ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Similarity in Sequences of COVID-19 Genetic Codes over Galois Field

Rapin Sunthornwat[1], Yupaporn Areepong[2] and Prathum Prommi[3]

ABSTRACT

The Coronavirus disease 2019 (COVID-19) outbreak has caused the economic and health problems for all countries. The origin based on genetic codes of its spreading is a significant key for identification and solution of the outbreak. The purpose of this research is to study the relationships based on similarity measurement over Galois field amongst genetic codes of COVID-19. A Galois field is an abstract structure for converting genetic codes to binary codes derived from polynomials and then similarity is measured by examing the binary codes. The application is the investigation of the relationships amongst the sequences of genetic codes of COVID-19 particles contaminated from waste water in Brazil, Spain, Italy and the sequences of COVID-19 genetic codes in Thailand and China over Galois field. The finding shows that the similarity of COVID-19 genetic code sequences between China and Brazil is the maximum similarity, 99.9746%. In addition, the relationships amongst the sequences' genetic COVID-19 codes from Wuhan markets, SARS and bats are also investigated over a Galois field. The finding found that the similarity of COVID -19 genetic codes sequences between Bat coronavirus RaTG13-MN996532.1 and Wuhan market- LR757995.1 is the maximum similarity, 55.8548%. In conclusion, the sequence of COVID-19 genetic codes in Brazil is possibly significant and related to the sequence in China, and vice versa. The sequence of COVID-19 genetic codes at Wuhan market- LR757995.1 is possibly transmitted from Bat coronavirus RaTG13 genetic code to humans in China.

Article information:

Keywords: Coronavirus Disease 2019, Galois Field, Cosine Similarity, Genetic Code, Algebraic Statistics

Article history:
Received: February 18, 2021
Revised: June 21, 2021
Accepted: September 4, 2021
Published: June 4, 2022
(Online)

DOI: 10.37936/ecti-cit.2022162.245353

1. INTRODUCTION

The Coronavirus disease 2019 (for short, COVID-19) is a novel virus which is the second generation of Severe Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). The original source for COVID-19 around the end of year 2019 is at Wuhan, China. The first infectious COVID-19 case was about at a seafood market in Wuhan in the middle of December, 2019. The Wuhan market sells seafood but also it sells wildlife meat such as bat, snake, bird, fox, etc. which are disease carriers. However, the Coronavirus in exotic wildlife may have been transmitted to humans, especially in bats of Rhinolophus malayanus [1, 2]. The outbreak of COVID-19 has spread all regions in the world. Thailand is also affected by the COVID-19 outbreak, both economy and in the health system. On January 13, 2020, the Ministry of Public Health, Thailand, announced that the first COVID-19 infectious case was a 61-year-old Chinese woman as a tourist coming from Wuhan, China. Several researchers have discussed the exact the origin of COVID-19 in an article published in the archive medRxiv which is a medical journal [3, 4]. A research group from the Federal Univer-

[1]The author is with Industrial Technology and Innovation Management Program, Faculty of Science and Technology, Pathumwan Institute of Technology, Bangkok, Thailand, 10330. Email: rapin@pit.ac.th
[2]The author is with Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand, 10800. Email: yupaporn.a@sci.kmutnb.ac.th
[3]The author is with Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand, 10800. Email: prathum.p@sci.kmutnb.ac.th
[3]Corresponding author: prathum.p@sci.kmutnb.ac.th

sity of Santa Catarina (UFSC) in Brazil studied two waste water samples from Florianopolis of the Brazilian state of Santa Catarina. The particles of COVID-19 from the samples were found in Florianopolis on November 27, 2019. Researchers at the University of Barcelona detected the contaminated novel coronavirus in waste water sampled on March 12, 2019. This research about the management of the water cycle in Barcelona's metropolitan area is a corporation between the researchers from the Enteric Virus Laboratory of the University of Barcelona and the public-private company Aigues de Barcelona [5]. In addition, the researchers at the National Institute of Health (ISS) of Italy found COVID-19 particles on 40 waste water samples collected between October 2019 and February 2020. These samples are compared with 24 control samples collected between September 2018 and June 2019. The results showed that 24 control samples are contaminated with COVID-19 the newer 40 samples [6]. COVID-19 can be easily transmitted via respiratory droplets from human to human. It is a rapidly spreading disease in all countries of the world. The control of its spread is difficult. To study the causes of emerging and evolution of COVID-19, molecular biology of COVID-19 and COVID-19 genetic code sequence studies have been investigated to develop vaccines and to discover drugs for control and to suppress the spreading of COVID-19.

