An Efficient Computational Method of Motif Finding in the Zika Virus Genome

Pushpa Susant Mahapatro[1], Jatinderkumar R. Saini[2*]
Vidyalankar School of Information Technology, Wadala, Mumbai, India[1]
Symbiosis Institute of Computer Studies and Research, Pune, India[2]
Symbiosis International (Deemed University), Pune, India[1, 2]

Abstract—**The Zika virus (ZIKV) outbreak and spread is a global health emergency declared by World Health Organization. ZIKV rapidly spread across the world, causing neurological disorders. It is gaining public and scientific consideration. ZIKV genome biology and molecular structure are better understood with published papers. Genetic regulation is better understood by finding the motif in the DNA Genome sequence. The transcription factor binding sites need to be identified to understand the genetic code. There is diversity in gene expression. Motif-finding methods work towards efficiently identifying the repeated patterns in the genome. ZIKV genome sequence is used in the study. Identifying the motif is still a difficult task. There is a low probability of identifying the binding sites. Finding all possible solutions is challenging as it requires a lot of time and has high space complexity for finding long motifs. The Greedy search technique with pseudocount finds the motif in real-time. The count matrix is computed, and the profile matrix is constructed from the genome of the Zika virus. The calculated consensus string helps in calculating the score of the motif. The Greedy motif search technique is applied in this paper to find the motifs in the Zika virus Genome. This technique is not applied earlier to find the motifs in Zika Virus. The motifs are identified using a Greedy motif search without pseudocount and with pseudocount.**

Keywords—Consensus string; genome study; greedy search technique; motif search; pseudocount; regulatory proteins; ZIKV; Zika virus

I. Introduction

The Zika virus (ZIKV) was first separated from a rhesus macaque in Uganda in 1947. The study of ZIKV was not given importance till 2015. In 2015, Brazil's big epidemic of ZIKV infections was linked to intensification in microcephaly cases. ZIKV can spread sexually and is known to persevere in the male and female reproductive systems. West Nile virus (WNV), Dengue virus (DENV), Yellow fever virus (YFV), and Japanese encephalitis virus (JEV) are included in the family Flaviviridae, genus Flavivirus. These are mosquito-borne pathogens. Zika virus also belongs to the same family of viruses. The ZIKV lifespan has Aedes mosquitoes and monkeys, whereas a broader range of mosquito species transmits WNV.

World Health Organization (WHO) has acknowledged the Zika virus (ZIKV) as a public health crisis worldwide. ZIKV is a flavivirus. It has its place in the family of Flaviviridae. It is spread over several parts of Africa, Southeast Asia, and the Pacific island. In August 2016, the ZIKV outburst in Brazil was the major ever recorded, with a projected 165000 doubted cases. ZIKV is transmitted through monkeys and the Aedes genus [1]. ZIKV infection causes slight illness, headache, and rash. A recent study suggested that the virus causes neurological disorders such as Guillain-Barre syndrome. ZIKV is transmitted from mother to child and is also transmitted sexually. It has identical symptoms as compared to different arboviral diseases like DENV. Analysis based on symptoms is unpredictable for precise detection. Test center analysis is vital to obtain conclusive results. A suitable choice of molecular analysis is meaningful for routine ZIKV or flavivirus detection. No specific treatment for the infection is available. The course of injection and medicine development is extremely complicated. Anti-ZIKV vaccine development may take several years. Traditional drug development policies make the circumstances worse. Silico methods are helpful in enlightening possible vaccine candidates.

This fast rise of ZIKV cases and its consequence is severe. The disease has provoked the research community to produce interventions to battle Zika disease. The disease mechanism is presently unstated. Reviewing and analyzing ZIKV genome ecology and pathogenesis can provide awareness of ZIKV. microRNAs are also found to play a significant part in virus-related diseases and activation of the phenomenal immune response. Methodical genome-wide study of the ZIKV genome may assist in designing antiviral therapeutics. Zika virus has a 10.7 kb genome of single-stranded RNA. Multiscale examination of the genomic data produced throughout the widespread of infection and available in public databases was combined. The focus of the study was to inspect virus-related appearance at different scales [2] [3].

The research problem was understanding the regulatory mechanism of genes. This involves finding transcription factor binding sites. Identifying regulatory networks is also challenging. The objective of the research was to find the regulatory relations of genes. It is better considered with the identification of motifs. Motifs are known to have complex forms. An essential class of motifs is spaced motifs. It consists of short segments separated by nucleotides of different lengths. Locating motifs is a difficult task. Existing algorithms identify short contiguous motifs. Better algorithms identify spaced motifs with a different number of spaces in between [4]. So, this research is significant as it will help in identifying the spaced motifs with a mutation at one or more places in the genome.

In this Research paper, a literature review is conducted on various studies conducted on the Zika virus. The study is

*Corresponding Author.

conducted on phylogenetic analysis, ZIKV circulation in different countries, RNA-protein interaction, Viral RNA targets, Host cell-binding mechanism of ZIKV, Motif finding and Genetic behaviors of ZIKV, ZIKV infection, and the role of E-protein amino acids and many other aspects of Zika virus. The count matrix and profile matrix is calculated on the dataset of the Zika virus genome with pseudocount. The Greedy motif search method is applied to the dataset, and results are tabulated with and without pseudocount.

II. Literature Review

Zika virus has turned out to be a worldwide health problem as it is linked to potential congenital disabilities. The virus was revealed 70 years ago; still, the genomic construction and genetic variation are not fully known. The genome structure is compared with different other flaviviruses. Structural and functional similarity is found between the various flaviviruses' genome structures. The similarity is found in the conserved terminal structures. So, it is concluded that the Zika virus shares a high constitutional and functional similarity with other viruses of the Flavi family. It is known from the genomic comparison. Also, the prediction of motifs in viral proteins in the Zika virus with other viruses shows similarities. All Zika virus strains in America have similarities with the strains in Asia. Some conserved amino acids differentiate earlier African strains from Asian and American strains. Studies provide clues for different viruses' studies.

The Zika virus spread over more than fifty nations of the world. The evolution and spread are understood by studying the replication in the genome. Zika infection happens when an infected mosquito bites a person. It is also transmitted from one person to another person due to various reasons. ZIKV molecule needs to be studied to understand the infection in detail. It can also help in finding a solution to the problem. The neural progenitor cell growth is affected by ZIKV infection in monkeys [5] [6]. The same things happen in the case of humans. The virus damages the DNA of humans and monkeys. It also initiates DNA damage responses. The DNA damage response is then attenuated. The cycle of the growth of the virus is considered to determine the behavior of the virus during different times of the day. The virus may behave differently during the day and during the night. A different mutation of the Zika virus is reported in America. These mutations are different from the mutations found in Asia and Africa. The Aedes Aegypti mosquito was studied with mutation, and the results were compared with the studies performed without mutation. Fitness is increased for new mutations. It increases the risk of the spread of the virus. Fitness is reduced for original mutations. Zika virus infects the immunocompetent adult. It precipitates and increases brain damage due to antiviral responses. The immune system of mice is studied. The modifications stimulated by the Zika virus were found. A significant decline in micro-organisms like Actinobacteria and Firmicutes phyla was found due to the Zika virus infection. A boost of Spirochaetes and Deferribacteres was prominent in infected mice compared to healthy mice. The modulation caused the enhancement of white blood cells. The Zika virus induced the modulation of microbiota. Birth defects are caused due to utero exposure to the Zika virus. The ill effects of this virus were not prominent in the early stages of birth. The researchers concluded that if they studied the toddlers till twenty-four months, then the effects would be clear. It was concluded that the women who became pregnant during the year 2016 were to be studied. It was the time of the outbreak of the Zika virus in America. The study was conducted on the neurodevelopment stages of toddlers from birth to twenty-four months. The behavior of the child was normal before and after birth. The activities of the body seemed to be reduced as the child started growing. The child started showing the symptoms of birth defects. The connection of birth defects to Zika virus infection was later identified. As such, no vaccine was available for the disease, and treatment of the condition was not possible. People were not aware of such an infection. A vaccine named VacDZ was produced using the dengue vaccine as a reference [7]. The immune responses are seen to the Zika virus. The vaccine seems to be showing positive results in mice infected with the Zika virus. Blood samples were collected from Brazil to study the impact and spread of the virus. The people of Brazil were impacted by three viruses: Zika, chikungunya, and dengue. The situation was challenging to handle and trace the spread of the virus [8]. The review of the literature is presented in Table I.

