

Genetic Behaviour of Zika Virus and Identification of Motif

Pushpa Susant Mahapatro¹, Jatinderkumar R. Saini^{2*}

Ph.D. Scholar, Symbiosis International (Deemed University), Pune, India¹
Professor & Director, Symbiosis Institute of Computer Studies and Research²
Symbiosis International (Deemed University), Pune, India

Abstract—ZIKV is a mosquito-borne disease. It is known to cause neurological disorders and congenital disabilities in newborns. The Genome Sequence of the Zika virus is used for the study. The essential cell functionalities like circadian behavior and expression of genes are studied. Regulatory proteins are alternating functionalities during daytime and night time. Identifying motif is made by understanding the features of motifs, finding the count matrix, and formulating the profile matrix. The consensus string of the Zika virus is to be computed, and the score motif is to be calculated. Different techniques of motif finding like the Brute Force technique and Greedy Search techniques are proposed. In the Brute Force technique, each motif is selected, its score is to be calculated, and then the minimum score can be obtained. The Brute Force technique will take an enormous amount of time, but it is guaranteed to find a solution. The Greedy Search technique is not guaranteed to find motif like the Brute Force technique but can give a close answer in a realistic time. This paper presents the identification of motif in the Zika virus genome using programming techniques.

Keywords—*Circadian behaviour; consensus string; genome study; greedy search technique; motif search; regulatory proteins*

I. INTRODUCTION

Zika virus (ZIKV) was discovered in 1947. It is a mosquito-borne disease. In 2013, it spread in South Pacific. It continued to spread to all parts of America. Early studies revealed that the virus originated and remained in Africa for many decades. Zika virus study is interesting as it can solve many biological problems. Genome sequences are quite complex. It is not possible to explain by a probabilistic model. So low-order Markov models explain the properties quite well. The DNA k-mer frequencies in the genome sequence of the Zika virus provide an insight into the genome complexity. It is possible to study the k-mers with different lengths as segments of genome sequence from different animals are available. A study was carried out for k-spectra of different species from Archaea, Bacteria, and e-coil. It studied the modalities of distributions. Some species have multimodal spectra, whereas all other species have a unimodal k-mer spectrum [1].

As a patient falls ill, based on the medical signs and symptoms, Symptoms bases categorization is possible. A frequent association can be extracted based on association mining [2].

There is a need to extract a sequence of existing strains. Different disciplines are collaborating for combating the outbreak of the disease. The different methodologies used are

RNA extraction and material validation. Also, genome sequencing, consensus variation, and sequence analysis are done to understand the whole genome sequencing. The scientific method does not need to separately authenticate the virus reagents. The data is available through public repositories. It helps to study the pathogenesis, neurotropism, prevention, and possible spread [3]. Microcephaly, primers, and probes in ZIKV is detected. ZIKV is transmitted from a viremic host to normal people. Non-African mosquitoes have more potential to transmit the disease compared to African mosquitoes. Mosquito to mosquito transmission is possible and was evaluated in a study. In Africa, aegypti is less susceptible to ZIKV [4]. The spread of ZIKV is associated with neurological complications. The ZIKV spread and mode of transmission is studied carefully. Serological tests of ZIKV react with antibodies by other viral infection. Viral nucleic acids are present in polymerase chain reaction testing. Also, virus isolation is done for confirmation. It confirms the presence of ZIKV. It combines with global aedes vector distribution. Mother to the foetus and sexual transmission is very common person-to-person transmission modes [5].

The detection of the infection becomes difficult due to its close association with flaviviruses. The potential of antibody-dependent enhancement also increases as the cross reactivity to flaviviruses like the dengue virus and the West Nile virus. Serum samples were collected and tested. A study was conducted on the dengue virus and the West Nile virus to find whether these viruses can enhance or neutralize the ZIKV. The West Nile virus enhanced ZIKV, so it failed to neutralize [6].

People with ZIKV have a mild fever and fewer symptoms to identify the infection. Babies are born with birth defects for infected pregnant women. Deterministic models were designed, considering sexual transmission, mosquito-human transmission, and Wolbachia-infected male mosquito release. Disease-free equilibrium and its stability is studied. The study was performed on impact of parameters on reproduction. The intervals of the liberation of Wolbachia-infected male mosquitoes were studied. A bounded global solution was derived for the extinction of ZIKV. As per the numerical simulations, ZIKV may be destroyed when the amount of white noise reaches a threshold value. The wild mosquitoes may be extinct with the delivery of Wolbachia-infected mosquitoes [7].

ZIKV is a severe public health issue. Still, a little study is done to find the transmission of the Zika virus in sexual groups. A study to control the spread of virus between

*Corresponding Author

individuals by changing contact patterns is done. A heterosexual network-based model is designed based on the Costa Rica case study. A study is carried out to measure the effect of changing the degree of heterogeneity. It is measured by removing the sexual contact of persons with a limited number but a greater degree and at different places. A threshold time for Zika virus infection next to the peak time was devised [8].

II. LITERATURE REVIEW

The ZIKV genome is studied to better understand the evolution and spread of Zika virus infection in more than fifty countries of the world. The spread is caused due to infected mosquito bites and person-to-person transmission. This disease is better understood with molecular insights of ZIKV [9]. It can help to better combat the disease. In monkeys and humans, the neural progenitor cell growth is attenuated by infection. The DNA is damaged by the virus and activates DNA damage responses [10]. The biological cycle of the virus is studied to understand the behaviour of the virus during the daytime, night time, and replication process. With the introduction of ZIKV to the Americans, four mutations of ZIKV were reported. This represents direct evasions from earlier mutations during the spread from Africa to Asia. Studies were performed with and without mutation on the experimental infection of aedes aegypti mosquito and human cells. It was found that fitness is reduced for original mutation for urban human-amplified transmission, whereas the fitness was enhanced for new mutations increasing the risk. The findings include three adaptive mutations of ZIKV [11].

