2012 International Conference on Bioscience, Biochemistry and Bioinformatics
IPCBEE vol.31(2012) © (2012)IACSIT Press, Singapoore

Learning the Diversity and Evolutionary Pattern of the Dengue Virus

M. M. A. T. N. Mannapperuma[+], M. W. A. C. R. Wijesinghe, A. R. Weerasinghe

University of Colombo School of Computing, Colombo, Sri Lanka

Abstract. Dengue, a mosquito borne infection caused by a family of viruses, poses significant health threats throughout the developed and the developing world. There are four serotypes of Dengue virus that cause the disease. Phylogenetic studies have identified genetic subtypes within each of the four Dengue virus serotypes. The Dengue virus differs in each subtype, making the dynamics of the Dengue virus population complex.

The rapid evolution processes of the Dengue virus dramatically increase the occurrences of a diversity of subtypes of the virus. However, the process of evolution has left patterns of relationships between Dengue virus subtypes. Our goal in this study was to learn and understand Dengue virus' evolutionary pattern and to predict its future evolution.

To better understand the evolutionary mechanism of the Dengue virus, we examined a data set consisting of 1191 nucleotide sequences of the envelope protein gene of the Dengue virus. A Phylogenetic tree was used to trace the origin of the selected Dengue viruses from this data set. We used association rules, bigrams and trigrams to extract frequently occurring evolutionary patterns in each position of the gene. Evolutionary patterns extracted using the above mentioned techniques were then evaluated. The association rule technique achieved an accuracy rate of 0.994 in predicting upcoming evolutionary shifts. Both bigram and trigram techniques achieved approximately 0.778 accuracy rates for predicting these patterns. Our experimental results conclude that, there is an evolutionary pattern in the Dengue virus which could be exploited for treating patients and halting epidemics.

Keywords: Dengue; evolution; pattern extraction

1. Introduction

Dengue, also known as "break-bone fever" or "dandy fever" is an infectious disease transmitted by mosquitoes [8]. Dengue is caused by a family of viruses and is transmitted to humans through the bites of infective female "Aedes" mosquitoes [8]. Dengue poses significant health threats throughout the developed and developing world.

The Dengue virus is a single stranded, positive sense RNA virus [9]. The virus has a genome of about 11,000 bases (nucleotides). The Dengue virus codes for a single polyprotein and produces ten viral proteins; three structural proteins (Capsid-C, precursor membrane-prM, and Envelope-E), and seven non-structural proteins (NS1, NS2a, NS2b, NS3, NS4a, NS4b, and NS5) [9]. There are four antigenically distinct, but closely related serotypes of the Dengue virus referred to as DEN-1, DEN- 2, DEN-3, and DEN-4 [12] [8]. It is possible for one person to get Dengue fever multiple times as it could be caused by any one of these four serotypes of the virus [8] and each serotype is sufficiently different that there is no cross protection. Although these four serotypes are different strains of the same Dengue virus, they have approximately 60%-80% homology between them. Phylogenetic studies have identified genetic subtypes within each of the four Dengue virus serotypes.

[+] Corresponding author. Tel.: +94 71 8096766; fax: +94 11 2695650.
E-mail address: nmannapperuma@gmail.com.

The Dengue virus has a high evolutionary rate since positive sense RNA viruses evolve rapidly [4]. The occurrence of subtypes of the four serotypes of the virus has risen dramatically in recent years, increasing the genetic diversity of the virus. The evolution of the Dengue virus has a major impact on their increased virulence as a disease-causing organism in humans, and on the epidemiology of the Dengue disease around the world. Three main contributing factors to the genetic diversity of the Dengue virus are:

- Natural evolution of the virus
- Mutation of the virus
  - towards resistance against treatment
  - towards the state of the immune system of the patient
- Recombination

The Dengue viruses' natural RNA replication process allows production of many different subtypes of its kind.

