On spatial molecular arrangements of SARS-CoV2 genomes of Indian patients

Sk. Sarif Hassan[a], Atanu Moitra[b], Ranjeet Kumar Rout[c,*], Pabitra Pal Choudhury[d], Prasanta Pramanik[e], Siddhartha Sankar Jana[f,**]

[a]*Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram 721140, India*
[b]*CMO, Government of West Bengal, India.*
[c]*Department of Computer Science and Engineering, National Institute of Technology, Srinagar, Hazratbal, Kashmir-190006(J& K), India.*
[d]*Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, West Bengal, India.*
[e]*Finance Department, Government of West Bengal, India.*
[f]*School of Biological Sciences, Indian Association for the Cultivation of Science, West Bengal, 700032, India.*

Abstract

A pandemic caused by severe acute respiratory syndrome coronavirus-2(SARS-CoV2) is being experienced by the whole world since December, 2019. A thorough understanding beyond just sequential similarities among the protein coding genes of SARS-CoV2 is important in order to differentiate or relate to the other known CoVs of the same genus. In this study, three genomes namely MT012098 (India-Kerala), MT050493 (India-Kerala), MT358637 (India-Gujrat) from India and another one NC_045512 (China-Wuhan) as a reference genome from China are considered to view the spatial as well as molecular arrangements of nucleotide bases of all the genes embedded in these four genomes. Based on different features extracted for each genes embedded in these genomes, corresponding phylogenetic relationships have been built up. This study would help to understand the virulence factors, disease pathogenicity, origin and transmis-

*Corresponding author
**Co-corresponding author
Email addresses: sarimif@gmail.com (Sk. Sarif Hassan), sa-wb@hda.gov.in (Atanu Moitra), ranjeetkumarrout@nitsri.net (Ranjeet Kumar Rout), pabitra@isical.ac.in (Pabitra Pal Choudhury), pk.pramanik@nic.in (Prasanta Pramanik), bcssj@iacs.res.in (Siddhartha Sankar Jana)

sion of the SARS-CoV2.

Keywords: MT012098 (India-Kerala); MT050493 (India-Kerala); MT358637 (India-Gujrat); NC_045512 (China-Wuhan); Spatial arrangement; SARS-CoV2.

1. Introduction

The disease COVID-19 is caused by the SARS-CoV2 initiated in late December 2019 in Wuhan, China, and since then it has been impulsed various countries across the world [1]. Presently, this disease, a pandemic as announced by the WHO, is a major health concern [2]. The family of coronaviruses is enclosed by different CoVs which are a single-stranded, positive-sense RNA genome of size approximately 26-32 kb [3]. The CoVs are classified into four genera, the α-CoVs, β-CoVs, γ-CoVs and δ-CoVs [4]. One of most important genera of coronaviruses is the β-CoVs where the present SARS-CoV2 belongs [5]. The β-CoVs mainly infect humans, bats including other animals such as camels, and rabbits and so on [6]. Two-third of the SARS-CoV2 genomes from 5' end is conserved for the ORF1 gene which encodes sixteen polyproteins and the 3' ends contains various structural protein coding genes including surface (S), envelope (E), membrane (M), and nucleocapsid (N) proteins [7]. In addition there are six accessory protein coding genes such as ORF3a, ORF6, ORF7a, ORF7b, and ORF8 also present in the SARS-CoV2 genome [8]. The spatial arrangement of genes over the SARS-CoV2 genome is presented in the Fig.1. [9].

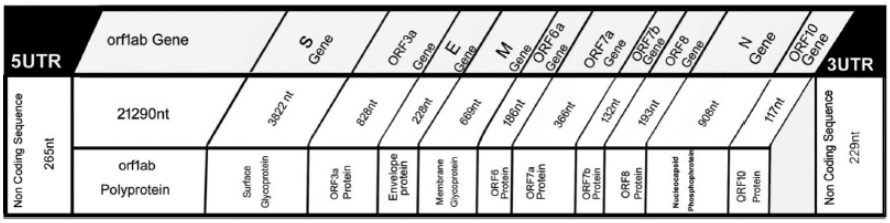

Figure 1: Spatial arrangement of genes over a typical SARS-CoV2 genome, Credit: [9].

The polyprotein ORF1ab encoded by the ORF1 gene play key roles in virus pathogenesis, cellular signalling, modification of cellular gene expression[10]. The envelope (E) proteins play multiple roles during infection, including virus morphogenesis [11, 12]. The N protein encoded by the gene N plays a vital role in the virus morphogenesis and assembly [13, 14]. The M gene encodes the M protein which plays a central role in virus morphogenesis and assembly via its interactions with other viral proteins [15, 16]. It does determine the shapes the virions, promotes membrane curvature. The M gene sequence of SARS-CoV2 is similar to that of SARS-CoV and MERS-CoV with 90.1% and 39.2% respectively [17]. The S protein (S gene) is one of most important structural proteins, which is used as a key that the virus uses to enter host cells. The spike protein attaches the virion to the cell membrane by interacting with host receptor and infects the host cell [18, 19]. In viral replication, the accessory proteins such as ORF3a, ORF6, ORF7a, ORF7b, ORF8 have a key role [20].

SARS-COV2 genome shows 79.6% identity with SRS-COV1 genome [21]. It is reported that the Spike glycoprotein of the Wuhan coronavirus is modified via homologous recombination [22]. The SARS-CoV2 is more phylogenetically related to SARS-CoV than to MERS-CoV [23]. Still, the proximal origin of COVID-19 transmission or evolutionary relationship of SARS-CoV2 and other coronaviruses is very much controversial. The outbreak and infectious behaviour of the SARS-CoVs and the lack of effective treatments for CoV infections de-

mand the need of detailed understanding of coronaviral molecular biology, with a specific focus on both their structural proteins as well as their accessory proteins.

From a molecular biology perspective, figuring out why the virus is so much virulent and infectious than other CoVs belonging to the genus β-CoVs is one of the most important aspects to look into [24]. The present SARS-CoV2 genomes are classified, based on SNPs, into two broad groups known as L and S [25]. The nature of virulence is also associated to the L type of CoV2 genomes.

Clearly, on having information of sequential similarity among genes and genomes of various CoVs is not enough to decipher the deep message regarding various characteristics viz. virulence, infection and transmission capacities, embedded in the RNA sequence. So in order to find out the genomic information of the two types (L and S) of SARS-CoV2, an attempt is made to discover the molecular and spatial organizations of each gene embedded over a sample of four genomes of which three of them are from India and one from China-Wuhan.

1.1. Datasets and Methods

In the NCBI virus database, as on 30th April, 2020, there are three complete genomes viz. MT012098 (India-Kerala), MT050493 (India-Kerala) and MT358637 (India-Gujrat) of SARS-CoV2 from Indian patients are available, which we consider for this present study. As a reference genome, NC_045512 (China-Wuhan) is taken. Note that, the genomes MT012098, MT050493 and NC_045512 belong to the S-type and other genome MT358637 from India belongs to L-type as per classification made based on SNPs[25]. An information regarding the lengths and names (followed strictly by NCBI database) of all eleven genes embedded in the four genomes is presented in the Table 1. Note that, in the Table 1, $'*'$ denotes absence of the gene in the respective genome.

Table 1: Information of the eleven genes of the India's and Wuhan's Genomes

Genes	NC_045512 (China)	MT012098 (India)	MT050493 (India)	MT358637 (India)
ORF1	21290	21291	21291	21291
S (ORF2)	3822	3819	3822	3822
ORF3a	828	828	828	828
E (ORF4)	228	228	228	228
M (ORF5	669	669	669	669
ORF6	186	186	186	186
ORF7a	366	366	366	366
ORF7b	132	*	*	132
ORF8	366	366	366	366
N (ORF9	1260	1260	1260	1260
ORF10	117	117	117	117

From the sequence based similarity, a phylogenetic relations is given in the Fig.1 which describe that the genomes NC_045512 from Wuhan and MT012098 from India are very close to each other with 99.98% sequential similarity as mentioned in the article by Yadav P.D. et.al. [26].

Figure 2: Phylogeny among the four genomes based on sequential similarities. Credit: NCBI

From the Fig.2 it is quite clear that the two genomes namely NC_045512 (China-Wuhan) and MT012098 (India-Kerala) are sequentially similar to the genome MT050493 (India-Kerala) while the other genome MT358637 (India-Gujrat) is not as sequentially similar to the other three genomes as shown. It is to be noted from Table-1 above that the gene ORF7b is absent in the genomes MT012098 and MT050493 both found in India.

Prior to proceed further, each sequence to a binary sequence of $1's$ and $0's$ as per the definition 2 is transformed to a binary representation. Here purine (A,G) and pyrimidine (T,C) bases are represented as $'1'$ and $'0'$ respectively. This binary representation is named as purine-pyrimidine representation [27, 28, 29, 30, 31].

$$\begin{aligned} A/G &\rightarrow 1 \\ T/C &\rightarrow 0 \end{aligned} \tag{1}$$

Also each sequence is transformed to a binary sequence with respect to a nucleotide base B of $1's$ and $0's$ as per the following definition **??**.

$$\begin{aligned} X &\rightarrow 1 \quad if\ X = B \\ X &\rightarrow 0 \quad if\ X \neq B \end{aligned} \tag{2}$$

Hence four binary representations for each nucleotide $B \in \{A, T, C, G\}$ would be derived for a given nucleotide sequence. These binary representations are actually the spatial template of each nucleotides. Each of these spatial templates are to be analysed using various methods as mentioned in the following.

Binary Shannon Entropy: The Shannon entropy (SE) measures information entropy of a Bernoulli process with probability p of the two outcomes (0/1) [32, 33]. It is defined as

$$SE = -\sum_{i=1}^{2} p_i log_2(p_i)$$

where $p_1 = \frac{k}{2^l}$ and $p_2 = \frac{l-k}{2^l}$; here l is the length of the binary sequence and k is the number of 1's in the binary sequence of length l [**?**]. The binary Shannon entropy is a measure of the uncertainty in a binary sequence. If the probability $p = 0$, the event is certainly never to occur, and so there is no uncertainty, leading to an entropy of 0. Similarly, if the probability $p = 1$, the result is certain, so the entropy must be 0. When $p = 0.5$, the uncertainty is at

a maximum and consequently the SE is 1.

Nucleotide Conservation Shannon Entropy: Shannon entropy is a measure of the amount of information (measure of uncertainty. Conservation of each of the four nucleotides has been determined using Shannon entropy [34]. For a given RNA sequence, the conservation SE is calculated as follows:

$$SE = -\sum_{i=1}^{4} p_{N_i} log_2(p_{N_i})$$

where p_{N_i} represents the occurrence frequency of a nucleotide N_i in a RNA sequence.

