Hindawi Publishing Corporation
Computational Biology Journal
Volume 2015, Article ID 839692, 10 pages
http://dx.doi.org/10.1155/2015/839692

Research Article

Mining Association Rules in Dengue Gene Sequence with Latent Periodicity

Marimuthu Thangam[1] and Balamurugan Vanniappan[2]

[1]*Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu 627012, India*
[2]*Department of Information Technology, AMET University, Chennai, Tamil Nadu 603112, India*

Correspondence should be addressed to Marimuthu Thangam; mastersvksmca@gmail.com

Received 19 August 2014; Revised 30 November 2014; Accepted 2 January 2015

Academic Editor: Clifford Shaffer

The mining of periodic patterns in dengue database is an interesting research problem that can be used for predicting the future evolution of dengue viruses. In this paper, we propose an algorithm called Recurrence Finder (RECFIN) that uses the suffix tree for detecting the periodic patterns of dengue gene sequence. Also, the RECFIN finds the presence of palindrome which indicates the possibilities of formation of proteins. Further, this paper computes the periodicity of nucleic acid and amino acid sequences of any length. The periodicity based association rules are used to diagnose the type of dengue. The time complexity of the proposed algorithm is $O(n^2)$. We demonstrate the effectiveness of the proposed approach by comparing the experimental results performed on dengue virus serotypes dataset with NCBI-BLAST algorithm.

1. Introduction

Periodicity is the tendency where the sequences of events or values recur at particular intervals [1]. Periodicity plays an important role in discovering interesting frequent patterns in any sequence including genomic sequence that is made of amino acids present in the human cells. Latent periodicity refers to the presence of hidden or reverse subsequence in the given sequence during the particular interval. Finding the latent periodicities or regularities among gene sequences will be helpful for the drug designers in predicting the future evolution of viruses that cause the particular disease.

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 [2].

Although the proposed work can be applied on any gene sequence such as *ebola* and *chikungunya*, with suitable modification, the main focus is shown on the dengue gene sequence alone owing to its significance in the recent years.

The incidence of dengue has grown dramatically around the world in recent decades. Over 2.5 billion people, 40% of the world's population, are now at risk on account of dengue. World Health Organization (WHO) currently estimates that there may be 50–100 million dengue infections worldwide every year [3]. As per the medical record of Government

of Tamil Nadu, India, 15,535 persons were affected and 96 expired in the year 2009. The outbreak of dengue in India in the year 2012 was the worst in the previous six years [4].

Under these circumstances, the research on dengue virus genome sequence plays a vital role in the diagnosis of the disease. Therefore, it is necessary to predict the presence of cooccurrence patterns which are similar elements present in dengue gene sequences. This work derives the periodic association rules (PAR) that will reveal the possibilities of occurrence of similar disease pattern using a novel technique called *Periodic Association Rule Mining (PARM)*.

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.

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 RECFIN 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 affects the formation of proteins [6].

The dengue virus belongs to Flaviviridae family that is transmitted to people through the bite of the *Aedes aegypti* or *Aedes albopictus* mosquitoes. There are four types of dengue virus serotypes that cause the disease [7]. Serotypes refer to the subdivisions of a virus that are classified based on their cell surface. They are listed in Table 1.

Table 1: Types of dengue virus serotypes.

Virus type	Name of the virus
DEN1	*Strain Hawaii*
DEN2	*Strain New Guinea C*
DEN3	*Strain H87*
DEN4	*Strain H241*

There are three main types of dengue infection, namely, classic dengue fever (CD), dengue hemorrhagic (DH) fever, and dengue shock syndrome (DSS) [8]. All the types of dengue fever begin with noticeable symptoms within four to seven days after the *Aedes aegypti* mosquito's bite. The symptoms of CD include headache, pain behind the eyes, joints, and muscles, vomiting, and body rash. It also reduces the count of white blood cells (WBC). DH fever includes all the classic symptoms with higher fever and sharp decrease in the number of platelets in the blood. Platelets are small, disk shaped fragments that are the natural source of growth factors. They are circulated in the blood and involved in the formation of blood clots. As a result of this, victims bleed from the nose, gums, and skin. DSS is the most severe form of the disease which causes massive bleeding and fall in the blood pressure [9]. Each virus type has its own characteristics.