The adult with moderate immunocompetent features may get infected by ZIKV. It triggers and enhances antiviral responses and brain damage. The neuroendocrine functions, inflammation, and immune reaction for different pathogens can

get modulated due to gut microbiota composition. The modification stimulated by ZIKV in the belly microbiome of immune-capable mice was studied. It was found that the infection caused a considerable decline in microbes like Actinobacteria and Firmicutes phyla; compared to healthy mice. A significant boost of Deferribacteres and Spirochaetes was identified. Intestinal harm and extreme white blood cells recruitment were caused due to modulation of microbiota induced by the Zika virus [12]. The birth defects are associated with utero exposure to ZIKV. In early childhood, the impact remained unclear. The study of neurodevelopment and impact of ZIKV to 24-month toddlers born to pregnant women infected with ZIKV was conducted. These women were pregnant during the 2016 ZIKV outbreak in America. There was no abnormal transfontanelle cerebral ultrasound finding before and after delivery. But later, the child had reduced brain activity and birth defects [13]. There are no approved vaccines, i.e., antiviral treatments, available. Using the dengue vaccine as a reference, a chimeric dengue/ZIKV named VacDZ was created. It is a live diluted inoculation to ZIKV. It reveals key markers of dilution of pathogenicity in interferon deficient adult mice. The vaccine shows an immune response to ZIKV. It neutralizes the virus and shows a successful shot against ZIKV in mice [14]. The mobilization of the health test center network to detect COVID-19 patients was prompted by the emergence of SARS-Co V-2. It started tracing the contacts; identify the hot spot area prone to active community transmission. The Brazilian public health system faced difficulties amid triple epidemics, i.e., dengue, chikungunya, and Zika virus. Various samples were collected from Brazil and tested. An inter-disciplinary response to health gained importance. A need to search for an effective vaccine became important as no vaccine is 100% effective to any virus [15]. The Literature Review is shown in Table I.

TABLE I. LITERATURE REVIEW

Sr. No.	Year	Ref.	Topic	Concept/ Theoretical Model	Paradigm/ Method	Context/ Setting/ Sample	Findings	Future Research
1	2016	[5]	Origin and spread of ZIKV	Paper illustrates the neurological complications and spread of Zika virus.	The Zika virus spread and its mode of transmission.	The virus is separated, the viral nucleic acid detection is done using polymerase chain reaction testing	Viruses mostly spread due to international travel. Person-to-person transmission both vertically and horizontally is possible.	Reasons for the spread of the virus
3	2018	[3]	Genome sequencing and variant analysis, phylogenetics, and profound sequencing of Zika virus strains	The existing strains of the Zika virus are validated.	RNA Extraction and Material Validation, Genome Sequencing, 3' Race, Read Assembly, Code Availability, and many more are done.	Sequence Comparison of Various Stocks, Deep Sequencing for Minor Variants, Recombination Analysis,	Minor variants were detected. Separate authentication is not needed. It represents the viral population diversity of ZIKV.	To understand better immunology, dissemination, possible cure, and avoidance
4	2019	[1]	Models of Genomic DNA k-mer spectra	The genomic complexity is studied in whole genome sequences using frequencies of DNA k-mers	k-mer sequences and modalities of around 100 species of Archea, Bacteria, and viruses are studied. A few species have multimodal spectra, whereas others have unimodal	To study complicated Genomic sequences, a probabilistic model is not sufficient. A low-order Markov model is discovered for study.	Dinucleotide suppression like C+G and CpG happens in Multimodal spectra. Tetrapods were analyzed. The sensor system was identified in the Human genome	Another genome like Entamoeba histolytica is to be studied.

5	2019	[2]	Apriori-based Frequent Symptoms Association Mining in Medical Databases	Symptom-based categorization, the symptoms are described by patients to a medical consultant	A large volume of the data repository is used, and using association rule mining; frequent associations are extracted.	Using the symptoms, the database is scanned to count the minimum support of each candidate.	Patients are found to be infected by malaria as most have the same set of symptoms.	The classification algorithm suggested mining important information health care databases.
6	2019	[16]	Flavivirus replication stimulated by circadian clock components	The invulnerable reactions of viruses are regulated by the circadian clock. It also affects pathogen replication. The underlying molecular mechanisms is difficult to understand.	Some components like BMAL1 and REV-ERB α are identified as circadian components.	Replication of flaviviruses, dengue, and ZIKV is inhibited by REV-ERB	Highlights the circadian clock component in regulating flavivirus replication	Fragment entry into hepatocytes and RNA reproduction to be performed
7	2020	[4]	Susceptibility of Enhanced Zika virus	ZIKV is transmitted from an infected host. Non-African mosquitoes have more potential to transmit.	Transmission from mouse to mosquito is assessed in immune-compromised	Day Post Mouse infection	The population outside Africa are more susceptible to the Zika virus as compared to native populations of aegypti in Africa	More sophisticated experiments to be performed
8	2020	[17]	ZIKV peripheral blood Gene expression responses	Congenital ZIKV is transmitted from mother to fetus. Tangential blood cells are unfettered to Zika virus and carry the infection	The gene expression of the Profile of diseased and healthy Peripheral blood mononuclear cells is compared. Samples were collected from expectant and non-pregnant women.	More expression by M1 shifted pro-inflammatory responses, less expression by M2-shifted anti-inflammatory	Shaping neonatal pathology in Pregnancy-induced immune dysregulation	Records participating in osteoclast diversity and cardiac myopathies
9	2020	[18]	Locating Motif in DNA sequences	Discovering motif or short replicating forms in genetic strings	Genomic monitoring systems of living creatures	Exhibited the motif hunt challenge	The actual motif is obtained from a set of candidate motif	Locate motif with better precision
10	2021	[6]	Enhancement of ZIKV infection by antibodies from West Nile	Owing to quiet antigenic similarity, exposure to ZIKV is difficult. It raises the potential of antibody-dependent development.	The capacity to improve or deactivate ZIKV infection is examined from the serum trials acquired with suggestive or symptomless WNV disease.	Sero-investigation information showed a 7% occurrence for WNV antibodies	WNV antibodies in the sera substantially improve Zika virus in Fc receptor constructive cells	Additional appropriate versions of ADE
11	2021	[7]	Spread of ZIKV mitigated due to release of Wolbachia diseased mosquitos	Deterministic and stochastic versions are created	Disease-free balance and its strength are explored. The effect of the factors on basic breeding is studied.	The stochastic prototype has a distinctive and bordered overall solution	The randomness may force the disease to be eliminated when the strength of the white noise is significant enough.	Wild mosquitos turn out to be destroyed as Wolbachia-diseased mosquitos reduce the time
12	2021	[8]	Costa Rica analysis on effect of contact patterns of sexual groups	Diseased mosquitos and sexual relations are primary reasons for Zika spread	Altering contact patterns among persons to manage Zika virus propagation	Build heterosexual network-based version to find threshold time of Zika infection	Analysis of the impact of altering the degree of heterogeneity in the sexually active places at distinct period	Upper limit later than the peak time of ZIKV infected cases
13	2021	[11]	Spread of ZIKV due to mutational reversions and health restoration	Zika caused four alterations prior to propagation in the Americas,	Diseases of Zika virus with and without alterations reveal that the original transformations decreased strength for human-intensified spread.	Classification of transmission adaptive Zika virus alterations	Early alterations and drift when the virus was announced	Pandemic rise due to reversions restored fitness

