\mu_{1(ixy)} &: (T,x,y) \to C \quad Rule\ 1(ixy) \\
\mu_{2(ixy)} &: (T,x,y) \to A \quad Rule\ 2(ixy) \\
&\vdots \qquad \cdots \quad \cdots \qquad \vdots \\
\mu_{15(ixy)} &: (G,x,y) \to G \quad Rule\ 15(ixy)
\end{aligned}$$

These combinations of 16 CA rules can further be classified into three sets which depict No Mutation, Transition and Transversion of nucleotides irrespective of the position where the point mutation occurs.
According to the rules for point mutation with respect to constant nucleotides x and y we get:

$$Rule0(T \to T), Rule5(C \to C), Rule10(A \to A), Rule15(G \to G)$$

representing No Mutations;

$$Rule1(T \to C), Rule4(C \to T), Rule11(A \to G), Rule14(G \to A)$$

representing Transitions where point mutations occur due to substitutions between any two purine (A or G) bases or pyrimidine bases (T or C);

$$Rule2(T \to A), Rule3(T \to G), Rule6(C \to A), Rule7(C \to G),$$

$$Rule8(A \to T), Rule9(A \to C), Rule12(G \to T), Rule13(G \to C)$$

representing Transversions where point mutations occur due to substitution of a purine (A or G) base by a pyrimidine base (T or C) or vice-versa.
These classifications have been tabulated in Table 2.

2.4 Amino Acids Arising due to Point Mutations Represented by Equivalent Rules

Any two local transition functions for an i^{th} cell denoted by $\mu_{R(xyi)}$ and $\mu_{R'(xyi)}$ are equivalent if both the rules produce same output. Thus

$$\mu_{R(xyi)}(x,y,c_i) = \mu_{R'(xyi)}(x,y,c_i)$$

Table 2: Classification of CA rules

CA RULES	CLASSIFICATION
$0(T \to T)$ $5(C \to C)$ $10(A \to A)$ $15(G \to G)$	4*NO MUTATION
$1(T \to C)$ $4(C \to T)$ $11(A \to G)$ $14(G \to A)$	4*TRANSITION
$2(T \to A)$ $3(T \to G)$ $6(C \to A)$ $7(C \to G)$ $8(A \to T)$ $9(A \to C)$ $12(G \to T)$ $13(G \to C)$	8*TRANSVERSION

where $R(xyi)$ and $R'(xyi)$ are rule numbers for $R, R' \in \{0, 1, 2, ..., 15\}$. Any two equivalent rules belong to same class of $\mathbb{Z}_4$ where $\mathbb{Z}_4 = \{[0], [1], [2], [3]\}$.
From the list of rules we get that, if point mutation occurs at the third position then :

- $[0](xyi) = \{Rule\ 0(xyi), Rule\ 4(xyi), Rule\ 8(xyi), Rule\ 12(xyi)\}$
- $[1](xyi) = \{Rule\ 1(xyi), Rule\ 5(xyi), Rule\ 9(xyi), Rule\ 13(xyi)\}$
- $[2](xyi) = \{Rule\ 2(xyi), Rule\ 6(xyi), Rule\ 10(xyi), Rule\ 14(xyi)\}$
- $[3](xyi) = \{Rule\ 3(xyi), Rule\ 7(xyi), Rule\ 11(xyi), Rule\ 15(xyi)\}$

Correspondingly, amino acids produced from codon having nucleotide base

- T in the 3rd position is obtained by applying $[0](xyi)$
- C in the 3rd position is obtained by applying $[1](xyi)$
- A in the 3rd position is obtained by applying $[2](xyi)$
- G in the 3rd position is obtained by applying $[3](xyi)$

If point mutation (substitutions) occurs at second or first position then similar rules are equivalent. Also amino acids produced from corresponding point mutations of codons can be obtained similarly. Thus the Table 3 shows all possible changes in codons due to point mutations and possible changes in amino acids due to it through the light of CA rules.

3 Results and Discussion

3.1 Collection of Genomic Sequences

To establish the novelty of the methodologies discussed in previous section, it is necessary to apply the same into a given dataset. To carry out the experiment, mutated genomic sequences of three types of genes SARS-CoV-1, SARS-CoV-2 and H1N1 are taken. The information about collected dataset are summarized at Table 4. The 39680 numbers of genomic sequences of SARS-CoV-2 reported for Asian countries are collected from https://covidcg.org/, which is an open resource to track SNVs (single-nucleotide variations). For SARS-CoV-1, 54 mutated genomic sequences are considered. 35008 numbers of patients' data of H1N1 type A are collected from NCBI influenza virus database (https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi#mainform).

