International Journal of Research in Advent Technology, Vol.7, No.5S, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Analysis of Protein Formation on Dengue Virus Gene Sequences Using Context Free Grammars

R. Vengatesh Kumar*[1], T. Marimuthu[2], R. Lawrance[3]
[1]*Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India.*
[2,3]*Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India.*
e-mail: vengatesh.kumar@gmail.com[1], mastersvksmca@gmail.com[2] , lawrancer@yahoo.com[3]

Abstract -The formation of protein is analyzed on dengue virus gene sequences by using context free grammars along with the novel periodicity mining algorithm. The process of protein formation plays a vital role in any disease creating cell especially dengue viruses directly affects the proteins of blood cells. Hence, the identification of protein is essential in the discovery of drugs for dengue fever. In this research article, traditional context free grammars are used to predict the protein formation with the proposed *Periodicity Analysis in Context Free Grammars (PACFG)* algorithm. The concepts of context free grammars are applied in mutation or motif detection in gene sequences. The periodicity based context free grammars provide optimal solution to the biological sequences with their specified characteristics. We demonstrate the effectiveness of the proposed approach by comparing the experimental results performed on dengue virus dataset with online databases.

Keywords – Context Free Grammars, Periodicity Analysis, Motif, Mutation, Dengue Virus

1. INTRODUCTION

Cells of the human body have a central core called nucleus, which is packaged in units known as chromosomes. Humans have 23 pairs of chromosomes, which are together known as genome. Genes are a specific region of the genomes, which is the molecular unit of heredity of a living organism. Gene sequence contains a sequence of nucleic and amino acids. Nucleic acid consists of a chain of linked units called nucleotide. Nucleic acid sequence has the combination of nucleotide bases within deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). DNA is a chain of four types of molecules adenine (A), cytosine (C), guanine (G), and thymine (T). A sample DNA sequence may be like *TCCTGAT AAGTCAG TGTCTCCT*. RNA is represented as the combination of four nucleotide bases adenine (A), cytosine (C), guanine (G), and uracil (U). RNA sequence may be like *UCCUGAU AAGUCAG UGUCUCCU*. DNA and RNA play a major role in the formation of proteins. The constituents of proteins are amino acids which are represented using 20 English letters except for B, J, O, U, X, and Z. A sample protein sequence may look alike *CFPUEQGHILDCLKSTFEWEGHILDWES*. Protein sequences are shorter than DNA sequences [1].

New gene sequencing technologies are producing large amount of DNA, RNA or protein sequences. Indeed, the discovery of the double helix structure of DNA in 1953 showed that the genetic information contained in this biological macromolecule can be represented by two long complementary sequences over a four-letter alphabet *viz.* {*A;C;G;T}* which are the *nucleotides*, the complementary letters (called Watson-Crick *base pairs*) being *A–T* and *C–G*. This genetic information is used to construct and operate a living organism by the *transcription* when needed of pieces of DNA sequences, named *genes*, into RNA single strand macromolecules which can also be represented by a sequence on almost the same four-letter alphabet {*A;C;G;U}*, where *T* has been replaced by *U*. Sequences of RNAs coding for proteins are in turn *translated* into protein sequences of *amino acid residues*, over the 20 amino acid's alphabet {*A;C;D;E;F;G;H;I;K;L;M;N;P;Q;R;S;T;V;W;Y}*, that determine their three-dimensional conformations and functions in the cells [2].

The theory of formal languages deals with the systematic analysis, classification, and construction of sets of words generated by finite alphabets. The key ideas of formal languages originate in linguistics. Linguistic objects are structured objects. A computational device which infers structure from grammatical strings of words is known as a '*parser'*.

An alphabet $\sum$ is a nonempty finite set of symbols. With $\sum^0 := \{\epsilon\}$ denoting the set of the empty word ϵ, and $\sum^1 := \sum$ we define

An element w in $\sum^*$ is called a '*word*' over the alphabet $\sum$. The length of a word w, written |w|, is the number of symbols it contains. A subset L ϵ $\sum^*$ is called a language over the alphabet $\sum$.

In bioinformatics, important alphabets are $\sum_{DNA}$ = {*A, C, G, T*} representing the four DNA nuclebases, $\sum_{RNA}$ ={*A, C, G, U*} representing the four RNA nuclebases, or $\sum_{Amin}$ = { *A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y* } representing the 22 amino acids. The words over these alphabets are usually called '*sequences*'.