14	2021	[12]	Immunocompetent rats stomach microbiota modulation induced by ZIKV	Neuroendocrine functions are modulated by stomach microbiota composition. It includes contamination, cellular and immunologic reactions	Next-generation sequencing is used to analyze stomach microbiota structure	Compared to uninfected mice, the Zika virus activated a substantial illness in the microbes in the infected mouse.	A significant rise in the microbes from the Spirochaetaceae heredity in the stomach microbiota is observed.	Colon soft tissue homeostasis in adult immunocompetent rats is diseased by ZIKV.
15	2021	[13]	Exposure of ZIKV and neurodevelopment in toddlers normocephalic at birth	In French regions, newborn to Women pregnant in 2016 is studied. Impact of utero ZIKV exposure on neurodevelopment among toddlers	Inhabitants-based mother-child associate study, pregnancies overlapped with the 2016 ZIKV epidemic	Due to ZIKV infection during pregnancy, exposure of the toddler to ZIKV was identified. Different stages of questionnaires can be created for the assessment of Brain development	15.4% of ZIKV-revealed toddlers and 25.3% ZIKV-unexposed toddlers were found. ASQ result of 2SD cut-off (P = 0.10) is obtained.	Population-based cohorts of in utero ZIKV-exposed normocephalic newborns to be analyzed.
16	2021	[14]	Single-dose live-attenuated chimeric vaccine	VacDZ vaccine is proposed against ZIKV. PDK-53 dengue virus vaccine to be used as a backbone	It is studied as the key makers of mitigation. Plaque phenotype, heat sensitivity, mitigation of neurovirulence in suckling mice is studied.	VacDZ to be controlled as a conventional live virus inoculation, or as a DNA-introduced inoculation that emits living VacDZ.	Vaccine expressions induce an impervious defensive reaction to ZIKV in AG129 mice. It neutralizes antibodies and a strong Th1 reaction.	VacDZ to be an effective vaccine
17	2021	[15]	Spread of DENV, CHIKV, ZIKV, and SARS-CoV-2 in Brazil	Viruses like SARS-CoV-2 have triggered the deployment of the network. Detect COVID-19 affected role, locate interactions and classify regions	The incursion of SARS-CoV-2 during ongoing triple arboviral pandemics produced by dengue, ZIKV, and Chikungunya	Sample collected from Brazil.	Fitness methodology is an efficient way to handle to lessen the destructive impact triggered by pathogens.	Successful antiviral serum treatment for diseased patients and vaccination for non-infected

In the biological sequences, motifs are short repeating patterns. Motifs are least conserved, so; it is a challenging task to identify the motifs. Motifs are important to study the genetic behavior of the Zika Virus. The candidate motif of the dataset can be computed to identify the real motif using the computation method.

III. MATERIALS AND METHODS

A. Data Collection: Zika Virus Genome

ZIKV is family of Flaviviridae. It is a type of virus family. The entire genome sequence is available at NCBI. The functionalities of DNA were studied and discussed by Watson and Crick [19]. The filename of the dataset is ZikaVirus.fasta. It is stored as a nucleotide sequence, and fasta defines the file format. The size of the dataset is 11 KB. The genome of the Zika virus is stored in the file.

B. Circadian behavior in Zika Virus

The daily activities of any virus or any living organism are controlled by an internal clock called the circadian clock. Animals also follow the daily routine work based on the circadian clock. The clock maintains a 24-hour activity cycle. When it starts malfunctioning due to disorder, then many organisms face genetic diseases. It is called a delayed sleep-phase syndrome. The circadian clock has its base at the molecular level. Because of the malfunction of the circadian clock, people become prone to the many diseases. Heart attack is more common in the daytime, whereas asthma attack is more common in the night time.

Scientists Ron Konopka and Seymour Benzer identified abnormal circadian patterns in mutant flies and traced their causes. They found that the mutation in a single gene. Later after many years, a similar clock gene in mammals was discovered. Then many circadian genes were discovered. These genes display a high degree of evolutionary conservations across different species. Maintaining the circadian clock in a plant is very important as its entire life cycle depends on it. It is a matter of life and death for plants. More than a thousand plant genes are circadian. Such genes include the genes that control photosynthesis, photoreception, budding and flowering. Circadian behavior of the Zika virus is studied [16]. The immune system is regulated by the circadian clock. The immune system reacts to microbes, and pathogen replication is affected. BMAL1 and REV-ERB α are circadian components related to flaviviruses in dengue and Zika. The replication of flavivirus is regulated by the circadian clock.

C. Representations of Genes

DNA makes RNA which makes proteins. It is composed of four ribonucleotides, namely adenine, cytosine, guanine, and uracil. Thymine is replaced by Uracil in RNA. RNA transcript is translated to the amino acids sequence of a protein. These proteins regulate the function inside the cell.

DNA replication happens at the origin of replication called ori. Finding the position of ori is a complicated task even for biologists. The process of transcription and transpiration is also a complicated task happening inside the cells [20]. During transcription, all occurrences of Thymine (T) in DNA is replaced with Uracil (U). The RNA strand is then translated into an amino acid sequence. The RNA strand is partitioned into 3-mers. These 3-mers are called codons. Each codon takes the form of one of the 20 amino acids. During this, it follows the genetic code. Each of the 64 codons encodes an amino acid. Out of 64 codons, 3 codons are stop codons which halt the translation. For example, the DNA string "ATATCGAAA" transcribes into the RNA string "AUAUCGAAA" which translates into the amino acid "ISK".

Cells can transcribe different genes and can form RNA. The rates may be different for other genes. This is known as gene transcripts or gene expression. That is the reason why brain cells and skin cells behave in various manners. Both have different functionalities and vary greatly in their features. These variations help the cells to understand the time and keep track of it. Pregnancy-associated variations in reactions to ZIKV were identified using DNA expression of samples of different women. ZIKV infected pregnant showed pro-inflammatory responses [17].

IV. REGULATORY PROTEINS

The dataset contains the nucleotide sequence of the Zika virus. The length of the string can be found using the python program. It was found to be 10780. Each cell in the plant keeps track of day and night. There are three master cells, which are called clock masters. These are CCA1, TOC1 and LHY. These genes are controlled by external factors like sunlight and the availability of nutrients in the soil. This helps the organism to adjust to the gene expression.

The regulatory protein TOC1 regulates the expression of LHY and CCA1. The expression of TOC1 is suppressed by LHY and CCA1. It basically works in a negative feedback loop. Sunlight activates the transcription of LHY and CCA1. This deactivates the TOC1 transcription. At night time, TOC1 peaks and starts promoting the transcription of LHY and CCA1. LHY and CCA1 repress the transcription of TOC1, and the loop continues. The Condon usage is controlled by biased nucleotide composition in the Zika virus [21].

The transcription regulates a gene by binding to a specific short DNA. It is called a regulatory motif [22, 23]. It is also called as the transcription binding site. It is the upstream gene region which is 600-1000 nucleotide long, also the start of the gene. CCA1 can bind to "AAAAAATCT" in the upstream region. It will be helpful for bioinformaticians if the regulatory motifs can be in the gene. An algorithm to find motif will be useful.