Hurst Exponent The Hurst Exponent (HE) is used to interpret the trend of a sequence, which could be positive or negative [35]. The HE belongs to the unit interval (0, 1). If HE lies within (0, 0.5) then the sequence possesses a negatively trending. Otherwise if the HE belongs to (0.5, 1) then the sequence is positively trending. If the HE is turned out to be 0.5 then the sequence must possesses randomness.

The HE of a sequence b_n (length: n) is defined as

$$\left(\frac{n}{2}\right)^{HE} = \frac{X(n)}{Y(n)} \tag{3}$$

where

$$Y(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - m)^2}$$

$Z_t = \sum_{j=1}^{t} S_t$ for $t = 1, 2, 3, \ldots, n$ where $S_t = b_t - m$ for $t = 1, 2, 3, \ldots, n$ and $X(n) = max(Z_i) - min(Z_i)$ for $i = 1, 2, 3, \ldots, n$

In addition to these two parameters Shannon entropy and Hurst exponent, some basic derivative features such as nucleotide frequency, double nucleotide frequency, codon usage frequency, GC content, purine-pyrimidine density are

obtained for a given nucleotide sequence [29, 31]. Also based on nucleotide densities, a decreasing order (density order) is obtained for a given sequence. It is worth noting that first positive frame has been considered to determine codons and double nucleotides over a given gene.

2. Results

By using the above methods features for all the genes viz. ORF1, S, ORF3a, E, M, ORF6, ORF7a, ORF7b, ORF8, N and ORF10 of the four genomes, are found and analysed.

2.1. Quantifications of genes over the four SARS-CoV2 genomes

For a given gene, we define a feature vectors as (length, frequency of individual nucleotides, GC content, % of purines and pyrimidines, frequency of each codon usage, frequencies of each double nucleotides, Shannon entropy (SE) and Hurst Exponent (HE) of the spatial representations of each nucleotides, purine and pyrimidine).

Here we briefly state the findings based on the feature vectors derived for every gene embedded in the four genomes and accordingly based on the findings some discussions are made. Before we proceed to make specific observations about codon and double nucleotide usages we present a table (Table-2) below describing the molecular information of the each genes with associated remarks.

Table 2: The molecular descriptions of the genes of four genomes with remarks

Genome_id	Gene	Length	A	T	C	G	GC Content	Pyrimidines	*Remarks*
NC_045512	**E**	228	49	92	45	42	38.1579	60.0877	*Pyrimidine-Rich; T-A-C-G*
MT012098	**E**	228	49	92	45	42	38.1579	60.0877	
MT050493	**E**	228	49	92	45	42	38.1579	60.0877	
MT358637	**E**	228	49	92	45	42	38.1579	60.0877	
NC_045512	**M**	669	171	213	146	139	42.6009	53.6622	*Pyrimidine-Rich; T-A-C-G*
MT012098	**M**	669	171	213	146	139	42.6009	53.6622	
MT050493	**M**	669	171	213	146	139	42.6009	53.6622	
MT358637	**M**	669	171	213	146	139	42.6009	53.6622	
NC_045512	**N**	1260	400	265	315	280	47.2222	46.0317	*Purine-Rich; A-C-G-T*
MT012098	**N**	1260	400	265	315	280	47.2222	46.0317	
MT050493	**N**	1260	400	265	315	280	47.2222	46.0317	
MT358637	**N**	1260	401	266	314	279	47.0635	46.0317	
NC_045512	**ORF1**	21290	6425	6891	3744	4230	37.4542	49.9530	*Purine-Rich; T-A-C-G*
MT012098	**ORF1**	21291	6425	6893	3743	4230	37.4477	49.9554	
MT050493	**ORF1**	21291	6424	6894	3742	4231	37.4477	49.9554	
MT358637	**ORF1**	21291	6424	6893	3743	4231	37.4524	49.9554	
NC_045512	**ORF10**	117	35	42	21	19	34.1880	53.8462	*Pyrimidine-Rich; T-A-C-G*
MT012098	**ORF10**	117	35	42	21	19	34.1880	53.8462	
MT050493	**ORF10**	117	35	42	21	19	34.1880	53.8462	
MT358637	**ORF10**	117	35	42	21	19	34.1880	53.8462	
NC_045512	**ORF3a**	828	225	276	174	153	39.4928	54.3478	*Pyrimidine-Rich; T-A-C-G*
MT012098	**ORF3a**	828	225	276	174	153	39.4928	54.3478	
MT050493	**ORF3a**	828	225	276	174	153	39.4928	54.3478	
MT358637	**ORF3a**	828	225	276	174	153	39.4928	54.3478	
NC_045512	**ORF6**	186	68	66	26	26	27.9570	49.4624	*Purine-Rich; A-T-C-G*
MT012098	**ORF6**	186	68	66	26	26	27.9570	49.4624	
MT050493	**ORF6**	186	68	66	26	26	27.9570	49.4624	
MT358637	**ORF6**	186	68	66	26	26	27.9570	49.4624	
NC_045512	**ORF7a**	366	108	118	79	61	38.2514	53.8251	*Pyrimidine-Rich; T-A-C-G*
MT012098	**ORF7a**	366	108	118	79	61	38.2514	53.8251	
MT050493	**ORF7a**	366	108	118	79	61	38.2514	53.8251	
MT358637	**ORF7a**	366	108	118	79	61	38.2514	53.8251	
NC_045512	**ORF8**	366	101	134	64	67	35.7923	54.0984	*Pyrimidine-Rich;T-A-G-C*
MT012098	**ORF8**	366	101	134	64	67	35.7923	54.0984	
MT050493	**ORF8**	366	101	133	65	67	36.0656	54.0984	
MT358637	**ORF8**	366	101	134	64	67	35.7923	54.0984	
NC_045512	**ORF7b**	132	31	60	24	17	31.0606	63.6364	*Pyrimidine-Rich: T-A-C-G*
MT358637	**ORF7b**	132	31	60	24	17	31.0606	63.6364	
NC_045512	**S**	3822	1125	1271	723	703	37.3103	52.1716	*Pyrimidine-Rich: T-A-C-G*
MT012098	**S**	3819	1124	1270	723	702	37.3134	52.1864	
MT050493	**S**	3822	1125	1272	722	703	37.2841	52.1716	
MT358637	**S**	3822	1123	1271	723	705	37.3626	52.1716	

In Table-3 below, SEs and HEs of spatial representations of the four nucleotide bases as well as of the purine-pyrimidine representations over each genes of four different genomes, are presented.

Table 3: The SEs and HEs of the spatial representations of the four nucleotide bases as well as of the purine-pyrimidine representations over the genes of four different genomes

Genome_id	Gene	HE_A	HE_C	HC_T	HC_G	SE_A	SE_C	SE_T	SE_G	B_HE	B_SE	SE-Conv
NC_045512	**E**	0.63928	0.54726	0.55748	0.57277	0.75077	0.71663	0.97297	0.68920	0.62451	0.97044	0.97673
MT012098	**E**	0.63928	0.54726	0.55748	0.57277	0.75077	0.71663	0.97297	0.68920	0.62451	0.97044	0.97903
MT050493	**E**	0.63928	0.54726	0.55748	0.57277	0.75077	0.71663	0.97297	0.68920	0.62451	0.97044	0.97903
MT358637	**E**	0.63928	0.54726	0.55748	0.57277	0.75077	0.71663	0.97297	0.68920	0.62451	0.97044	0.98152
NC_045512	**M**	0.63822	0.62440	0.66460	0.57028	0.82004	0.75693	0.90262	0.73720	0.63096	0.99613	0.97735
MT012098	**M**	0.63822	0.62440	0.66460	0.57028	0.82004	0.75693	0.90262	0.73720	0.63096	0.99613	0.97903
MT050493	**M**	0.63822	0.62440	0.66460	0.57028	0.82004	0.75693	0.90262	0.73720	0.63096	0.99613	0.97903
MT358637	**M**	0.63822	0.62440	0.66460	0.57028	0.82004	0.75693	0.90262	0.73720	0.63096	0.99613	0.98152
NC_045512	**N**	0.61742	0.50535	0.53219	0.55515	0.90160	0.81128	0.74209	0.76420	0.57043	0.99545	0.97929
MT012098	**N**	0.61742	0.50535	0.53219	0.55515	0.90160	0.81128	0.74209	0.76420	0.57043	0.99545	0.97903
MT050493	**N**	0.61742	0.50535	0.53219	0.55515	0.90160	0.81128	0.74209	0.76420	0.57043	0.99545	0.97903
MT358637	**N**	0.61484	0.51273	0.53645	0.55913	0.90247	0.81002	0.74360	0.76277	0.57043	0.99545	0.98152
NC_045512	**S**	0.59665	0.59186	0.61746	0.55111	0.87427	0.69973	0.91751	0.68861	0.58340	0.99864	0.97644
MT012098	**S**	0.59687	0.59037	0.61752	0.55202	0.87423	0.70004	0.91751	0.68836	0.58340	0.99862	0.97903
MT050493	**S**	0.59665	0.59070	0.61623	0.55111	0.87427	0.69918	0.91777	0.68861	0.58340	0.99864	0.97903
MT358637	**S**	0.59926	0.59186	0.61746	0.54722	0.87361	0.69973	0.91751	0.68973	0.58340	0.99864	0.98152
NC_045512	**ORF1**	0.62093	0.56846	0.66817	0.59449	0.88346	0.67093	0.90833	0.71929	0.64206	1.00000	0.97627
MT012098	**ORF1**	0.62093	0.56652	0.66840	0.59449	0.88346	0.67072	0.90843	0.71929	0.64206	1.00000	0.97903
MT050493	**ORF1**	0.62075	0.56849	0.66828	0.59536	0.88340	0.67062	0.90848	0.71938	0.64206	1.00000	0.97903
MT358637	**ORF1**	0.62075	0.56800	0.66793	0.59427	0.88340	0.67072	0.90843	0.71938	0.64206	1.00000	0.98152
NC_045512	**ORF10**	0.66535	0.53126	0.62087	0.52204	0.88024	0.67895	0.94183	0.64000	0.65199	0.99573	0.97648
MT012098	**ORF10**	0.66535	0.53126	0.62087	0.52204	0.88024	0.67895	0.94183	0.64000	0.65199	0.99573	0.97903
MT050493	**ORF10**	0.66535	0.53126	0.62087	0.52204	0.88024	0.67895	0.94183	0.64000	0.65199	0.99573	0.97903
MT358637	**ORF10**	0.66535	0.53126	0.62087	0.52204	0.88024	0.67895	0.94183	0.64000	0.65199	0.99573	0.98152
NC_045512	**ORF3a**	0.65733	0.60293	0.64026	0.54415	0.84395	0.74176	0.91830	0.69043	0.65731	0.99454	0.97677
MT012098	**ORF3a**	0.65733	0.60293	0.64026	0.54415	0.84395	0.74176	0.91830	0.69043	0.65731	0.99454	0.97903
MT050493	**ORF3a**	0.65733	0.60293	0.64026	0.54415	0.84395	0.74176	0.91830	0.69043	0.65731	0.99454	0.97903
MT358637	**ORF3a**	0.65733	0.60293	0.64026	0.54415	0.84395	0.74176	0.91830	0.69043	0.65731	0.99454	0.98152
NC_045512	**ORF6**	0.64357	0.51174	0.61783	0.64681	0.94723	0.58368	0.93832	0.58368	0.60386	0.99992	0.97703
MT012098	**ORF6**	0.64357	0.51174	0.61783	0.64681	0.94723	0.58368	0.93832	0.58368	0.60386	0.99992	0.97903
MT050493	**ORF6**	0.64357	0.51174	0.61783	0.64681	0.94723	0.58368	0.93832	0.58368	0.60386	0.99992	0.97903
MT358637	**ORF6**	0.64357	0.51174	0.61783	0.64681	0.94723	0.58368	0.93832	0.58368	0.60386	0.99992	0.98152
NC_045512	**ORF7a**	0.59207	0.58179	0.57340	0.52244	0.87520	0.75251	0.90698	0.65002	0.60385	0.99577	0.97710
MT012098	**ORF7a**	0.59207	0.58179	0.57340	0.52244	0.87520	0.75251	0.90698	0.65002	0.60385	0.99577	0.97903
MT050493	**ORF7a**	0.59207	0.58179	0.57340	0.52244	0.87520	0.75251	0.90698	0.65002	0.60385	0.99577	0.97903
MT358637	**ORF7a**	0.59207	0.58179	0.57340	0.52244	0.87520	0.75251	0.90698	0.65002	0.60385	0.99577	0.98152
NC_045512	**ORF8**	0.56982	0.58887	0.63231	0.59225	0.84988	0.66871	0.94765	0.68673	0.59024	0.99515	0.97693
MT012098	**ORF8**	0.56982	0.58887	0.63231	0.59225	0.84988	0.66871	0.94765	0.68673	0.59024	0.99515	0.97903
MT050493	**ORF8**	0.56982	0.58962	0.62382	0.59225	0.84988	0.67479	0.94546	0.68673	0.59024	0.99515	0.97903
MT358637	**ORF8**	0.56982	0.58887	0.63231	0.59225	0.84988	0.66871	0.94765	0.68673	0.59024	0.99515	0.98152
NC_045512	**ORF7b**	0.71222	0.54143	0.72032	0.52686	0.78637	0.68404	0.99403	0.55410	0.71541	0.94566	0.91796
MT358637	**ORF7b**	0.71222	0.54143	0.72032	0.52686	0.78637	0.68404	0.99403	0.55410	0.71541	0.94566	0.98152