In molecular biology, the genetic code consists of 4 bases: Adenine, Thymine, Cytosine, and Guanine. A group of three bases is called a codon. There are 64 codons representing 20 amino acids. Sequences of genetic codons are important in a lot of medical research such as drug discovery, marker design, detection, classification, etc. Application of finite-state automata theory, regular expressions, and partially ordered sets of genetic codes was used for detecting and classifying mutations in sequences of Beta-Thalassaemia genetic codes in Thailand [7]. Genomewide analysis of single nucleotide polymorphism data of Crohn's disease was analysed by nonparametric and kernel machine regression [8]. The genetic code sequences of COVID-19 have been studied for identification of the SARS-CoV-2 coronavirus and protecting its spread. Identification of coronavirus isolated from a patient in Korea with COVID-19 was studied by full genome sequencing and electron microscopy. The finding revealed that the virus genome exhibited sequence homology of SARS-CoV-2 isolated from patients from other countries [9]. The source, evolution, and mutation in genetic sequence of COVID-19 for attracting and retaining knowledge to control spreading of COVID as well as to learn the relationships of bats and human coronaviruses [10, 11]. Genome sequencing of the SARS-CoV-2 virus which exists in wildlife animals was revealed and discussed to fill in the gaps of understanding of its origins [12]. The short section of COVID-19 found by the bioinformatics tool was used as the suitable peptide for creating a peptide synthetic vaccine and a peptidomimetic therapeutic method [13]. Genomewide association of 1980 patients with Covid-19 and severe disease was conducted and cross-replicating associations could be detected [14]. Bat populations which may be an ancestor of coronaviruses in 15 provinces of China were sampled and phylogenetic relations were analysed by sequencing full genomes of coronaviruses from sampled bats [15].

Moreover, the analysis of the sequences of genetic codes needs background knowledge of mathematics, statistics, and computers. Binary codes for representing bases of 64 codons based on partially order sets were constructed and established to build Hasse diagrams for genetic code structure with the relation in the same position [16]. A Galois Field was constructed as an abstract algebra structure for a finite field of polynomials [17]. The various architectures, methods, and techniques for execution cryptographic operations on Galois field arithmetic with realization of cryptographic algorithms were proposed [18]. Genetic similarities between individuals from different populations and the same populations were studied [19]. A classification which is based on number of loci, allele frequency, populations sampled, and polymorphism ascertainment strategy was conducted. The similarity between viral codon usage and the codon usage of the individual genes of a host genome in in the cell cycle was analysed [20].

However, this research is an application of Galois fields to genetic codes of COVID-19. To describe the genetic codes of COVID-19, the polynomials and binary code patterns of genetic code are constructed over a Galois field. Furthermore, the similarity measurement of two sequences of genetic codes is presented to identify the relationship. The relationship on sequences of COVID-19 genetic codes contaminated in waste water in Brazil, Spain, and Italy as well as the sequences of COVID-19 genetic codes in Thailand and China were studied over Galois fields. Furthermore, the relationship of sequences of COVID-19 genetic codes at Wuhan markets, SARS coronavirus genetic codes, and bat coronavirus genetic codes is investigated over Galois fields. The next section provides mathematical and statistical background of the methods for this research. In addition, the results, discussion, and conclusions are given.

2. MATERIALS AND METHODS

This section presents the mathematical and biological backgrounds for this research (e.g. [17, 21-24]).

2.1 Mathematical Definitions and Theorems

Definition 2.1 Let F be a field which is an abstract structure in mathematics. A polynomial $f(x)$ over F is an expression of the from

$$f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n = \sum_{i=0}^{n} a_ix^i;\ a_i \in F$$

where a_i is a coefficient of x^i for $0 \leq i \leq n$,
x is the indeterminate,
n is a non-negative integer, the degree of $f(x)$.

In addition, $F[x]$ is the set of all polynomials over the field F. The degree of $f(x)$ is denoted as $\deg(f(x))$.

Theorem 2.1 If $f(x), g(x) \in F[x]$ with $g(x) \neq 0$, then there exist polynomials $q(x)$ and $r(x)$ in $F[x]$ such that $f(x) = g(x)q(x) + r(x)$ and either $r(x) = 0$ or $\deg(r(x)) < \deg(g(x))$.

In addition, $q(x)$ is called a quotient and $r(x)$ is called a remainder. If $r(x) = 0$, then $f(x) = g(x_q(x)$ or $g(x)$ is a factor of $f(x)$, written $f(x)|g(x)$ or $f(x)$ is divisible by $g(x)$.

Definition 2.2 $f(x) \in F[x]$ is called a *reducible polynomial* over F if there exist polynomials $a(x), b(x) \in F[x]$ of $\deg(a(x))$ and $\deg(b(x))$ less than $\deg(f(x))$ such that $f(x) = a(x)b(x)$. Otherwise, $f(x) \in F[x]$ is called an *irreducible polynomial* over F.

In other words, $f(x) \in F[x]$ is called an irreducible polynomial over F if $f(x) = a(x)b(x)$ implies that either $a(x)$ or $b(x)$ is a non-zero constant polynomial.

Definition 2.3 $f(x) \in F[x], \beta$ is a root or solution of polynomial $f(x)$ if and only if $f(\beta) = 0$.

Let Z_p refer the integer $\{0, 1, 2, \ldots, p-1\}$ using modulo p arithmetic. Z_p is a field if and only if p is a prime number. A finite field is sometimes called a *Galois field* with a finite number of p^n elements and can be denoted by F_q or $GF(q)$ where $q = p^n$ for any prime number p and any positive integer n. In other words, $GF(p^n)$ comprises all of the elements of degree n over the field Z_p.

Let $GF(p^n)[x] = F_q[x]$ be comprised all of the polynomials of degree n over the field Z_p. The coefficients of these polynomial take on values in the field Z_p.