TABLE I. Literature Review

Sr. No.	Year	Ref.	Topic	Concept/ Theoretical Model	Paradigm/ Method	Context/ Setting/ Sample	Findings	Research Gap
1	2016	[2]	ZIKV circulation in America and phylogenetic analysis	ZIKV is a global public health problem. The genetic alterations and genomic structure are constantly under study.	The comparison of the genome organization of mosquito-borne diseases and the ZIKV genome was made to understand to features of the Zika virus genome. The conserved terminal structure is studied.	The similarity of structural and functional components was found between the Zika virus genome and other viruses. It was clear with sequence comparison and prediction of functional motifs.	American strains of the Zika virus form a unique clade with the Asian lineage. A sequence of preserved amino acids was identified that discriminated the Asian strains from the African strains. Studies provide an overview of the characters of the ZIKV genome.	Critical study to be conducted in the virological and epidemiological domain.
2	2018	[3]	RNA-protein interaction	RNA-protein interactions control	RaPID has usage in different applications.	A new mutant BirA has been introduced to progress the	A direct study of Protein-RNA interaction is enabled	The RNA protein connections can be further studied

			detection in living cells	cellular functionalities and diseases. Proximity-dependent protein labeling is used by RNA-protein interaction detection (RaPID).	It helps in assessing protein binding to mutant RNA motifs in human hereditary illnesses in finding potential post-transcriptional networks in breast cancer.	BirA-labelling component of RaPID.	in living cells. The timescale of interaction is as short as 1 min.	to understand more about cellular functionalities.
3	2018	[5]	small-molecule ligands in Viral RNA targets	Conserved structured motifs are contained in RNA genomes. These are attractive targets for small molecules of viral infection.	Conformational states are affected by ligand binding. These play an important role in the infection process.	HIV and Hepatitis C virus are studied. Inhibition process, RNA functions ligands are focused. The RNA targets in other viral pathogens are also studied.	The structural viral RNA motif is targeted by small molecules called ligands.	Ligand's discovery guides therapeutic opportunities. It guides the participation in key processes of infection as well as high conservation of pathogens.
4	2019	[6]	Host cell-binding mechanism of ZIKV	Two reputed binding mechanisms of inherited and emerging Zika viruses were identified. It features the envelope protein residue ASN154.	ASN 154 was visible to neuronal cells and fibroblasts to classical Zika virus protein and cell communications.	Peptides meaningfully expressed Vero cell contamination by ZIKV strains. A putative binding mechanism of ancestral African ZIKV strains and emergent Western Hemisphere strains is represented.	Western hemisphere strains may additionally be capable of utilizing PS-mediated entry to infect host cells. The region surrounding E protein ASN154 is capable of binding fibroblasts and primary neuronal cells	PS-mediated entry may be a secondary mechanism for infectivity utilized by Western Hemisphere strains.
5	2021	[4]	Motif finding and Genetic behaviors of ZIKV	The genome of ZIKV is used for the study. Circadian behavior and the appearance of genetic factors are considered.	Motif finding is done using a count matrix and profile matrix. The score of the consent string is calculated.	The Brute force method is certain to find the motif, but it will take a huge amount of time. So other techniques are to be used.	The motif finding is done using Greedy search techniques, and results are analyzed.	The score can be enhanced using other techniques, such as Gibbs Sampler.
6	2021	[9]	ZIKV infection and the role of E-protein amino acids	Learning of E-proteins in ZIKV infection, as it is the furthest proteins in the constitution of ZIKV	ZIKV uses covering proteins to bind to cell addition receptors.	Many sites are situated in an unusual position of the protein construction, such as the α-helix in the stem area.	Cover protein amino acids that contribute to the flavi virus initial contamination process are presented.	A detailed study of E-protein assistance can accomplish an efficient antigen strategy.
7	2021	[10]	Dengue virus whole genome of structural landscape	Protein binding domains and secondary structure is to be used for categorization	A prediction score of 85% is to be achieved with a secondary structure. The consensus secondary structure of profiles is computed. For computation, silico models are used.	Viruses of Dengue and Ebola are less structured. Viruses show structural patterns.	A correlation between the number of interaction sites with human proteins and the secondary structure of 89% is achieved in ZIKV.	Further classification of complex viruses to be done using the given approaches
8	2021	[11]	Dependent immune response genes of DENV provoke the countenance of PAF1	To establish infection, DENV disturbance of innate excepted responses is critical. DENV non-structural protein five plays a dominant part in such distress.	PAF1, LEO1, CTR9, and CDC73 are primary members of PAF1C. It helps in the immune reaction. It encourages the appearance of	Knockout of PAF1 enhances the DENV infection. NS5 atomic localization and the C-terminal area of the methyl transferase field are compulsory for its	PAF1C plays a significant role in restricting DENV reproduction.	A study can be performed on PAF1C role in dengue virus replication.

					antiviral, antimicrobial, and inflammatory responses.	communication with PAF1C.		
9	2021	[12]	Viral and molecular factors on ZIKV entry	Flavivirus are arthropod-borne pathogens for a wide range of mortality.	Around seventy viruses are contained in the genus. Different viruses present different clinical presentations even though the genomic and structural similarity is seen.	New methods to control virus infection are searched in order to control the disease.	Different steps present in the arboviruses entry process are attachment, internalization, endosomal escape, and capsid uncoating. Cellular machinery, host factor, and cellular pathways are taken over by events.	These are required as possible approaches for developing broad-spectrum antiviral drugs.
10	2021	[13]	The search for pseudoknots and three-layered associations in the genome of ZIKV RNA	The Flavivirus RNA genome covers a number of functionally significant constructions in 30 untranslated regions.	The genome sequence has diversity, and its identification is difficult. Flavivirus replication is well understood with the prediction of structure and helps in the development of antiviral strategies.	An algorithm is developed for structure design examination in the RNA sequence. Secondary constructions, pseudoknots, and triple base interactions are also identified. The structural descriptor is constructed by means of the data on preserved flavivirus 30UTR constructions. It enclosed a variety of designs in these motifs.	The construction of a database of flavivirus 30UTR structures is done. A total of areasidenticalto the overall design of exoribonuclease Xrn1-resistant RNAs in the rising collection of bug-specific flaviviruses is recognized.	Diagnostics, genotyping, mutational study, and drug design can be done with the record of flavivirus RNA.
11	2022	[14]	The common protein - Binding Motif of ZIKV and Dengue is PEX 19	ZIKV capsid (C) protein regulates peroxisomes. It causes reduced production of interferons as it interacts with protein PEX19	The usage and interaction of C interaction with PEX19 in the Zika virus are studied. The importance and usage of peroxisomes are also studied.	Peroxisomes are required for the replication of DENV. A conserved PEX19 binding motif is used for DENV and Zika virus C binding. This motif commonly occurs in cellular peroxisomal membrane proteins (PMPs).	This paper find identifies and questions how flavi virus C might downregulate peroxisomal abundance	Study to be conducted on the unknown role of peroxisomes in ZIKV.
12	2022	[15]	RNA requisite proteins used for interior Initiation of mRNA conversion	Viruses utilize the host's protein synthesis mechanism for interpreting their mRNAs. Viruses are obligate intracellular parasites.	The virus-related mRNA (vRNA) contends with the host mRNA together with ribosomes, tRNAs, and the eukaryotic translation initiation factor (eIFs) pool.	Viruses targets host eIFs, and RNA essentials reprogram cellular gene expression, guaranteeing the favored transformation of mRNAs.	vRNA IRES facilitated conversion beginning is highlighted. The role of RNA-required proteins (RBPs), other than the established conversion initiation factors, in modifying their movement is also discussed.	Accepting how the PTMs of ITAFs impact viral IRES activity also arises as a striking study domain.
13	2022	[7]	Computational method to cure ZIKV infection by the screening of phytochemicals against NS5	ZIKV study provides stimulation for new and capable anti-ZIKV agents. RNA-dependent RNA polymerase (RdRp) is found to be significant among all. It is considered a valuable drug target.	The cellular dock approach is used to reasonably screen the files of 5000 phytochemicals to find inhibitors against NS5 RdRp.	Polydatin, Dihydrogenistin, Liquiritin, Rhapontin, and Cichoriinwere investigated with NS5 RdRp. The replicated complexes showed steadiness.	Natural and low price medicines are to be produced against the Zika virus	The efficiency of can be obtained by further study and research

14	2022	[8]	Tick-Borne Encephalitis Viral disease and the contribution of IFITM Proteins	A study is conducted on the Tick-borne encephalitis virus. The factors that inhibit the multiplication are Interferon-induced transmembrane proteins,	The part of IFITM proteins in the reticence of TBEV infection is studied. IFITM3 plays the important role of including analysis of useful motifs.	The role and additive action of endogenous IFITMs in TBEV dominance is confirmed.	TBEV might, to some extent, run off interferon and IFITM interceded containment throughout high-density co-culture infection.	Cell-to-cell reach may form an approach for viruses to break out from native host defenses.