For the alphabet $\sum$ = {a, t} we have
$\sum^0$ = { ϵ }
$\sum^1$ = {a, t}
$\sum^2$ = {aa, at, ta, tt}
.
.
.
$\sum^+$ = {a, t, aa, at, ta, tt, aaa, aat, ata, . . .}
$\sum^*$= { ϵ, a, t, aa, at, ta, tt, aaa, aat, ata, . . .}
Then L = {aa, at} is a language over S.

A grammar G is a four-tuple (V, $\sum$, P, S) consisting of a finite set of variables V, a finite alphabet $\sum$ of terminals, a finite set P of productions, or substitution rules, a start variable S $\sum$ V [3].

Periodicity in genome sequence can be classified into two types, namely, element periodicity and subsequence periodicity. Element periodicity deals with the repetition of individual elements of gene sequence during a particular period whereas subsequence periodicity deals with the periodicity of the entire sequence or some portion of the given sequence [4].

A palindrome is a sequence of letters or words such as racecar and madam I madam which are read the same in forward as well as in reverse direction [5]. The PACFG finds the presence of palindrome in the given sequence which will be helpful in identifying the formation of protein. Each protein adopts a unique 3-dimensional structure, which is decided by its amino acid sequence. A slight change in the sequence can drastically change the functioning of the protein. In case of dengue gene sequences the presence of latent regularities affect the formation of proteins [6].

In Section 2, the work related to the analysis of dengue gene sequences are outlined. Section 3 demonstrates the methodologies related to the prediction of the protein formation in dengue gene sequences. Section 4 exhibits the experimental results that were obtained using dengue virus dataset. Finally, Section 5 describes the conclusion and future scope of this research work.

II. RELATED WORK

Basic research includes a wide range of studies focused on learning how the dengue virus is transmitted and how it infects cells and causes disease. Further many research works investigate several aspects of dengue viral biology that includes exploration of the interactions between the virus and humans as well as the repetition of dengue virus serotypes. Researchers have also been studying the dengue viruses to understand the factors that are responsible for transmitting the virus to humans. They found that specific viral sequences are associated with severe dengue symptoms.

In a similar direction, we propose here an approach to find the latent periodicities in dengue virus gene sequences and predict the formation of proteins on dengue viruses in order to diagnose the dengue syndrome. The major works related to the identification of the latent periodicities in the time series and biological sequences are described below.

Indyk et al. [7] presented periodic trends algorithm that finds the subsequence periodicity alone, by analyzing the recurrence of a sequence of elements in a given time series. Time series is a sequence of values observed over certain time intervals. They developed an algorithm whose time complexity was *O(n log 2n)*, where n is the length of the time series. They used the linear distance measure for finding latent periods.

Elfeky et al. [8] presented two algorithms to find symbol and segment periodicities in the time series. The complexity of their algorithm was *O(n log n)*. They used the fast Fourier transformation and convolution for discovering element and subsequence periodicities.

The algorithm of Ma and Hellerstein [9] computed the symbol periodicity with time tolerance window which is used to accommodate various types

of noise in the data. They used the edit distance measure for discovering periods of the element's occurrence. The result of the element periodicity was used to find the approximation of subsequence periodicity.

Rasheed et al. [10] proposed an algorithm that considers the periodicity of alternative substrings and introduced the concept of relaxed range window (RRW) for detecting periodic occurrences in biological sequences. This approachprovides equal treatment for A and T and also for C and G. For example, the sequence TTACGAATGGTAGT has the periodicity for alternative string group (TT, AA, and TA) with period 4. The strings TA, TT, and AA are parts of an alternative group and the presence of any of these is counted as valid repetition. Another example for RRW concept is in the sequence abdadbacc. Here, "a" is periodic with period 3 starting from position "0" with periodic strength of 100%. They combined the results of the periodicity of individual symbols and combined them by considering their starting positions. They used the suffix tree representation for detecting the periodicities in DNA sequence by modifying the algorithms of Elfeky et al. [8] and Ma and Hellerstein [9].

Huang and Chang [11] presented their algorithm for finding similar periodic patterns, by varying the time limit of the sequence. They used the dynamic time warping (DTW) method for discovering the periods. DTW is a technique for measuring similarity between two temporal sequences which may vary in time or speed. DTW has been applied to temporal sequences of audio, video, and graphics data. The warping function was used to compute the distance between any two elements.

Apart from the above works, there are many research works in the field of biological science that are related to the dengue sequence. Some of the works that are relevant to the current work are furnished below.

Kececioglu and DeBlasio [12] developed a software tool for searching the similarity based on sequence alignment algorithms (SAA). SAA include local, global, and multiple sequence alignment for providing accurate results while analyzing the sequence.

Prada-Arismendy and Castellanos [13] presented a technique called Forensic Investigation Analysis which uses the information related to existing protein structure and predicts the formation of proteins by using visualization techniques.