If $f(x)$ is a polynomial with degree n in $F_q[x]$, then $F_q[x]/f(x)$ is a set of q^n polynomials with degree less than n and each coefficient in F_q.

Theorem 2.2 $F_q[x]/f(x)$ is a finite field if and only if $f(x)$ is an irreducible polynomial.

Therefore, the finite field $F_q[x]/f(x)$ can be constructed.

Definition 2.4 An irreducible polynomial $f(x)$ of degree n over field F_q is called a *primitive polynomial* if $f(x)$ is divisible by $(x^k - 1)$ for $k = q^m - 1$ and $f(x)$ is not divisible by $(x^k - 1)$ for $k < q^m - 1$.

2.2 Genetic Code and Cosine Similarity Measurement

Genetic code is a code comprised of the 4 alphabet symbols which are called 4 bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). A sequence of three bases is called codon which contains the genetic information. It is possible to form $4^3 = 64$ patterns from 4 bases – A, G, C, and T – for forming codons with starting codon ATG and final codons TAA, TGA, and TAG (e.g. [21]). Thus, there are 64 possible codons corresponding to 20 amino acids. For example, Alanine={GCT, GCC, GCA, GCG}, Asparagine={AAT, AAC}, Cysteine ={TGT, TGC}, etc.

The similarity measurement of two sequences of genetic code which form the two vectors on vector space is based on inner products and norms of vectors (e.g. [23, 25]).

Let $\vec{x} = (x_1, x_2, \ldots, x_n$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$ be vectors of codes in vector space with dimension n. Similarity between two vectors $\vec{x}$ and $\vec{y}$ can be defined by using an inner product and norm as

$$\text{Similarity}=\cos(\theta) = \frac{<\vec{x}, \vec{y}>}{\|\vec{x}\|\|\vec{y}\|} = \frac{x_1y_1 + x_2y_2 + \ldots + x_ny_n}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

where θ is an angle between two vectors $\vec{x}$ and $\vec{y}$.

Next, the algorithm for grouping the vectors is also provided.

Algorithm 1. Algorithm for grouping based on similarity (e.g. [24])

Step1: Compute the similarity matrix $\mathbf{S} = (s_{ij})$ for $i, j = 1, 2, 3, \ldots, n$.

Step2: Find two vectors which are the most similar and collect them into one group.

Step3: Calculate the similarity matrix between the remaining $n - 1$ groups.

Step 4: Find two groups with the most intergroup similarity and join them.

Step 5: Repeat step 4 until all vectors are combined in groups.

Step 6: Construct a dendrogram.

2.3 Construction of Polynomials and Binary Codes of Genetic Codes with Application to COVID-19

In order to identify the similarity of two sequences of genetic codes, the sequences of genetic codes are collected from GenBank and other NCBI repositories which are from the National Institutes of Health (NIH) genetic sequence database on the website (e.g. [26]). The GenBank and NCBI are the tools for sampling of the genomic sequences of COVID-19 because they are well-known and generalized databases of the genomic sequences. Therefore, the sequences of genetic codes are selected from GenBank and other NCBI repositories as the samples for this research.

To identify the similarity which is a relationship amongst sequences of COVID-19 genetic codes for

contamination in waste water discussed in papers (e.g. [3-6]) of COVID-19 genetic codes in Thailand and China, the sequences of genetic codes of COVID-19 are selected via the purposive sampling technique which corresponds to the purposes of this research as follows.

Sequence of COVID-19 genetic codes in Thailand

Accession: **MT447176.1**

Definition: Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/THA/SI206917-NT/2020, complete genome

Nucleotide:

AACTTTCGATCTCTTGTAGATCTGTTCTCT
AAACGAACTTTAAAATCTGTGT
... ...
TAGTGCTATCCCCATGTGATTTTAATAGCT
TCTTAGGAGAATGACAAAAAAA

Sequence of COVID-19 genetic codes in China

Accession: **MT079851.1**

Definition: Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/WHUHnCoV011/2020, complete genome

Nucleotide:

ATTAAAGGTTTATACCTTCCCAGGTAACAA
ACCAACCAACTTTCGATCTCTT
... ...
AGCCTCAGCAGCAGATTTCTTAGTGACAGT
TTGGCCTTGTTGTTGTTGGCCT

Sequence of COVID-19 genetic codes in Brazil

Accession: **MT738101.1**

Definition: Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/BRA/HIAE-SP03/2020, complete genome

Nucleotide:

ATTAAAGGTTTATACCTTCCCAGGTAACAA
ACCAACCAACTTTCGATCTCTT
... ...
TTCTTAGGAGAATGACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAA

Sequence of COVID-19 genetic codes in Spain

Accession: **MT359866.1**

Definition: Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ESP/VH198152683/2020, complete genome

Nucleotide:

TATACCTTCCCAGGTAACAAACCAACCAAC
TTTCGATCTCTTGTAGATCTGT
... ...
TTCTTAGGAGAATGACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAA

Sequence of COVID-19 genetic codes in Italy

Accession: **MT531537.2**

Definition: Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/ITA/Siena-1/2020, complete genome

Nucleotide:

TAAAGGTTTATACCTTCCCAGGTAACAAAC
CAACCAACTTTCGATCTCTTGT
... ...
TTCTTAGGAGAATGACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAA

To identify the similarity in cases to show the relationship amongst the sequences of genetic code of COVID-19 at Wuhan market, SARS, and Bats which are discussed in papers (e.g. [10,11]), the following sequences of genetic codes were selected.