From the detailed literature review, it is identified that different categories of algorithms are applied to find the motif in DNA sequence. The limitations found in the previous search are that the Greedy motif search without pseudocount and with pseudocount is not applied to find motifs in the Zika virus [16]. Heuristics algorithms worked for solving combinatorial, but bit did not work for large datasets [17]. The greedy motif search algorithm is proposed to find the motif in RNA sequences [18]. A greedy mixture learning technique is proposed for finding motifs in already known motifs in real proteins by using the PRINTS datasets [19]. Time series-based data of different lengths are aligned and joined using the Greedy search technique [20]. The Greedy search algorithm is used to discover motifs in hm03r, yst04r, and yst08r. The results show that the algorithm is effective in finding motifs in the DNA sequences of hm03r, yst04r, and yst08r [21]. DNA motif discovery is made using the Greedy motif search method over the datasets - GATA1, SOX2, OCT4, STAT3, and KLF1 [22]. So, a research gap is identified that the Greedy search technique for motif finding is not applied to the genome of the Zika virus.

III. Materials and Methods

A. Data Collection: Zika virus genome

ZIKV genome structure data from openly existing catalogs are collected and used in this study. The Zika virus genome sequence is available at NCBI [23]. ZikaVirus.fasta is the file name of the dataset. Fasta is a file format; the genome sequence is stored as a nucleotide sequence. The size of the dataset is 10.7 KB.

B. Gene Representations

RNA is formed using DNA. RNA further gets converted into proteins. Four ribonucleotides from the RNA. These four nucleotides are namely Adenine, Cytosine, Guanine, and Uracil. DNA replaces Thymine with Uracil. Amino acid sequences of proteins are formed by RNA transcripts [9]. Proteins regulate the function of the cell. Ori is the replication for the origin of DNA, so DNA replication starts at ori. Ori has some specific properties. Biologists find it difficult to identify the position of replication. Some other complicated tasks happening inside the cell are transcription and transpiration. The transcription process replaces what happens inside the cells. Thymine (T) that occurs in DNA is getting converted into Uracil (U) during the transcription process. The amino acids sequence is formed from RNA. There are a total of twenty different amino acids in RNA [13]. Three different nucleotides form these amino acids, also called 3-mers or codons. Each combination of these 3-mers forms different amino acids following a genetic code. Due to transcription, different genes can form RNA. Different genes may transcribe at different rates. This property is also called gene expression. Due to gene expression, different cell at different parts of the body of any living being behaves differently. Brain cells behave differently compared to skin cells. They differ in features and functionality. Cells with different variations know how to keep track of time. The variation in cell functionalities is known to occur in people infected with ZIKV. Pro-inflammatory reactions are prominent in women infected with the Zika virus.

IV. Computation of Motif in Zika Virus Genome

Zika virus genome nucleotide sequence has a length of 10780. The length is calculated using the program. Based on the profile matrix of the Zika virus, the probability of a string can be calculated. The regulatory motif binds to specific short DNA. It regulates the gene. The site of binding is generally the upstream region and is important to identify. A method to identify the motif is useful for gene study.

A. Importance of Motif

Motifs are important to identify and study. Motifs have finite lengths. These are short sequences in DNA. Sequence motifs are used to signify transcription factor binding sites. Transcription regulations are better assumed with motif sequences. Dynamic sites of enzymes and proteins are characterized by motifs. The individual instances of the motif is calculated and scored using the ideal motif. An ideal motif is not known and can only be predicted. To recognize motif, a k-mer string is selected from each string. Based on identical nucleotides, each motif is scored. A list of t strings is created. The length of the string in each list is n. a motif collection is created by selecting k-mer nucleotides from each string. A t X k motif matrix is formed. From the t X k matrix created, the nucleotides are counted and stored in an array. There are four different types of nucleotides, so four rows are created. The first row represents nucleotide A, the second row represents C, the third row represents G, and the fourth row represents T. Now in the matrix, the columns are viewed to find the nucleotide with the highest count. So for that column, the nucleotide with the highest count is represented in the uppercase letter. Different motif matrices for DNA strings if formed using different values of k. The aim is to obtain the most conserved motif matrix. The conserved matrix also means the matrix has more capital letters. The minimum score is to be obtained for the collection of k-mers.

B. Calculating the Count Matrix

A count matrix is formed for the Motifs. It is 4 X k matrix. It is abbreviated as count (Motifs). It is the sum of each nucleotide column-wise. The element (I, j) represents I nucleotide in jth column. The count matrix obtained is further used for the calculation in the next steps [4].

C. Finding the Count Matrix with Pseudocount

Pseudocount is a small number that is added to zeroes. This improves the unfair scoring. This method is named Laplace's Rule of Succession. In motifs, pseudocount method, one or a small number is added to the count matrix. The different matrices are formed for calculations. These are the motif, count, and profile matrices. A count matrix of 4 X k is formed for a given matrix. This count matrix's (I, j) element represents nucleotide I of column j. Pseudocounts one is added to each element of this count matrix.

{'A': [46, 38, 40, 42, 43, 37, 34, 47, 46, 51, 42, 42, 43, 47, 39, 46, 47, 37, 40, 50, 39, 43, 41, 46, 37, 40, 43, 42, 41, 31, 37, 41, 37, 45, 41, 47, 37, 46, 41, 49, 38, 47, 42, 42, 48, 41, 45, 52, 39, 48, 42, 47, 30, 50, 42, 52, 40, 41, 40, 47, 34, 51, 53, 37, 51, 41, 46, 59, 48, 51],

'C': [24, 40, 35, 30, 25, 35, 38, 34, 37, 31, 28, 35, 24, 36, 42, 39, 35, 44, 37, 33, 47, 32, 38, 35, 36, 43, 39, 29, 41, 34, 39, 36, 44, 38, 35, 34, 33, 42, 27, 36, 37, 37, 32, 37, 42, 35, 40, 25, 33, 34, 33, 30, 50, 43, 32, 25, 36, 34, 33, 34, 43, 27, 27, 31, 33, 37, 35, 27, 34, 32],

'G': [42, 54, 47, 46, 50, 50, 43, 36, 47, 46, 49, 48, 58, 39, 46, 51, 39, 51, 44, 43, 34, 50, 41, 43, 57, 44, 42, 47, 42, 46, 49, 49, 40, 38, 42, 46, 43, 39, 58, 43, 48, 46, 51, 48, 42, 47, 39, 44, 57, 50, 50, 53, 51, 36, 49, 42, 41, 47, 48, 43, 53, 43, 41, 51, 49, 49, 49, 36, 46, 39],

'T': [46, 26, 36, 40, 40, 36, 43, 41, 28, 30, 39, 33, 33, 36, 31, 22, 37, 26, 37, 32, 38, 33, 38, 34, 28, 31, 34, 40, 34, 47, 33, 32, 37, 37, 40, 31, 45, 31, 32, 30, 35, 28, 33, 31, 26, 35, 34, 37, 29, 26, 33, 28, 27, 29, 35, 39, 41, 36, 37, 34, 28, 37, 37, 39, 25, 31, 28, 36, 30, 36]}

D. Framing the Profile Matrices with Pseudocount

Finding a perfect motif is a challenging task. It binds the finest to the transcription site. A motif discovery problem is solved by finding the similarity to the ideal motif. A motif similar to an ideal motif is calculated, as an ideal motif is not known. Capital letters are used to denote the most common nucleotide in each column. Motif[i][j] represent the ith row and jth column. If the matrix has more capital letters, it means the matrix is more conserved. The genome of the Zika virus is used to create different motif matrices for different values on k. The count of the small letter is noted. The sum of this count gives the score of the matrix. Then a k-mer is assumed, which will reduce the value of the score. The profile of the motif matrix is calculated. The elements of the count are divided by the number of rows. The sum of any column of the profile matrix is unity. The profile matrix for the Zika virus genome is given below.