2.1.1. Findings and Discussions on the gene E

Here we present codon and double nucleotides usages of the gene E across the all four genomes.

Codon usages:

Table 4: Codon usages and their corresponding frequencies

Sr No	Codons used in E	No of Codons	Freq.
1	CTT, GTT	2	7
2	TCT,TAC,AAT	3	4
3	TTC, GTA	2	3
4	TTT, TTG, CTA, CTG, GTG, CCT, ACA, GCG, AAA & TGC	10	2
5	ATG, TTA, ATT, ATC, ATA, GTC, TCA, TCG, ACT, ACG,	10	1
6	GCT, GCC, TAA, AAC, GAT, GAA, GAG, TGT, CGT, CGA, AGT, AGC,AGA and GGT.	14	1
7	CTC, TCC, CCC, CCA, CCG, ACC, GCA, TAT, TAG, CAT,	10	0
8	CAC, CAA, CAG, AAG, GAC, TGA, TGG, CGC, CGG, AGG, GGC, GGA and GGG.	13	0

It is observed that, from the Table-4 above that out of the six possibilities of the codons which code for L amino acid, only CTT has been chosen seven times in the primary protein sequence. The primary protein sequences contain the three amino acids say S, Y and N with frequency 4, which are coded by TCT, TAC and AAT respectively. In contrast, it is found that the amino acid V is encoded in the primary protein sequence by three different codons such as GTA, GTC and GTT are used with different frequency viz. 3, 1, 7 respectively. The amino acid Tryptophan (W) which is encoded by TGG only, does not appear in the amino acid sequence induced by the gene E of the four genomes.

Double nucleotide usages: The frequency of usage of the double nucleotides CA, CC, GA, GC, CG, AA, AG, TC, AT, AC, TG, TA, CT, GT, TT are 1, 3, 4, 4, 5, 6, 7, 7, 8, 8, 8, 9, 14, 14 and 16 respectively. The double nucleotide GG is not used at all in the gene E over the said genomes. It seems there is no bias of using double nucleotides unlike in the case of codon usages.

HEs and SEs of spatial representations

- From the Table 3 it follows that all the spatial representations of the nucleotides A, C, T and G of the gene E are positively trending. Based

on the HEs obtained the positive trend of the bases A, C, T and G can be ordered as A, G, T and C. That is the spatial presence of the nucleotide A is positively trending most as compared to others. Therefore it is observed that the spatial arrangements of purine bases are more positively trendier than that of the pyrimidine bases which is apparently clear from the HE of the purine pyrimidine representations too.

- While keeping the positive trending behaviour of the spatial representations of nucleotide bases, the uncertainty (amount of information) of the spatial representations are turned out to be high specially for the nucleotide T.

- From the Table 3, it follows that the Shannon entropy of the nucleotide density over the gene E in the four genomes NC_045512, MT012098, MT050493 and MT358637 are 0.97673, 0.97903, 0.97902 and 0.68920 respectively. The nucleotide conservation entropy is turned out to be significant close to 1 for the gene E of the three genomes except MT358637. It depicts the uncertainty of spatial arrangement of the nucleotide bases over the gene E.

- The conservation entropy of the gene E in the genomes MT012098, MT050493 are extremely close whereas that of the gene E in the genomes NC_045512 is little different. This feature essentially discovers the conservation of A, T, C and G in the gene E of the genomes NC_045512, MT012098 are different though the sequential similarity of the genomes is 99.98%.

2.1.2. Findings and Discussions on the gene M

Codon Usages: The codons CTT and GCT are used in the gene M with highest frequency which is 12. The codon ATT used 11 times in the gene M over the four genomes. The gene M uses the codons AAC and TGG seven times.

The codons TTC, CTC, ATC, GTA, GAA and GGA used in the gene M six times. The gene M over the four genomes uses the codons TTT, CTA, ACT, TAC, GAC, CGT and GGT five times. The codon ATG, TTA, TTG, CTG, GCA, TAT, CAT, AAT, AAA, TGT and AGT are used four times in the gene M over the genomes. The gene M uses the codons ATA, GTT, GTG, TCC, TCA, CCA, ACC, ACA, AAG, AGA, AGG and GGC three times. The frequency of usage of the codons TCT, ACG, GCC, CAA, CAG, CGC and AGC in the gene M is two. These codons TCG, CCT, CCG, GCG, TAA, CAC, GAT, GAG and CGA present in the gene M over the genomes, once. The gene M of the genomes NC_045512, MT012098, MT050493 and MT358637 does not use the codons GTC, CCC, TAG, TGC, TGA, CGG and GGG.

It is found that the start codon ATG is used four times where the only one stop codon TAA is used once in the gene M over the four genomes. The amino acid (L) is encoded by six possible codons out of which only one CTT is used in the gene M with highest frequency. This is clearly showing the usage of codon bias. On the other side, all the three codons ATT, ATC and ATA which encode the amino acid (I) are used in the gene M with different frequency of usages as mentioned above. This shows there is no choice bias in selecting the codons in encoding (I) in the gene M. It is noted that the only codon TGG which encodes the amino acid Tryptophan (W) is used seven times in the gene M over the four genomes.

Double Nucleotide Usages: The frequencies of usage of double nucleotides CG, CC, AG, GA, GG, GT, GC, AC, CA, AT, CT, AA, TA, TC, TG and TT are 10, 12, 13, 16, 17, 18, 18, 21, 21, 22, 23, 25, 27, 29, 30 and 32 respectively. It is noted that all possible double nucleotides are used with at least frequency greater than or equals to ten. The double nucleotide TT is used maximum with frequency 32. Note that the TT also has the highest frequency (16) if usage in

the gene E over the four genomes. Clearly, there is not at all any bias exist in choosing the double nucleotides in constructing the gene M. It is worth noting that the unused double nucleotide in the gene E, is used seventeen times in the gene M of the four genomes.

HEs and SEs of spatial representations

- The positive trend of each nucleotide base over the gene M is exactly same in the four different genomes. The most positive trendiest nucleotide over the gene E is T and then nucleotide A comes in place and then C and G come in the decreasing order. It is worth noting that the order of trendiness is changed in the case of the gene M from the gene E. The HE of the purine-pyrimidine representation in the gene M over the genomes is 0.63095549 which signifies that the representations is positively trending.

- The SE of the binary representation of purine-pyrimidine bases over the gene E in the four genomes is 0.9961267687 which implies the amount of uncertainty is at almost maximum, i.e. the spatial arrangement of purine and pyrimidine bases are equally probable in the sequence of the gene M over all the four genomes. As in the case of the gene E, the amount of uncertainty of spatial arrangement of the nucleotide bases is highest for T.

- The nucleotide conservation entropy of the gene M over the genomes are getting varied from one to another. From the Table 3 it follows that the amount of uncertainty of conservations of nucleotides is found to be very high and close to 1. From the different SEs of the M genes respective to the four genomes, it can be concluded that the spatial conservation of nucleotide bases over the gene M is different. The conservation entropies of nucleotides over the genomes MT012098, MT050493 are found to be almost same whereas that of the genomes NC_045512 and MT358637 are

not so close. Like in the case of the gene E, the microscopic difference among the M genes is found in the conservation of the nucleotide bases.