Sequence of COVID-19 genetic codes Wuhan market, China

Accession: **LR757998.1**

Definition: Wuhan seafood market pneumonia virus genome assembly, chromosome whole genome

Nucleotide:

AACAAACCAACCAACTTTCGATCTCTTGTA
GATCTGTTCTCTAAACGAAC TT
... ...
TGATTTTAATAGCTTCTTAGGAGAATGAC
AAAAAAAAAAAAAAA AAAAAA

Sequence of COVID-19 genetic codes Wuhan market, China

Accession: **LR757995.1**

Definition: Wuhan seafood market pneumonia virus genome assembly, chromosome: whole genome.

Nucleotide:

TTTCCCAGGTAACAAACCAACCAACTTTCG
ATCTCTTGTAGATCTGTTCTCT
... ...
CCATGTGATTTTAATAGCTTCTTAGGAGA
ATGACAAAAAAAAAAAAAAAAA

Sequence of COVID-19 genetic codes Wuhan market, China

Accession: **LR757996.1**

Definition: Wuhan seafood market pneumonia virus genome assembly, chromosome: whole genome

Nucleotide:

CAGGTAACAAACCAACCAACTTTCGATCTC
TTGTAGATCTGTTCTCTAAACG
... ...
ATGTGATTTTAATAGCTTCTTAGGAGAAT
GACAAAAAAAAAA AAAAAAAA

Sequence of COVID-19 genetic codes Wuhan, China

Accession: **NC_045512.2**

Definition: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

Nucleotide:

ATTAAAGGTTTATACCTTCCCAGGTAACAA
ACCAACCAACTTTCGATCTCTT
... ...
TTCTTAGGAGAATGACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAA

Sequence of coronavirus genetic codes in SARS

Accession: **MK062180.1**

Definition: SARS coronavirus Urbani isolate icSARS-MA, complete genome

Nucleotide:

ATATTAGGTTTTTACCTACCCAGGAAAAGCC
AACCAACCTCGATCTCTTGT

... ...

ATTTTAGTAGTGCTATCCCCATGTGATTT
TAATAGCTTCTTAGGAGAATGAC

Sequence of coronavirus genetic codes in Bat

Accession: **MN996532.1**

Definition: Bat coronavirus RaTG13, complete genome

Nucleotide:

CTTTCCAGGTAACAAACCAACGAACTCTCG
ATCTCTTGTAGATCTGTTCTCT

... ...

TGTGATTTTAATAGCTTCTTAGGAGAATG
ACAAAAAAAAAAAAAAAAAAAA

The genetic code sequences are constructed as polynomials over Galois fields based on abstract algebra and converted to binary codes based on coding theory. Binary codes for each codon of genetic code were developed from the Boolean algebra, Boolean lattices, partially order sets, and Hasse diagrams. The relation for construction of polynomials and binary codes of genetic codes is that the same position of base pair for each codon is related. Namely, 00(G) $<$ 01(A) $<$ 11(C) and 00(G) $<$ 10(T) $<$ 11(C) are ordered or comparable. On the other hand, 01 (A) and 10 (T) are not ordered and are not comparable (e.g. [7, 16]).

Next, the construction of a Galois field for the genetic code is shown. Let F be a field. The polynomial of degree n over field is expressed as

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots + a_nx^n = \sum_{i=0}^{n} a_ix^i$$

where $a_i \in F$ and $a_n \neq 0$.

To correspond to genetic code of $64(2^6)$ codons, the polynomial of degree $n = 6$ over Galois field F_2 is constructed as in the expression $f(x) = a_0+a_1x+a_2x^2+a_3x^3+\ldots+a_6x^6$ where $a_i \in F_2$ and $a_6 \neq 0$. There are $2^6 = 64$ expressions of polynomial $f(x)$. For example, $f(x) = 1 + x^2 + x^3 + x^6$ as $a_1 = a_4 = a_5 = 0$. Irreducible polynomials of degree 6 over field F_2 are 9 expressions : $1 + x + x^6$, $1 + x + x^2 + x^4 + x^6$, $1+x+x^2+x^5+x^6$, $1+x^3+x^6$, $1+x^2+x^3+x^5+x^6$, $1+x+x^3+x^4+x^6$, $1+x^5+x^6$, $1+x^2+x^4+x^5+x^6$, $1 + x + x^4 + x^5 + x^6$.

Besides, the primitive polynomials of degree 6 over field F_2 are 6 expressions : $1+x+x^6$, $1+x+x^2+x^5+x^6$, $1+x^5+x^6$, $1+x^2+x^3+x^5+x^6$, $1+x+x^3+x^4+x^6$, $1 + x + x^4 + x^5 + x^6$.