Mairiang et al. [14] focused on the combined analysis of protein interactions. They tested each identified host protein against the proteins of all four serotypes of dengue and identified the interactions that are conserved across serotype. Their contribution was useful in understanding the interplay between dengue and its hosts.

Bletchly [15] proposed the pathogen analysis which helps to explore the human immune response to dengue virus infection and to analyze the antigen and structure of the protein. Pathogen is an infectious agent that causes disease or illness to its host. This analysis examines both the human immune response system and the circulation of the serum of infected patients.

Though, there are various techniques available to find the periodic patterns in time series and other sequences, the works related to the biological sequences are very limited. Further, the existing works concentrate mainly on element periodicity or subsequence periodicity. Therefore, there is a need for holistic approach that computes all kinds of periodicities. In the current work, we propose an approach called PACFG to compute several periodicities including latent periodicity using context free grammars. The result of the PACFG produces the probability of the kind of protein in dengue virus gene sequence.

III. METHODOLOGY

The PACFG algorithm reads the given input sequence and identifies the occurrences of each nucleic acid bases like A, C, G and T. CFGs are used to generate the language of given gene sequence which is denoted as L(G) like A→T and C→G or TTT → F. The next step is the order of nucleic bases which are very important to identify the formation of proteins by using the universal genetic code. Finally the PACFG algorithm divides the given gene sequence into required protein codons viz. 3-letter or 8-letter or 16-letter. The divided codon clearly depicts the protein formation and its structure. The flow of algorithm is shown in figure 3.1. The universal genetic code is demonstrated in the table 3.1 along with the corresponding proteins. The implementation of PACFG is illustrated as transition diagram in figure 3.2. Further, the PACFG algorithm is incorporated with our own tool named as '*Sequence Miner*' which is the Java based bio-

computational tool for diagnosing the dengue fever. Let assume the sample sequence from the above table, TTT, TTC, TTA and TTG for the implementation of transition diagram.

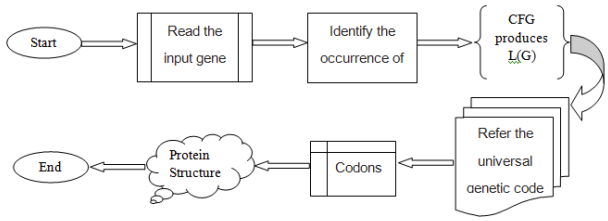

Figure3.1. Workflow of PACFG algorithm

For the above sequence corresponding CFG is T → TTT | TTC | TTA | TTG. The first productions are producing the protein F and rests of them are producing the protein L. Similarly, the entire sequence is mapped through their respective proteins.

Table3.1. Universal Genetic Code

Codon (5′ → 3′) ↦ amino acid							
TTT	F	TCT	S	TAT	Y	TGT	C
TTC	F	TCC	S	TAC	Y	TGC	C
TTA	L	TCA	S	TAA	*	TGA	U
TTG	L	TCG	S	TAG	O	TGG	W
CTT	L	CCT	P	CAT	H	CGT	R
CTC	L	CCC	P	CAC	H	CGC	R
CTA	L	CCA	P	CAA	Q	CGA	R
CTG	L	CCG	P	CAG	Q	CGG	R
ATT	I	ACT	T	AAT	N	AGT	S
ATC	I	ACC	T	AAC	N	AGC	S
ATA	I	ACA	T	AAA	K	AGA	R
ATG	M	ACG	T	AAG	K	AGG	R
GTT	V	GCT	A	GAT	D	GGT	G
GTC	V	GCC	A	GAC	D	GGC	G
GTA	V	GCA	A	GAA	E	GGA	G
GTG	V	GCG	A	GAG	E	GGG	G

Amino acid abbreviations			
A: Ala	alanine	N: Asn	asparagine
C: Cys	cysteine	O: *Pyl*	*pyrrolysine*
D: Asp	aspartic acid	P: Pro	proline
E: Glu	glutamic acid	Q: Gln	glutamin
F: Phe	phenylalaline	R: Arg	arginine
G: Gly	glycine	S: Ser	serine
H: His	histidine	T: Thr	threonine
I: Ile	isoleucine	U: *Sec*	*selenocysteine*
K: Lys	lysine	V: Val	valine
L: Leu	leucine	W: Trp	tryptophane
M: *Met*	*methionine (start)*	Y: Tyr	tyrosine
		*: (Stop)	—

Here, the occurrence of palindrome string is also plays a vital role to form the respective proteins. The latent periodic intervals are also calculated through our Recurrence Finder (RECFIN) algorithm. The entire PACFG algorithm is illustrated in table 3.2.

