Subtyping of Dengue Viruses Using Return Time Distribution Based Approach

Pandurang S Kolekar[1], Mohan M Kale[2], Urmila Kulkarni-Kale[1]

[1]Bioinformatics Centre, University of Pune, Pune 411 007, India.
[2]Department of Statistics, University of Pune, Pune 411 007, India.

Abstract

Dengue virus (DENV) is the causative agent of Dengue Hemorrhagic Fever and Dengue Shock Syndrome, and continuing to represent major public health hazard. DENVs are antigenically classified in four serotypes and each serotype is - further divided into respective genotypes. The association between DENV subtypes and the kind & severity of disease caused by them is known. Experimental and computational approaches for subtyping are routinely used for the purpose of diagnosis and treatment of DENV, in addition to study of phylodynamics. All virus-specific molecular subtyping tools make use of sequence alignments at backend. But as the volume of molecular data increases, alignment dependent methods become computationally intensive. Hence, the need of alternative efficient approaches for subtyping of viruses becomes apparent. Recently, the concept of Return time distribution (RTD) was proposed and validated for alignment-free clustering and molecular phylogeny. The RTD-based approach is extended here for the subtyping of DENVs.

Subtyping methodology involves compilation of curated genomic data of known subtypes, computing RTD of these sequences at different levels of k-mers, derivation a distance matrix and clustering. The subtype of the unknown is predicted based on its clustering with known ones.

Dataset consisting of 1359 DENV genomes with sequence identity (>92%) were clustered using RTD based approach at k=5. Serotype specific clades, despite of the geographical and temporal variation, in the dataset, were observed with 100% accuracy. The method was also found to be efficient in terms of time and implementation, apart from accuracy in subtyping of DENV.

Background

- Genotyping of viruses is of clinical importance

The diagnosis and treatment of viral diseases is driven by the genotype of infected virus e.g. Dengue, Hepatitis etc. [1].

- Existing methods of viral genotyping

In addition to experimental methods, current computational methods for genotyping involves multiple sequence alignment, bootstrapping and subsequent phylogeny analysis.

- Limitations of existing methods

Computationally intensive with increasing data and time consuming

- Challenges: Availability of genomic sequence data due to next generation sequencing technology [2]

- Thus, there is a need of alternative methods

Background ...

- Global threat of Dengue

According to the estimation by the World Health organization, each year 500 000 people are suffered from Dengue infection, in more than 100 countries, and about 2.5% of those affected die.

- Taxonomy of Dengue virus

Dengue virus (DENV), a single stranded positive sense RNA virus belonging to the genus *Flavivirus* of the family *Flaviviridae* with a genome of approximately 11 kb Antigenically there are four serotypes of Dengue virus (1-4) and each serotype is further genetically divided into their respective genotypes.

- Host and vector

Humans are the major mammalian host for the DENV transmitted via peridomestic mosquito species, *Aedes aegypti*.

DENV infection causes dengue fever (DF), lifethreatening dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS).

Return Time Distribution (RTD)

- What is "Return time distribution (RTD)"?
  - The time required for the reappearance of particular state without its epoch in between
  - RTD in the context of nucleotide sequence: time required for the reappearance of particular base or k-mer
- Computation of RTDs of mononucleotides

Sample nucleotide sequence

CTACACAACTTTGCGGGTAGCCGGAAACATTGTGAATGCGGTGAACA

RTD of "A" and "T" at k=1 (mononucleotide)

Return time for A (X)	Frequency (F)
0	5
1	4
5	2
7	1
10	1

Return time for T (X)	Frequency (F)
0	3
1	1
3	1
4	1
5	1
7	1
11	1

- Similarly RTDs for k-mers can be calculated

Method

- Deriving Statistical parameters of RTDs

For each RTD two statistical parameters, mean (μ) and standard deviation (σ) were computed.

In general, at k-level each sequence will be represented by a numeric vector of size $2*4^k$ i.e. at k=2 (di-nucleotides) the size of the vector would be 32 and so on for other integer values of k.

- Distance function between genomes based on parameters of RTDs

$$D_{ij} = (\sum [G_{ir\mu} - G_{jr\mu}]^2 + \sum [G_{ir\sigma} - G_{jr\sigma}]^2)^{1/2} \qquad (1)$$

Where,

r stands for the RTD of particular k-word, μ and σ are the mean and standard deviation of respective RTDs. For k=1 there are four (4^1) possible RTDs; r belongs to {A, T, G, C} RTDs.