{'A': [0.29, 0.24, 0.25, 0.27, 0.27, 0.23, 0.22, 0.3, 0.29, 0.32, 0.27, 0.27, 0.27, 0.3, 0.25, 0.29, 0.3, 0.23, 0.25, 0.32, 0.25, 0.27, 0.26, 0.29, 0.23, 0.25, 0.27, 0.27, 0.26, 0.2, 0.23, 0.26, 0.23, 0.28, 0.26, 0.3, 0.23, 0.29, 0.26, 0.31, 0.24, 0.3, 0.27, 0.27, 0.3, 0.26, 0.28, 0.33, 0.25, 0.3, 0.27, 0.3, 0.19, 0.32, 0.27, 0.33, 0.25, 0.26, 0.25, 0.3, 0.22, 0.32, 0.34, 0.23, 0.32, 0.26, 0.29, 0.37, 0.3, 0.32],

'C': [0.15, 0.25, 0.22, 0.19, 0.16, 0.22, 0.24, 0.22, 0.23, 0.2, 0.18, 0.22, 0.15, 0.23, 0.27, 0.25, 0.22, 0.28, 0.23, 0.21, 0.3, 0.2, 0.24, 0.22, 0.23, 0.27, 0.25, 0.18, 0.26, 0.22, 0.25, 0.23, 0.28, 0.24, 0.22, 0.22, 0.21, 0.27, 0.17, 0.23, 0.23, 0.23, 0.2, 0.23, 0.27, 0.22, 0.25, 0.16, 0.21, 0.22, 0.21, 0.19, 0.32, 0.27, 0.2, 0.16, 0.23, 0.22, 0.21, 0.22, 0.27, 0.17, 0.17, 0.2, 0.21, 0.23, 0.22, 0.17, 0.22, 0.2],

'G': [0.27, 0.34, 0.3, 0.29, 0.32, 0.32, 0.27, 0.23, 0.3, 0.29, 0.31, 0.3, 0.37, 0.25, 0.29, 0.32, 0.25, 0.32, 0.28, 0.27, 0.22, 0.32, 0.26, 0.27, 0.36, 0.28, 0.27, 0.3, 0.27, 0.29, 0.31, 0.31, 0.25, 0.24, 0.27, 0.29, 0.27, 0.25, 0.37, 0.27, 0.3, 0.29, 0.32, 0.3, 0.27, 0.3, 0.25, 0.28, 0.36, 0.32, 0.32, 0.34, 0.32, 0.23, 0.31, 0.27, 0.26, 0.3, 0.3, 0.27, 0.34, 0.27, 0.26, 0.32, 0.31, 0.31, 0.31, 0.23, 0.29, 0.25],

'T': [0.29, 0.16, 0.23, 0.25, 0.25, 0.23, 0.27, 0.26, 0.18, 0.19, 0.25, 0.21, 0.21, 0.23, 0.2, 0.14, 0.23, 0.16, 0.23, 0.2, 0.24, 0.21, 0.24, 0.22, 0.18, 0.2, 0.22, 0.25, 0.22, 0.3, 0.21, 0.2, 0.23, 0.23, 0.25, 0.2, 0.28, 0.2, 0.2, 0.19, 0.22, 0.18, 0.21, 0.2, 0.16, 0.22, 0.22, 0.23, 0.18, 0.16, 0.21, 0.18, 0.17, 0.18, 0.22, 0.25, 0.26, 0.23, 0.23, 0.22, 0.18, 0.23, 0.23, 0.25, 0.16, 0.2, 0.18, 0.23, 0.19, 0.23]}

E. Calculating a Consensus String for the Genome of the Zika Virus and Score

The consensus string of the Zika virus is calculated from the genome and is given below.

AGGGGGGAGAGGGAGGAGGACGAAGGAGGTGGCAGATAGAGAGGAGAAGGGGGAGAGGGAGAAGAGGAAA. The score of the consensus string is calculated to be 7444.

V. Discussion of Results

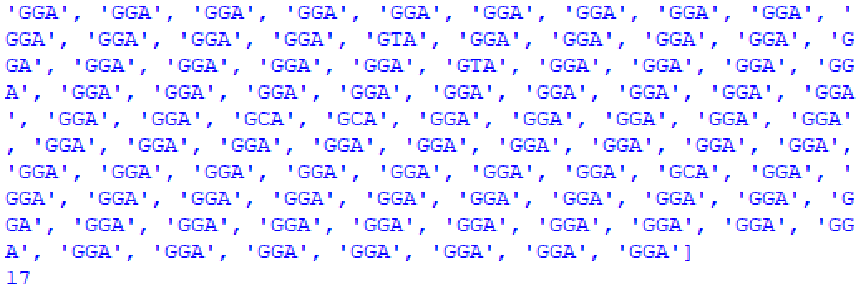
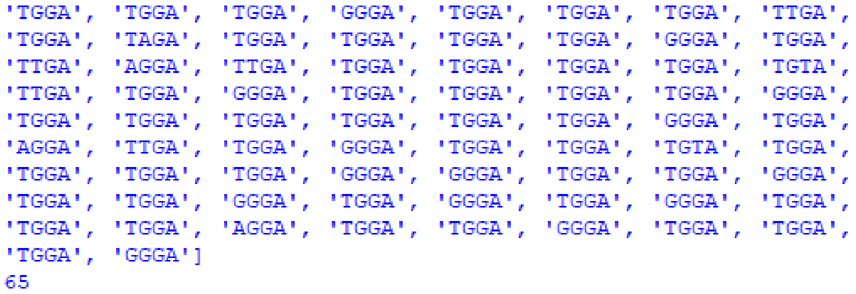
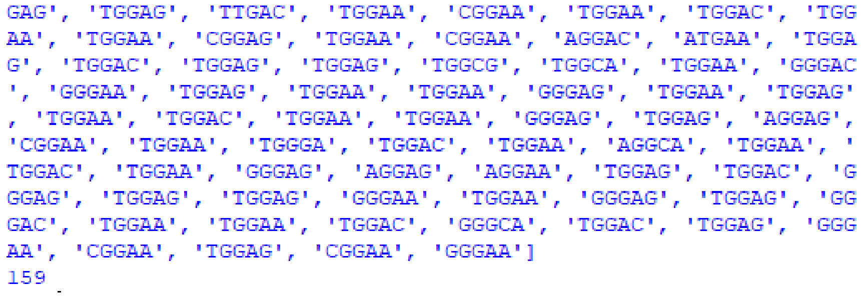
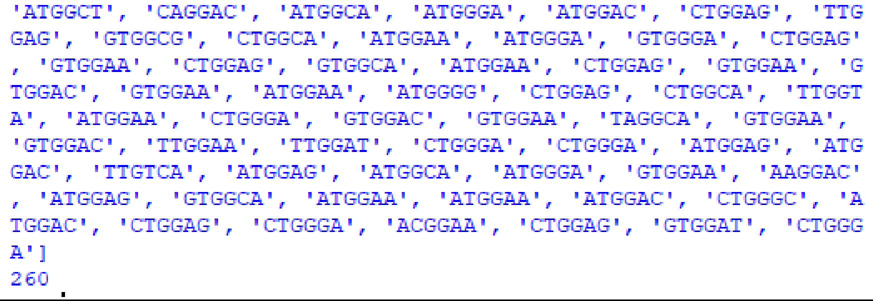
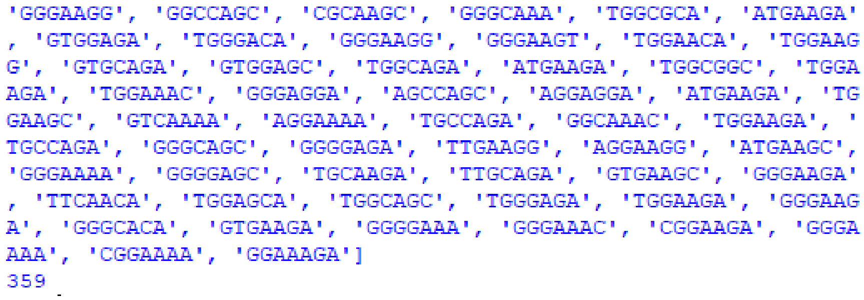
A. Using Profile Matrix for Calculation