3.2 Derive Degree of Mutation of each dataset

It is observed that mutations occurred at different positions of codon throughout the dataset. Here in this section we have tried to find out the highest occurrence of codon transitions. The degree of mutations for each mutation is analysed. It has been observed that mutations are majorly taken place of degree 3 for all datasets, which has been shown in Figure 1.

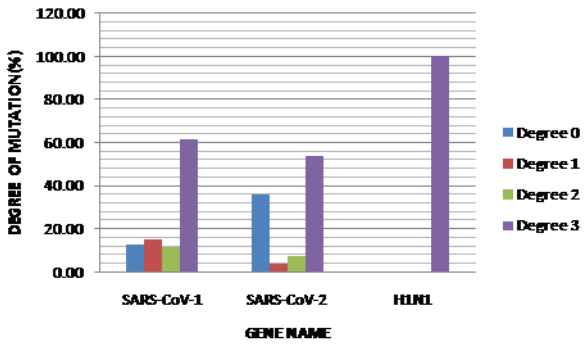

Figure 1: Percentage-wise analysis of degree of mutation

3.3 Position-wise analysis of mutations occur at 3 different nucleotide positions of codon.

Codons are triplets comprising 3 nucleotides at its three positions. Mutations may occur at any of those three positions. In this sub section percentage wise calculations have been made on mutations occurred at those three different positions of codons for the SNVs of the 3 sets of genomic sequences taken. As shown in Figure 2, it has been observed that in the SNVs of SARS-CoV-1 (41.18%) the mutations majorly took place at 2^{nd} positions. In the SNVs of H1N1 Type A mutations occurred in equal percentage at 1^{st} and 2^{nd} positions of codons. In SARS-CoV-2 (41.18%) the maximum mutations occurred at 1^{st} positions of codon.

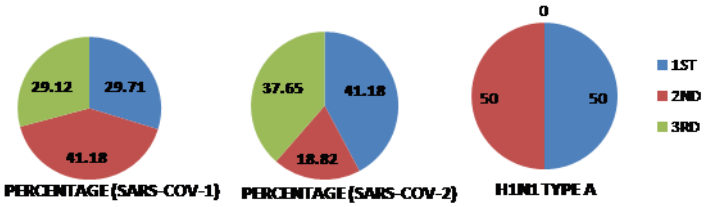

Figure 2: Representation of mutations occur in different positions of codon

3.4 Representation of mutations occur in different codon positions based on the rules of cellular automata

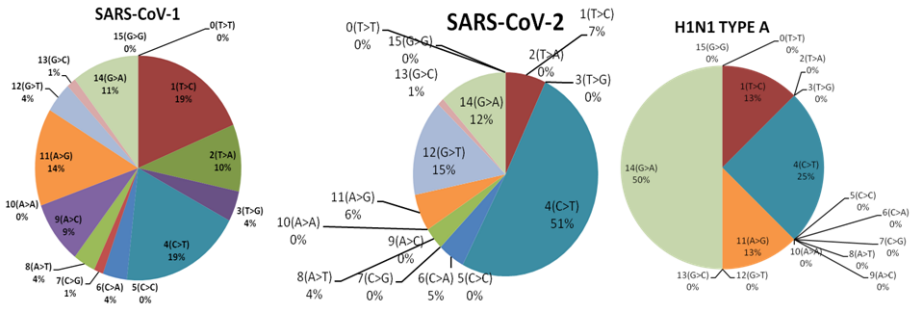

Figure 3: Percentage-wise mutations occur in three virus datasets according to 16 rules of cellular automata