The RECFIN evaluates the element and subsequence periodic patterns including palindrome among the given dengue sequences. RECFIN comprises three parts. The first part deals with the formation of the suffix tree to find the periodic patterns. In the second part, a recurrence identification procedure is proposed to find the periodic patterns and, in the third part, a novel palindrome detection procedure is presented to find the presence of palindrome in the given sequence. Based on the resultant patterns, the periodic association rules are generated using PARM. These rules are used to classify the type of dengue. The support threshold is defined as the extent to which the periodic patterns have the periodic repetition of values when compared to the given sequence. The support threshold is measured in terms of percentage. The experimentation is performed on the dengue virus serotype dataset. The entire dataset has the name, gene ID, description, and location of chromosome details. The classification accuracy is compared with the National Center for Biotechnology Information (NCBI) database [10].

In Section 2, the work related to the dengue and periodicity detection is outlined. Section 3 demonstrates the methodologies related to the prediction of the periodic pattern and palindrome in dengue gene sequences. Section 4 exhibits the experimental results that were obtained using dengue virus serotype dataset [10]. Section 5 illustrates the comparative analysis of the results. Finally, Section 6 describes conclusion.

2. Review of Related Works

The causes and effects of dengue have been focused on by the research community for the past two decades. Current research on dengue aims to provide better surveillance to limit the effect of dengue outbreak. 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 [11].

In a similar direction, we propose here an approach to find the latent periodicities and periodical associations in dengue virus serotypes 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 [9] are described below.

Indyk et al. [12] 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. [13] 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.

Rasheed et al. [14] 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 approach provides 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. [13] and Ma and Hellerstein [15].

The algorithm of Ma and Hellerstein [15] 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.

Huang and Chang [16] 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.

Pujeri and Karthik [17] proposed the constraint-based periodicity mining (CBPM) algorithm that uses frequent pattern growth (FPG) tree in time series databases. For constraint-based association rule mining, the user can specify various types of constraints which include constraints based on knowledge, data, dimension, level, interestingness, and rule. By specifying CBPM, the user can evaluate the one-dimensional rule such as *buy (school bag)* $\rightarrow$ *buy (uniform)* where the dimension is *buy*. Also, the user can evaluate the rule such as *occupation (student)* $\rightarrow$ *buy (textbook)* which has two dimensions *occupation* and *buy.* Further, multidimensional rules can be evaluated in a similar manner. The time complexity of CBPM algorithm is $O(kN)$, where N is the length of input sequence and k is the length of periodic pattern.

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 [18] 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 [19] 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. [20] 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 [21] 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 and their associations.

In the current work, we propose an approach called RECFIN to compute several periodicities including latent periodicity. RECFIN algorithm adopts the suffix tree technique. PARM generates the periodic association rules from frequent item sets. Though our algorithm follows the worst case of time complexity of $O(n^2)$, it is helpful in predicting the future evolution of dengue virus types accurately.

3. Dengue Virus Detection Problem

3.1. Notation. The genome sequence of any living organism consists of nucleic acid and amino acid. DNA sequence comprises of the values *A, C, G,* and *T* which are used to form a protein. Consider a dengue gene sequence $D = A, C, G, T$. The set of values in the DNA sequence of dengue can be denoted by $\sum = \{A, C, G, T, \ldots\}$. In the current work, the DNA sequence of four different dengue virus serotypes, namely, DEN1, DEN2, DEN3, and DEN4, is considered.

For the gene sequence of given length, RECFIN algorithm computes the element and subsequence periodicities. In addition, it finds the presence of palindrome which will be helpful in predicting the formation of protein. The proposed approach utilizes the suffix tree (ST) data structure. Based on the occurrence of element and subsequence periodicities, the PAR is generated.

3.2. Element Periodicity. In a DNA sequence D, an element x is said to be element periodic with a period p if x exists for almost every p periodic intervals. For example, in the DNA sequence $D_1 = ACGACCACGC$, the symbol C is periodic with period 4 since C exists every four periodic intervals (i.e., in positions 1, 5, and 9). Moreover, the element A is periodic with period 3 since A exists almost every three time intervals

(i.e., in positions 0, 3, and 6 not 9). The element periodicity is defined as follows.