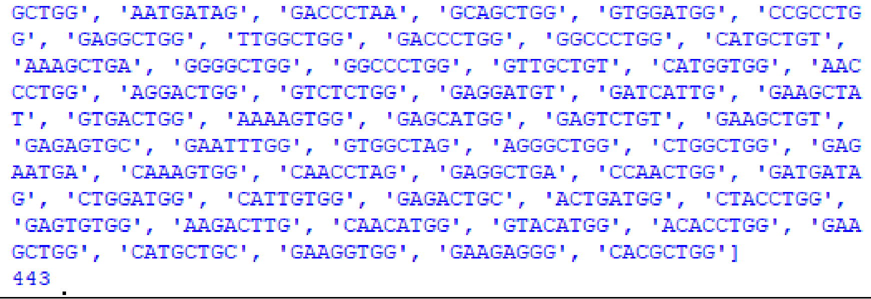
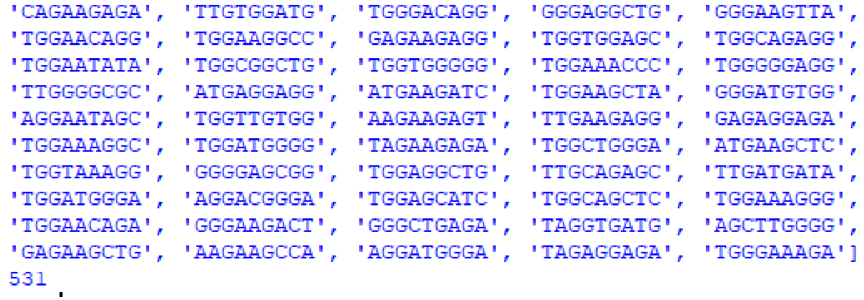
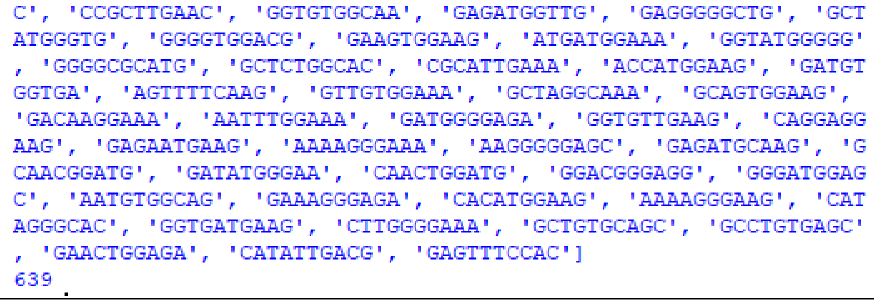
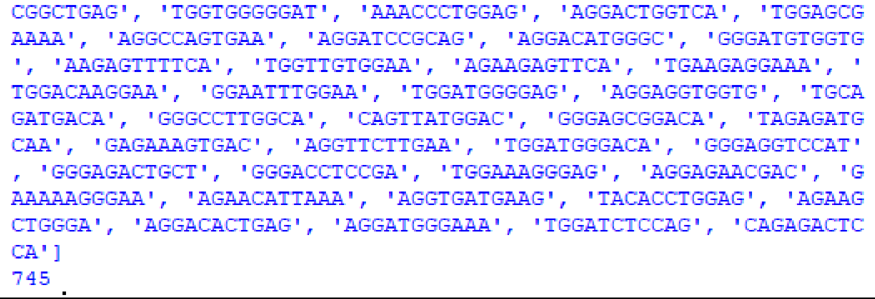
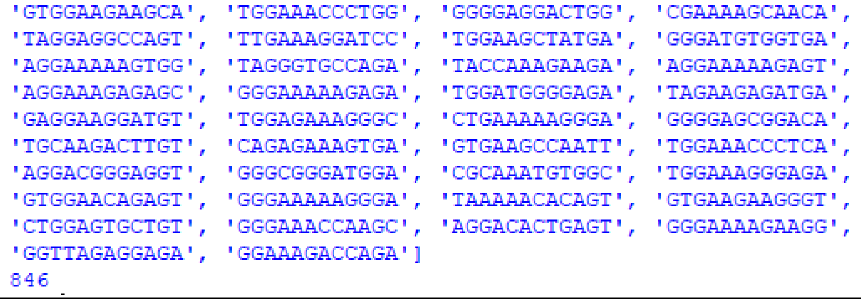
Iterative algorithms use select different alternatives during each iteration. The greedy search technique selects the most attractive alternative in each iteration. The profile matrix of the Zika virus is calculated in the preceding section and is used the same. The likelihood of any string can be intended. The probability of the consensus string is also calculated.

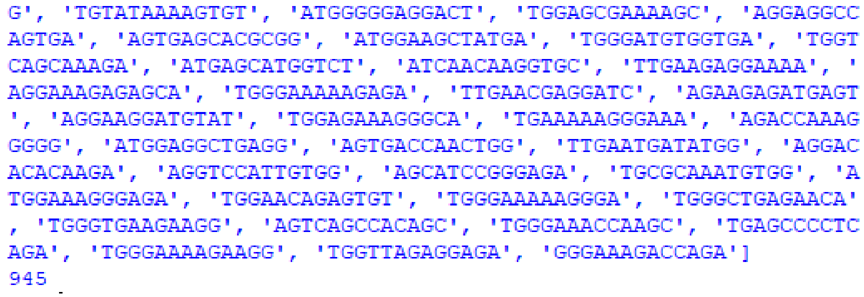
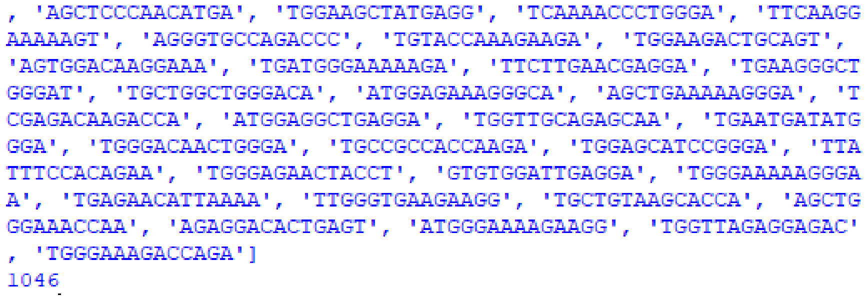
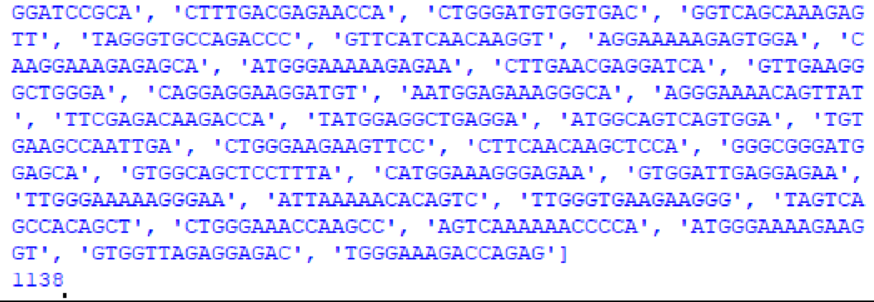
B. The Search of Binding Sites

The Greedy motif algorithm is implemented with pseudocount. The results of the Greedy motif search with pseudocount are summarized in Table II. The table contains values from 3-mer strings to 15-mer strings. The string with one nucleotide and two nucleotides have little significance and are hence ignored.