Here in this subsection it is tried to make a mapping between rules defined

by genetic code and 16 rules of Cellular Automata. The model is applied on the all three datasets taken. The Figure 3 shows percentage of mutations occurred according to CA rule. It has been observed that the SNVs of SARS-CoV-2 has a trend to mostly follow the rules 4 (51%), whereas, CA rule 4 (19.01%) and CA rule 1 (18.71%) have approximately equal contributions in of SARS-CoV-1. The SNVs of H1N1 has the trend of rule 14 (51%). The rule 4 indicates the substitution of nucleotide C by T and rule 1 specifies substitution of T by C, i.e. between pyrimidines and rule 14 indicates the substitution of neucleotide G by A, i.e. between purines. Further microscopic view has been given on the codon position wise degree of mutations occurred in SARS-CoV-1 and SARS-CoV-2 where rule 4 ($C \rightarrow T$) is applied and in H1N1 rule 14 ($G \rightarrow A$) is applied maximum. The result is shown in Figure 4. It is remarkable that in both the datasets of SARS-CoV-1 and SARS-CoV-2 maximum mutations took place at 2^{nd} position of codons and they are of degree 3. In H1N1 TYPE A virus all the mutations of degree 3 are taken places equally at the 1^{st} and 2^{nd} position of codons. Few transversions (15.29%) are also taken place in SARS-CoV-2. In these case base G of codons are substituted by T. In CA rule this substitution comes under rule 12.

Further analysis has been carried out with the mutations occurred under rule 4 for the datasets of SARS-CoV-1 and SARS-CoV-2 and under rule 14 for H1N1 Type A virus respectively (shown in Table 5). It has been found that some mutations have dominance over the others and codon position wise they have commonalities between SARS-CoV-1 and SARS-CoV-2. In both the datasets $L \rightarrow L$, $L \rightarrow F$ are majorly found mutations at the first positions of codons and $T \rightarrow I$ at second positions. Individually frequently found mutations IN SARS-CoV-1 are $L \rightarrow L$, $A \rightarrow V$, $T \rightarrow I$, $P \rightarrow L$, $Y \rightarrow Y$. In SARS-CoV-2 they are $L \rightarrow L$, $F \rightarrow F$ and $Q \rightarrow STOP\ \ CODON$. In H1N1 type A $S \rightarrow N$ and $H \rightarrow Y$ are found the most.

In SARS-CoV-1, few transversions are found, where substitutions are taken place between A and T, which are defined by CA rule 2 ($T \rightarrow A$) and rule 8 ($A \rightarrow T$). It is reported that the most harmful mutations due to substitutions take place between A and T. These kind of mutations change the hydropathy and polarity of amino acids. Hence, next point of investigation is carried out with it. It has been observed that according to CA rules, 9.61% and 3.51% of total SNVs found in the data set of SARS-CoV-1 are following the rule 8 ($T \rightarrow A$) and rule 2 ($A \rightarrow T$) respectively (shown in Table 6).

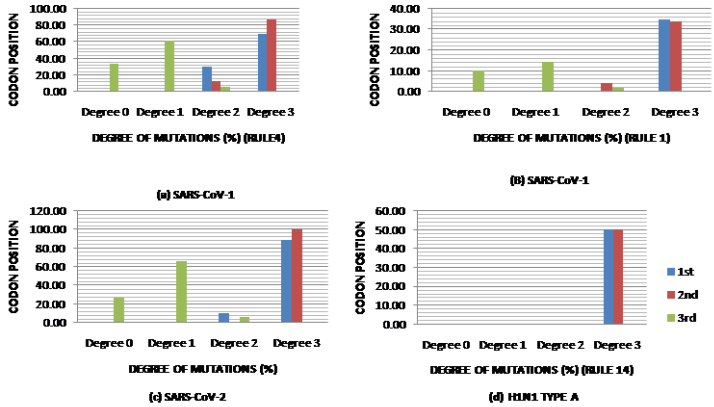

Figure 4: Codon position wise degree of mutations occurred in SARS-CoV-1 and SARS-CoV-2 where rule 4 is applied and in H1N1 where rule 14 is applied maximum. a) SARS-CoV-1 (b) SARS-CoV-2 (c) H1N1 type A