A. Importance of Motif

Motifs are short sequence patterns. It has a finite length. It is used to study the features of DNA, RNA, and Proteins. Transcription factor binding sites are represented using sequence motifs. Finding the motif sequences of motifs can help in understanding the transcription regulation [24, 25]. Motifs represent active sites of enzymes and proteins structures

and stability. Study of DNA Arrays is done to identify the genes that are active during the daytime in plants. The upstream region of nearly 500 genes was extracted to find the circadian behaviour. The frequently appearing pattern in the upstream region was identified. Suppose it was found that "AAAATATCT" is the most frequent word that appears more than 40 times. It was named as an upstream region evening element. The gene loses its circadian behaviour if the gene is muted. In plants, the evening element is quite conserved. It is easy to find the evening element in the plant whereas in animals, finding the evening element is quite difficult because of many mutations. If a fly is infected with a bacterium, its immunity genes will get activated to fight with the bacterium. The immunity gene has elevated expression levels as the fly gets infected. The most common 12-mers is "TCGGGGATTTC" in the upstream region of many genes. It is the binding of the transcription factor NF-kB that activates the various genes in flies. The biological challenge of finding a regulatory motif is to be converted into a computational problem.

Depending on the similarity with the ideal motif, it will score individual instances of motifs. An ideal motif is the transcription factor binding site that best binds to the transcription factor. An attempt is made to select a k-mer from each string, as the ideal motif is not known. Each motif is scored depending on their similarity to each other. A list of t DNA string DNA is taken. Each string is of length n. k-mer from each string is selected to form a collection of motifs. It represents a (t X k) motif matrix. The motif matrix of the Zika virus is formulated. The most frequent Nucleotide in each column is identified and denoted by upper case. By using different values of k-mers in each string, a different motif matrix from each DNA string is created. The most conserved motif matrix is to be obtained. It also means matrix with most uppercase characters or few lower-case characters. The goal is to compute a collection of k-mers that minimizes the score.

B. Finding the Count Matrix

A 4 X k count matrix can be created for a given Motifs. It is denoted by count (Motifs). It represents the count of each Nucleotide in each column of the motif matrix. The element (I, j) represents the count of Nucleotide I in column j of Motifs.

{'A': [45, 37, 39, 41, 42, 36, 33, 46, 45, 50, 41, 41, 42, 46, 38, 45, 46, 36, 39, 49, 38, 42, 40, 45, 36, 39, 42, 41, 40, 30, 36, 40, 36, 44, 40, 46, 36, 45, 40, 48, 37, 46, 41, 41, 47, 40, 44, 51, 38, 47, 41, 46, 29, 49, 41, 51, 39, 40, 39, 46, 33, 50, 52, 36, 50, 40, 45, 58, 47, 50],

'C': [23, 39, 34, 29, 24, 34, 37, 33, 36, 30, 27, 34, 23, 35, 41, 38, 34, 43, 36, 32, 46, 31, 37, 34, 35, 42, 38, 28, 40, 33, 38, 35, 43, 37, 34, 33, 32, 41, 26, 35, 36, 36, 31, 36, 41, 34, 39, 24, 32, 33, 32, 29, 49, 42, 31, 24, 35, 33, 32, 33, 42, 26, 26, 30, 32, 36, 34, 26, 33, 31],

'G': [41, 53, 46, 45, 49, 49, 42, 35, 46, 45, 48, 47, 57, 38, 45, 50, 38, 50, 43, 42, 33, 49, 40, 42, 56, 43, 41, 46, 41, 45, 48, 48, 39, 37, 41, 45, 42, 38, 57, 42, 47, 45, 50, 47, 41, 46, 38,

43, 56, 49, 49, 52, 50, 35, 48, 41, 40, 46, 47, 42, 52, 42, 40, 50, 48, 48, 48, 35, 45, 38],

‘T’: [45, 25, 35, 39, 39, 35, 42, 40, 27, 29, 38, 32, 32, 35, 30, 21, 36, 25, 36, 31, 37, 32, 37, 33, 27, 30, 33, 39, 33, 46, 32, 31, 36, 36, 39, 30, 44, 30, 31, 29, 34, 27, 32, 30, 25, 34, 33, 36, 28, 25, 32, 27, 26, 28, 34, 38, 40, 35, 36, 33, 27, 36, 36, 38, 24, 30, 27, 35, 29, 35]}.

C. Formulating the Profile Matrices

An ideal motif is a transcription factor binding site that binds the best to the transcription factor. A motif finding problem is would score instances of motifs depending on the similarity to the ideal motif as the ideal motif is not known to us. Our aim is to find a k-mer from each string of the array and find the score depending on similarity.

The most frequent Nucleotide in each column is identified and denoted in the upper case. If two nucleotides are most frequent, then randomly, one Nucleotide is selected. The motif matrix is represented as a string of motif matrices. The i-th row and j-th column can be accessed by using the motif[i][j]. A conserved matrix is a matrix with a smaller number of lower-case characters or more uppercase characters. A most conserved motif matrix is to be selected from several different motif matrices. From a given sample of DNA string, using different values of k-mer, a different motif matrix can be created. The score of the motif matrix is found by counting the number of lower-case letters in the motif matrix. Then we can find a set of k-mer that reduces the score. To find it, all elements of the count matrix is divided by the number of rows in the motif i.e., t. The resultant matrix is the Profile of the motif matrix. The element (I, j) is the i-th nucleotide frequency in the j-th column of the motif matrix. The sum of any column is 1 in the profile matrix.

D. Finding a Consensus String for Zika Virus

A consensus string for the Zika virus is derived by identifying the most common Nucleotide present in the column of the motif matrix. If two nucleotides have the same frequency, anyone is selected at random. If the motif is selected correctly, the consensus matrix provides a candidate regulatory motif.

The most frequent Nucleotide in each column i.e. the Consensus (Motifs) of the Zika virus genome is:

AGGGGGGAGAGGGAGGAGGACGAAGGAGGTGGCAG
ATAGAGAGGAGAAGGGGGAGAGGGAGAAGAGGAAA
. So, the consensus string of the Zika virus is known.

E. Score Motif

The score motif of the Zika virus can be calculated using the consensus matrix. The number of symbols in the j-th column that does not match with the symbol at position j of the consensus matrix is added. The score of the Zika virus genome is 7444.

V. FINDING THE BINDING SITES

The motif finding problem is to be solved using a collection of strings. A set of k-mers for each string to be identified minimizes the score of the resulting motif. The input to the

problem is the DNA string and an integer k. The output is k-mer collection motif for each DNA. The output will minimize the score motif for any choice of k-mers. A general problem-solving technique like Brute Force algorithm can find a solution that will take a lot of time to execute for a large genome. The brute force algorithm will consider each possible k-mers Motifs and gives a solution as motifs with the least score.