TABLE II. RESULTS TABULATED FOR GREEDY MOTIF SEARCH WITHOUT PSEUDOCOUNT AND WITH PSEUDOCOUNT FOR DIFFERENT K-MER STRINGS

Sr. No.	k-mer string	The score of the Greedy motif search without pseudocount search	The score of the Greedy motif with pseudocount search	Snapshot of the output obtained k-mer string
1	3-mer	10	17	 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', ' GGA', 'GGA', 'GGA', 'GGA', 'GTA', 'GGA', 'GGA', 'GGA', 'GGA', 'G GA', 'GGA', 'GGA', 'GGA', 'GGA', 'GTA', 'GGA', 'GGA', 'GGA', 'GG A', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA ', 'GGA', 'GGA', 'GCA', 'GCA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA' , 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GCA', 'GGA', ' GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'G GA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GG A', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA', 'GGA'] 17
2	4-mer	85	65	 'TGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'TGGA', 'TTGA', 'TGGA', 'TAGA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'TTGA', 'AGGA', 'TTGA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'TGTA', 'TTGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'AGGA', 'TTGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'TGTA', 'TGGA', 'TGGA', 'TGGA', 'TGGA', 'GGGA', 'GGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'GGGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'TGGA', 'AGGA', 'TGGA', 'TGGA', 'GGGA', 'TGGA', 'TGGA', 'TGGA', 'GGGA'] 65
3	5-mer	166	159	 GAG', 'TGGAG', 'TTGAC', 'TGGAA', 'CGGAA', 'TGGAA', 'TGGAC', 'TGG AA', 'TGGAA', 'CGGAG', 'TGGAA', 'CGGAA', 'AGGAC', 'ATGAA', 'TGGA G', 'TGGAC', 'TGGAG', 'TGGAG', 'TGGCG', 'TGGCA', 'TGGAA', 'GGGAC ', 'GGGAA', 'TGGAG', 'TGGAA', 'TGGAA', 'GGGAG', 'TGGAA', 'TGGAG' , 'TGGAA', 'TGGAC', 'TGGAA', 'TGGAA', 'GGGAG', 'TGGAG', 'AGGAG', 'CGGAA', 'TGGAA', 'TGGGA', 'TGGAC', 'TGGAA', 'AGGCA', 'TGGAA', ' TGGAC', 'TGGAA', 'GGGAG', 'AGGAG', 'AGGAA', 'TGGAG', 'TGGAC', 'G GGAG', 'TGGAG', 'TGGAG', 'GGGAA', 'TGGAA', 'GGGAG', 'TGGAG', 'GG GAC', 'TGGAA', 'TGGAA', 'TGGAC', 'GGGCA', 'TGGAC', 'TGGAG', 'GGG AA', 'CGGAA', 'TGGAG', 'CGGAA', 'GGGAA'] 159
4	6-mer	288	260	 'ATGGCT', 'CAGGAC', 'ATGGCA', 'ATGGGA', 'ATGGAC', 'CTGGAG', 'TTG GAG', 'GTGGCG', 'CTGGCA', 'ATGGAA', 'ATGGGA', 'GTGGGA', 'CTGGAG' , 'GTGGAA', 'CTGGAG', 'GTGGCA', 'ATGGAA', 'CTGGAG', 'GTGGAA', 'G TGGAC', 'GTGGAA', 'ATGGAA', 'ATGGGG', 'CTGGAG', 'CTGGCA', 'TTGGT A', 'ATGGAA', 'CTGGGA', 'GTGGAC', 'GTGGAA', 'TAGGCA', 'GTGGAA', 'GTGGAC', 'TTGGAA', 'TTGGAT', 'CTGGGA', 'CTGGGA', 'ATGGAG', 'ATG GAC', 'TTGTCA', 'ATGGAG', 'ATGGCA', 'ATGGGA', 'GTGGAA', 'AAGGAC' , 'ATGGAG', 'GTGGCA', 'ATGGAA', 'ATGGAA', 'ATGGAC', 'CTGGGC', 'A TGGAC', 'CTGGAG', 'CTGGGA', 'ACGGAA', 'CTGGAG', 'GTGGAT', 'CTGGG A'] 260
5	7-mer	412	359	 'GGGAAGG', 'GGCCAGC', 'CGCAAGC', 'GGGCAAA', 'TGGCGCA', 'ATGAAGA' , 'GTGGAGA', 'TGGGACA', 'GGGAAGG', 'GGGAAGT', 'TGGAACA', 'TGGAAG G', 'GTGCAGA', 'GTGGAGC', 'TGGCAGA', 'ATGAAGA', 'TGGCGGC', 'TGGA AGA', 'TGGAAAC', 'GGGAGGA', 'AGCCAGC', 'AGGAGGA', 'ATGAAGA', 'TG GAAGC', 'GTCAAAA', 'AGGAAAA', 'TGCCAGA', 'GGCAAAC', 'TGGAAGA', ' TGCCAGA', 'GGGCAGC', 'GGGGAGA', 'TTGAAGG', 'AGGAAGG', 'ATGAAGC', 'GGGAAAA', 'GGGGAGC', 'TGCAAGA', 'TTGCAGA', 'GTGAAGC', 'GGGAAGA' , 'TTCAACA', 'TGGAGCA', 'TGGCAGC', 'TGGGAGA', 'TGGAAGA', 'GGGAAG A', 'GGGCACA', 'GTGAAGA', 'GGGGAAA', 'GGGAAAC', 'CGGAAGA', 'GGGA AAA', 'CGGAAAA', 'GGAAAGA'] 359