4 Discussion

In this article, point mutation (substitution type) of a codon has been associated with local transitions of Cellular Automata (CA) having 3-celled local configurations. Clearly, 16 substitutions are possible with four DNA bases A, C, T and G, which can be represented in terms of combinations of 16 local transition functions of CA. The experiment has been carried out with three sets of SNVs of three different viruses but the symptoms of the diseases caused by them are to some extent similar to each other. They are SARS-CoV-1, SARS-CoV-2 and H1N1 Type A viruses. The aim is to understand the impact of nucleotide substitutions in different codon positions on mutations occurred in a particular disease phenotype. It is to be noted that although the size of genomic sequences taken for all three viruses are huge, but H1N1 type A virus has comparatively very few variants. Codon usage bias is observed in all organisms even in viruses too. The reason behind may be either pressure of natural selection or due to biases in the mutation process. According to the origin and evolution theory of genetic code, codons are selected in such a way so that it can minimize the adverse effect of point mutations and translation errors. It has been observed that in all the datasets maximum mutations have taken place at the codons having degree of mutation 3. The codons having degree of mutation 3 are capable to change up to 3 amino acids due to substitution of nucleotides at a particular position. It

has been observed that in the SNVs of SARS-CoV-1 the mutations majorly took place at 2nd positions but in SNVs of H1N1 type A 1st and 2nd positions of codons are equally affected. In SARS-CoV-2 the maximum mutations occurred at 1st positions of codon. The second position of codon is the most functionally constrained position and causes non-synonymous change. According to the nature of mutations, 16 CA rules of substitutions are classified into 3 classes namely, 'No Mutation', 'Transition' and 'Transversion'. Experimental results find substitutions from CA class 'Transition' more than the other two classes. Transition mutations are more likely than transversions, because transversions make substitutions of nucleotides between purine (having 2 rings in it structure) and pyrimidine (having 1 ring). Hence, substitution of a single ring structure with another single ring structure is more likely than substitution of a double ring with a single ring. Transitions are more certain to change amino acids. Harmful substitutions from CA class 'Transversion' (rule 2 and rule 8) are noticed (13.16% in total) between bases A and T in some SNVs of SARS-CoV-1, which are responsible to make huge structural changes in existing proteins.

5 Conclusion

In this article a Cellular Automaton has been used to model substitutions of four DNA bases A, C, T and G at different positions of codons. Considering codon as a triplet, substitution of nucleotides may take place in any one of the three positions of a codon and cause point mutations. All possible point mutations have been represented here as functions of 16 CA transition rules. Point mutations may or may not make changes in the amino acids. The degree of mutation at a particular position of a codon defines the number of amino acids change due to substitution of nucleotides at each position of the codon, when other two positions of that codon are fixed. Hence, the degree of mutation specifies the capability of nucleotide substitutions in a particular position of a codon to produce new amino acids and their impacts in a particular disease pathogenesis. Thus, the aim of this work is investigating the codon alteration patterns due to nucleotide substitutions and their impact during mutations of a gene responsible for a particular disease. Hence, signature of a particular disease could be portrayed in the light of CA transition rules and codon alteration patterns.

Table 3: CA rules to identify changes in codons due to point mutations and possible changes in amino acids due to it