Let D be a sequence. Then, $X_{p,l}(D)$ will be the projected sequence that contains the periodic values of element d which starts at position l in which period p can be shown as

$$X_{p,l}(D) = d_l, d_{l+p}, d_{l+2p}, \ldots, d_{l+(m-1)p}, \quad (1)$$

where $0 \leq l < p$, $m = |(n-l)/p|$, and n is the length of D. For example, if $D_1 = ACGACCACGC$, then $X_{4,1}(D_1) = CCC$ and $X_{3,0}(D_1) = AAAC$. Naturally, the ratio of the number of occurrences of an element x in certain $X_{p,l}(D_1)$ to the length of this projection indicates how often this element occurs after every p periodic intervals.

3.3. Subsequence Periodicity. Unlike element periodicity that focuses on the elements where different elements may have different periods, the subsequence periodicity focuses on the repetition of sequence of values. The DNA sequence D is said to be periodic with a period p if D can be divided into equal-length subsequences, each of length p, which are almost similar. For example, the DNA sequence $D_2 = ACGACGACG$ is clearly periodic with a period 3; likewise, the DNA sequence $D_3 = ACGACTACG$ is partially periodic with a period 3 for the same subsequence despite the fact that its second subsequence is not identical to other subsequences.

3.4. Latent Periodicity. The detection of hidden regularity patterns like palindrome in DNA sequences plays a major role in deciding the classification of dengue virus serotypes such as DEN1 and DEN2. Consider, the DNA sequence $D_4 = CAGGAC$, which has the palindrome sequence. The rearranged sequence of D_4 is $D_{4r} = GACGAC$. The periodic interval of each element in D_{4r} is 3. The D_{4r} is said to be a complicated palindrome [2].

3.5. Periodicity Detection with RECFIN Algorithm. The RECFIN algorithm has four steps as described below.

3.5.1. Suffix Tree Based Representation. A suffix tree (ST) is a nonlinear data structure that has been proved to be very useful in string processing [14]. It is useful in searching a substring of the original string. Also, it is useful in finding the frequent substring. Each of the branches of the suffix tree represents a suffix of the original string. Hence, a suffix tree for a string of length n has n branches and, thus, n leaf nodes.

Each leaf node in the tree has an integer value showing the starting position of the substring achieved through the path from the root to that of leaf in the original string. Since there are exactly n suffixes for a string, each starting at one of the index positions, there are n leaf nodes in the tree. Each internal node has the value representing the length of the substring so far achieved while traversing from the root to the node. In a suffix tree, each node contains a unique field called index. It identifies the starting index of a substring in the multiple sequences.

Consider the DNA sequence $t = CAGTCAGG$. The sequence can be written based on its index. A symbol \$ is added with t being a termination indicator. The construction of ST is illustrated in Figure 1, where the non-leaf nodes are generated based on the first occurrence of the subsequence in reverse order of t. The leaf nodes are generated based on the occurrence of parent node as well as in the indexed order till the end of the sequence. Therefore, the ST is useful in the identification of all subsequences such as *G\$, GG\$, AGG\$, CAGG\$, TCAGG\$, GTCAGG\$, AGTCAGG\$,* and *CAGTCAGG\$*.

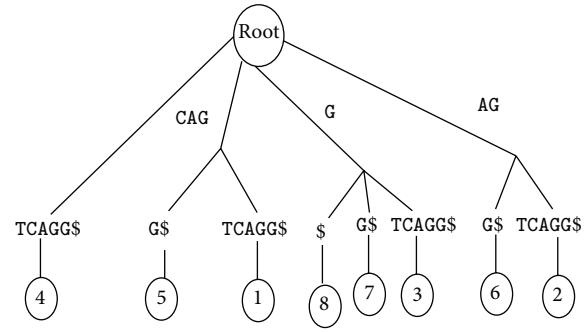


Figure 1: Construction of a suffix tree.