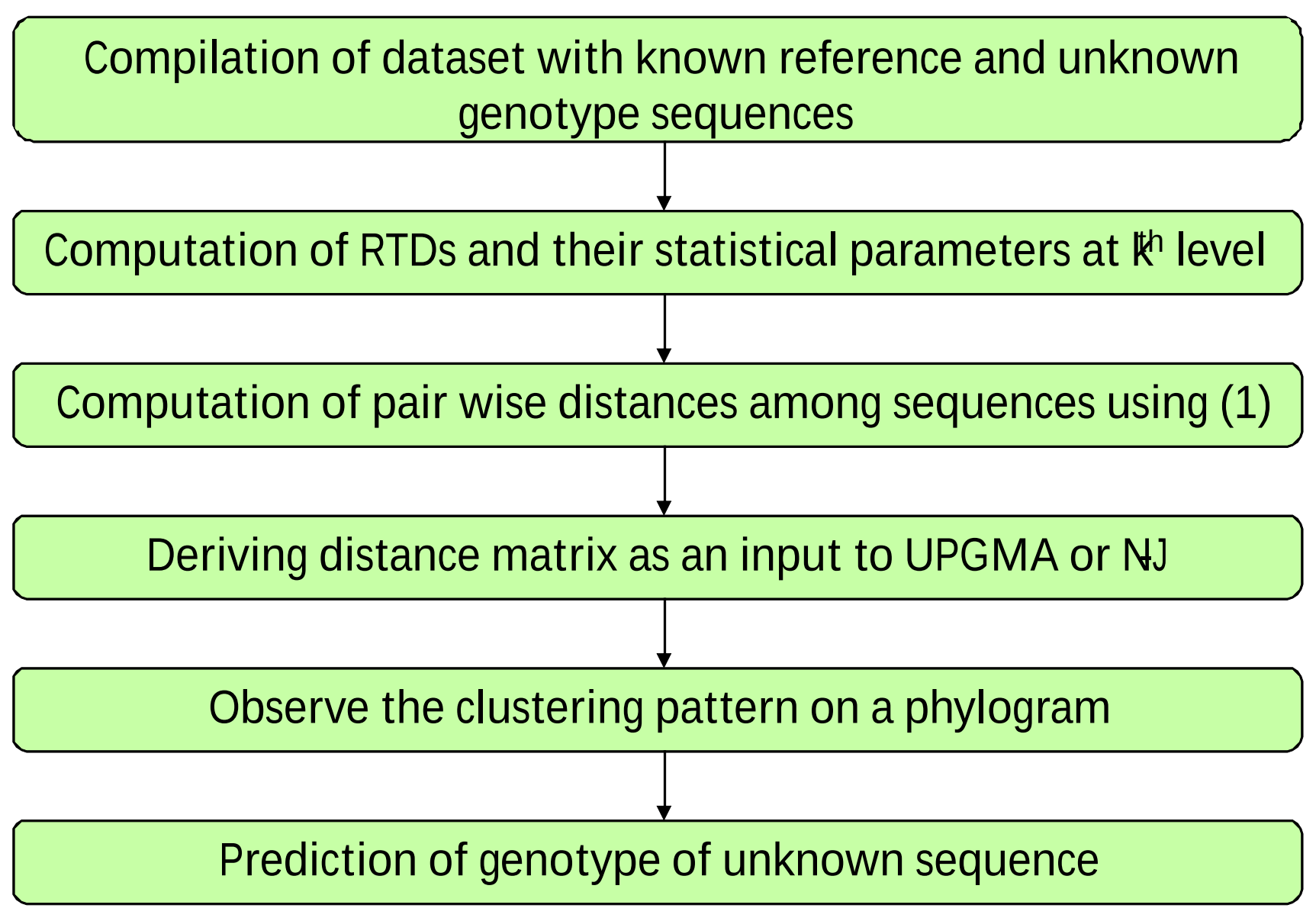
- Computation of Distance matrix

Pair wise distances among the sequences in dataset under study can be computed using (1).

The distance matrix thus obtained can be given as an input to distance based UPGMA or NJ method for the inference of phylogenetic tree [3].

The clustering of unknown sequence with known reference genotypes on a tree is used for its genotyping.

Flowchart



Compilation of dataset with known reference and unknown genotype sequences

Computation of RTDs and their statistical parameters at k^{th} level

Computation of pair wise distances among sequences using (1)

Deriving distance matrix as an input to UPGMA or NJ

Observe the clustering pattern on a phylogram

Prediction of genotype of unknown sequence

Materials

- Datasets

Reference dataset consists of 30 representative genome sequences of Dengue viruses (4 Serotype and their respective genotypes).