A. Comparing the Working of Brute Force Motif Finding

The Brute Force motif finding technique identifies all possible solutions. These algorithms may be easy to design. It will be guaranteed to find a solution as it will verify each and every possible solution and identify the best solution or the motif with the lowest score. These algorithms will take an enormous amount of time as it has to check all possible solutions to discard a motif. The number of candidate motif will be too large to verify.

In the brute force algorithm, n-k+1 choice of k-mers is possible. There is a number of ways to form motif are (n-k+1)^t. The algorithm can calculate the score in k X t steps. The running time of the algorithm is of the order ((n-k+1)^t) X k X t. This value is too high to be calculated using even the fastest computer. If the value of k is already known, then it may be a little easy, but this is not possible. So, another method needs to be explored.

VI. RESULTS AND DISCUSSION

A. Use of Profile Matrix

Iterative procedures are used in many algorithms that select different alternatives during the iterations. Some of these iterations are correct, whereas some are not. The most attractive alternative is selected by greedy search algorithms. In a chess game, the Greedy search algorithm at every move tries to capture valuable piece. Greedy may not find the best solution but can quickly predict the approximate solution in many cases. So, The Greedy search is to be applied to biological problems to approximate a solution. So, this algorithm is applied for motif finding. A collection of k-mers from a DNA string is motif. The columns of the profile matrix are viewed as four-sided dice. Each Nucleotide {A, C, G, T} is present on each side. The first column of the profile matrix has the data (0.2922077922077922, 0.14935064935064934, 0.2662337662337662, 0.2922077922077922).

The sum of all probabilities is 1 for any column. So, it means that the probability of generating A is 0.2922077922077922, C is 0.14935064935064934, G is 0.2662337662337662, and T is 0.2922077922077922. The profile matrix for the Zika virus is given in the previous section. The probability of any selected string can be calculated using the entry in the i-th column of the Nucleotide. Say, for example, the probability of the series “ACGG” is found to be 0.006458989565084851. A higher probability k-mer is achieved when it is more like the consensus string, “AGGGGGGAGAGGGAGGAGGACGAAGGAGGTGGCA GATAGAGAGGAGAAGGGGGAGAGGGAGAAGAGGAA A”.

Prob(“AGGGGGGAGAGGGAGGAGGACGAAGGAGGTG
GCAGATAGAGAGGAGAAGGGGGAGAGGAGAAGAG
GAAA”) = 1.757001027053479e-36

B. The Search for Binding Sites

Search for Binding Sites or Greedy Motifs is done. The best motif is set to the first k-mer from each string in Deoxyribonucleic acid (DNA). The DNA string is represented using the abbreviation DNA. These strings will be helpful for study. It ranges over all possible k-mers in DNA[0]. It finds a value for each motif [0]. The algorithm builds a profile matrix for the k-mer. Then motif [1] is set equal to Profile most probable k-mer in DNA[1]. Greedy motif search is iterated by

updating Profile. To generalize, to find k-mers motifs in the i-strings of DNA, greedy motif search constructs a profile matrix and sets motif[i] equal to Profile most probable k-mer from DNA[i]. k-mer from each string in DNA is obtained as a collection of strings. Greedy motif search compares whether the motif score is greater than the best scoring collection of motifs. If it is greater than the best score motif is updated, otherwise ignored. The execution moves to the beginning of the loop, and the next symbol in the DNA[0] is selected. The results of the Greedy motif search can be for different k-mer strings and summarized in Table II. The 1-mer and 2-mer string have less significance, so the results are demonstrated for 3-mer till 15-mer string.

TABLE II. RESULTS OF GREEDY MOTIF SEARCH FOR DIFFERENT K-MER STRINGS

Sr. No.	k-mer	Score	Snapshot of the k-mer string
1	3-mer	10	['ACA', 'GGT', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', , 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', , 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGT A', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'AGA', 'GGA', 'GGA', 'GG GA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GCA', 'A GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'AGA', 'GGA', 'A 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA',
2	4-mer	85	['TGAC', 'GGTT', 'GGAT', 'GGAT', 'GGTT', 'AAGA', 'GGAA', 'GGAC', , 'GGAA', 'GGAC', 'GGAA', 'GGAA', 'GGAT', 'GGAA', 'GGAC', 'GGAC', , 'GGAT', 'GGAA', 'GGAT', 'GGAA', 'GGAC', 'GGAA', 'GGAA', 'GGA A', 'GGAA', 'GGAT', 'GGAA', 'GGAA', 'GGTT', 'GGAT', 'GGAA', 'GG AC', 'GGAC', 'GGAA', 'TGAA', 'TGAC', 'GGAA', 'GGAA', 'GGAA', 'G GAA', 'GGAA', 'GGAC', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAC', 'A GGAA', 'AGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAT', 'GGAT', 'GGAA', 'AGAA', 'GGAC', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAA', 'GGAA']
3	5-mer	166	['GAAAG', 'GGTTT', 'GGATT', 'GGATG', 'GAAAT', 'GGAAG', 'GGAAT', TG', 'GGAAG', 'GGATG', 'GGAAG', 'TGAAT', 'GGAGT', 'GGTAG', 'GG 'GGAAG', 'GGAAG', 'GGAGT', 'GGAAG', 'GGATG', 'GGAGG', 'GGAGG', G', 'GGTTG', 'GGAGA', 'TGAAG', 'TGAGG', 'GGAAG', 'GGAGG', 'GGT TGAAG', 'GGAAG', 'GAAAT', 'TGTAT', 'GGATG', 'GATGG', 'GGAAG', , 'GGAAG', 'GGAGA', 'GGAAG', 'GGAAA', 'GGTGG', 'GGATG', 'GGAG GAGG', 'GGAAA', 'GGAGT', 'GGAGG', 'GGTGG', 'GGAAG', 'GGAAA', 'A , 'GGATG', 'GGAGG', 'GGAAG', 'TGAAG', 'GGAAA', 'GGGAG', 'GGAAC AAA', 'GGAAA', 'GGAAG', 'GGATG', 'GGAAA', 'GGAAA']
4	6-mer	288	['TCAGAC', 'GGTTTT', 'CGGATT', 'TGGGTC', 'TGGGTC', 'AAGAAG', 'A C', 'TCAGAC', 'TGGATG', 'GGGAAG', 'TGGGAT', 'TGGAAAG', 'TGTATT', GAAT', 'TGGAAC', 'GGGGAG', 'AGGGAC', 'GGGAAG', 'TGGAAAG', 'TGGG 'TGGATC', 'AGGAAG', 'TGGGAG', 'TGGGTG', 'TGGTAT', 'TGAAG', 'T , 'TGGGAG', 'AGGGAG', 'TGGAAAG', 'AGGAAC', 'AGGAAC', 'TGGGAG', GAC', 'AGAGAG', 'AGGAAT', 'TGGGTG', 'ACAAAG', 'TGGAAC', 'CGGAA TCTATG', 'AGTGAC', 'TGGGAC', 'ACAATT', 'GAGAAC', 'TGGAAAT', 'TG , 'GGAAAC', 'GGGGAG', 'TGTGTC', 'TGAAT', 'TGAAG', 'TGGAAAG', AG', 'ACAAAG', 'GGGGAG', 'AGGAAG', 'TGGGAT', 'TGAATG', 'TGGGAC GGAAG', 'AGGATG', 'TGGATC', 'GGAAAG']