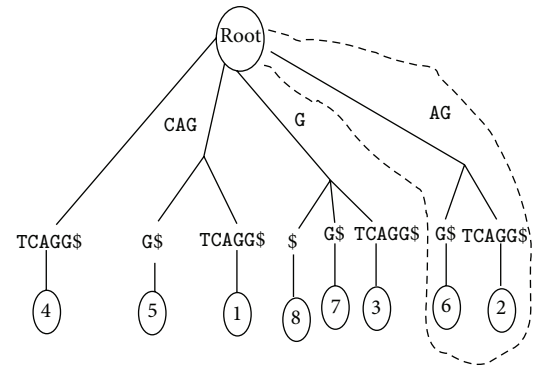


Figure 2: Traversal in suffix tree.

3.5.2. Element and Subsequence Periodicity. After the construction of the suffix tree, the tree traversal process is performed in the bottom up fashion. During the traversal, each leaf node passes its value to its parent. A subsequence starting with position p can be found by traversing the corresponding leaf node that contains the value p and its parent nodes till the root is reached. Consider the sequence starting with index 2, that is, *AGTCAGG\$*. To get the sequence, the traversal is performed from the leaf node 2 towards its root through the parent nodes as shown in Figure 2. Similarly the traversal for the subsequence *AGG\$* is performed from the starting leaf node 6. The resultant sequence must be reversed in order to get the required subsequence. The traversal process from leaf node to root needs to be performed recursively and is known as recurrence calculation. In the algorithm, *reccal* procedure is used for this.

In a suffix tree, a leaf can represent more than one parent. The total number of parents can be calculated as $(n+1) - i$, where n is the length of t and i is the index value. For $i = 3$ and

$n = 8$ the possible parent values can be calculated as $(n+1)-i$; that is, $(8+1)-3=6$. Therefore, six combinations such as *G, GT, GTC, GTCA, GTCAG,* and *GTCAGG* are possible. Hence, the value of *reccal* is incremented by 1. The count represents the frequency of the occurrence of a sequence [22]. Thus, the suffix tree based representation helps us to find the element and subsequence periodicities simultaneously for the given sequence.

3.5.3. Latent Periodicity in Suffix Tree. Apart from finding the subsequence periodicity in the forward direction, the occurrence of palindrome can also be found.

If we calculate the reverse of the string, it provides the reverse of the first half; then, it is said to be the latent periodicity. For example, the DNA sequence $D_4 = CAGGAC$ has the palindrome sequence which contains rearranged values of first half in the second half. Thus, the presence of palindrome is found. In the algorithm, the procedure *polycheck* is used for this purpose.

3.5.4. Periodic Association Rules. A further step in this direction is the prediction of cooccurrence patterns among the dengue gene sequences. This can be done by evaluating the rules that can reveal the occurrence of an element or subsequence. Such rules are called periodic association rules, and the corresponding technique is called *Periodic Association Rule Mining*. The PARM is similar to market basket analysis. In PARM terminology, the nucleic or amino acids may be considered as items and the gene subsequences as the baskets that contain the items. In the traditional association rules, only the number of frequent items is calculated whereas PARM calculates the occurrence order of frequent item sets along with its periodic position.

To obtain periodic association rule, the frequencies of nucleic or amino acids are computed in each dengue gene sequence. The rule can be expressed as $A \rightarrow C$, where A and C are the associated items. The rules state that if a nucleic acid A is present in a given sequence with $f1$ periodicity, then there will be another nucleic acid C that will have similar periodicity with respect to their respective initial positions. The PARM procedure enables finding the periodicity $f1$ along with its starting positions.

Let $I = \{i_1, \ldots, i_k\}$ be a set of k elements, called items. Let $I_s = \{b_1, \ldots, b_n\}$ be a set of n subsets of I. We call each b_i a set of transaction. In the market basket application [22], the set I denotes the items stocked by a retail outlet and each basket b_i is the set of items of a transaction. Similarly, in case of gene sequence, the set I denotes the elements of nucleic or amino acid and the basket b_i is the orderly subsequences. The order and frequency of the elements can be evaluated using the suffix tree. The PAR is intended to capture the orderly dependence among the elements of dengue virus dataset and the rule can be represented as $i_1 \rightarrow i_2$ along with the period and starting position of i_1 and i_2, provided that the following conditions hold good:

(1) i_1 and i_2 occur at regular intervals in the sequence for at least s% of the n baskets where s is the support and n is the number of subsequences;

```
Objective: To Mine PAR
Input: Gene sequence of Dengue
       virus D, minimum support s and confidence c.
Output: Periodic Association Rules
Method:
  (1) Construction of Suffix tree
    (a) Read the given input.
    (b) Affix the $ symbol at the end of
        the sequence. ($ = n + 1) where n is the
        number of elements in the sequence.
    (c) Call Suffixtree(D, $);
  (2) Call reccal(sp, pi); where sp = starting
        position, pi = periodic interval
  (3) Call polycheck();
  (4) Generate the periodic association rules
        for predicting the type of Dengue virus
        serotype. Call PAR();
Procedure Suffixtree(D, $)
    (1) Initialize root.
    (2) For each child node with element e till $
    (3) If e is already marked
            Goto reccal(sp, pi)
        Else
        Create new node
    (4) Mark the Index value.
Procedure reccal(sp, pi)
    (1) Count the element and subsequence pattern.
    (2) Increment the count value by 1 when
        the new pattern is arrived.
    (3) Maintain the minimum support &
        confidence threshold.
Procedure polycheck()
    (1) For each occurrence pattern find the
        presence of palindrome.
    (2) Find the reverse of the occurrence pattern.
    (3) If palindrome Mark as "Palindrome".
Procedure PAR()
    (1) Calculate the frequent patterns with s and c.
    (2) Generate PAR.
```

Algorithm 1: RECFIN algorithm.

(2) for all the subsequences containing i_1, at least c% of subsequences contains i_2 where c is the confidence [22].

The above definition can be extended to form multidimensional periodic association rule such as $AC \rightarrow GT$, where AC and GT are element of nucleic acid with periodic dependence. The association rules are considered to be interesting if they satisfy both minimum support and confidence thresholds. The threshold values are set by users based on their domain expertise [22].

To evaluate the PAR, we propose the RECFIN algorithm. The following steps are involved in the RECFIN algorithm:

(1) based on the occurrence positions, the elements are mapped into integers;

(2) based on the support threshold, the element periodicity is found; the set of elements that satisfies the

```
AGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCAAATCGGAA
GCTTGTTAACGTAGTTCTAACAGTTTTTTATTAGAGAGCAGATCT
CTGATGAACAACCAACGGAAAAAGACGGGTCGACCGTCTTTCAAT
ATGCTGAAACGCGCGAGAAACCGCGTGTCAACTGTTTCACAGTTG
GCGAAGAGATTCTCAAAAGGATTGCTTTCAGGCCAAGGACCCATG
AAACTGGTGATGGCTTTTATAGCATTCCTAAGATTTCTAGCCATA
CCTCCAACAGCAGGAATTTTGGCTAGATGGGGCTCATTCAAGAAG
AATGGAGCGATCAAAGTATTACGGGGTTTCAAGAAAGGAATCTCA
AACATGTTAAACATAATG
```

Box 1: Partial DNA sequence of DEN 4.

minimum support threshold is called the frequent item set;

(3) the frequent item sets are used to generate association rules; for example, consider the item set $\{A, C, G\}$; the following rules can be evaluated using the given item set:

Rule 1 is as follows: $A \wedge C \rightarrow G$;
Rule 2 is as follows: $C \wedge G \rightarrow A$;
Rule 3 is as follows: $A \wedge G \rightarrow C$;
Rule 4 is as follows: $G \wedge A \rightarrow C$;
Rule 5 is as follows: $C \wedge A \rightarrow G$;
Rule 6 is as follows: $G \wedge C \rightarrow A$.

In the above rules the element that appears in left hand side is called antecedent and that of the right hand side is called consequent; the confidence is computed using the conditional probability of antecedent. For example, the confidence of the rule 1 is computed as follows:

Confidence = support $\{A, C, G\}$/support $\{A, C\}$;

if the confidence is equal to or greater than a given confidence threshold, the rule is considered to be interesting rule;

(4) based on the support and confidence, the PAR is generated.

3.5.5. RECFIN Algorithm. In this section, we describe the pseudocode of the *RECFIN* algorithm in Algorithm 1 which covers the entire processes, element, subsequence periodicities, palindrome checking, and the generation of periodic association rules.

4. Experimental Results

To demonstrate the functionality of the RECFIN algorithm, dengue gene sequences datasets of NCBI have been used [10]. 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 with support threshold 50% and confidence threshold 70%. The partial DNA sequence of DEN4 is shown in Box 1.