Mutation is a permanent change in the RNA sequence. Viruses have high mutation rates and so can adapt to changing environments very well. Environmental parameters include external influences such as the presence of drugs and the state of the immune system of the patient. These factors have an important impact on the evolution of the virus.

Many RNA viruses have the capacity to exchange genetic material with one another. Recombination is the genetic exchange between different Dengue virus strains [1]. A single sequence can be produced by genetic exchange between viruses belonging to several different subtypes. As a result, new types of biologically important Dengue viruses may be generated.

The Dengue virus differs in each subtype. Thus, there is a complex dynamics at play in the Dengue virus population. Variants in viruses cause the dynamics in the disease. Some virus subtypes cause more disease than other subtypes. Therefore, a better understanding about the variants of the Dengue virus, and about how variants are caused, will help in detecting the type of virus in each case. This in turn will help to control the virus and to create better treatments such as drug and vaccine formulation.

Identifying Dengue virus subtypes is important. Identifying and learning about Dengue viruses' evolutionary shifts is also important. Knowledge of Dengue virus evolution will help to predict future evolutions of the Dengue virus, which will be very important in treating patients infected by Dengue virus and in controlling such virus infection. It will also help to detect new virus types and to create treatments in a proactive way.

In summary, we observed that Dengue virus sequences can result from three different evolutionary processes; natural evolution, resistance mutation and recombination. Our goal in this study was to learn this diversity and to understand the Dengue viruses' evolutionary pattern so as to predict its future evolution.

2. Methodology

There are four sufficiently different serotypes of the Dengue virus. Phylogenetic studies have identified subtypes within each of these Dengue virus serotypes. Owing to its high rate of evolution, the occurrence of the four serotypes of Dengue virus has risen dramatically. Viruses evolve and create many different variants of its kind, making the dynamics in the virus population very complex. In summary, its inherent evolutionary process increases the diversity of the Dengue virus.

In this study, we were interested in understanding the Dengue virus's life through time - not just at one time in the past or present, but over long periods of time. That is to understand the connections between all Dengue virus strains and subtypes as evident from ancestor/descendant relationships in its evolution. Hence we deal with the problem of identifying and understanding the relationships between the many different kinds of Dengue viruses recorded.

A. Hypothesis

"There exists an evolutionary pattern in the Dengue virus."

The above hypothesis is an assumption about the Dengue virus population. This assumption may or may not be true. We determine whether the hypothesis is true or not by examining the sample of the Dengue virus population available in public databases.

B. Approach

To test whether there is any pattern in Dengue virus evolution; the virus's RNA strains were analyzed. The Analysis consisted of four main steps. We collected a set of Dengue virus sequences from the database. Next we obtained the evolution history of the Dengue virus from the collected data. Then we analyzed the evolution history to identify the pattern of evolution. Finally, we predict future evolutions if any evolutionary pattern exists.

C. Data

To better understand the mechanisms underlying the pathogenicity of the Dengue virus, we examined a data set consisting of nucleotide sequences of the envelope gene of the Dengue virus. The Primary data set was retrieved from the Dengue Virus Database (DengueDB) [11]. The total data set consists of 1666 complete nucleotide (RNA) sequences of the Dengue virus envelope gene.

The data file had sequences with different IDs but with the same nucleotide sequence. It was therefore filtered to remove duplicate sequences based on the nucleotide sequence, rather than on the sequence ID. 475 duplicate sequences were removed from the primary data set. The filtered data file with 1191 nucleotide sequences was used for further processing.

Two samples (sample I and sample II) were drawn from filtered primary data set. 596 sequences were included in sample I, with remaining sequences marked as sample II.

D. Evolution

Evolution is technically defined as *a gradual process in which something changes into a different and usually more complex or better form*. The process of evolution produces patterns of relationships between living organisms. As lineages evolve and split, and modifications are inherited, their evolutionary paths diverge. This produces a branching pattern of evolutionary relationships.