Therefore, there are 64 expressions of polynomials and binary codes which correspond to 64 amino acid codons. Because any two Galois fields of the same order are isomorphic, one polynomial of degree less than 6 from primitive polynomial is selected to be the generator for construction of polynomials over the Galois field. Here, $f(x) = 1 + x + x^6$ is selected as the primitive polynomial for construction of $GF(2^6) = F_2[2]/(1 + x^2 + x^6)$. If β is a root of $f(x) = 1 + x + x^6$, then $f(\beta) = 1 + \beta + \beta^6 = 0$. That is, $1 + \beta + \beta^6 = 0$ or $\beta^6 = 1 + \beta$. These constructed polynomials correspond to binary codons with the relation 00 (G) $<$ 01 (A) $<$ 11 (C) and 00 (G) $<$ 01 (T) $<$ 11 (C) and are shown in Table 1.

Furthermore, the similarity based on statistics is also calculated to identify the relationships and sources of the COVID-19 disease spread. The results are shown in the next section.

3. RESULTS AND DISCUSSION

In this section, the results of this research are given and interpreted.

3.1 Constructed Polynomials and Binary Codes of Genetic Code over Galois Field

A Galois field is an abstract structure in mathematics. It is a finite field with 2^n elements or with order 2^n. A Galois field with order 2^n can be denoted as $GF(2^n)$. The polynomials and binary codons over Galois fields are enumerated in Table 1.

Table 1: *Binary Codes, Genetic Codes, and Polynomials over Galois Field $GF2^6$.*

No.	Binary codon	Genetic Codon	Polynomial
1	000000	GGG	0
2	100000	TGG	1
3	010000	AGG	β
4	001000	GTG	β^2
5	000100	GAG	β^3
6	000010	GGT	β^4
7	000001	GGA	β^5
8	110000	CGG	$1 + \beta$
9	011000	ATG	$\beta + \beta^2$
10	001100	GCG	$\beta^2 + \beta^3$
11	000110	GAT	$\beta^3 + \beta^4$
12	000011	GGC	$\beta^4 + \beta^5$
13	110001	CGA	$1 + \beta + \beta^5$
14	101000	TTG	$1 + \beta^2$
15	010100	AAG	$\beta + \beta^3$
16	001010	GTT	$\beta^2 + \beta^4$
17	000101	GAA	$\beta^3 + \beta^5$
18	110010	CGT	$1 + \beta + \beta^4$
19	011001	ATA	$\beta + \beta^2 + \beta^5$
20	111100	CCG	$1 + \beta + \beta^2 + \beta^3$
21	011110	ACT	$\beta + \beta^2 + \beta^3 + \beta^4$
22	001111	GCC	$\beta^2 + \beta^3 + \beta^4 + \beta^5$

23	110111	CAC	$1+\beta+\beta^3+\beta^4+\beta^5$
24	101011	TTC	$1+\beta^2+\beta^4+\beta^5$
25	100101	TAA	$1+\beta^3+\beta^5$
26	100010	TGT	$1+\beta^4$
27	010001	AGA	$\beta+\beta^5$
28	111000	CTG	$1+\beta+\beta^2$
29	011100	ACG	$\beta+\beta^2+\beta^3$
30	001110	GCT	$\beta^2+\beta^3+\beta^4$
31	000111	GAC	$\beta^3+\beta^4+\beta^5$
32	110011	CGC	$1+\beta+\beta^4+\beta^5$
33	101001	TTA	$1+\beta^2+\beta^5$
34	100100	TAG	$1+\beta^3$
35	010010	AGT	$\beta+\beta^4$
36	001001	GTA	$\beta^2+\beta^5$
37	110100	CAG	$1+\beta+\beta^3$
38	011010	ATT	$\beta+\beta^2+\beta^4$
39	001101	GCA	$\beta^2+\beta^3+\beta^5$
40	110110	CAT	$1+\beta+\beta^3+\beta^4$
41	011011	ATC	$\beta+\beta^2+\beta^4+\beta^5$
42	111101	CCA	$1+\beta+\beta^2+\beta^3+\beta^5$
43	101110	TCT	$1+\beta^2+\beta^3+\beta^4$
44	010111	AAC	$\beta+\beta^3+\beta^4+\beta^5$
45	111011	CTC	$1+\beta+\beta^2+\beta^4+\beta^5$
46	101101	TCA	$1+\beta^2+\beta^3+\beta^5$
47	100110	TAT	$1+\beta^3+\beta^4$
48	010011	AGC	$\beta+\beta^4+\beta^5$
49	111001	AGC	$1+\beta+\beta^2+\beta^5$
50	101100	TCG	$1+\beta^2+\beta^3$
51	010110	AAT	$\beta+\beta^3+\beta^4$
52	001011	GTC	$\beta^2+\beta^4+\beta^5$
53	110101	CAA	$1+\beta+\beta^3+\beta^5$
54	101010	TTT	$1+\beta^2+\beta^4$
55	010101	AAA	$\beta+\beta^3+\beta^5$
56	111010	CTT	$1+\beta+\beta^2+\beta^4$
57	011101	ACA	$\beta+\beta^2+\beta^3+\beta^5$
58	111110	CCT	$1+\beta+\beta^2+\beta^3+\beta^4$
59	011111	ACC	$\beta+\beta^2+\beta^3+\beta^4+\beta^5$
60	111111	CCC	$1+\beta+\beta^2+\beta^3+\beta^4+\beta^5$
61	101111	TCC	$1+\beta^2+\beta^3+\beta^4+\beta^5$
62	100111	TAC	$1+\beta^3+\beta^4+\beta^5$
63	100011	TGC	$1+\beta^4+\beta^5$
64	100001	TGA	$1+\beta^5$

3.2 Similarity amongst COVID-19 genetic codes sequences over Galois Field

The relationships based on the similarity measurement to identify the origin of COVID-19 amongst Thailand-MT447176.1, China-MT079851.1, and the contamination of COVID-19 particles in waste water of Brazil-MT738101.1, Spain-MT359866.1, and Italy-MT531537.2 are displayed in this section. Moreover, the similarity of sequences of COVID-19 genetic codes from Wuhan market, SARS coronavirus genetic codes, and bat coronavirus genetic codes are shown.