Amino acid	Codon	Point mutation	Fixed nucleotides	Rules
3*Phe	3*TTT/TTC	3rd	TT	[0](TTi), [1](TTi)
		2nd	TT/TC	[0](TiT), [0](TiC)
		1st	TT/TC	[0](iTT), [0](iTC)
3*Leu	3*TTA/TTG	3rd	TT	[2](TTi), [3](TTi)
		2nd	TA/TG	[0](TiA), [0](TiG)
		1st	TA/TG	[0](iTA), [0](iTG)
	2*CTT/CTC/CTA/CTG	3rd	CT	[0](CTi), [1](CTi), [2](CTi), [3](CTi)
		2nd	CT/CC/CA/CG	[0](CiT), [0](CiC), [0](CiA), [0](CiG)
		1st	TT/TC/TA/TG	[1](iTT), [1](iTC), [1](iTA), [1](iTG)
3*Ile	3*ATT/ATC/ATA	3rd	AT	[0](ATi), [1](ATi), [2](ATi)
		2nd	AT/AC/AA	[0](AiT), [0](AiC), [0](AiA)
		1st	TT/TC/TA	[2](iTT), [2](iTC), [2](iTA)
3*Met	3*ATG	3rd	AT	[3](ATi)
		2nd	AG	[0](AiG)
		1st	TG	[2](iTG)
3*Val	3*GTT/GTC/GTA/GTG	3rd	GT	[0](GTi), [1](GTi), [2](GTi), [3](GTi)
		2nd	GT/GC/GA/GG	[0](GiT), [0](GiC), [0](GiA), [0](GiG)
		1st	TT/TC/TA/TG	[3](iTT), [3](iTC), [3](iTA), [3](iTG)
3*Ser	3*TCT/TCC/TCA/TCG	3rd	TC	[0](TCi), [1](TCi), [2](TCi), [3](TCi)
		2nd	TT/TC/TA/TG	[1](TiT), [1](TiC), [1](TiA), [1](TiG)
		1st	CT/CC/CA/CG	[0](iCT), [0](iCC), [0](iCA), [0](iCG)
	2*AGT/AGC	3rd	AG	[0](AGi), [1](AGi)
		2nd	AT/AC	[3](AiT), [3](AiC)
		1st	GT/GC	[2](iGT), [2](iGC)
3*Pro	3*CCT/CCC/CCA/CCG	3rd	CC	[0](CCi), [1](CCi), [2](CCi), [3](CCi)
		2nd	CT/CC/CA/CG	[1](CiT), [1](CiC), [1](CiA), [1](CiG)
		1st	CT/CC/CA/CG	[1](iCT), [1](iCC), [1](iCA), [1](iCG)
3*Thr	3*ACT/ACC/ACA/ACG	3rd	AC	[0](ACi), [1](ACi), [2](ACi), [3](ACi)
		2nd	AT/AC/AA/AG	[1](AiT), [1](AiC), [1](AiA), [1](AiG)
		1st	CT/CC/CA/CG	[2](iCT), [2](iCC), [2](iCA), [2](iCG)
3*Ala	3*GCT/GCC/GCA/GCG	3rd	GC	[0](GCi), [1](GCi), [2](GCi), [3](GCi)
		2nd	GT/GC/GA/GG	[1](GiT), [1](GiC), [2](GiA), [3](GiG)
		1st	CT/CC/CA/CG	[3](iCT), [3](iCC), [3](iCA), [3](iCG)
3*Tyr	3*TAT/TAC	3rd	TA	[0](TAi), [1](TAi)
		2nd	TT/TC	[1](TiT), [1](TiC)
		1st	AT/AC	[0](iAT), [0](iAC)
3*His	3*CAT/CAC	3rd	CA	[0](CAi), [1](CAi)
		2nd	CT/CC	[2](CiT), [2](CiC)
		1st	AT/AC	[1](iAT), [1](iAC)
3*Gln	3*CAA/CAG	3rd	CA	[2](CAi), [3](CAi)
		2nd	CA/CG	[2](CiA), [2](CiG)
		1st	AA/AG	[1](iAA), [1](iAG)
3*Asn	3*AAT/AAC	3rd	AA	[0](AAi), [1](AAi)
		2nd	AT/AC	[2](AiT), [2](AiC)
		1st	AT/AC	[2](iAT), [2](iAC)
3*Lys	3*AAA/AAG	3rd	AA	[2](AAi), [3](AAi)
		2nd	AA/AG	[2](AiA), [2](AiG)
		1st	AA/AG	[2](iAA), [2](iAG)
3*Asp	3*GAT/GAC	3rd	GA	[0](GAi), [1](GAi)
		2nd	GT/GC	[2](GiT), [2](GiC)
		1st	AT/AC	[3](iAT), [3](iAC)
3*Glu	3*GAA/GAG	3rd	GA	[2](GAi), [3](GAi)
		2nd	GA/GG	[2](GiA), [2](GiG)
		1st	AA/AG	[3](iAA), [3](iAG)
3*Cys	3*TGT/TGC	3rd	TG	[0](TGi), [1](TGi)
		2nd	TT/TC	[3](TiT), [3](TiC)
		1st	GT/GC	[0](iGT), [0](iGC)
3*Trp	3*TGG	3rd	TG	[3](TGi)
		2nd	TG	[3](TiG)
		1st	GG	[0](iGG)
3*Arg	3*CGT/CGC/CGA/CGG	3rd	CG	[0](CGi), [1](CGi), [2](CGi), [3](CGi)
		2nd	CT/CC/CA/CG	[3](CiT), [3](CiC), [3](CiA), [3](CiG)
		1st	GT/GC/GA/GG	[1](iGT), [1](iGC), [1](iGA), [1](iGG)
	3*AGA/AGG	3rd	AG	[2](AGi), [3](AGi)
		2nd	AA/AG	[3](AiA), [3](AiG)
		1st	GA/GG	[2](iGA), [2](iGG)
3*Gly	3*GGT/GGC/GGA/GGG	3rd	GG	[0](GGi), [1](GGi), [2](GGi), [3](GGi)
		2nd	GT/GC/GA/GG	[3](GiT), [3](GiC), [3](GiA), [3](GiG)
		1st	GT/GC/GA/GG	[3](iGT), [3](iGC), [3](iGA), [3](iGG)
3*Stop Codon	3*TAA/TAG	3rd	TA	[2](TAi), [3](TAi)
		2nd	TA/TG	[2](TiA), [2](TiG)
		1st	AA/AG	[0](iAA), [0](iAG)
	3*TGA	3rd	TG	[2](TGi)
		2nd	TA	[3](TiA)
		1st	GA	[0](iGA)