5	7-mer	412	['AGTTTGAA', 'GGTTTTAT', 'CGGATTGT', 'AGGATGGT', 'AATCAAGC', 'GGGTTGAT', 'GAGCTGGT', 'GGTAAGAT', 'AAGGTGAA', 'GATGTGGA', 'CAGATGGA', 'AGGTTGAT', 'GAGAAGAA', 'CATTGAA', 'CAGTTGGA', 'T', 'AGGTTAGA', 'AGTGTGAT', 'AGGCTGAA', 'CAGATGGA', 'TAGCTGGG', 'TT', 'TGGTTGTT', 'GCTCTGAC', 'AGTGTGAA', 'AGGCTGGT', 'GAGCTGGC', 'GGT', 'AATATTTA', 'CAGCAGGA', 'TGGGAGTT', 'AGGGAGGA', 'GAGCTGGT', 'TGGG', 'TGGCTGGA', 'AGGCAGGA', 'TGTATGGA', 'AAGATGGC', 'AGGAAGCAAAG', 'GGGCTTGA', 'GGGAAGGA', 'CAGCTTGG', 'AAGCTGGA', 'GGTATGTGTC', 'GAGATGGT', 'GGGCTGGA', 'TGGGTGGA', 'GGGGTGGG', 'TGGC', 'GTCAGGT', 'AGTCTGTA', 'TATTTGAA', 'TGTGTGTA', 'AATTTGGA', 'GGG
6	8-mer	508	['AGTTTGAA', 'GGTTTTAT', 'CGGATTGT', 'AGGATGGT', 'AATCAAGC', 'GGGTTGAT', 'GAGCTGGT', 'GGTAAGAT', 'AAGGTGAA', 'GATGTGGA', 'CAGATGGA', 'AGGTTGAT', 'GAGAAGAA', 'CATTGAA', 'CAGTTGGA', 'T', 'AGGTTAGA', 'AGTGTGAT', 'AGGCTGAA', 'CAGATGGA', 'TAGCTGGG', 'TT', 'TGGTTGTT', 'GCTCTGAC', 'AGTGTGAA', 'AGGCTGGT', 'GAGCTGGC', 'GGT', 'AATATTTA', 'CAGCAGGA', 'TGGGAGTT', 'AGGGAGGA', 'GAGCTGGT', 'TGGG', 'TGGCTGGA', 'AGGCAGGA', 'TGTATGGA', 'AAGATGGC', 'AGGAAGCAAAG', 'GGGCTTGA', 'GGGAAGGA', 'CAGCTTGG', 'AAGCTGGA', 'GGTATGTGTC', 'GAGATGGT', 'GGGCTGGA', 'TGGGTGGA', 'GGGGTGGG', 'TGGC
7	9-mer	603	['AGTTTGAAG', 'GGTTTTATT', 'CGGATTGTC', 'AGGATGGTC', 'AATCAAGC', 'GGTCATGAT', 'GGGTTGATG', 'GAGCTGGTT', 'GGTAAGATC', 'AAGGTGAT', 'TGGATAAAC', 'CGGCTGAAA', 'CAGATGGAC', 'AGGTTGATA', 'GGGGGGAG', 'CATGTGGAG', 'CGGCTGGAA', 'TGGATGGTG', 'AGGTTAGAG', 'AGCTGAAG', 'TGGGAGGAT', 'GAGATGTAG', 'GAGCTAATT', 'TGTCTTTTG', 'AGGTGGTC', 'CAGCTGGAG', 'CAGATGGAG', 'AAGTTGGAG', 'GGGGAGATG', 'AATCTGTAT', 'TGGATGAGG', 'ACGCATTTT', 'GCTTTGATT', 'AGTTTGTT', 'GGTGTGGAC', 'TGGATGGAC', 'CGGCTTTTG', 'GGGCAGAGA', 'GGGGTTGC', 'CATTITGCT', 'CAGCTGGCA', 'TGGATGGAA', 'TAGCAGTAG', 'AGGCT
8	10-mer	697	['GCGAAAGCTA', 'GGTTTTATTT', 'CGGATTGTCA', 'CCGGACTTCT', 'AATC', 'GGTTTCGCTT', 'GGTCATGATA', 'AGGTATGTCA', 'GAGCTGGTTA', 'GAC', 'AAGGCCAACT', 'GAGCAGTTCA', 'CCTCTGGCCA', 'GGGACAGTCA', 'GCT', 'GGGACAGGTA', 'AAGCAAGCCT', 'ACGCAATCCT', 'GGGATCTGTA', 'CC', 'GAGATCACCC', 'GATAAGGCC', 'AGTAAGGTCA', 'AGGAAGGGCT', 'ACC', 'GAGCTGGCCC', 'GGGTTGCGCA', 'AGTGTGGACA', 'GCGGAAGTCA', 'TTCA', 'AGTTATGTTA', 'AGAGACTCCT', 'GGGAAAACCA', 'GAGGAAGCCC', 'CAACT', 'AAGCCGGTCA', 'ACACATGCCA', 'ACGAAGACCA', 'AAGATGGCCT
9	11-mer	806	['GAGTTTGAAGC', 'GGTTTTATTTT', 'CGGATTGTCAA', 'CCGGACTTCTG', 'GCGGTCGCAAA', 'GGTTTCGCTTT', 'GGTCATGATAC', 'AGGTATGTCAG', 'GACACAGGCCCTT', 'CGGAACCTCCAC', 'GCACATGCCAA', 'GAGCAGTTCCAC', 'GGGCTGTCTG', 'GGTCTGAACAC', 'GGAGTGTGAT', 'GGGGTGTTCGT', 'GGGAAGGACAC', 'GATCTGATCAT', 'GAGAGGGCTAC', 'GAGCTTGAAAT', 'AGGGTGATCGC', 'GCGATGGTTGT', 'GGGCACACTGC', 'GGGGGGTTTAT', 'AGGCTGGTCC', 'GAGGGGGTCTT', 'GGGGAGATGTC', 'GATCTGGTGTG', 'GCGAGGAACAT', 'CAAGAGGATAC', 'CGTGACGCATT', 'GAAGTGAAGT', 'AATGAGATCGC', 'GGAATAACCTA', 'GCAGAGGTGTG', 'GAGGTGGATGG', 'GGAGTGTGGA', 'C