Table 2: *Similarity Matrix for COVID-19 Genetic Code Sequences amongst Thailand, China, Brazil, Spain, and Italy.*

MT	447176.1	079851.1	738101.1	359866.1	531537.2
447176.1	100	48.9845	48.9837	49.8278	49.3306
079851.1	48.9845	100	99.9746	48.9872	48.9931
738101.1	48.9837	99.9746	100	48.9931	49.8425
359866.1	49.8278	48.9872	48.9931	100	48.8441
531537.2	49.3306	49.8366	49.8425	48.8441	100

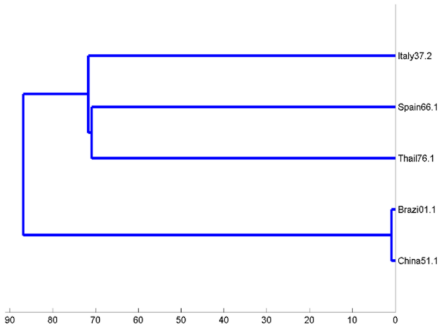

Fig. 1: *Dendrogram for COVID-19 codes amongst Thailand, China, Brazil, Spain, and Italy.*

Table 2 demonstrates the similarity amongst the sequences of COVID-19 genetic codes in Thailand, China, Brazil, Spain, and Italy. Fig. 1 is the dendrogram showing the hierarchical relationship amongst genetic codes sequences for COVID-19 in contaminated waste water in Brazil, Spain, Italy as well as COVID-19 genetic code sequences in Thailand and China. The finding shows that the similarity of COVID-19 genetic code sequences between China and Brazil is the maximum similarity, 99.9746%. That is, it is may be possible that the COVID-19 spreading in Brazil derives from COVID-19 in China, and vice versa. Besides, the similarity of COVID-19 genetic code sequences between the two other countries is approximately 50%.

Table 3: *Similarity Matrix in COVID-19 Genetic Codes Sequences for Wuhan Markets, SARS, and Bat.*

	LR 757998.1	LR 757995.1	LR 757996.1	NC 045512.2	MK 062180.1	MN 996532.1
LR 757998.1	100	48.9625	48.4386	48.8811	50.3000	48.7348
LR 757995.1	48.9625	100	48.4303	50.3610	49.3858	55.8548
LR 757996.1	48.4386	48.4303	100	49.1295	49.3926	49.5754
NC 045512.2	48.8811	50.3610	49.1295	100	49.1768	50.2568
MK 062180.1	50.3000	49.3858	49.3926	49.1768	100	49.1509
MN 996532.1	48.7348	55.8548	49.5754	50.2568	49.1509	100

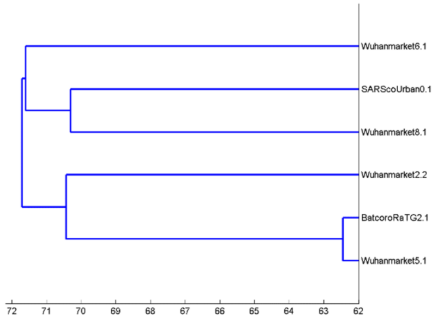

Fig.2: *Dendrogram for Wuhan Markets, SARS, and Bats.*

Table 3 demonstrates the similarity amongst sequences of COVID-19 genetic codes of Wuhan markets, SARS, and bats. Fig. 2 is the dendrogram showing the hierarchical relationship amongst genetic sequences from Wuhan market-LR757998.1, Wuhan market- LR757995.1, Wuhan market-LR757996.1, Wuhan market-NC_045512.2, SARS coronavirus Urbani-MK062180.1, and Bat coronavirus RaTG13-MN996532.1. The results found that the similarity of COVID -19 genetic codes sequences between Bat coronavirus RaTG13-MN996532.1 and Wuhan market- LR757995.1 is the maximum similarity, 55.8548%. That is, it is might be possible that the origin of COVID-19 in Wuhan market- LR757995.1 were transmitted from Bat coronavirus RaTG13 to humans in China. In addition, the similarity between the two other genetic codes sequences is approximately 50%.