The length of the input sequence is 10,735 characters.

Table 2: Latent periodicity results.

Latent periodicity results of DEN4		
Length of palindrome	Starting position	Palindrome sequence
2	21	AA
	45	CC
	68	GG
	44	TT
3	4, 13	CAC
	6, 9, 62	CGC
	11, 29, 31, 80	CTC
	16, 34	GTG
	24	AGA
	30	TCT
	44	ATA
	45, 89	TAT
	59, 89, 90, 95	TTT
	63	GCG
	66	CCC
6	6	CGCCGC
	53	GCAACG
	88	ATTTTA
	91	TTAATT
20	5960	CCTCCTCCTC CCTCCTCCTC
25	4774, 4776 4778, 4780 4782, 4784 4786, 4788 4790, 4792 4794, 4796 4798, 4800 4802, 4804 4806, 4808	ATATATATAT ATATATATAT ATATA
	4775, 4777 4779, 4781 4783, 4785 4787, 4789 4791, 4793 4795, 4797	TATATATATA TATATATATA TATAT
	5974	TCCTCCCCCT CCCCCTCCCC CTCCT

For the demonstration, consider the following given sequence: *CATCATGG*. The suffix tree of the given sequence is illustrated in Figure 3.

Figure 4 illustrates the periodic occurences of the given sequence *AGAA*.

Table 2 displays the latent periodicity results of RECFIN algorithm for the DEN4 virus serotype. In addition, the algorithm detects many more periods, some of which are quite interesting. However, for the entire sequence (10, 735), it is difficult to explain and the results are highlighted for

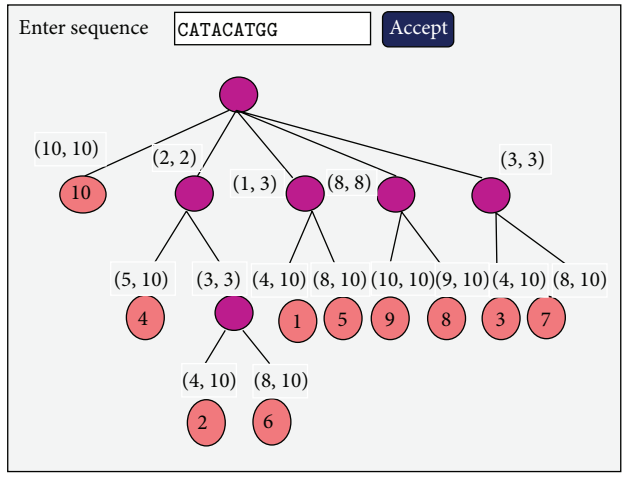


Figure 3: Construction of a suffix tree.

Figure 4: Periodic occurrences.

the partial number of output periods along with the latent periodicity.

The final step of the RECFIN algorithm is to evaluate the PAR. The PAR that is generated by RECFIN contains interesting as well as extraneous patterns. Therefore, pruning is necessary to extract the useful patterns. The interesting pattern is the pattern that has the strong periodic dependence with high support and confidence. The interesting PAR covers the rules of similar periodic intervals among the different dengue virus serotypes shown in Table 3.

The PAR contains the elements along with their starting position, periodicity values, and their dependence with support and confidence. Rule 1 of Table 3 reveals the element periodicity. Further, the occurrence of elements *A*, *G*, and *C* with periodicity 21 reveals the periodic dependence of element *T*.

5. Comparative Analysis

The NCBI-GenBank database is used for the comparative analysis which has 171 million sequences as of February, 2014 [10]. For the comparative analysis of the algorithm, we have used DNA sequences of four different dengue virus serotypes. This dataset varies in the length of the characters. The varying length of each dengue virus serotypes is listed in Table 4.

Table 3: PAR obtained from RECFIN algorithm.

Periodic association rules (PAR)	Support (%)	Confidence (%)
$A_{1,21} \wedge G_{3,21} \wedge C_{5,21} \rightarrow T_{7,21}$	50	70
$TT_{45,18} \wedge GG_{32,18} \wedge CC_{12,18} \rightarrow A_{21,18}$	40	60
$GAG_{25,50} \wedge CAG_{98,50} \wedge TAG_{255,50} \rightarrow AGA_{405,50}$	25	50
$GCAACG_{83,25} \wedge ATTTTA_{88,25} \wedge TTAATT_{91,-25} \rightarrow CGCCGC_{101-25}$	30	40

Table 4: Length of dengue virus serotypes.