2.1.3. Findings and Discussions on the gene N

Codon Usages: The frequency of codon usages in the gene N are strictly same in three genomes NC_045512, MT012098 and MT050493 where as the that of the gene N in the genome MT358637 is slightly varied. The codons TAG, TGT, TGC and TGA are unused in the gene N across the four genomes. The single use of the codons ATA, GTA, TAA, CAC and AGG are seen in the gene N over the four genomes. The gene N contains the codons TTA, CTC, GTT, GTG, TCG, CCG, ACG and TAT twice over the four genomes. The codon CGG is used twice in the gene N over the three genomes NC_045512, MT012098 and MT050493 whereas the CGG is used only once in the gene N over MT358637. Three times usage of the codons TTT, CTA, CTG, GTC, TCC, GCG and CAT are observed in the gene N over the four genomes. The codons ATC, GAG and GGG present over the gene N four times. The gene N contains the codons TGG and CGC five times over all the genomes. The frequency of usage of codon CGA over NC_045512, MT012098 and MT050493 is five whereas the codon is used six times in the gene N over the genome MT358637. The codons ACC, AAC, CGT and AGC are present over the gene N across the four genomes, six times. The three codon ATG (start codon), CCC, GCC used seven times in the gene N over the four genomes. The gene N, across all the four genomes, contains the codons CTT, TCT, CCT, ACA, GCA, CAG and GAA, eight times. The frequency of usage of TTG, ATT, TCA, TAC and AGT over the gene N across the three genomes NC_045512, MT012098 and MT050493 is 9. Note that the codon ATT is used in gene N of the genome MT358637 ten times. The codons TTC, AAG, GAC, AGA and GGT are used ten times in the gene N of the four genomes. The only codon CCA which is present eleven times over the gene N

across the four genomes. The codons GGA and GAT present over the gene N across the four genomes thirteen and fourteen times respectively. The gene N in the three genomes NC_045512, MT012098 and MT050493 contains the codons ACT, AAT and GGC sixteen times. It is noted that the codon ACT is used fifteen times in the gene N over the genome MT358637. The frequency of usage of the codons GCT, AAA and CAA in the gene N are 19, 21 and 27 respectively.

It is noticed that the codon CAA and CAG which encode the amino acid Glutamine (Q) are used respectively 27 and 8 times in the gene N. So there is no choice bias in selecting the codons for encoding the amino acid (Q). The codons CGT, CGC, CGA, CGG, AGA and AGG (encode the amino acid Arginine(R)) are all used in the gene N over the four genomes. Clearly there is no bias is observed in selecting the codons which encode the amino acid R. There is no codon bias is observed for the codons (six) which encode the amino acid Leucine(L). Also there is no codon bias, in the gene N over the four genomes, is observed for the codons (six) which encode the amino acid Serine(S). Note that the codon TGG is used five times in the gene N over the four genomes. The only stop codon which is used in the gene N once is TAA.

Double Nucleotide Usages: The frequency usages of the double nucleotides TA, CG, GT, AT, GG, TC, CC, CT, TT, AG, GA, TG, GC, AC, CA and AA in the gene N over the four genomes are 17, 17, 17, 34, 34, 35, 35, 37, 40, 43, 43, 45, 47, 51, 58 and 77 respectively. The frequency usages of the only double nucleotides AT and GC in N of the genome MT358637 are 35 and 46 respectively. It is noted that the highest frequency of usage is attained by the double nucleotide AA unlike in the previous cases. Clearly, there is bias of choices of use of the all sixteen double nucleotides.

HEs and SEs of spatial representations

- From the Table 3, it is seen that the HE of the spatial representations

of the nucleotides A, C, T and G in the gene N over the three genomes NC_045512, MT012098 and MT050493 are 0.61742, 0.50535, 0.53218 and 0.55515 respectively whereas that of the gene N over the genome MT358637 is turned out to be slightly different and they are 0.61484, 0.51273, 0.53644 and 0.55913 respectively for the A, C, T and G spatial representations. The most positively trending spatial representations is of the nucleotide A. The spatial representations of C is nearly random as the HE is turned out to be very close to 0.5. The binary spatial representation of the purine and pyrimidine bases over the gene N over the four genomes is surprisingly invariant and that is 0.57043. Clearly the spatial representation of purine and pyrimidine bases is positively trending.

- From the Table 3, it is observed that the SE of the spatial representations of the nucleotides A, C, T and G in the gene N over the three genomes NC_045512, MT012098 and MT050493 are 0.90160, 0.81128, 0.74209 and 0.90262 respectively. The SE of the binary representations of the nucleotides A, C, T and G in the gene N of the genome MT358637 are 0.90247, 0.81001, 0.74360 and 0.90262 respectively. The uncertainty of presence and absence of the purine and pyrimidine bases in the gene N over the four genomes is almost close to 1. This says the amount of purine and pyrimidine bases are equally likely to appear in the gene sequence although the trend of this purine-pyrimidine representation is positively trending as mentioned earlier.

- It is observed from the data presented in the Table 3 that the highest amount of uncertainty is present in the conservation of the nucleotide bases over the genome MT358637. Clearly the the uncertainty of nucleotides conservation in the gene N is getting varied from one to another in a very minute scale. Though the SEs of the gene N over the genomes MT012098,

MT050493 based in Kerala are extremely near to each other.

2.1.4. Findings and Discussions on the gene S

Codon usages: Here we list all the codon with their respective frequencies in the gene S over the four genomes NC_045512 (G_1), MT012098(G_2), MT050493(G_3) and MT358637 (G_4) in the Table 5.

Table 5: Frequency of codon usages in the gene S over the four genomes

Codon	G_1	G_2	G_3	G_4	Codon	G_1	G_2	G_3	G_4
CCG	0	0	0	0	**CAG**	16	16	16	16
TAG	0	0	0	0	**AGT**	17	17	17	17
TGA	0	0	0	0	**GGA**	17	17	17	17
CGA	0	0	0	1	**TTC**	18	18	18	18
TAA	1	1	1	1	**ATA**	18	19	18	18
CGC	1	1	1	1	**GAC**	19	19	19	19
TCG	2	2	2	2	**TTG**	20	20	20	20
GCG	2	2	2	2	**AGA**	20	19	20	20
CGG	2	2	2	2	**GTC**	21	21	21	21
CTG	3	3	3	3	**AAG**	23	23	23	23
ACG	3	3	3	3	**CCA**	25	25	25	25
GGG	3	3	3	3	**TCA**	26	26	26	26
CCC	4	4	4	4	**GCA**	27	27	27	27
CAC	4	4	4	4	**TTA**	28	28	28	28
AGC	5	5	5	5	**TGT**	28	28	28	28
GCC	8	8	8	8	**CCT**	29	29	29	29
CTA	9	9	9	9	**AAC**	34	34	34	34
CGT	9	9	9	9	**GAA**	34	34	34	34
ACC	10	10	10	10	**CTT**	36	36	36	36
AGG	10	10	10	10	**TCT**	37	37	37	37
CTC	12	12	12	12	**AAA**	38	38	38	38
TCC	12	12	12	12	**ACA**	40	40	40	40
TGC	12	12	12	12	**TAT**	40	39	40	40
TGG	12	12	12	12	**GCT**	42	42	41	42
GTG	13	13	13	13	**GAT**	43	43	43	42
CAT	13	13	13	13	**ATT**	44	44	44	44
ATG	14	14	14	14	**ACT**	44	44	44	44
ATC	14	14	14	14	**CAA**	46	46	46	45
TAC	14	14	14	14	**GGT**	47	47	47	48
GAG	14	14	14	14	**GTT**	48	48	49	48
GTA	15	15	15	15	**AAT**	54	54	54	54
GGC	15	15	15	15	**TTT**	59	59	59	59

It is found that the the codons are CCG, TAG, TGA and CGA unused in the gene S over the four genomes. The codon CGA is present in the gene S over the genome MT358637. The codon AGA is present in the gene S over the genome MT012098 nineteen times whereas it is present over the same gene S in the other three genomes twenty times. Also it is observed that the frequency of the codon TAT is present in the gene S over the genome MT012098 thirty nine times while it is present 40 times in the same gene of the other three genomes. The gene S of the three genomes NC_045512 MT012098 and MT358637 contains GCT forty two times whereas the gene S of the genome MT050493 contains forty one

times. Similar observations are depicted in the Table 2 for the codons GAT, CAA, GGT and GTT. The codon TTT is present in the gene S over all the four genomes with highest frequency which is 59.

Note that the codons TTT and TTC which encode the amino acid Phenylalanine (F) are present in the gene S with frequencies 59 and 18 respectively. It shows there is no bias in selecting the codons. it is also found that the only stop codon TAA is used once in the gene S over the four genomes. The amino acid Leusine(L) encoded by the six amino acids TTA, TTG, CTT, CTC, CTA, CTG which are present over the gene S in the four genomes with different frequency of usages. Also the codons CGT, CGC, CGG, AGA and AGG which encode the amino acid Arginine (R), are all present in the gene S over the four genomes. The amino acid (R) coding codon CGA present only in the gene S of the genome MT358637. This shows no codon bias is observed in the gene S over the four genomes. The codon TGG which encodes the amino acid (W), is present in the gene S over the four genomes with frequency 12 as depicted in the Table 2. There is slight bias in the gene S is observed for the codon CCG which encodes the amino acid (R). The codon CCG is never used in the gene S but the other codons which encode the amino acid (P), are all present over the S gene of the four genomes.

Double nucleotide usages: The double nucleotide frequency usages over the gene S of the four genomes are presented in the following Table 6.

Table 6: Double nucleotides frequencies in the gene S over the four genomes

DN	G_1	G_2	G_3	G_4	**DN**	G_1	G_2	G_3	G_4
CG	15	13	15	16	**CT**	116	144	115	116
GC	61	78	61	61	**AC**	118	131	118	118
CC	71	63	71	71	**AT**	138	164	138	137
GG	75	68	75	75	**TA**	139	124	139	139
GA	90	96	90	90	**CA**	155	141	155	154
AG	107	97	107	107	**TG**	166	166	166	166
GT	114	116	114	115	**AA**	189	185	189	189
TC	116	90	116	116	**TT**	241	233	242	241

It is found that all the sixteen double nucleotide are used in the S gene over

the four genomes. The TT is used with highest frequency and CG is present with lowest frequency over the gene S of the four genomes. It is noticed that the only double nucleotide TG is present in the gene S over the four genomes, with equal frequency (166).

HEs and SEs of spatial representations

- The HEs as well as the SEs of the spatial arrangements of the nucleotide bases A, C, T and G in the gene S over the four genomes, are given in the Table 7.

Table 7: HEs and SEs of the spatial representations of nucleotides over the gene S of the four genomes.