Table 4: Dataset Specification

Gene Name	# Isolates	# SNVs
SARS-CoV-2	39680	85
SARS-CoV-1	54	342
H1N1	35008	8

Table 5: Amino acid changes due to mutations occurred under rule 4 for the datasets of SARS-CoV-1 and SARS-CoV-2 and under rule 14 for H1N1 Type A virus

Substitution	Codon Position	Possible AA Changes	SARS-CoV-1	SARS-CoV-2
3*$C \rightarrow T$	1st	L→F, P→S, H→Y, R→W, R→C, Q→STOP CODON	H→Y, Q→STOP CODON, P→S, L→F, L→L	R→C, H→Y, L→F, L→L, P→S, Q→STOP CODON
	2nd	S→F, S→L, P→L, T→I, T→M, A→V	T→I, T→M, P→L, A→V, S→L, L→F	T→I, A→V, P→L, S→F
	3rd	Synonymous Changes (F→F, L→L, I→I, V→V, S→S, P→P, T→T, Y→Y, H→H, N→N, D→D, C→C, R→R, G→G)	N→N, T→T, S→S, I→I, Y→Y, L→L, D→D, A→A, G→G, V→V	Y→Y, C→C, F→F, H→H, I→I, L→L, N→N, S→S, T→T
Substitution	Codon Position	Possible AA Changes	H1N1 TYPE A	
3*$G \rightarrow A$	1st	V→M, V→I, A→T, D→N, E→K, G→R, G→S	A→T, E→K	
	2nd	G→D, G→E, R→K, S→N, R→Q, R→H, W→STOP CODON, Y→C	S→N, G→E	
	3*3rd	Synonymous Changes (L→L, S→S, STOP CODON→STOP CODON, L→L, P→P, Q→Q, R→R, T→T, K→K, V→V, A→A, E→E, G→G) Non-synonymous Changes (M→I, W→STOP CODON)	4*None	

Table 6: Substitutions in SARS-CoV-1 according to CA rule 8 and rule 2.

CA Rule (Substitution)	POSITION	AA CHANGED
3*Rule 2 ($T \rightarrow A$)	1st	STOP CODON→K, Y→N, S→T, C→ S, L→I, L→M
	2nd	I→K, V→E, M→K, I →N, V→D, V→E, L→STOP CODON, F→rY
	3rd	I→I, Y→STOP CODON, A→A, V→V, F→L
3*Rule 8 ($A \rightarrow T$)	1st	N→Y, I→L, K→ STOP CODON, I→F, N→Y, STOP CODON→ L
	2nd	STOP CODON→ L, D→V,Q→L
	3rd	G→G, A→A, L→F

References

[1] Négadi T. The genetic code multiplet structure, in one number. arXiv preprint, arXiv:07072011.

[2] Zhou Z, Dang Y, Zhou M, Li L, Yu Ch, and Fu J. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proceedings of the National Academy of Sciences.113(41):E6117-25.

[3] LaBella A.L, Opulente D.A, Steenwyk J.L, Hittinger C.T, Rokas A. Correction: Variation and selection on codon usage bias across an entire subphylum. PLoS Genetics. 2021;17(9):e1009824.

[4] Knight RD, Freeland S.J, Landweber L.F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome biology. 2001;2(4):1-13.

[5] Huang W, Guo Y, Li N, Feng Y, Xiao L. Codon usage analysis of zoonotic coronaviruses reveals lower adaptation to humans by SARS-CoV-2. Infection, Genetics and Evolution. 2021;89:104736.

[6] Hormoz S. Amino acid composition of proteins reduces deleterious impact of mutations. Scientific reports. 2013;3(1):1-10.

[7] Bofkin L, Goldman N. Variation in evolutionary processes at different codon positions. Molecular Biology and Evolution. 2007;24(2):513-21.