Virus type	Length of the DNA sequence (in characters)
DEN1	10,073
DEN2	10,069
DEN3	10,017
DEN4	10,735

Table 5: Average periodic intervals.

Name of the periodicity	Periodic intervals
Element periodicity	21
Subsequence periodicity	124
Latent periodicity	423

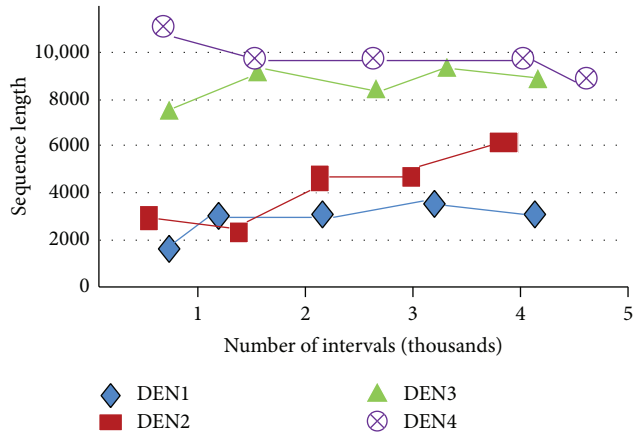


Figure 5: Periodic intervals.

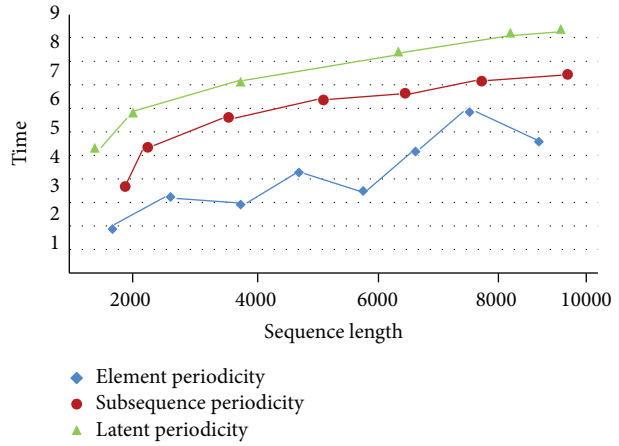


Figure 6: Sequential characteristics of the RECFIN algorithm.

The result of the NCBI-Basic Local Alignment Search Tool (BLAST) algorithm is compared with our proposed algorithm through the experiments. The most important aspect is the accuracy with respect to the discovered periods of the proposed algorithm as discussed in Section 5.1. Then, the time performance of the proposed algorithm is displayed in Section 5.2.

5.1. Accuracy. The accuracy measure is the ability of the algorithm to detect the periodicities in the given sequence. To accurately discover a period, the periods discovered with a high periodicity threshold value are better candidates than those discovered with a lower periodicity threshold value. Therefore, we examine the accuracy by measuring the closeness of the periodic values estimated by the algorithm. The accuracy is measured by the average periodic intervals. Table 5 shows the average periodic interval of element, subsequences, and latent periodicities in the given sequence (DEN4).

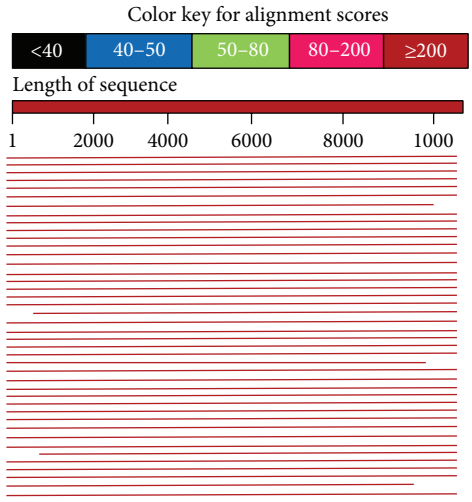
Figure 5 shows that the periodicity is increasing in level when more intervals are included.