Genome	**HE_A**	**HE_C**	**HC_T**	**HC_G**	**B_HE**
NC_045512	0.5967	0.5919	0.6175	0.5511	0.5834
MT012098	0.5969	0.5904	0.6175	0.5520	0.5834
MT050493	0.5967	0.5907	0.6162	0.5511	0.5834
MT358637	0.5993	0.5919	0.6175	0.5472	0.5834
Genome	**SE_A**	**SE_C**	**SE_T**	**SE_G**	**B_SE**
NC_045512	0.8743	0.6997	0.9175	0.7278	0.9986
MT012098	0.8742	0.7000	0.9175	0.7199	0.9986
MT050493	0.8743	0.6992	0.9178	0.7275	0.9986
MT358637	0.8736	0.6997	0.9175	0.7272	0.9986

- It is seen that all the spatial representations of the nucleotides as well as the purine-pyrimidine bases are positively trending as the HEs are coming out to be greater than 0.5. The HEs of each binary representations of A, T, C and G and in the purine-pyrimidine level of the gene S over the two genomes NC_045512 and MT050493 are almost same. The most positively trendiest spatial arrangement, of the nucleotide in the gene S over the four genomes, is base T as observed from the Table 5. It is noted that the HE of the spatial binary representation of the purine and pyrimidine bases over the gene S of the genomes are identical although the individual spatial representations of the nucleotides are not identical.

- The amount uncertainty of occurrence of purine and pyrimidine bases over the gene S of the four genomes are is at maximum and that implies the

equal probability of occurrence of the purine and pyrimidine bases in the gene S. Also the SEs are turned out to be same for the spatial binary representations of the purine-pyrimidine bases.

- The SE of the nucleotide conservation entropy of the gene S over the four genomes NC_045512, MT012098, MT050493 and MT358637 are 0.9764, 0.9790, 0.9790 and 0.9815 respectively. The amount of uncertainty of conservation of nucleotides is high in the gene S over the four genomes. It is observed that the conservation of nucleotides are very much close for the gene S in the genomes MT012098 and MT050493.

2.1.5. Findings and Discussions on the gene ORF1

Codon usages: The frequency of codon usages in the gene ORF1 over the four genomes are given in the Table 6. it is found that all the sixty four codons are used over the gene ORF1 in the four genomes. It is noticed that the frequency of usages in the gene ORF1 of the genome NC_045512 is significantly different from that of the other three genomes. In most of the codons the frequency of usages in the gene ORF1 over the three genomes are identical as observed from the Table 8.

Table 8: frequency of codon usages in the gene ORF1 over the four genomes.

Codon	G_1	G_2	G_3	G_4	Codon	G_1	G_2	G_3	G_4
CGG	8	7	7	7	**GAC**	108	138	138	138
CGA	9	10	10	10	**AGT**	110	119	120	119
CCG	12	10	10	10	**GCA**	111	131	131	131
CCC	16	17	17	17	**CAG**	116	76	76	76
CGC	19	25	25	25	**CTG**	122	28	28	28
GCG	22	16	16	16	**AAC**	122	116	116	116
TCG	23	5	5	5	**TCA**	129	128	128	128
GGG	23	10	10	10	**GTA**	140	137	137	137
ACG	24	24	24	24	**TCT**	141	155	154	154
AGC	27	23	22	23	**TAC**	141	127	127	127
TCC	35	27	27	27	**GAT**	144	251	251	251
TGA	42	0	0	0	**GTG**	145	88	88	88
CGT	42	59	59	59	**CTA**	146	77	78	77
GGC	48	63	63	63	**TAT**	148	208	208	208
GCC	51	70	71	71	**CAA**	149	163	163	163
AGG	52	31	31	31	**ATA**	150	117	118	117
TGC	53	42	42	42	**TGT**	155	184	184	184
TAA	58	1	1	1	**AAG**	171	153	153	154
GGA	59	76	76	76	**CTT**	180	185	185	186
CTC	66	61	61	61	**GGT**	185	263	263	263
TGG	68	78	78	78	**GCT**	186	269	269	269
GTC	70	81	81	81	**GAA**	197	248	248	248
ACC	70	47	47	47	**ATT**	199	168	168	169
TAG	72	0	0	0	**ACT**	201	233	232	232
ATC	75	57	57	57	**TTG**	210	116	116	116
AGA	76	112	112	112	**TTT**	217	254	254	255
TTC	78	95	95	94	**AAT**	218	268	268	268
CAC	83	43	43	43	**ACA**	219	224	223	224
CCA	87	111	110	111	**TTA**	220	201	201	201
GAG	93	92	92	92	**AAA**	243	281	281	280
CCT	99	135	136	135	**GTT**	245	293	293	292
CAT	102	102	102	102	**ATG**	266	168	168	168

Clearly there is no specific codon bias in the gene ORF1 over the four genomes since all the codons . It is noted that only TAA is used in the ORF1 for the four genomes. The frequency of usages of TAA in the genome NC_045512 is 58 whereas in the gene oRF1 of the other three genomes the frequency of usage of TAA is one. The other two stop codons TAG and TGA are not at all used by the ORF1 in the three genomes whereas the gene ORF1 of NC_045512 contains TAG and TGA with frequencies respectively 71 and 42. Usage of the all stop codons in the gene ORF1 makes the genome different from other three genomes.

Double nucleotide usages: The frequencies of double nucleotides (DN) over the gene ORF1 of the four genomes are presented in the Table 9. It is seen that all the sixteen double nucleotides are present over the gene ORF1 across the four genomes.

Table 9: frequencies of double nucleotides (DN) over the gene ORF1 of the four genomes.

DN	G_1	G_2	G_3	G_4	DN	G_1	G_2	G_3	G_4
CG	149	132	132	132	**CA**	722	705	705	706
CC	302	301	299	300	**CT**	734	739	741	741
GG	395	415	415	415	**GT**	756	728	727	727
GC	403	379	380	380	**AT**	869	820	821	821
TC	415	484	485	483	**TA**	914	906	905	905
GA	586	547	547	548	**TG**	937	960	961	960
AG	609	654	654	654	**AA**	1004	1045	1045	1044
AC	717	702	701	701	**TT**	1133	1128	1127	1128

It is observed that the most frequently used double nucleotides is TT in the gene ORF1 across the four genomes.

HEs and SEs of spatial representations:

- From the Table 3, it is quite clear that the all the spatial representations are positively trending. Each of the four nucleotide spatial representation has its own positive autocorrelations (trending behaviour) as the HEs are different significantly from one to another. The order of positive trendiness of each nucleotide spatial of Importantly, the spatial organization of the purine-pyrimidine bases of the gene ORF1 is positively trending identically across the four genomes.

- The SE of each of the spatial representations of nucleotides are almost identical of the gene ORF1 across the genomes. The uncertainty is at maximum of the presence and absence of purine bases over the binary representation of the purine-pyrimidine bases of the gene ORF1 over the four genomes.

- From the SE of the nucleotide conservations as presented in the Table 3, it is observed that the conservation of nucleotides over the gene ORF1 of the genomes MT012098 and MT050493 are close enough. The highest amount of uncertainty of information of conservations of nucleotides appears in the gene ORF1 over the genome NC_045512.

2.1.6. Findings and Discussions on the gene ORF10

Codon usages: The codon TTC, TTG, CTA, TCT, CCG, ACA, ACG, GCT, GCA, TAC, TAG, CAA, GAT, TGC, CGT, AGT, AGA and GGC present once over the gene ORF10 across the four genomes. The gene ORF10 contains the codons ATG, CTC, GTT, GTA, TAT and AAT twice. And the codons TTT, ATA and AAC are used thrice in the gene ORF10 across the genomes. The other codons are absent in the gene. The amino acid sequence corresponding to the gene ORF10 does not contain the amino acid (W) as the codon (TGG) which encodes it, does not present in the gene ORF10. The amino acid Glutamin(Q) does not present in the amino acid sequence corresponding to ORF10 as the codons CAA and CAG are absent from the gene sequence. As the codons CAT and CAC are absent from the gene ORF10 then the primary protein sequence corresponding to ORF10 would not contain the amino acid Histidine. For the same reason the amino acid Lysine is absent from the amino acid sequence encoded by the gene ORF10. Out of four possibilities of codon choices among CCC, CCG, CCT and CCA, only the CCG has been chosen to encode the amino acid Proline(P). So here a bias in choosing a codon is seen in the gene ORF10 across the four genomes.

Double nucleotide usages: The double nucleotides GC, GG, AA, AC, GT, AG, TG, GA, TC, CG, TT, CA, CT, AT and TA are present once in the gene ORF10 across the four genomes. The double nucleotide CC is thoroughly absent from the gene ORF10 across the four genomes. It is noted that the TA is having highest frequency in the gene ORF10 unlike in the previous cases. It is worth noting that in the aforementioned genes double nucleotides such as AA and TT were having highest frequencies in the corresponding gene.

HEs and SEs of spatial representations

- From the HEs of genes ORF10 across four genomes, it is observed that

the spatial representations are turned out to be positively auto-correlated. The spatial binary representation of the nucleotide A is the most auto-correlated representation whereas the least auto-correlated spatial representation is of the nucleotide G.

- As in the case of others genes, the uncertainty is at maximum of the presence and absence of pyrimidine bases in ORF10 across the four genomes.
- The SE of conservation of nucleotides over the gene ORF10 across the four genomes NC_045512, MT012098, MT050493 and MT358637 are turned out to be 0.97648, 0.97903, 0.97903 and 0.98152. This illustrates that the nucleotide conservations in the gene ORF10 over the genomes MT012098, MT050493 are identical. The uncertainty of conservation of nucleotides over the gene ORF10 is high as the SEs are closed to 1.

2.1.7. Findings and Discussions on the gene ORF3a

Codon usages: The codons CTA, GTG, TAA, GAG, CGT, CGC and AGG are present only once over the gene ORF3a across the genomes. The codons CTG, CCG, ACC and AGC present twice over the gene ORF3a. The frequency of codon usages of codons TTA, GTC, TCT, CCA, ACG, GCC, GCA, TGT, AGA and GGC over the gene ORF3a is 3. The codons ATG, TCC, CAT, CAC, CAG, AAT, AAC, AAG, TGC and GGA present four times over the genes ORF3a across the four genomes. The frequency of codons CTC, ATC, CAA and AGT in the gene ORF3a is five. The codon GTT is present 14 times over the gene ORF3a across the four genomes. Note that that the codons TCG, CCC, GCG, TAG, TGA, CGA, CGG and GGG do not appear in the gene ORF3a. The stop codons TAG and TGA are not at all used in the gene ORF3a over the four genomes. From the presence of the codons in the ORF3a it is found that all the twenty amino acids are present in ORF3a protein.

Double nucleotide usages: All the sixteen double nucleotides CG, GG, CC, AG, AT, TA, TC, GC, GT, GA, AC, CA, TG, CT, AA and TT are present with frequencies respectively 7, 10, 11, 13, 24, 24, 24, 25, 26, 28, 29, 29, 34, 38 39 and 53. The highest frequency is attained the double nucleotide TT in the gene ORF3a over the four genomes. So there is no bias of choice of double nucleotides in ORF3a across the four genomes.