10	12-mer	922	['AGACTGCGACAG', 'GGTTTTATTTTG', 'CGGATTGTCAAT', 'AGGATGGTCTTG ACGGAAC', 'AGAATACACAAA', 'TGGATATTCAGG', 'GGGAAGCTCAAC', 'AGG' , 'AGGTTGAGATAA', 'GGACAGGCCTTG', 'ACGCTGGGGCAG', 'CCAAAAGGCA AAAAGCATT', 'TCATTGGGCAAG', 'AGGAATGTCCTG', 'ACACAAAGAATG', 'G' TG', 'AGGTTAGAGAAG', 'AGTGTGATCCAG', 'AGGCTGAAGAGG', 'TGAATGGC CTATGATCCTG', 'GGAAATGAACAC', 'AGTCAGACCAGC', 'AGCATGCTGCTG', CAGG', 'AAGATGCGGAAG', 'AGAGAGATCATA', 'AGACTGGAAAAA', 'GGGATG' 'AGGAAGAGACTC', 'GGGAAAACCAGG', 'TCGCTGCTGAAA', 'ACTCTGGAACAG' GGACTT', 'GGAGATGCCTAA', 'GCGCATAGGCAG', 'AGACTGACGAAG', 'CGAC
11	13-mer	1008	['TTGAAGCGAAAG', 'GGTTTTATTTTGG', 'CGGATTGTCAATA', 'CCGGACTTC' A', 'ACGGAGATCTAGA', 'TCGCAAACCTGGT', 'TGGATATTCAGGA', 'TACTTG CTCC', 'GATGAGAATAGAG', 'TTGGAAGCCTAGG', 'TTGTATTACTTGA', 'TCC TCTAAGA', 'GGGGAGAAGAAGA', 'TGAAAAGCATT', 'TGGATCAGTTGGA', 'i GAGTTCAACT', 'TGGAAGGCTTGGG', 'AAGAATGCCACT', 'TGGAGGATCATGG', GGGCTAAAGATGG', 'GGTACTGCAGGA', 'TTGGAGTCTTGT', 'TGGCAGTCTG , 'TGGTGGGACTGCT', 'CTGTTGGCCTGAT', 'GGTTTCGCCAAGG', 'AGTGTGGA GA', 'CTGTGGTCCATGG', 'AGGAACATCCAGA', 'AGGAACCTCAGGA', 'GGGT
12	14-mer	1113	['TTGAAGCGAAAGCT', 'GGTTTTATTTTGGGA', 'CGGATTGTCAATAT', 'CCGGAC CTATGCTGGAT', 'ACGGAGATCTAGAA', 'TCGCAAACCTGGTT', 'TGGATATTCAG TTTGGC', 'GGGAAGAGCATCCA', 'ATGATTGTTAATGA', 'TGGAAGCCTAGGAC', T', 'AGGGACAGATGGAC', 'GAGAACTCTAAGAT', 'GGGGACTCTTACAT', 'TGG AAGCAAGCCTGGGA', 'ACGCAATCCTGGAA', 'GGGATCTGTAaaaa', 'TGGAAGGC CCCAAATGA', 'GTGAAGAGCTTGAA', 'ACCCACTGCAAGCG', 'ATGGAGATAAGGC GCAA', 'GTTCTCATCAATGG', 'TGGCAATCCTGGCT', 'TCTCTCTGAAGGGA', 'i , 'TGGTACGTATACGT', 'GACCACAGATGGAG', 'TGGGAGTTATGCAA', 'GCGCT GAAAACCAGGAG', 'TGTCGCTGCTGAAA', 'TGGAACAGAAATCG', 'ACTATAATCT GATGGGC', 'GGTCATACTTGATG', 'TGTCACACATGCCA', 'ACGAAGACCATGCA',
13	15-mer	1190	['CTGTTGCTGACTCAG', 'GGTTTTATTTTGGAT', 'CGGATTGTCAATATG', 'CCGG TG', 'CAGATGACGTCGATT', 'ACGGAGATCTAGAAG', 'TCGCAAACCTGGTTG', ' TACCTT', 'AAGGGAGCCTGGTGA', 'CAGAGAATCTGGAGT', 'CAGGACATGAAACTG CGCCTGAAAA', 'CTGAAACACTGCACG', 'GCGGTGGACATGCAA', 'CGTAATCACTG TGGACTTCTCAAAG', 'GTTGAAGCCTGGAGG', 'AAGCAAGCCTGGGAA', 'ACGCAAT , 'AAGAATGACACATGG', 'ATGGAATAGAAGAGA', 'CACAATACCAGAGAG', 'GTG TAG', 'AATGAACACTGGAGG', 'TTCAAAGTCAGACCA', 'TGCGATCTCCGCCTT', CTGCTAA', 'CGTGGTCTCAGGAAA', 'GCGGAAGTCACTGGA', 'AGGTGGTCTGATG GAGAGAGCGAG', 'CGTTGCGCTGGATT', 'GTGGGAGAGTGATAG', 'CTGAAGAAGA

The results are obtained for various k-mers. The 15-mer has a score of 1190. The score obtained using the Greedy method may not be the optimal score, as there may exist a motif with a minimum score which can be obtained by finding all possible solutions. But this method provides 15-mer motif in a reasonable time. This score can be improved using other algorithms.

VII. CONCLUSION AND FUTURE WORK

The study is conducted to understand the behaviour of the Zika virus. ZIKV shows circadian expression, which also regulates the day-to-day functions in genes. Zika virus infection causes birth defects like neurological disorders in babies, and no proper cure or vaccine is available. This paper attempts to find the probability of every k-mer for a given profile matrix. The Profile most probable k-mer is calculated. This k-mer is most likely to be generated by Profile compared to all k-mers in the text. For example, if the size of k-mer string selected is 12, then the highest probability k-mer is found using the probabilities values is “AGGGGGGAGAGG”. Similarly, if the size of k-mer string selected is 15, then the highest probability k-mer is found using the probabilities values is “AGGGGGGAGAGGGAG”. Greedy motif Search is done as

it is better compared to Brute Force search in real-time. The best motif is set to the first k-mer from each string in DNA. k-mer from each string in DNA is obtained as a collection of strings. Greedy motif search compares whether the motifs score is greater than the best scoring collection of motifs. If it is greater than the best score motif is updated, otherwise ignored. Python programming is used to study the genetic behaviour of the Zika virus genome.

If any value in the profile matrix is zero, then the entire probability of the string becomes zero. If a string is obtained for which the profile matrix value is zero, the string is completely rejected. Such results can be improved using other methods like the Laplace Rule of succession. The score obtained can be further improved using this Laplace Rule of succession.