6	8-mer	508	443	 GCTGG', 'AATGATAG', 'GACCCTAA', 'GCAGCTGG', 'GTGGATGG', 'CCGCCTG G', 'GAGGCTGG', 'TTGGCTGG', 'GACCCTGG', 'GGCCCTGG', 'CATGCTGT', 'AAAGCTGA', 'GGGGCTGG', 'GGCCCTGG', 'GTTGCTGT', 'CATGGTGG', 'AAC CCTGG', 'AGGACTGG', 'GTCTCTGG', 'GAGGATGT', 'GATCATTG', 'GAAGCTA T', 'GTGACTGG', 'AAAAGTGG', 'GAGCATGG', 'GAGTCTGT', 'GAAGCTGT', 'GAGAGTGC', 'GAATTTGG', 'GTGGCTAG', 'AGGGCTGG', 'CTGGCTGG', 'GAG AATGA', 'CAAAGTGG', 'CAACCTAG', 'GAGGCTGA', 'CCAACTGG', 'GATGATA G', 'CTGGATGG', 'CATTGTGG', 'GAGACTGC', 'ACTGATGG', 'CTACCTGG', 'GAGTGTGG', 'AAGACTTG', 'CAACATGG', 'GTACATGG', 'ACACCTGG', 'GAA GCTGG', 'CATGCTGC', 'GAAGGTGG', 'GAAGAGGG', 'CACGCTGG'] 443
7	9-mer	603	531	 'CAGAAGAGA', 'TTGTGGATG', 'TGGGACAGG', 'GGGAGGCTG', 'GGGAAGTTA', 'TGGAACAGG', 'TGGAAGGCC', 'GAGAAGAGG', 'TGGTGGAGC', 'TGGCAGAGG', 'TGGAATATA', 'TGGCGGCTG', 'TGGTGGGGG', 'TGGAAACCC', 'TGGGGGAGG', 'TTGGGGCGC', 'ATGAGGAGG', 'ATGAAGATC', 'TGGAAGCTA', 'GGGATGTGG', 'AGGAATAGC', 'TGGTTGTGG', 'AAGAAGAGT', 'TTGAAGAGG', 'GAGAGGAGA', 'TGGAAAGGC', 'TGGATGGGG', 'TAGAAGAGA', 'TGGCTGGGA', 'ATGAAGCTC', 'TGGTAAAGG', 'GGGGAGCGG', 'TGGAGGCTG', 'TTGCAGAGC', 'TTGATGATA', 'TGGATGGGA', 'AGGACGGGA', 'TGGAGCATC', 'TGGCAGCTC', 'TGGAAAGGG', 'TGGAACAGA', 'GGGAAGACT', 'GGGCTGAGA', 'TAGGTGATG', 'AGCTTGGGG', 'GAGAAGCTG', 'AAGAAGCCA', 'AGGATGGGA', 'TAGAGGAGA', 'TGGGAAAGA'] 531
8	10-mer	697	639	 C', 'CCGCTTGAAC', 'GGTGTGGCAA', 'GAGATGGTTG', 'GAGGGGGCTG', 'GCT ATGGGTG', 'GGGGTGGACG', 'GAAGTGGAAG', 'ATGATGGAAA', 'GGTATGGGGG' , 'GGGGCGCATG', 'GCTCTGGCAC', 'CGCATTGAAA', 'ACCATGGAAG', 'GATGT GGTGA', 'AGTTTTCAAG', 'GTTGTGGAAA', 'GCTAGGCAAA', 'GCAGTGGAAG', 'GACAAGGAAA', 'AATTTGGAAA', 'GATGGGGAGA', 'GGTGTTGAAG', 'CAGGAGG AAG', 'GAGAATGAAG', 'AAAAGGGAAA', 'AAGGGGGAGC', 'GAGATGCAAG', 'G CAACGGATG', 'GATATGGGAA', 'CAACTGGATG', 'GGACGGGAGG', 'GGGATGGAG C', 'AATGTGGCAG', 'GAAAGGGAGA', 'CACATGGAAG', 'AAAAGGGAAG', 'CAT AGGGCAC', 'GGTGATGAAG', 'CTTGGGGAAA', 'GCTGTGCAGC', 'GCCTGTGAGC' , 'GAACTGGAGA', 'CATATTGACG', 'GAGTTTCCAC'] 639
9	11-mer	806	745	 CGGCTGAG', 'TGGTGGGGGAT', 'AAACCCTGGAG', 'AGGACTGGTCA', 'TGGAGCG AAAA', 'AGGCCAGTGAA', 'AGGATCCGCAG', 'AGGACATGGGC', 'GGGATGTGGTG ', 'AAGAGTTTTCA', 'TGGTTGTGGAA', 'AGAAGAGTTCA', 'TGAAGAGGAAA', ' TGGACAAGGAA', 'GGAATTTGGAA', 'TGGATGGGGAG', 'AGGAGGTGGTG', 'TGCA GATGACA', 'GGGCCTTGGCA', 'CAGTTATGGAC', 'GGGAGCGGACA', 'TAGAGATG CAA', 'GAGAAAGTGAC', 'AGGTTCTTGAA', 'TGGATGGGACA', 'GGGAGGTCCAT' , 'GGGAGACTGCT', 'GGGACCTCCGA', 'TGGAAAGGGAG', 'AGGAGAACGAC', 'G AAAAAGGGAA', 'AGAACATTAAA', 'AGGTGATGAAG', 'TACACCTGGAG', 'AGAAG CTGGGA', 'AGGACACTGAG', 'AGGATGGGAAA', 'TGGATCTCCAG', 'CAGAGACTC CA'] 745
10	12-mer	922	846	 'GTGGAAGAAGCA', 'TGGAAACCCTGG', 'GGGGAGGACTGG', 'CGAAAAGCAACA', 'TAGGAGGCCAGT', 'TTGAAAGGATCC', 'TGGAAGCTATGA', 'GGGATGTGGTGA', 'AGGAAAAAGTGG', 'TAGGGTGCCAGA', 'TACCAAAGAAGA', 'AGGAAAAAGAGT', 'AGGAAAGAGAGC', 'GGGAAAAAGAGA', 'TGGATGGGGAGA', 'TAGAAGAGATGA', 'GAGGAAGGATGT', 'TGGAGAAAGGGC', 'CTGAAAAAGGGA', 'GGGGAGCGGACA', 'TGCAAGACTTGT', 'CAGAGAAAGTGA', 'GTGAAGCCAATT', 'TGGAAACCCTCA', 'AGGACGGGAGGT', 'GGGCGGGATGGA', 'CGCAAATGTGGC', 'TGGAAAGGGAGA', 'GTGGAACAGAGT', 'GGGAAAAAGGGA', 'TAAAAACACAGT', 'GTGAAGAAGGGT', 'CTGGAGTGCTGT', 'GGGAAACCAAGC', 'AGGACACTGAGT', 'GGGAAAAGAAGG', 'GGTTAGAGGAGA', 'GGAAAGACCAGA'] 846

11	13-mer	1008	945	 G', 'TGTATAAAAGTGT', 'ATGGGGGAGGACT', 'TGGAGCGAAAAGC', 'AGGAGGCC AGTGA', 'AGTGAGCACGCGG', 'ATGGAAGCTATGA', 'TGGGATGTGGTGA', 'TGGT CAGCAAAGA', 'ATGAGCATGGTCT', 'ATCAACAAGGTGC', 'TTGAAGAGGAAAA', ' AGGAAAGAGAGCA', 'TGGGAAAAAGAGA', 'TTGAACGAGGATC', 'AGAAGAGATGAGT ', 'AGGAAGGATGTAT', 'TGGAGAAAGGGCA', 'TGAAAAAGGGAAA', 'AGACCAAAG GGGG', 'ATGGAGGCTGAGG', 'AGTGACCAACTGG', 'TTGAATGATATGG', 'AGGAC ACACAAGA', 'AGGTCCATTGTGG', 'AGCATCCGGGAGA', 'TGCGCAAATGTGG', 'A TGGAAAGGGAGA', 'TGGAACAGAGTGT', 'TGGGAAAAAGGGA', 'TGGGCTGAGAACA' , 'TGGGTGAAGAAGG', 'AGTCAGCCACAGC', 'TGGGAAACCAAGC', 'TGAGCCCCTC AGA', 'TGGGAAAAGAAGG', 'TGGTTAGAGGAGA', 'GGGAAAGACCAGA'] 945
12	14-mer	1113	1046	 , 'AGCTCCCAACATGA', 'TGGAAGCTATGAGG', 'TCAAAACCCTGGGA', 'TTCAAGG AAAAAGT', 'AGGGTGCCAGACCC', 'TGTACCAAAGAAGA', 'TGGAAGACTGCAGT', 'AGTGGACAAGGAAA', 'TGATGGGAAAAAGA', 'TTCTTGAACGAGGA', 'TGAAGGGCT GGGAT', 'TGCTGGCTGGGACA', 'ATGGAGAAAGGGCA', 'AGCTGAAAAAGGGA', 'T CGAGACAAGACCA', 'ATGGAGGCTGAGGA', 'TGGTTGCAGAGCAA', 'TGAATGATATG GGA', 'TGGGACAACTGGGA', 'TGCCGCCACCAAGA', 'TGGAGCATCCGGGA', 'TTA TTTCCACAGAA', 'TGGGAGAACTACCT', 'GTGTGGATTGAGGA', 'TGGGAAAAAGGGA A', 'TGAGAACATTAAAA', 'TTGGGTGAAGAAGG', 'TGCTGTAAGCACCA', 'AGCTG GGAAACCAA', 'AGAGGACACTGAGT', 'ATGGGAAAAGAAGG', 'TGGTTAGAGGAGAC' , 'TGGGAAAGACCAGA'] 1046
13	15-mer	1190	1138	 GGATCCGCA', 'CTTTGACGAGAACCA', 'CTGGGATGTGGTGAC', 'GGTCAGCAAAGAG TT', 'TAGGGTGCCAGACCC', 'GTTCATCAACAAGGT', 'AGGAAAAAGAGTGGA', 'C AAGGAAAGAGAGCA', 'ATGGGAAAAAGAGAA', 'CTTGAACGAGGATCA', 'GTTGAAGG GCTGGGA', 'CAGGAGGAAGGATGT', 'AATGGAGAAAGGGCA', 'AGGGAAAACAGTTAT ', 'TTCGAGACAAGACCA', 'TATGGAGGCTGAGGA', 'ATGGCAGTCAGTGGA', 'TGT GAAGCCAATTGA', 'CTGGGAAGAAGTTCC', 'CTTCAACAAGCTCCA', 'GGGCGGGATG GAGCA', 'GTGGCAGCTCCTTTA', 'CATGGAAAGGGAGAA', 'GTGGATTGAGGAGAA', 'TTGGGAAAAAGGGAA', 'ATTAAAAACACAGTC', 'TTGGGTGAAGAAGGG', 'TAGTCA GCCACAGCT', 'CTGGGAAACCAAGCC', 'AGTCAAAAAACCCCA', 'ATGGGAAAAGAAG GT', 'GTGGTTAGAGGAGAC', 'TGGGAAAGACCAGAG'] 1138