3.3 Discussion

In this research, converting COVID-10 genetic codes to binary code is carried out by the Galois field. The main method for this research is to evaluate the similarity of the binary code to identify the COVID-19 genetic code amongst the sample countries: Thailand, China, Brazil, Spain, and Italy. The result from this research is that China is mostly related to Brazil. Moreover, the similarity amongst the COVID genetic code from Wuhan Markets, SARS, and Bat is investigated. The result shows that the genetic code from Bat coronavirus RaTG13-MN996532.1 and Wuhan market- LR757995.1 is mostly related. This research was done to evaluate the similarity for studying the origin of viral sequences of SARS-CoV at Wuhan market- LR757995.1. However, other research works have studied several aspects of the coronavirus genetic sequences such as identification of the origin, diversification, evolution, monitoring, tracking, and vaccination. Genetic similarity and diversification among coronavirus strains is investigated to gain the origin, evolutionary history, and phylogeny of targets of SARS-CoV species for developing vaccines are compared and investigated [27]. The origin of the severe acute respiratory syndrome outbreak, based on genetic similarities of human coronavirus OC43 (HCoV-OC43) and bovine coronavirus (BCoV) that allowed a nonhuman coronavirus to adapt to a human host is studied [28]. The method of genome annotation for Genome-Wide Analysis of SARS-CoV virus strains was employed to monitor and track the pandemic situation. The complete genome sequence SARS-CoV Wuhan-Hu-1 strain (Accession NC_045512, Version NC_045512.2) was selected as the referenced genome [29]. The search for the origin of the SARS-CoV outbreak based on the similarity and phylogenetic analysis was carried out and the result found that the virus is closely related to bat coronavirus RaTG13, approximately 96.3% [30].

In contrast to other research work [27-30], the method for this research is based on the similarity and the relationship between two sequences of COVID-19 genetic codes. This result from this research corresponds to the investigation of the origin of SARS-CoV [30] in that the original SARS-CoV and bat coronavirus are related to each other. Namely, the result from the research [30] showed that the sequence of the genetic code of the origin SARS-CoV virus at Wuhan market- LR757995.1 was most similar to the sequence of bat coronavirus RaTG13. The result from this research showed also that the sequence of the genetic code of Bat coronavirus RaTG13 and Wuhan market-LR757995.1 returned the maximum similarity.

4. CONCLUSIONS

This research constructed the polynomials and measured similarity of genetic code sequences over a Galois field with application to sequences of COVID-19. The polynomials of COVID-19 genetic codes are the abstract structure for measuring similarity of COVID-19 genetic code sequences. The origin of COVID-19 is unknown. Brazil, Spain, and Italy are original countries of COVID-19 emerging by the report of the archive medRxiv medical journal in cases of discussion on the origin of COVID-19 contaminated from waste water, but COVID-19 is generally known to have emerged at Wuhan market, China, at the end of year 2019. This research is the application of similarity measurement over a Galois field to identify the relationships of COVID-19 genetic sequences amongst China, Brazil, Spain, Italy, as well as Thailand. In addition, the relations of genetic sequences amongst Wuhan markets COVID-19, SARS coronavirus, and bat coronavirus were identified by similarity measurements. This research found that the sequence of COVID-19 genetic code in Brazil is significant and related to COVID-19 in China, and vice versa because the similarity between Brazil and China is 99.9746%. Moreover, COVID-19 at Wuhan market- LR757995.1 was probably transmitted from Bat coronavirus RaTG13 to humans in China. However, the similarity is about 50%. It implies that the COVID-19 genetic code may be mutated until the se-

quences of COVID-19 genetic codes are not similar. Therefore, the mutation of COVID-19 with similarity should be focused in future research. Moreover, this research investigates the similarity in sequences of COVID-19 genetic codes based on algebraic structure for disease genetic codes. The polynomials of disease genetic code sequences are also established for the contribution to coding theory research. The method from this research can be applied to genetic code sequences of other diseases such as H1N1 Flu Virus (Swine Flu).

ACKNOWLEDGMENT

The authors have to thank Pathumwan Institute of Technology and King Mongkut's University of Technology North Bangkok for support and encouragement during this research.

References

[1] Z. Wu et al., "Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases," *ISME Journal*, Vol. 10, No.3, pp. 609–620, 2016.

[2] P. Zhou et al., "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, Vol. 579, pp. 270–273, 2020.

[3] F. Gislaine et al., "SARS-CoV-2 in human sewage in Santa Catalina, Brazil," *medRxiv*, November 2019, 2020.06.26.20140731, 2019.

[4] Z. Lara, and M. M. Marcelo, "The COVID-19 pandemic in Brazil: An urge for coordinated public health policies," ⟨hal-02881690v2⟩, 2020.

[5] C. M. Gemma et al., "Sentinel surveillance of SARS-CoV-2 in wastewater anticipates the occurrence of COVID-19 cases," *medRxiv*, 2020.06.13.20129627, 2020.

[6] L. R. Giuseppina et al., "First detection of SARS-CoV-2 in untreated waste waters in Italy," *Science of the Total Environment*, Vol. 736, pp.139652, 2020.

[7] R. Sunthornwat, E. J. Moore, and Y. Temtanapat, "Detecting and Classifying Mutations in Genetic Code with an Application to Beta-Thalassemia," *Science Asia*, Vol. 37, No. 1, pp. 51-61, 2011.

[8] P. Kirdwichai and M. Baksh, "The analysis of genomewide SNP data using nonparametric and kernel machine regression," *The Journal of Applied Science*, Vol. 18, pp. 20-30, 2019.

[9] J. M. Kim et al., "Identification of Coronavirus Isolated from a Patient in Korea with COVID-19," *Osong Public Health and Research Perspectives*, Vol. 11, pp. 3-7, 2020.