REFERENCES

- [1] Benny, H. David, G. Nick, L. Yaron and T. Massingham, "Genomic DNA k-mer spectra: models and modalities," *Genome Biology*, 2019.
- [2] K. R. P. Ram, R. Jayakumar and A. Sankaridevi, "Apriori-based Frequent Symptomset Association Mining in Medical Databases," *International Journal of Recent Technology and Engineering (IJRTE)*, ISSN: 2277-3878, Volume-7, Issue-5C, 2019.
- [3] S. Susmita and P. Vinita, "Whole genome sequencing, variant analysis, phylogenetics, and deep sequencing of Zika virus strains," www.nature.com/scientificreports, 2018.
- [4] A. Fabien, D. Stéphanie, M. Caroline, F. Igor, R. Noah, M. Elliott, M. Daria and B. Artem, "Enhanced Zika virus susceptibility of globally invasive *Aedes aegypti* populations," sciencemag.org, VOL 370 ISSUE 6519, 2020.
- [5] K. K. Mary, A. Tomas, F. Veronika, S. S. Ravi and D. Christopher, "Zika: the origin and spread of a mosquito-borne virus," *Bull World Health Organ* 2016;94:675–686C, 2016.
- [6] H. Garg, R. Yeh, W. Douglas M, T. Mehmetoglu-Gurbuz, R. Resendes and B. Parsons, "Enhancement of Zika virus infection by antibodies from West Nile virus seropositive individuals with no history of clinical infection," Garg et al. *BMC Immunology* (2021) 22:5 <https://doi.org/10.1186/s12865-020-00389-2>, 2021.
- [7] X. Ling, C. Xinru and W. Hui, "Releasing *Wolbachia*-infected mosquitoes to mitigate the transmission of Zika virus," *Journal of Mathematical Analysis and Applications*, Volume 496, Issue 1, 2021, 124804, ISSN 0022-247X, <https://doi.org/10.1016/j.jmaa.2020.124804>, 2021.
- [8] L. Xiao-Feng, J. Zhen, H. Daihai and L. Li, "The impact of contact patterns of sexual networks on Zika virus spread: A case study in Costa Rica," *Applied Mathematics and Computation*, Volume 393, 2021, 125765, ISSN 0096-3003, <https://doi.org/10.1016/j.amc.2020.125765>, 2021.
- [9] A. Wang, S. Thurmond, L. Islas, K. Hui and R. Hai, "Zika virus genome biology and molecular pathogenesis. *Emerging microbes & infections*," <https://doi.org/10.1038/emi.2016.141>, 2017.
- [10] C. Hammack, S. C. Ogden, J. Madden, A. X. C. Medina, P. E. Y. Son, A. Cone, S. Giovanazzi, R. Didier, D. Gilbert, H. Song, M. G. Z. Wen, M. Brinton, A. Gunjan, T. H and M. Heise, "Zika Virus Infection Induces DNA Damage Response in Human Neural Progenitors That Enhances Viral Replication," *Journal of Virology*, 10.1128/JVI.00638-19, 2019.
- [11] L. Jianying, L. Yang, S. Chao, T. D. N. Bruno, R. Yun, L. H. Sherry, H. R. Grace, R. A. Sasha, R. A. Clark, P. Kenneth, V. Nikos, S. Pei-Yong and C. W. Scott, "Role of mutational reversions and fitness restoration in Zika virus spread to the Americas," <https://doi.org/10.1038/s41467-020-20747-3>, 2021.
- [12] C. Rafael, O. S. Igor de, A. B. Heloísa, P. d. S. Livia, d. N. Raquel, P. Gabriel, S. P. Paulo, P. K. Gary, F. M. Corinne and M. KellyGrace, "Gut microbiota modulation induced by Zika virus infection in immunocompetent mice," <https://www.nature.com/srep/>, 2021.
- [13] G. Rebecca, F. Olivier, T. Benoît, D. Mama, E. Narcisse, M. Nicolas, M. Adeline, H. Jean-Christophe, L. Noémie, C. Elvire, H. Bruno and F. Arnaud, "In utero Zika virus exposure and neurodevelopment at 24 months in toddlers normocephalic at birth: a cohort study," *BMC Medical*, 2021.
- [14] C. Wei-Xin, C. H. L. Regina, K. Parveen, S. L. Tian, Y. Thinesswary, Y. K. Hao, T. Zi-Yun, K. S. Cyrill, Z. Rong-Rong, L. Xiao-Feng, A. Sylvie, Q. Cheng-Feng and J. H. C. Justin, "A single-dose live attenuated chimeric vaccine candidate against Zika virus," *Sealy Institute for Vaccine Sciences*, 2021.
- [15] J. R. d. S. Severino, J. F. d. M. Jurandy and P. Lindomar, "Simultaneous Circulation of DENV, CHIKV, ZIKV and SARS-CoV-2 in Brazil: an Inconvenient Truth," *Science Direct, One health*, 2021.
- [16] X. Zhuang, A. Magri, M. Hill, A. Lai, A. Kumar, S. B. Rambhatla, C. L. Donald, Lopez-Clavijo, F. R. A. P. K. S, W. H. Chang, P. A. C. Wing, R. Brown, X. Qin, P. Simmonds, T. F. Baumert, D. Ray, A. Loudon, P. Balfe and M. Wakelam, "The circadian clock components BMAL1 and EVERB α regulate flavivirus replication," *Nature Communications* Volume 10, Issue 1, 1 December 2019, Article number 377, Dec 2019.
- [17] S. J. Lim, A. Seyfang, S. Dutra, B. Kane and M. Groer, "Gene expression responses to Zika virus infection in peripheral blood mononuclear cells from pregnant and non-pregnant women," *MicrobiologyOpen*, Volume 9, Issue 12, 2020.
- [18] B. A. Faisal and S. R. S. Md, "MFEA: An evolutionary approach for motif finding in DNA sequences," *Informatics in Medicine Unlocked*, 2020.
- [19] L. Pray, "Discovery of DNA structure and function: Watson and Crick.," *Nature Education* 1(1):100, 2008.
- [20] G. D. Stormo, "DNA binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16-23, 2000.
- [21] V. H. Formijn and B. Ben, "Nucleotide composition of the Zika virus RNA genome and its codon usage," *Virology Journal*, 2016.
- [22] X. Peng, C. Xingyu and R. Sanguthevar, "EMS3: An Improved Algorithm for Finding Edit-Distance Based Motifs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, 2021.
- [23] X. Peng, C. Xingyu and R. Sanguthevar, "Efficient Algorithms for Finding Edit-Distance Based Motifs," *Algorithms for Computational Biology*, pp. 212-223, 2019.
- [24] K. S. Geir and D. Finn, "A survey of motif discovery methods in an integrated framework," *Biology Direct*, 2006.
- [25] K. D. Modan and D. Ho-Kwok, "A survey of DNA motif finding algorithms," *BMC Bioinformatics*, vol. 8, 2007.