This dataset was also used as a reference dataset for genotyping of Dengue viruses at Viral Bioinformatics Resource Center [4].

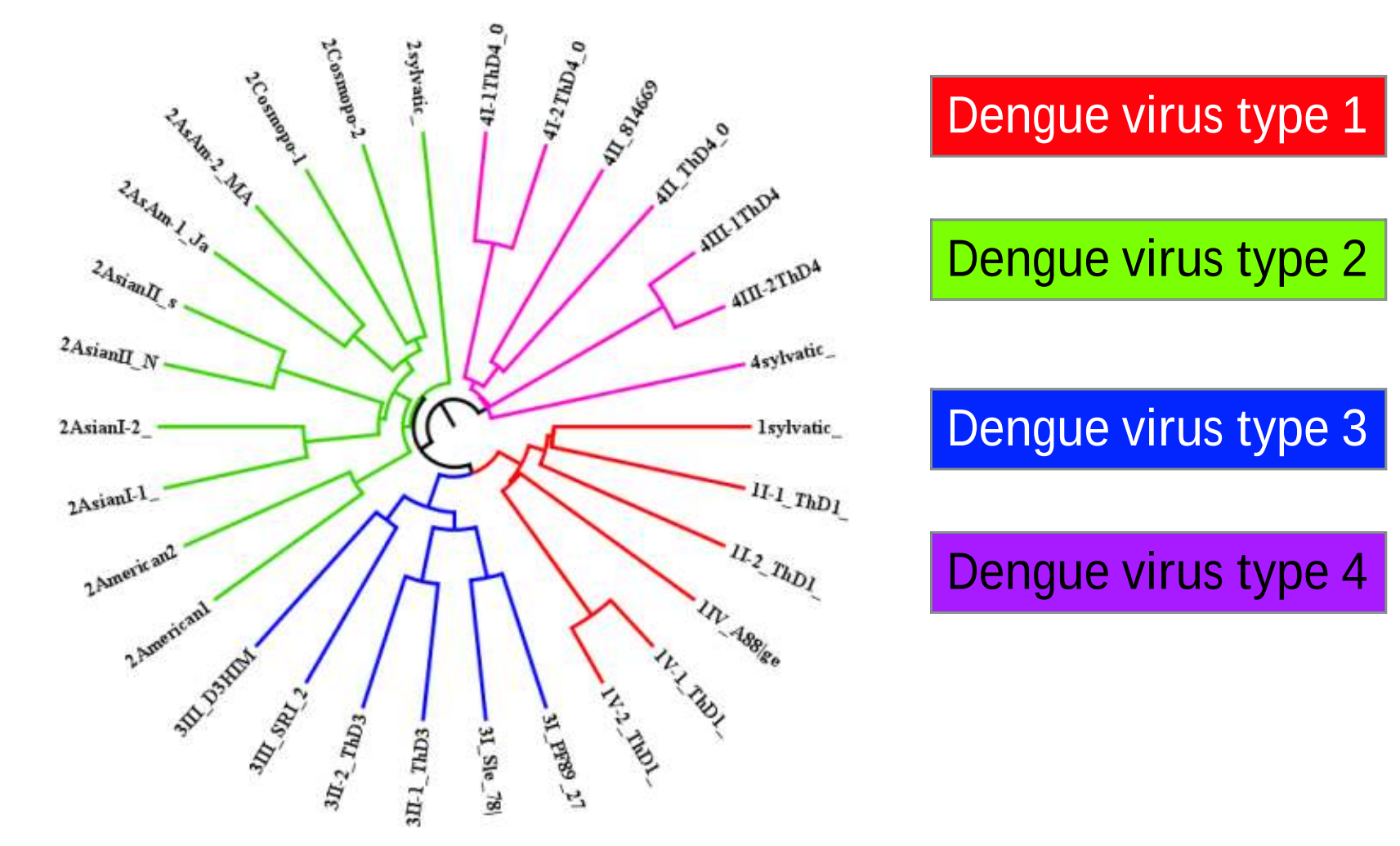
All available, 1359, Dengue virus genomes (~11 kb each) were downloaded from Dengue Virus resource at NCBI [5]. Out of them 956 were found to be non-recombinant as predicted with 100% confidence by Dengue genotype determination tool at VBRC.