5.2. Time Performance. To evaluate the time performance of the proposed RECFIN algorithm, Figure 6 exhibits the sequential characteristics of algorithms with respect to the sequence length.

5.3. Real Data Experiments. The above results are compared with National Centre for Biotechnology Information (NCBI) Basic Local Alignment Search Tool (BLAST) that shows the most relevant results of our RECFINalgorithm. The NCBI-BLAST has enormous amount of datasets which compare the given sequence with the existing online dataset.

The output of the proposed RECFIN algorithm shows the most similar result to NCBI-BLAST result. Also the alignments of the sequences are also compared and are shown in Figures 7(a) and 7(b).

The entire length of sequences is marked as Query. The color key can be used to show the score of alignment. In this case, the entire sequence will be aligned exactly and is represented as color key for alignment score ≥200 in Figure 7(a). The attributes, score, and identities show the alignment of the entire sequence in Figure 7(b).



(a)

Range: 1 to 10,735

Score	Identities	Gaps
19824 bits(10735)	10735/10735	0/10735

Sequence: TCGGAAGCTTGTTAACGTAGTTCTAACAGTTTTTTATTAGAGAGCAGATCTTCA
Alignment: TCGGAAGCTTGTTAACGTAGTTCTAACAGTTTTTTATTAGAGAGCAGATCTTCA

Sequence: AATGAGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCAAA
Alignment: AATGAGTTGTTAGTCTACGTGGACCGACAAGAACAGTTTCAAA

Sequence: AAAAGACGGGTCGACCGTCTTTCAATTCAAATCGGAAATCGG
Alignment: AAAAGACGGGTCGACCGTCTTTCAATTCAAATCGGAAATCGG

Sequence: CGGGTCGACCGTCTTTCAATTCAAATCGGAATAATTCAAATCGGAAT
Alignment: CGGGTCGACCGTCTTTCAATTCAAATCGGAATAATTCAAATCGGAAT

(b)

Figure 7: (a) NCBI BLAST alignment results. (b) RECFIN alignment results.

PAR is generated based on the occurrence of both element and subsequence periodicities along with latent periodicity. After the analysis of the results, we have obtained some of the interesting frequent patterns based on the periodic intervals.

6. Conclusion

In this paper, we have derived PAR to predict the dengue serotype and to define three types of periodicities. The element periodicity addresses the periodic intervals among the elements; the subsequence periodicity addresses the periodic intervals among the subsequences along with the latent periodic patterns. The proposed RECFIN algorithm for detecting each type of periodicity in $O(n^2)$ time is based on suffix tree method, for a gene sequence of length n. Finally, our algorithm is used to define periodic association rules for each dengue virus serotype with the interestingness measures of support and confidence thresholds which helps to predict the future evolution of dengue virus serotypes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] W. J. H. McBride and H. Bielefeldt-Ohmann, "Dengue viral infections; pathogenesis and epidemiology," *Microbes and Infection*, vol. 2, no. 9, pp. 1041–1050, 2000.

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

[3] Health Map Details, http://www.healthmap.org/dengue/en/.

[4] http://articles.timesofindia.indiatimes.com/.

[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 Biomedical Circuits and Systems*, vol. No, pp. 20–23, December 2004.

[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] S. Nimmannitya, "Dengue and dengue hemorrhagic fever diseases," in *Proceedings of the 1st International Conference on Bioinformatics*, vol. 2, pp. 765–772, Singapore, 2003.

[8] S. Fahri, B. Yohan, and Hidayat, "Molecular surveillance of dengue virus serotype-1," *International Journal on Molecular Biology*, vol. 2, no. 1, pp. 345–349, 2013.

[9] M. Ahdesmaki, H. Lahdesmaki, and O. Yli-Harja, "Robust Fisher's test for periodicity detection in noisy biological time series," in *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics*, pp. 1–4, Tuusula, Finland, June 2007.

[10] http://www.ncbi.nlm.nih.gov/guide/dna-rna/.

[11] http://www.nature.com/news/2002/020415/full/news020415-10.html.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] V. Pujeri and G. M. Karthik, "Constraint based periodicity mining in time series databases," *International Journal on Computer Network and Information Security*, vol. 4, no. 10, pp. 37–46, 2012.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Elsevier, 2nd edition, 2007.