To learn the evolution of the Dengue virus, we need to keep track of evolutionary relationships among various Dengue virus types. The Phylogenetic or Evolutionary tree is a branched diagram that helps in visualizing the evolutionary relationships among different entities [5]. In a Phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants. A Phylogenetic tree was used in this study to trace the origin of the envelope protein gene of the Dengue virus.

Any phylogenetic tree construction method should follow three main steps [2]. Multiple sequence alignment is the essential preliminary step to tree construction. After an accurate alignment of homologous sequences is discovered, the alignment data needs to be converted into a phylogenetic tree. Finally we need to assess the accuracy of the reconstructed tree.

1) Multiple Sequence Alignment: The Dengue virus sequences of the primary data set (sample I and sample II) were aligned using ClustalW (version 2.0.12) [3] to reveal the degree of sequence conservation.

2) Phylogenetic Tree Construction: The Phylogenetic tree should be a rooted tree, in which the path from the root to a node represents an evolutionary path. The DNAMLK (DNA Maximum Likelihood program with molecular clock) program in the PHYLIP (version 3.69) [6] package was used to generate the Maximum Likelihood rooted phylogenetic trees from the aligned data sets.

The maximum likelihood rooted phylogenetic tree was reconstructed using the DNAMLK program for sample I, sample II and for the entire primary data set.

In order to extract the evolutionary pattern we need to represent the inferred common ancestors. We inferred ancestral sequences in the phylogenetic trees using the maximum likelihood method.

- The Phylogenetic tree constructed using sample I resulted 1470 sites in the final sequences
- The Phylogenetic tree constructed using sample II resulted 1474 sites in the final sequences
- The Phylogenetic tree constructed using the full primary data set resulted 1468 sites in the final sequences

3) Assess the accuracy of constructed phylogenetic tree: Several phylogenetic studies done earlier have categorized Dengue viruses into four serotypes [12] [8]. We used this fact to assess the accuracy of the estimated phylogenetic trees. Each Dengue virus sequence of the same serotype was grouped together in all three trees obtained, implying that we have obtained the correct tree topologies for all data samples.

3. Pattern Extraction

The term pattern often refers to *recognizable regularity*. The idea of frequent pattern discovery is to find *recurring events that occur in a predictable manner*. Data mining techniques can be used as a model or template in predicting upcoming events. In this study, we attempted to extract frequently occurring evolutionary events in the Dengue virus to predict its upcoming evolutionary events.

Mutation is a permanent change in the nucleotide sequence. Mutations in the Dengue virus are caused as a result of its evolutionary process, which creates many different types of its kind. The primary goal of this study is to identify mutations that appear to play a major role in the initiation or propagation or survival of the Dengue virus. In simple terms, we try to identify mutations contributing to the development of new Dengue virus strains.

The basic assumption of many phylogenetic tree construction models is that each site in the sequence evolves independently from every other site in the sequence [7]. This assumption simplifies our project problem. We assume that the above assumption is valid; therefore we conducted our analysis independently for each site in the sequence.

We first examined the nucleotide substitution frequencies (number of nucleotide substitutions per site). The substitution rates are not uniform among sites, but nor is a high substitution rate variation observed among sites. We can conclude that the Dengue virus demonstrates relatively similar rates of evolution at different sites in the nucleotide sequence.

1) Association Rules for Pattern Discovery: Association rules were used to extract evolutionary patterns. We examined the behavior of the Dengue virus in terms of the relationship between ancestors and descendants at a given site. For example, associations rule A→CT (P) states that the ancestor A has mutated to C and T descendants with a probability of P at this site. We computed the probability (P) for each association rule. The Phylogenetic tree constructed using the primary data set was used in computing the probabilities for association rules. We extracted an association rule for each ancestor that resulted in the highest probability of occurrence, as a pattern of the Dengue virus evolution at that site in the sequence.