[10] J. C. Perez, "Wuhan COVID-19 Synthetic Origins and Evolution," *International Journal of Research – GRANTHAALAYAH*, Vol. 8, pp. 285-324, 2020.

[11] J. C. Perez and L. Montagnier, "COVID-19, SARS and Bats Coronaviruses Genomes Unexpected Exogeneous RNA Sequences," OSF Preprints, 2020.

[12] Z. Z. Yong and C. H. Edward, "A Genomic Perspective on the Origin and Emergence of SARS-CoV-2," *Cell*, Vol. 181, pp. 223-227, 2020.

[13] B. Robson, "Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus," *Computers in Biology and Medicine*, Vol.119, pp. 103670, 2020.

[14] D. Ellinghaus et al. "Genomewide Association Study of Severe Covid-19 with Respiratory Failure," *New England Journal of Medicine*, Vol. 383, pp. 1522-1534, 2020.

[15] X. C. Tang et al., "Prevalence and genetic diversity of coronaviruses in bats from China," *Journal of Virology*, Vol. 80, pp. 7481-7490, 2006.

[16] S. Robert, M. Eberto and G. Ricardo, "A Genetic code Boolean structure I. The meaning of Boolean deductions," *Journal of Mathematical Biology*, Vol. 67, pp. 1-14, 2005.

[17] A. G. Joseph, *Contemporary abstract algebra*, 4th ed, Houghton Mifflin, Boston, 1998.

[18] E. Savas and C. Koc, "Finite Field Arithmetic for Cryptography," *Circuits and Systems Magazine-IEEE*, Vol. 10, pp. 40 - 56, 2010.

[19] D. J. Witherspoon et al., "Genetic Similarities Within and Between Human Populations," *Genetics*, Vol. 176, pp. 351–359, 2007.

[20] K. Jitobaom et al., "Codon usage similarity between viral and some host genes suggests a codon-specific translational regulation," *Heliyon*, Vol. 6, pp. e03915, 2020.

[21] P. S. Eldra, R. B. Linda and W. M. Diana, *Biology*, 6th ed, Brooks/Cole, Australia, 2002.

[22] R. Lewis, *Human Genetics: Concepts and Applications*, 6th ed, McGraw-Hill, Boston, 2005.

[23] J. Han, M. Kamber and J. Pei, *Data mining: Concepts and techniques*, 3rd ed, Morgan Kaufmann Publishers, Waltham Mass, 2012.

[24] W. K. Härdle and H. Zdenek, *Multivariate statistics exercises and solutions*, London, Springer, 2015.

[25] W. K. Härdle and L. Simar, *Applied multivariate statistical analysis*, 4th ed, Springer, London, 2019.

[26] National Centre of Biotechnology Information, "NCBI SARS-CoV-2 Resources," 2020. [online] Available at: `https://www.ncbi.nlm.nih.gov/sars-cov-2`

[27] N. Kaur et al., "Genetic comparison among various coronavirus strains for the identification of

potential vaccine targets of SARS-CoV2," *Infection, Genetics and Evolution*, Vol. 89, pp. 1-15, 2021.

[28] L. Vijgen et al., "Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event," *Journal of Virology*, Vol. 79, pp. 1595-1604, 2005.

[29] M. R. Islam et al., "Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity," *Scientific Reports*, Vol. 10, pp. 1-9, 2020.

[30] Y. A. Helmy et al., "The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control," *Journal of Clinical Medicine*, Vol. 9, pp. 1-29, 2020.

Rapin Sunthornwat received the B.S. degree in Applied Mathematics from King Mongkut's University of Technology North Bangkok, B.Acc. degree in Accounting and Finance from Ramkhamhaeng University, Bangkok, Thailand, the M.S. degree in Applied Mathematics, and the Ph.D. degree in Applied Statistics from King Mongkut's University of Technology North Bangkok, Bangkok, Thailand. At present, he is an Assistant Professor of the program in industrial technology and innovation management, faculty of science and technology, Pathumwan Institute of Technology, Bangkok, Thailand. He is an expert in biostatistics, algebraic statistics, statistical process control, and forecasting modelling. Nowadays, his fields of interest are application growth curve, stochastic model to predict COVID-19 outbreak, and financial statistics.

Yupaporn Areepong received the B.S. degree in Statistics from Chiang Mai University, Chiang Mai, Thailand, the M.S. degree in Statistics from Chulalongkorn University, Bangkok, Thailand, and the Ph.D. degree in Mathematical Science from University of Technology Sydney, Sydney, Australia. She is currently a Professor at department of applied statistics in King Mongkut's University of Technology North Bangkok, Thailand. She is an expert in statistical process control, sequential change-point analysis, and time series analysis. Nowadays, her fields of interest are application statistical process control for monitoring and controlling COVID-19 outbreak.

Prathum Prommi received the B.Ed. degree in Mathematics from Srinakharinwirot University, Bangkok, Thailand and the M.S. degree in Mathematics from Chiang Mai University, Chiang Mai, Thailand. She is an Associate Professor at department of mathematics in King Mongkut's University of Technology North Bangkok, Thailand. She is an expert in abstract algebra and its applications.