HEs and SEs of spatial representations

- The spatial representations of A, T, C and G as well as of the purine-pyrimidine bases are found to be positively auto-correlated.
- SEs of the binary representations of the nucleotides A, C, T and G over the gene ORF3a across the genomes are invariant as found in the Table 3. The SE of the binary spatial representation of the purine and pyrimidine bases of the gene ORF3a across the four genomes is 0.99454 which is very closed to 1 and that represents maximum uncertainty .
- From the nucleotide conservation SEs of ORF3a as found in the Table 3, it is found that the nucleotide conservations over ORF3a across the two genomes MT012098 and MT050493 are similar.

2.1.8. Findings and Discussions on the gene ORF6

Codon usages: The codons CTT, ATC, TCC, TCA, CCA, GCA, TAT, TAC, TAA, CAT, CAG, AAC, AAG, GAC, GAA, TGG and AGG are present once in the gene ORF6 over the four genomes. the frequency of usage of codons CTC, CTA, TCT and CAA over ORF6 is 2. The gene ORF6 contains the codons ATG, TTT, TTA, GTT, ACT,AAT and AAA three times. The frequencies of the codons GAT, ATA, GAG and ATT in ORF6 are 3, 4, 4, and 5 respectively. The frequency of each codon in ORF6 is remain invariant across the genomes. The rests codons such as TTC, TTG, CTG, GTC and so on are

absent throughout the gene ORF6 across the four genomes.

It is found that the codon GCA has been chosen by the gene ORF6 across the four genomes, among the four codons which encode the amino acid Alanine(A). Out of six codons only AGG has been considered in ORF6 to encode the amino acid Arginine(R). Out of four possible codons only CCA is chosen by ORF6 to encode the amino acid Proline(P). therefore there is a strong codon bias is seen in ORF6 across the four genomes. In the first positive frame of ORF6 the codons TGT and TGC are absent and hence the corresponding amino acid Cysteine(C) cannot be made in the primary protein sequence. For the similar reason, the amino acid Glycine would not be present in the primary protein sequence encoded by the gene ORF6 in the four genomes.

Double nucleotide usages: The frequencies of the double nucleotides CC, GC, GG, GT, CT, AG, AC, TG, CA, TC, AT, GA, AA, TT and TA are 1, 1, 1, 3, 4, 5, 6, 6, 6, 7, 9, 9,10, 12 and 13 in the gene ORF6. The only double nucleotide CG is absent in the gene ORF6 thoroughly over the four genomes. It is noted that the highest frequency is obtained by TA although the frequency of A is highest in ORF6.

HEs and SEs of spatial representations

- Clearly all these five different spatial representations are positively auto-correlated.
- From the Table 3, it is seen that the SEs of the representation of A and T are almost same and hence it is concluded that the spatial representations of A (purine-base) and T (pyrimidine-base) are similar. The SE of the spatial representation of purine and pyrimidine bases in ORF6 over the four genomes are again found same and that is 0.99992 which is very nearer to one.
- The SEs of the conservation of nucleotides over ORF6 across the three

genomes NC_045512, MT012098, MT050493 is indentical adn that is 0.97703 whereas that in the genome MT358637 is 0.98152. So the amount of information of conservation of nucleotides are found to be identical in the three genomes except MT358637. In the genome MT358637, the SE of nucleotide conservation in ORF6 is found to be more uncertain as the SE is seen to be close enough to 1.

2.1.9. Findings and Discussions on the gene ORF7a

Codon usages: The codons ATG, TTG, CTA, ATC, GTC, GTG, GCC, GCG, CAT, CAG, AAT, AAC, AAG, GAT, GAC, TGA, CGT, AGC, GGT and GGA are present over the first postive frame of the gene ORF7a across the four genomes. The gene ORF7a uses the codon TTA, CTC, GTA, CCA, TAT, CAC and GGC twice. And there are codons which are present in ORF7a across four the genomes with frequencies 4, 6 and 7. The highest frequency seven is obtained by the codons TTT and ACA. The codons TCC, TCG, CCC, CCG, ACC, ACG, TAA, TAG, TGG, CGC, CGA, CGG, AGT, AGG and GGG are absent throughout ORF7a across the four genomes.

It is noted that the amino acid Leusine(L) is encoded by six different amino acids which are all represent with different frequencies over the gene ORF7a across the four genomes. Hence the bias of choices is not seen here. Out of the six codons only two codon AGG and CGT have been chosen to encode the amino acid Arginine (R) by the gene ORF7a. So some amount of bias is observed. The amino acid (W) would not be present in ORF7a protein since the necessary codon TGG is absent in ORF7a. It is noted that the one stop codon TGA is used once over the gene ORF7a across the four genomes.

Double nucleotide usages: The double nucleotides CC, CG, GG, GT, GC, AT, TA, AG, CA, TG, GA, TC, AA, CT, AC and TT are present in ORF7a across the four genomes with frequencies 2, 2, 4, 5, 8, 9, 9, 10, 11, 14,

14, 15, 17, 18, 21 and 24 respectively. So all the double nucleotides are used in ORF7a. Hence there is no choice bias is found.

HEs and SEs of spatial representations

- The spatial representation of A is found to be most positively trending as compared to others since the HE is maximum among all. The HE of the purine-pyrimidine arrangements is 0.60385 which depicts the positive auto-correlation over the representation.

- From the SEs of nucleotide bases in ORF7a as mentioned in Table 3, it is observed that the amount of uncertainty of presence of the nucleotide G over its binary representations is lowest as compared to others. The SE of binary representation of the purine and pyrimidine bases over the gene ORF7a across the genomes also same and that is 0.99577. As usual the highest amount of uncertainty is observed in the presence and absence of purine bases over the gene ORF7a across the four genomes.

- The SE of conservation of nucleotides in the gene ORF7a over the four genomes NC_045512, MT012098, MT050493 and MT358637 are 0.97710, 0.97903, 0.97903 and 0.98152. The conservation of nucleotide bases are found to be similar in ORF7a over the two genomes MT012098, MT050493 as the SEs in these two cases are found to be identical. As previously seen, in the genome MT358637, the SE of nucleotide conservation in ORF7a is found to be more uncertain as the SE is seen to be close enough to 1.

2.1.10. Findings and Discussions on the gene ORF8

Codon usages: The codon frequencies in the gene ORF8 across the four genomes are given in the following Table 10.

Table 10: Frequencies of codon usages in ORF8 across the four genomes.

Codon	G_1	G_2	G_3	G_4	Codon	G_1	G_2	G_3	G_4
ATG	1	1	1	1	**TAT**	6	6	6	6
TTT	4	4	4	4	**TAC**	1	1	1	1
TTC	4	4	4	4	**TAA**	1	1	1	1
TTA	6	6	5	6	**TAG**	0	0	0	0
TTG	2	2	2	2	**CAT**	2	2	2	2
CTT	2	2	2	2	**CAC**	2	2	2	2
CTC	0	0	0	0	**CAA**	3	3	3	3
CTA	0	0	0	0	**CAG**	3	3	3	3
CTG	0	0	0	0	**AAT**	2	2	2	2
ATT	5	5	5	5	**AAC**	0	0	0	0
ATC	5	5	5	5	**AAA**	5	5	5	5
ATA	0	0	0	0	**AAG**	0	0	0	0
GTT	6	6	6	6	**GAT**	4	4	4	4
GTC	0	0	0	0	**GAC**	3	3	3	3
GTA	4	4	4	4	**GAA**	4	4	4	4
GTG	2	2	2	2	**GAG**	2	2	2	2
TCT	2	2	2	2	**TGT**	5	5	5	5
TCC	1	1	1	1	**TGC**	2	2	2	2
TCA	3	3	4	3	**TGA**	0	0	0	0
TCG	1	1	1	1	**TGG**	1	1	1	1
CCT	4	4	4	4	**CGT**	2	2	2	2
CCC	1	1	1	1	**CGC**	0	0	0	0
CCA	1	1	1	1	**CGA**	0	0	0	0
CCG	1	1	1	1	**CGG**	0	0	0	0
ACT	2	2	2	2	**AGT**	2	2	2	2
ACC	0	0	0	0	**AGC**	0	0	0	0
ACA	3	3	3	3	**AGA**	2	2	2	2
ACG	0	0	0	0	**AGG**	0	0	0	0
GCT	3	3	3	3	**GGT**	3	3	3	3
GCC	0	0	0	0	**GGC**	0	0	0	0
GCA	2	2	2	2	**GGA**	2	2	2	2
GCG	0	0	0	0	**GGG**	0	0	0	0

The codons CTC, CTA, CTG, ATA, GTC, ACC, ACG, GCC, GCG, TAG, AAC, AAG, TGA, CGC, CGA, CGG, AGC, AGG, GGC and GGG are absent in the gene ORF8 across the four genomes. The preferred stop codon in ORF8 is TAA across the four genomes and it is used once only. It is worth mentioning that the same length gene ORF7a uses only the stop codon TGA. It is found that all the twenty amino acid are present ORF8 protein.

Double nucleotide usages: A list of frequency of double nucleotides over the gene ORF8 across the four genomes is given in the Table 11. It is found that all the sixteen double nucleotides are used in ORF8 across the four genomes. The highest frequency of TT in the gene ORF8 is turned out to be 24.

Table 11: Frequency of double nucleotides over the gene ORF8 across the four genomes

DN	G_1	G_2	G_3	G_4	DN	G_1	G_2	G_3	G_4
GG	3	3	3	3	**AA**	13	13	13	13
CG	5	5	5	5	**GA**	13	13	13	13
GC	5	5	5	5	**TG**	14	14	14	14
CC	6	6	6	6	**GT**	14	14	14	14
AC	7	7	7	7	**AT**	16	16	16	16
CT	8	8	8	8	**TC**	16	16	16	16
AG	10	10	10	10	**TA**	18	18	17	18
CA	11	11	12	11	**TT**	24	24	24	24

It is found that the double nucleotide CA is present twelve times in the gene ORF8 over the genome MT050493 whereas in the other three genomes the gene ORF8 contains CA only eleven times uniformly. Also it is observed that TA is present in ORF8 over the genome MT050493 seventeen times whereas in the rests genomes it is present eighteen times in the gene ORF8.

HEs and SEs of spatial representations

- The highest amount of autocorrelation is observed in the spatial representation of T as found from the Table 3. The HE of the spatial representation of the purine and pyrimidine bases in ORF8 across the four genomes is 0.59024.