2) Bigrams for Pattern Discovery: Bigrams were used to represent patterns among lineages. Bigrams are groups of *two* adjacent nucleotide symbols in the evolution of the virus in a given site. We used bigrams to simulate dependency relationship between ancestors and descendants. For example the bigram TC was used to represent the relationship between ancestor T with its descendant C. The conditional probability of each bigram was calculated for the phylogenetic tree constructed using sample I dataset and for the phylogenetic tree constructed using sample II data set separately.

3) Trigrams for Pattern Discovery: We successfully used trigrams for evolutionary pattern extraction. Trigrams are groups of *three* symbols. We used trigrams to simulate the behavior of the Dengue virus in terms of the relationship between ancestors with its immediate descendents. For example trigram ATC was used to represent the relationship between ancestor A with its immediate descendents T and C. The conditional probability of each trigram was calculated for the phylogenetic tree constructed using sample I dataset and for the phylogenetic tree constructed using sample II.

The pattern extraction algorithm for bigrams and trigrams was as follows:

```
while( Get the next site )
   while( Get the next pattern (bigram/trigram) group )
      patternOne = pattern (bigram/trigram) with highest conditional probability in sample I
      patternTwo = pattern (bigram/trigram) with highest conditional probability in sample II
      if(patternOne == patternTwo)
         Record pattern as an evolutionary pattern
      //end if
   //end while
//end while
```

The above pattern extraction algorithm identified 1885 bigrams (evolutionary patterns) at 1182 sites and 1885 trigrams (evolutionary patterns) at 1181 sites in the sequences.

4. Pattern Evaluation

A. Association Rule Evaluation

The length of the full phylogenetic tree was reduced by one before the association rules were extracted. The extracted association rules were then used in predicting the next evolution (which was removed in the previous step) of the Dengue virus. These predicted sequences were then compared with the original virus sequences (removed in the first step). The association rules correctly predicted upcoming evolutionary shifts in 355,819 cases, while the prediction was wrong in only 1,805 cases. Hence, the association rule technique can be said to have achieved an accuracy rate of 0.994 (or 99.4%).

B. Bigram/Trigram Evaluation

If a particular evolutionary pattern exists in both samples (sample I and sample II) we draw a conclusion that the pattern is also visible in the overall Dengue virus population. Thus, they should appear in the primary data set (publicly available Dengue virus population). The probability of occurrence of each extracted pattern (bigram or trigram), in the Dengue virus population was recorded.

Figure. 1 represents the probability of occurrence of each bigram in the primary data set. Figure. 2 represents the probability of occurrence of each trigram in the primary data set. Bigrams achieved an accuracy rate of 0.778 whereas trigrams achieved an accuracy rate of 0.775.

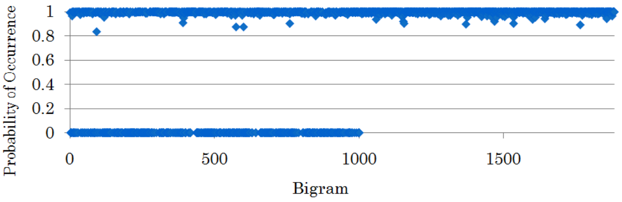

Fig.1 Probability of occurrence of each bigram in the primary data set.

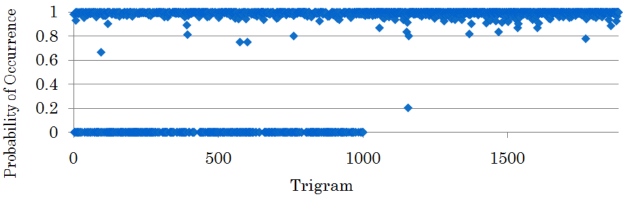

Fig 2. Probability of occurrence of each trigram in the primary data set.

5. Conclusion

Dengue fever is an infectious disease transmitted by mosquitoes. There are four serotypes of Dengue viruses that cause Dengue. Studies have identified genetic subtypes within each of the four Dengue virus serotypes. The Dengue virus differs in each subtype, which makes the dynamics in Dengue virus population very complex.