[8] Błażej P, Wnętrzak M, Mackiewicz D, Mackiewicz P. Correction: Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. Plos one. 2018;13(10):e0205450.

[9] Zeb A, Alzahrani E, Erturk VS, Zaman G. Mathematical model for coronavirus disease 2019 (COVID-19) containing isolation class. BioMed research international.2020;2020.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 1–21. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-1

[10] Jiang S, Li Q, Li C, Liu S, He X, Wang T. Mathematical models for devising the optimal SARS-CoV-2 strategy for eradication in China, South Korea, and Italy. Journal of translational medicine. 2020;18(1):1-11.

[11] Wu X, Cai Y, Huang X, Yu X, Zhao L, Wang F. Co-infection with SARS-CoV2 and influenza A virus in patient with pneumonia, China. Emerging infectious diseases. 2020;26(6):1324.

[12] Silvestro D, Antonelli A, Salamin N, Quental TB. The role of clade competition in the diversification of North American canids. Proceedings of the National Academy of Sciences. 2015;112(28):8684-9.

[13] Sengupta A, Hassan SS, Choudhury PP. Clade GR and clade GH isolates of SARSCoV-2 in Asia show highest amount of SNPs. Infection, Genetics and Evolution. 2021;89:104724.

[14] Kari J. Theory of cellular automata: A survey. Theoretical computer science.2005;334(1-3):3-33.

[15] Kiran Sree P, Babu IR, Usha Devi N S. Cellular Automata and Its Applications in Bioinformatics: A Review. arXiv e-prints. 2014:arXiv-1404.

[16] Burks C, Farmer D. Towards modeling DNA sequences as automata. Physica D: nonlinear phenomena. 1984;10(1-2):157-67.

[17] GCh S, Karafyllidis I, Ch M, Mardiris V, Thanailakis A, Tsalides P. A cellular automaton model for the study of DNA sequence evolution. Computers in Biology and Medicine. 2003;33(5):439-53.

[18] Ch M, GCh S, Mardiris V, Karafyllidis I, Glykos N, Sandaltzopoulos R. Reconstruction of DNA sequences using genetic algorithms and cellular automata: Towards mutation prediction? Biosystems. 2008;92:61-8.

[19] Chaudhuri PP, Ghosh S, Dutta A, Choudhury SP. A New Kind of Computational Biology: Cellular Automata Based Models for Genomics and Proteomics. Springer;2018.

[20] Dogaru R, Chua L.O. Mutations of the" Game of Life": A generalized cellular automata perspective of complex adaptive systems. International Journal of Bifurcation and Chaos. 2000;10(08):1821-66.

[21] Madain A, Dalhoum A.L.A, Sleit A. Application of local rules and cellular automata in representing protein translation and enhancing protein folding approximation. Progress in Artificial Intelligence. 2018;7(3):225-35.

[22] Ghosh S, Bhattacharya S. Computational model on COVID-19 pandemic using probabilistic cellular automata. arXiv preprint arXiv:200611270. 2020.

[23] Basu S, Ghosh S. Fuzzy Cellular Automata Model for Discrete Dynamical System Representing Spread of MERS and COVID-19 Virus. In: Internet of Medical Things for Smart Healthcare. Springer; 2020. p. 267-304.

[24] Sree P.K, Nedunuri SUD. A novel cellular automata classifier for COVID-19 trend prediction. Journal of Health Sciences. 2020;10(1).

[25] Beauchemin C, Samuel J, Tuszynski J. A simple cellular automaton model for influenza A viral infections. Journal of theoretical biology. 2005;232(2):223-34.

[26] Beauchemin C. Probing the effects of the well-mixed assumption on viral infection dynamics. Journal of theoretical biology. 2006;242(2):464-77.

[27] Beauchemin C, Forrest S, Koster F.T. Modeling influenza viral dynamics in tissue. In: International Conference on Artificial Immune Systems. Springer; 2006. p. 23-36.

[28] Crick F.H. The origin of the genetic code. Journal of molecular biology.1968;38(3):367-79.

[29] Ghosh S, Basu S. Some Algebraic Properties Of Linear Synchronous Cellular Automata. arXiv preprint arXiv:170809751. 2017.

[30] Ghosh S. Evolutions of Some One-Dimensional Homogeneous Cellular Automata. Complex Systems. 2021;30(1).