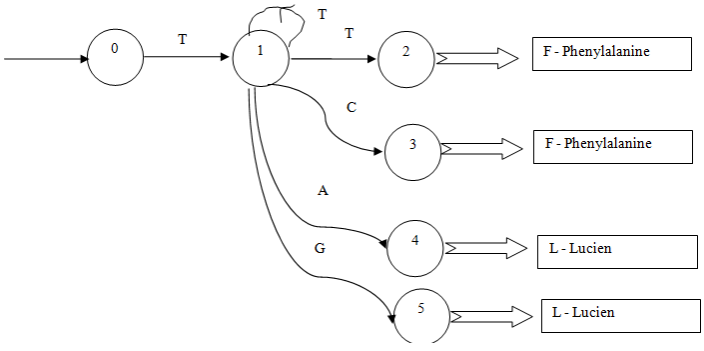

Figure3.2 Transition Diagram of Protein Formation

Table 3.2. PACFG Algorithm

Objective: Predict the Protein Formation using CFG

Input: Gene Sequences of Dengue Virus, G

Output: Codons with Protein Structure

Method:

(1) Find the occurrence of each nucleic base

(2) Generate the language for the given gene sequence L(G) using CFG

(3) Identify the order of elements in L(G) using RECFIN ()

(4) Compare with the universal genetic code

(5) Divide the sequence into codons (3', 5' 8')

(6) Visualize the structure of the protein

IV. RESULTS & DISCUSSION

To demonstrate the functionality of the PACFG algorithm, dengue gene sequences datasets of online databases have been used. These datasets contain four different dengue viruses, namely, DEN1, DEN2, DEN3, and DEN4.This experiment utilizes the DNA sequence of DEN4 as the input sequence. The length of the input sequence is 10,735 characters. The periodic patterns along with their intervals shown in figure 4.1. The output of the proposed PACFG algorithm shows the most similar result to National Centre for Biotechnology Information (NCBI) results.