The process of evolution results in patterns of relationships among living organisms. The Dengue virus has a high evolutionary rate and thus, the occurrence of the four subtypes of serotypes of Dengue virus has proliferated dramatically. Identifying and learning about the Dengue virus subtypes and their evolutionary shifts is important. Knowledge of Dengue virus evolution will help in predicting its future evolutions. Our goal in this study was to learn and understand the evolutionary pattern of Dengue virus in order to predict its future evolutions.

To test whether there is any pattern in Dengue virus evolution; nucleotide sequences of the envelope (E) gene of the Dengue virus were analyzed. The process consisted of four main steps. First, a set of Dengue virus sequences was collected from the DengueDB. Next, we generated a rooted phylogenetic tree to trace the origin of the Dengue virus from this data. Thirdly, we discovered frequently occurring evolutionary events from the Dengue virus' evolutionary process to find recurring events that occur in a predictable manner. We used association rules, bigrams and trigrams to examine the behavior of the Dengue virus in terms of the relationship between ancestors and descendants of each site in its Envelope gene nucleotide sequence to extract frequently occurring evolutionary patterns. Finally, the extracted evolutionary patterns were evaluated for their ability to accurately predict future evolutions. Patterns obtained using association rules obtained an accuracy rate of 0.994 in predicting the next generation of the Dengue virus' Envelope gene. Patterns extracted using the other two methods had approximately accuracy rates of 0.78 in such prediction.

From the experimental results obtained, we can conclude that there exists an evolutionary pattern in the Dengue virus.

6. References

[1] A. R. Michael Worobey and E. C. Holmes, “Widespread intra-serotype recombination in natural populations of dengue virus,” in PNAS: Proceedings of the National Academy of Science of the United States of America, vol. 96. National Academy of Sciences, June 1999, pp. 7352-7357.

[2] C. Notredame, “Recent progresses in multiple sequence alignment: a survey,” Pharmacogenomics, vol. 3, no. 1, pp. 131-144, January 2002.

[3] D. G. Higgins, Julie D. Thompson and T. J. Gibson, “Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice,” Nucleic Acids Research, vol. 22, no. 22, pp. 4673-4680, November 1994.

[4] Gerardo Perez-Ramirez, Alvaro Diaz-Badillo, Minerva Camacho-Nuez, Alejandro Cisneros, and Maria de Lourdes Munoz, “Multiple recombinants in two dengue virus, serotype-2 isolates from patients from oaxaca, mexico,” BMC Microbiology, vol. 9, no. 260, pp. 260+, December 2009.

[5] J. Felsenstein, Inferring Phylogenies. Sinauer Associates, 2003.

[6] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.69, Distributed by the author, 2005.s

[7] L. M. F. Merlo, M. Lunzer, and A. M. Dean, “An empirical test of the concomitantly variable codon hypothesis,” Proceedings of the National Academy of Sciences, vol. 104, no. 26, pp. 10 938-10 943, June 2007.

[8] National Institute of Allergy and Infectious Diseases, “Dengue fever,” Accessed: October. 2009. [Online].Available: http://www3.niaid.nih.gov/topics/DengueFever/

[9] R. J. Sugrue, Glycovirology Protocols (Methods in Molecular Biology). Humana Press, 2007.

[10] T. Gesell and A. von Haeseler, “In silico sequence evolution with site-specific interactions along phylogenetic trees,” Bioinformatics, vol. 22, no. 6, pp. 716-722, March 2006.

[11] Viral Bioinformatics Resource Center, “Information about dengue virus,” Accessed: January. 2010. [Online]. Available: http://www.denguedb.org/dengue info.asp

[12] World Health Organization, “Dengue and dengue haemorrhagic fever,” Accessed: October. 2009. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs117/en/