- The SEs of the binary representations of the nucleotides A, C, T and G in ORF8 over the three genomes except MT050493 are found to be same and they are 0.84988, 0.66871, 0.94765 and 0.68673 respectively whereas that in the gene ORF8 over the genome MT050493 are respectively 0.84988, 0.67479, 0.94546 and 0.68673. That is the the spatial template of the nucleotides C and T are different from that of the gene ORF8 over the other three genomes NC_045512, MT012098 and MT358637. As usual the binary SE of the spatial representation of the purine and pyrimidine bases over the gene OPR8 across the four genomes are found to be same and that is 0.99515.

- From the Table 3, it is found that the SEs of nucleotide conservations

over the gene ORF8 across the three genomes except MT358637 is same whereas that of the gene ORF8 in the genome MT358637 is 0.98152. As seen before the highest amount uncertainty is observed in conservation of nucleotides over the gene ORF8 of the genome MT358637. It is noted that the conservation of nucleotides of the gene ORF8 over the other three genomes are observed to be similar as the SEs are identical.

2.1.11. Findings and Discussions on the gene ORF7b

The presence of ORF7b gene makes the two genomes NC_045512 and MT358637 different from other two genomes where the ORF7b do not present as mentioned in the dataset section. It is worth noting that the absence and presence of ORF7b can be associated to the S and L type respectively. The features encountered for ORF7b over the two genomes NC_045512 and MT358637 are presented below:

Codon usages: The gene ORF7b contains the codon TTG, CTA, ATC, GTT, ACT, TAT, TAA, CAT, CAC, CAA, AAT, GAT, GAC, TGT, TGC and TGG once each. The codon ATG (start codon), CTG, TCA and GCC are present twice over the gene ORF7b. The frequency of the codon usages of the codons TTT, TTC, TTA and GAA is 3 in ORF7b. The codons CTT and ATT appears with highest frequency 4 in the gene ORF7b. It is noted that the TAA is the preferred stop codons among the three stop codons. The rest 38 codons are not at present in the gene ORF7b in the two genomes.

The amino acid Lysine(K) shall not be present in the amino acid sequence encoded by the first positive frame of the gene ORF7b since the necessary codons AAA and AAG for making the amino acid K is absent in ORF7b. For the similar reason, the amino acid Proline(P) will also be not present in the primary sequence of ORF7b protein. Among the six codons only one TCA has been chosen to code the amino acid Serine(S) in ORF7b. So there is clear codon

bias which is also apparently true because 59% of the codons are absent in the first positive frame of the ORF7b. The amino acid Arginine(R) encoded by six codons will be absent in the primary protein sequence of ORF7b protein as none of the necessary six codons is present in the ORF7b sequence. For similar reason the amino acid Glycine would not be present in the primary protein sequence of ORF7b protein.

Double nucleotide usages: The fourteen double nucleotides AG, CC, AA, CA, GT, TC, GA, GC, AC, TG, TA, CT, AT and TT with respective frequency 1, 1, 2, 2, 3, 4, 4, 4, 5, 5, 7, 7, 8 and 13. The highest frequency is obtained for TT in the gene ORF7b. It is noted the double nucleotides CG and GG are absent from the gene sequence of ORF7b.

HEs and SEs of the spatial representations

- The spatial representations are all positively trending as the HEs are found to be greater than 0.5.
- The SEs of the spatial representations of nucleotides A, C, T and G in ORF7b are found to be same in the two genomes and they are 0.78637, 0.68404, 0.99403 and 0.55410 respectively. The SE of the binary representation of the purine and pyrimidine bases are also same and that is 0.94566 which is significantly less as compared to other genes as observed.
- The nucleotide conservation SE of the gene ORF7b across the two genomes NC_045512 and MT358637 are found to be non-identical and they are 0.91796 and 0.98152 respectively. The uncertainty of nucleotide conservation over the gene ORF7b in the genome MT358637 high which implies the nucleotides in ORF7b of the genome NC_045512 is conserved more than that of the other.

2.2. Phylogenetic Relationships of the Genomes

Based on the features vectors obtained for each gene over the four genomes NC_045512, MT012098, MT050493 and MT358637, pairwise Euclidean distances have been enumerated. Based on the distance matrix with respect to the each gene of the four genomes, a corresponding phylogeny is obtained.

There are six different phylogenetic trees are derived from the features of respective genes of the four genomes. The phylogeny based on the features of the gene E, ORF3a, ORF6 and ORF7a are identical as shown in the Fig.3.

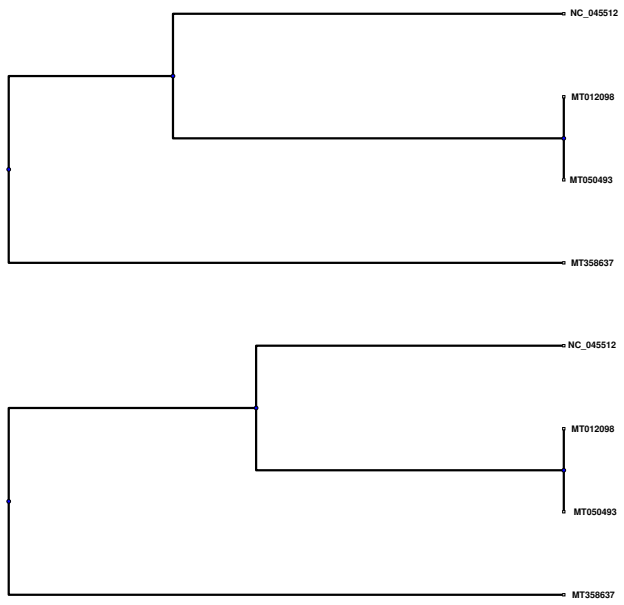

Figure 3: Phylogeny tree among based on the features of the genes E, ORF3a, ORF6 and ORF7a (left) and the genes M, N (right) over the four genomes.

It is observed from the phylogeny in the Fig.2, the genomes MT012098, MT050493 are most close to each other with respect to the features obtained for the genes E, ORF3a, ORF6 and ORF7a. These two genomes are close then with NC_045512 since they belong under the same binary branches. These three

genomes are distantly close to MT358637. It is worth noting that the phylogeny based on sequential similarities among the four genomes does not go simply with the phylogeny based on the spatial features.

The phylogenetic relationship (Fig.3) among the four genomes, it is derived that the genomes NC_045512, MT012098, MT050493 are close enough with respect to spatial and molecular organizations of the genes M and N as compared to the other genome MT358637 which belong to the other branch of the binary tree. Here again, the sequence-similarity based phylogeny (Fig.2) is not linearly reflected while phylogeny is derived by accounting the spatial and molecular organizations of the gene M.

The distinct phylogenetic relationship are developed in the Fig.4. by the features of the genes ORF1, ORF10, ORF8 and S.

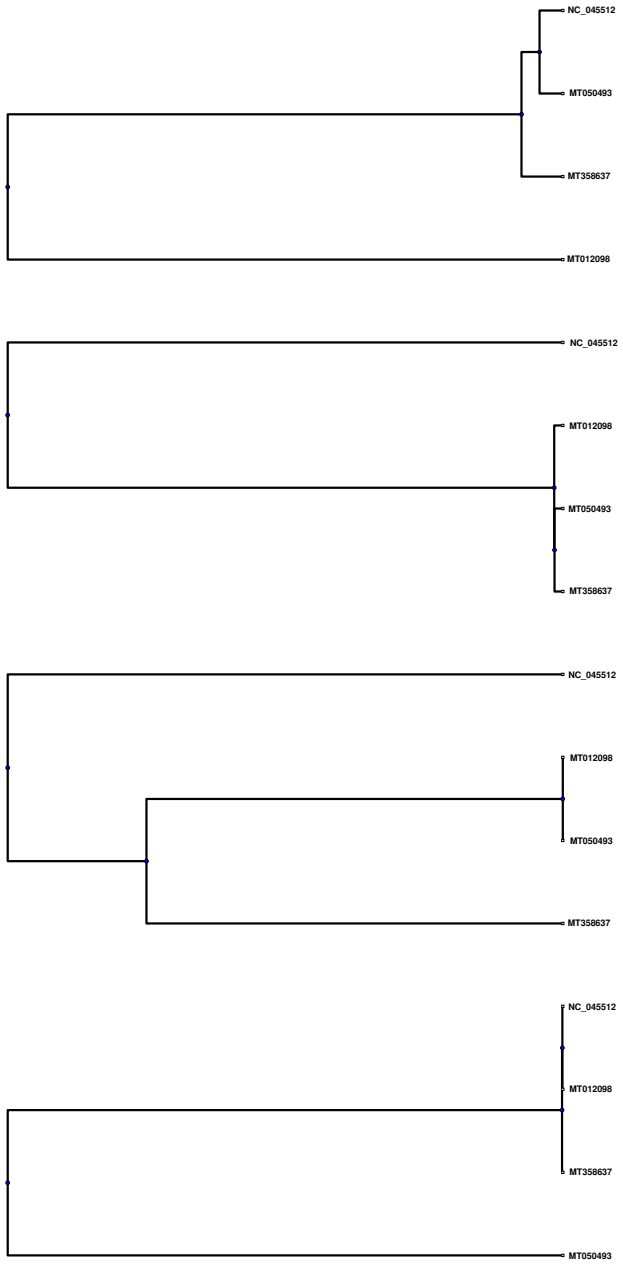

Figure 4: Phylogeny trees based on the features of the gene S (up-left), ORF1(up-right), ORF10 (down-left) and ORF8 (down-right) over the four genomes.

From the phylogeny based on the features of the gene S as shown in Fig.5, it is found that the genomes NC_045512 and MT050493 are most close to each other as belong to same level of the phylogeny. These two genomes are close to the genome MT358637. The genome MT012098 is distantly close to the other genomes since this genomes belong to a branch of primary binary level of the phylogeny. This phylogeny discriminates the genome MT012098 from NC_045512 according to the spatial and molecular organization of the gene S, although they are sequentially very close to each other as mentioned previously.

As we understand the codon usages over the gene ORF1 discriminate the genome NC_045512 from others and this is openly reflected in the phylogeny made in the Fig.7. The three genomes based in India are close enough to each other as they belong to a single branch and the Wuhan based genome NC_045512 belongs to the other branch of the binary tree.

As per the phylogenetic relationship based on the features of ORF10, it is found that the genome MT012098 and MT050493 are most close to each other as they below to a binary branches in the same level. Then the genome MT358637 is closed in the upper level of binary tree of the phylogeny. The genome NC_045512 is distantly related to the cluster of three other genomes. It is worth noting that the SEs of conservation of nucleotides in the gene ORF10 are the determining features of closeness among the genomes.