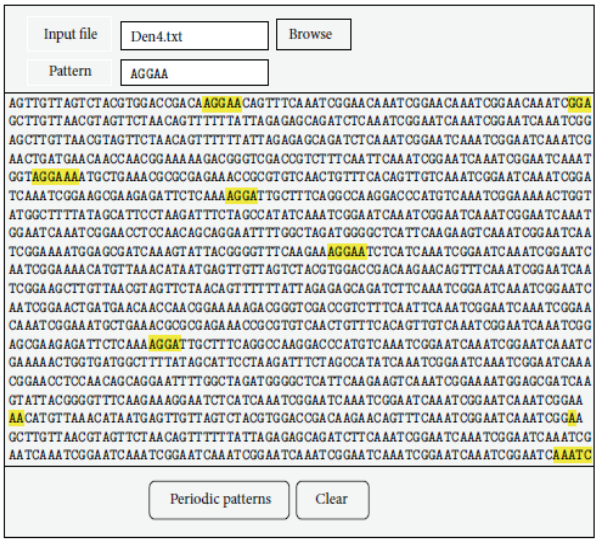

Figure 4.1. Periodic Patterns

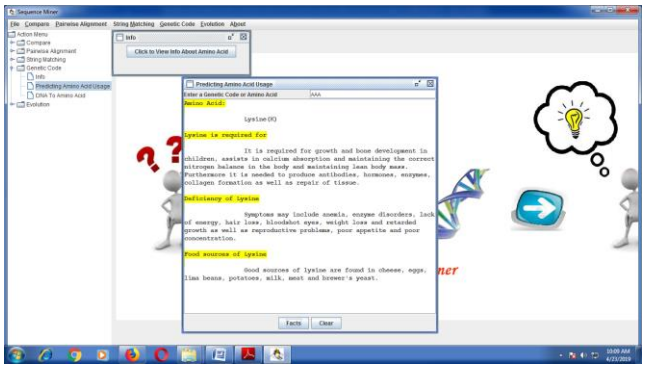

Figure 4.2. Predicting Amino Acid Usage

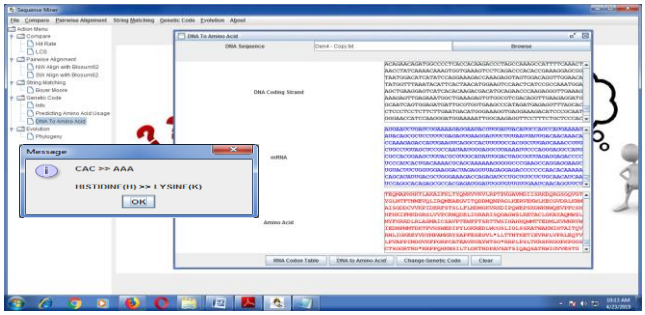

Figure 4.3. Conversion of DNA → RNA → Protein

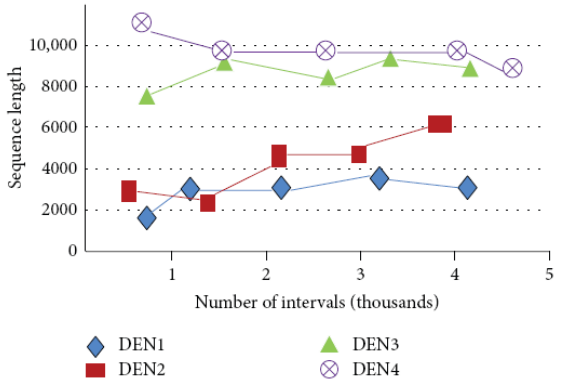

Figure 4.4 Periodic Intervals

V. CONCLUSION

In this paper, we have derived PACFG algorithm to predict the formation of proteins using context free grammars. The periodic intervals among the dengue DNA gene sequences define the mRNA and amino acid. Hence, this proposed research focuses on the periodicity mining in the dengue gene sequences. Finally, our algorithm shows the structure of the protein, mutation points.

The sequence length of four different dengue virus are DEN1 : 10, 073, DEN2: 10,069, DEN3: 10,017 and DEN4: 10,735. Among these 4 types DEN1 and DEN2 having the related interval periods. DEN3 and DEN4 are completely proportionate to previous types. The variations are mentioned in the figure 4.4.

REFERENCES

[1] W.K. Sung, "Algorithms in bio informatics," International Journal of Molecular Biology, vol. 2, no. 1, pp. 23–29, 2011.

[2] F.Coste, "Learning the Language of Biological Sequences", Inria Rennes Publisher, France, pp. 1-35, 2017.

[3] Andreas de Vries, "Genome Grammars and Formal Languages", pp.23-38, 2011.

[4] Marimuthu, T., and Balamurugan, V., "Mining Association Rules in Dengue Gene Sequence with Latent Periodicity", "Computational Biology Journal", Hindawi Publishing Corporation, vol.2015, pp.1-10, 2015.

[5] R. Gupta, A. Mittal, V. Narang, and W.-K. Sung, "Detection of palindromes in DNA sequences using periodicity transform," in Proceedings of the IEEE International Workshop on BiomedicalCircuits and Systems, vol. No, pp. 20–23, December 2014.

[6] F. Rasheed, M. Alshalalfa, and R. Alhajj, "Adapting machine learning technique for periodicity detection in nucleosomal locations in sequences," in Proceedings of the 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL '07), vol. No, pp. 870–879, Dubai, UAE, December 2007.

[7] P. Indyk, N. Koudas, and S. Muthukrishan, "Identifying representative trends in massive time series data sets using sketches,"
The International Journal on Very Large Data Bases, vol. 5, no. 2, pp. 123–128, 2000.

[8] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, "Periodicity detection in time series databases," IEEE

Transactions on Knowledge and Data Engineering, vol. 17, no. 7, pp. 875–887, 2005.

[9] S.Ma and J. L.Hellerstein, "Mining partially periodic event patterns in time series database," IEEE Transactions on Knowledgeand Data Engineering, vol. 2, no. 3, pp. 205–214, 2011.

[10] F. Rasheed, M. Alshalalfa, and R. Alhajj, "Efficient periodicity mining in time series databases using suffix trees," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 1, pp. 79–94, 2011.

[11] K.-Y. Huang and C.-H. Chang, "SMCA: a general model for mining asynchronous periodic patterns in temporal databases," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 774–785, 2005.

[12] J. Kececioglu and D. DeBlasio, "Parameter advising for dengue virus serotypes," in Proceedings of 2nd International Conference on Genomic Sequences, vol. 2, pp. 221–228, 2013.

[13] J. Prada-Arismendy and J. E. Castellanos, "Real time PCR application in dengue studies," International Journal on Proteomics Analysis, vol. 42, no. 2, pp. 89–96, 2010.

[14] D. Mairiang, H. Zhang, and A. Sodja, "Identification of new protein interactions between dengue fever virus and its hosts,"
International Journal of Biometrics and Bioinformatics Algorithms, vol. 25, no. 2, pp. 156–160, 2013.

[15] C. Bletchly, "Antigenic and structural analysis of the NS1 glyco protein of dengue virus," International Journal of Molecular Biology, vol. 5, no. 2, pp. 88–94, 2002.