The results of various k-mers are tabulated. The score of the Greedy motif search without pseudocount is compared with the score of the Greedy motif with pseudocount. The score of the Greedy motif search without pseudocount and with pseudocount is calculated for string of length three to sting of length fifteen. It is found that the score is less for calculations with pseudocount so these results are more promising compared to motifs obtained without pseudocount. For 15-mer string, the score is 1190 for the Greedy motif search and 1138 for the Greedy motif search with pseudocount. If the score is low means, the performance of the algorithm is good. For the fifteen nucleotides long string, a score of 1190 is obtained for the Greedy motif search without pseudocount. A score of 1138 is obtained for the Greedy motif search with pseudocount. So, the results have considerably improved by using the Greedy motif search with pseudocount.

VI. Conclusion and Future Work

The genome of ZIKV is considered for the study. In the case of an infected mother, ZIKV causes neurological disorders in babies. The causes and effects of ZIKV were not identified earlier as the symptoms of the disease are mild headache and fever. Later it was linked to reduced brain activities in babies. The diseases pass from the infected mother to the child. Medicines or vaccines for this infection are not available. This research paper studies the Zika virus genome to get more insight into the molecular structure. For a given profile matrix, the probability of every k-mer string is calculated and tabulated. The score is calculated with the Greedy motif search without pseudocount and with pseudocount. The results are computed and tabulated. The comparison shows the results are improved with the Greedy motif search with pseudocount. The aim is to reduce the score, and it is obtained with a Greedy motif search with pseudocount. The Greedy motif search for motif finding is applied to PRINTS datasets, hm03r, yst04r, and yst08r, in earlier research. It is also applied to datasets GATA1, SOX2, OCT4, STAT3, and KLF1. It is not applied to the Zika virus to identify the motifs in the Zika virus genome. It is concluded that the Greedy motif search with pseudocount performs better than the Greedy motif search without pseudocount as it gives a score of 1138 over a score of 1190.

References

[1] P. S. Mahapatro and J. R. Saini, "An Innovative Computer Programming based Analysis of Zika Virus for Identification of Genome Replication Location," IEEEXplorer, 2021.

[2] Q. Ye, Z.-Y. Liu, J.-F. Han, T. Jiang, X.-F. Li and C.-F. Qin, "Genomic characterization and phylogenetic analysis of Zika virus circulating in the Americas," Infection, Genetics and Evolution, vol. 43, no. 43, 2016.

[3] M. Ramanathan, K. Majzoub, D. Rao, P. Neela, B. Zarnegar, S. Mondal, J. Roth, H. Gai, J. Kovalski, Z. Siprashvili, T. Palmer, J. Carette and P. Khavari, "RN A-protein interaction detection in living cells," Nature Methods, vol. 15, no. 3, 2018.

[4] P. S. Mahapatro and J. R. Saini, "Genetic Behaviour of Zika Virus and Identification of Motif," International Journal of Advanced Computer Science and Applications, vol. 12, no. 9, 2021.

[5] T. Hermann, "Viral RNA targets and their small molecule ligands," Topics in Medicinal Chemistry, vol. 27, 2018.

[6] C. Rieder, J. Rieder, S. Sannajust, D. Goode, R. Geguchadze, R. Relich, D. Molliver, T. King, J. Vaughn and M. May, "A novel mechanism for zika virus host-cell binding," Viruses, vol. 11, no. 12, 2019.

[7] A. Rehman, U. Ashfaq, M. Javed, F. Shahid, F. Noor and S. Aslam, "The Screening of Phytochemicals Against NS5 Polymerase to Treat Zika Virus Infection: Integrated Computational Based Approach," Combinatorial Chemistry and High Throughput Screening, vol. 25, no. 4, 2022.

[8] A. Chmielewska, M. Gómez-Herranz, P. Gach, M. Nekulova, M. Bagnucka, A. Lipińska, M. Rychłowski, W. Hoffmann, E. Król, B. Vojtesek, R. Sloan, K. Bieńkowska-Szewczyk, T. Hupp and K. Ball, "The Role of IFITM Proteins in Tick-Borne Encephalitis Virus Infection," Journal of Virology, vol. 96, no. 1, 2022.

[9] T. Hu, Z. Wu, S. Wu, S. Chen and A. Cheng, "The key amino acids of E protein involved in early flavivirus infection: viral entry," Virology Journal, vol. 18, no. 1, 2021.

[10] R. Delli Ponti and M. Mutwil, "Structural landscape of the complete genomes of dengue virus serotypes and other viral hemorrhagic fevers," BMC Genomics, vol. 22, no. 1, 2021.

[11] M. Petit, M. Kenaston, O. Pham, A. Nagainis, A. Fishburn and P. Shah, "Nuclear dengue virus NS5 antagonizes expression of PAF1-dependent immune response genes," PLoS Pathogens, vol. 17, no. 11, 2021.

[12] C. Cordero-Rivera, L. De Jesús-González, J. Osuna-Ramos, S. Palacios-Rápalo, C. Farfan-Morales, J. Reyes-Ruiz and R. Del Ángel, "The importance of viral and cellular factors on flavivirus entry," Current Opinion in Virology, vol. 49, 2021.

[13] A. Zammit, L. Helwerda, R. Olsthoorn, F. Verbeek and A. Gultyaev, "A database of flavivirus RNA structures with a search algorithm for pseudoknots and triple base interactions," Bioinformatics, vol. 37, no. 7, 2021.

[14] M. Farelo, D. Korrou-Karava, B. K. T. Russell, M. Maringer and P. Mayerhofer, "Dengue and Zika Virus Capsid Proteins Contain a CommonPEX19-Binding Motif," Viruses, vol. 14, no. 253, 2022.

[15] B. López-Ulloa, Y. Fuentes, M. S. P. Ortega and M. López-Lastra, "RNA-Binding Proteins as Regulators of Internal Initiation ofViral mRNA Translation," Viruses, vol. 14, no. 188, 2022.

[16] T. Wang and G. D. Stormo, "Combining phylogenetic data with co-regulated genes to identify regulatory motifs," BIOINFORMATICS, vol. 19, no. 18, pp. 2369-2380, 2003.

[17] J. M. C. Garbelini, D. S. Sanches and A. T. R. Pozo, "Expectation Maximization based algorithm applied to DNA sequence motif finder," IEEE Congress on Evolutionary Computation, pp. 1-8, 2022.

[18] J. Gorodkin, L. J. Heyer and G. D. Stormo, "Finding the most significant common sequence and structure motifs in a set of RNA sequences," Nucleic Acids Research, vol. 25, no. 18, p. 3724–3732, 1997.

[19] K. Blekas, D. Fotiadis and A. Likas, "Greedy mixture learning for multiple motif discovery in biological sequences," BIOINFORMATICS, vol. 19, no. 5, pp. 607-617, 2003.

[20] M. Mollah, V. Souza and A. Mueen, "Multi-way Time Series Join on Multi-length Patterns," in IEEE International Conference on Data Mining, 2021.

[21] O. Gokalp, "DNA sequence motif discovery using greedy construction algorithm based techniques," in 5th International Conference on Computer Science and Engineering, UBMK 2020, 2020.

[22] C. Saad, L. Noé, H. Richard, J. Leclerc, M.-P. Buisine, H. Touzet and M. Figeac, "DiNAMO: highly sensitive DNA motifdiscovery in high-throughput sequencing data," BMC Bioinformatics, vol. 19, p. 223, 2018.

[23] National Center for Biotechnology Information, [Online]. Available: https://www.ncbi.nlm.nih.gov/. [Accessed 18 Feb 2022].