From the phylogenetic relationship among the four genomes based on the features extracted for the gene ORF8 it is found that the three genomes NC_045512, MT012098 and MT358637 are close enough each other as they all belong to a single branch of the binary phylogenetic tree whereas the genome MT050493 belongs to the other branch. It is observed that the Spatial arrangements of the nucleotide bases C and T in the gene ORF8 make the genome MT050493 different from other three.

Differences in phylogenetic tree arrangement with individual gene suggest that three genome of Indian have come from three different origin or evolution of viral genome is very fast process. Irrespective of evolution, biasness towards the usage of codon remains.

3. Conclusions

- There are several orders of nucleotide frequencies in different gene sequence have been observed. The pyrimidine-rich sequences E, M, S, ORF3a, ORF7a, ORF7b and ORF10 possess the order T-A-C-G. An exception happens in the case of purine-rich sequence ORF1. i.e. the order becomes T-A-C-G. The gene ORF8 being pyrimidine-rich is having the changed order as T-A-G-C. Although the N gene is purine-rich here we find a different ordering A-C-G-T while the purine-rich sequence ORF6 has the order A-T-C-G. Among all the eleven genes embedded in the four genomes the highest pyrimidine-rich (63.64%) sequence is of the gene ORF7b.

- The GC content for all these genes are widely varied over the closed interval [27.95, 47.20]. Note that both the genes, ORF6 and N having lowest and highest GC content respectively, present over all the SARS-CoV2 genomes. One may note that the GC content (47.2%) of N is significantly high which may distinctly characterized it from the other structural proteins E (38.15%), M (42.6%) and S(37.3%).

- The distribution of purine and pyrimidine bases over each gene across four genomes are found to be highly uncertain. That is the purine and pyrimidine bases are equally likely to appear in the sequences. Although it is noted that these purine-pyrimidine spatial organizations is positively trending.

- The most preferred stop codon in all the genes across the four genomes is TAA. But in exception most strikingly, it is noticed that the stop codons TGA and TAG are used 42 and 72 times respectively in ORF1 over the genome NC_045512 of China. It is worth nothing here that there are several other codons such as TCG, AGC, TCC, CCA, CCT, GAC, CTA, GTG and so on are present with abruptly different frequencies than that in ORF1 of the other three genomes based in India.

- In most of the genes all the sixteen double nucleotides are present. The genes E, ORF10, ORF6 do not contain the double nucleotides GG, CC and CG respectively. In exception, the gene ORF7b does not contain two of the double nucleotides viz. CG and GG. It is noted that the gene ORF7b does not belong to the genomes of S-type as it seems from the present data. The gene ORF7b (length:132 bases) is uniquely characterized by the absence of two double nucleotides CG and GG as observed on the other hand the ORF10 which is of smallest length (117 bases) is characterized by the absence of only one double nucleotide CC.

- From the Table 3, it is to be pointed out that the higher consistently nucleotide conservations (SE:0.98152) over the all genes is observed in the genome MT358637 (India-Gujrat).

The codon bias in the ORF1 gene in the genomes suggest the possibility of evolving of viral genome in these area. These change in codon usage might have huge clinical relevance which needs to be further validated in therapeutic approaches.

Author Contributions

SH and PPC conceived the problem. AM and RKR coded and produced the results. SH, PPC, SSJ, PP analysed the results. Finally SH wrote the

manuscript-draft and thoroughly checked by all the authors and approved.

Conflict of Interests

The authors do not have any conflicts of interest to declare.

References

[1] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, B. Berkhout, Identification of a new human coronavirus, Nature medicine 10 (4) (2004) 368–373.

[2] P. Yang, X. Wang, Covid-19: a new challenge for human beings, Cellular & Molecular Immunology (2020) 1–3.

[3] L. Mousavizadeh, S. Ghasemi, Genotype and phenotype of covid-19: Their roles in pathogenesis, Journal of Microbiology, Immunology and Infection (2020).

[4] K. V. Holmes, Sars-associated coronavirus, New England Journal of Medicine 348 (20) (2003) 1948–1951.

[5] J. F. Drexler, V. M. Corman, C. Drosten, Ecology, evolution and classification of bat coronaviruses in the aftermath of sars, Antiviral research 101 (2014) 45–56.

[6] Z.-R. Lun, L.-H. Qu, Animal-to-human sars-associated coronavirus transmission? (2004).

[7] Y.-Z. Zhang, E. C. Holmes, A genomic perspective on the origin and emergence of sars-cov-2, Cell (2020).

[8] W. Tai, L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang, Y. Zhou, L. Du, Characterization of the receptor-binding domain (rbd) of 2019 novel coronavirus: implication for development of rbd protein as a viral attachment inhibitor and vaccine, Cellular & molecular immunology (2020) 1–8.

[9] O. M. Khailany RA, Safdar M, Genomic characterization of a novel sars-cov-2., Gene Reports (2020).

[10] S. Angeletti, D. Benvenuto, M. Bianchi, M. Giovanetti, S. Pascarella, M. Ciccozzi, Covid-2019: the role of the nsp2 and nsp3 in its pathogenesis, Journal of medical virology (2020).

[11] W. Shang, Y. Yang, Y. Rao, X. Rao, The outbreak of sars-cov-2 pneumonia calls for viral vaccines, npj Vaccines 5 (1) (2020) 1–3.

[12] C. Verdiá-Báguena, J. L. Nieto-Torres, A. Alcaraz, M. L. DeDiego, L. Enjuanes, V. M. Aguilella, Analysis of sars-cov e protein ion channel activity by tuning the protein and lipid charge, Biochimica et Biophysica Acta (BBA)-Biomembranes 1828 (9) (2013) 2026–2031.

[13] M. Surjit, S. K. Lal, The sars-cov nucleocapsid protein: a protein with multifarious activities, Infection, genetics and evolution 8 (4) (2008) 397–405.

[14] S. Li, L. Lin, H. Wang, J. Yin, Y. Ren, Z. Zhao, J. Wen, C. Zhou, X. Zhang, X. Li, et al., The epitope study on the sars-cov nucleocapsid protein, Genomics, proteomics & bioinformatics 1 (3) (2003) 198–206.

[15] X. Fang, J. Gao, H. Zheng, B. Li, L. Kong, Y. Zhang, W. Wang, Y. Zeng, L. Ye, The membrane protein of sars-cov suppresses nf-κb activation, Journal of medical virology 79 (10) (2007) 1431–1439.

[16] Y. Hu, J. Wen, L. Tang, H. Zhang, X. Zhang, Y. Li, J. Wang, Y. Han, G. Li, J. Shi, et al., The m protein of sars-cov: basic structural and immunological properties, Genomics, proteomics & bioinformatics 1 (2) (2003) 118–130.

[17] H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang, C.-K. Tang, Coronavirus disease 2019 (covid-19): current status and future perspective, International Journal of Antimicrobial Agents (2020) 105951.

[18] L. Du, Y. He, Y. Zhou, S. Liu, B.-J. Zheng, S. Jiang, The spike protein of sars-cov—a target for vaccine and therapeutic development, Nature Reviews Microbiology 7 (3) (2009) 226–236.

[19] L. Du, Y. Yang, Y. Zhou, L. Lu, F. Li, S. Jiang, Mers-cov spike protein: a key target for antivirals, Expert opinion on therapeutic targets 21 (2) (2017) 131–143.

[20] Y. Yang, L. Zhang, H. Geng, Y. Deng, B. Huang, Y. Guo, Z. Zhao, W. Tan, The structural and accessory proteins m, orf 4a, orf 4b, and orf 5 of middle east respiratory syndrome coronavirus (mers-cov) are potent interferon antagonists, Protein & cell 4 (12) (2013) 951–961.

[21] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al., A new coronavirus associated with human respiratory disease in china, Nature 579 (7798) (2020) 265–269.

[22] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique, Covid-19 infection: origin, transmission, and characteristics of human coronaviruses, Journal of Advanced Research (2020).

[23] J. A. Jaimes, N. M. André, J. S. Chappie, J. K. Millet, G. R. Whittaker, Phylogenetic analysis and structural modeling of sars-cov-2 spike protein re-

veals an evolutionary distinct and proteolytically-sensitive activation loop, Journal of Molecular Biology (2020).

[24] M. M. Lai, D. Cavanagh, The molecular biology of coronaviruses, in: Advances in virus research, Vol. 48, Elsevier, 1997, pp. 1–100.

[25] X. L. Y. S. X. Y. X. W. Y. D. H. Z. Y. W. Z. Q. J. C. J. L. Xiaolu Tang, Changcheng Wu, On the origin and continuing evolution of sars-cov-2, in: National Science Review, Vol. 48, nwaa036, 1997, pp. 1–100.

[26] P. D. Yadav, V. A. Potdar, M. L. Choudhary, D. A. Nyayanit, M. Agrawal, S. M. Jadhav, T. D. Majumdar, A. Shete-Aich, A. Basu, P. Abraham, et al., Full-genome sequences of the first two sars-cov-2 viruses from india, The Indian journal of medical research (2020).

[27] C. Cattani, Fractals and hidden symmetries in dna, Mathematical problems in engineering 2010 (2010).

[28] S. S. Hassan, P. P. Choudhury, B. Daya Sagar, S. Chakraborty, R. Guha, A. Goswami, Quantitative description of genomic evolution of olfactory receptors, Asian-European Journal of Mathematics 8 (03) (2015) 1550043.

[29] S. S. Hassan, R. K. Rout, V. Sharma, A quantitative genomic view of the coronaviruses: Sars-cov2 (2020).

[30] S. S. Hassan, R. K. Rout, Spatial distribution of amino acids of the sars-cov2 proteins (2020).

[31] J. K. Das, P. P. Choudhury, A. Chaudhuri, S. S. Hassan, P. Basu, Analysis of purines and pyrimidines distribution over mirnas of human, gorilla, chimpanzee, mouse and rat, Scientific reports 8 (1) (2018) 1–19.

[32] P. Jacquet, G. Seroussi, W. Szpankowski, On the entropy of a hidden markov process, Theoretical computer science 395 (2-3) (2008) 203–219.

[33] P. Bernaola-Galván, J. L. Oliver, R. Román-Roldán, Decomposition of dna sequence complexity, Physical Review Letters 83 (16) (1999) 3336.

[34] F. Johansson, H. Toh, Relative von neumann entropy for evaluating amino acid conservation, Journal of bioinformatics and computational biology 8 (05) (2010) 809–823.

[35] A. Carbone, G. Castelli, H. E. Stanley, Time-dependent hurst exponent in financial time series, Physica A: Statistical Mechanics and its Applications 344 (1-2) (2004) 267–271.