This test dataset of 956 non-recombinant genomes was used to assess the validity of proposed alignment-free approach.

RTDs for the chosen datasets were computed for different values of k ranging from 1 to 5. For each value of k the distance matrix obtained using (1) was given as an input to UPGMA method assembled in Neighbor program of PHYLIP package [6]. Trees were visualized using FigTree [7].

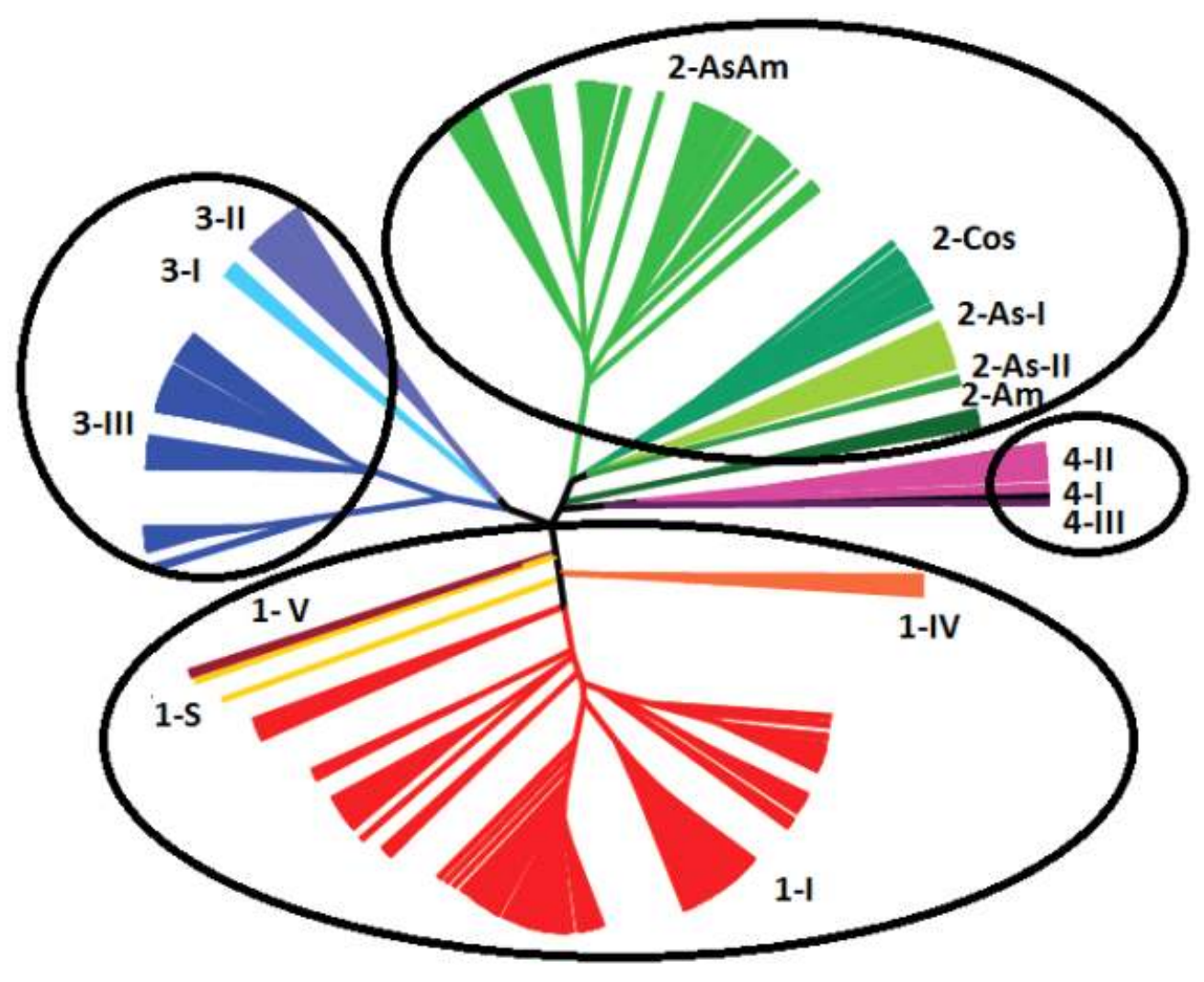
Results

- UPGMA phylogram obtained for reference dataset at k = 5.



Similar results were obtained using NJ method

UPGMA phylogram obtained for test dataset at k = 5.



Abbreviations

- S: Sylvatic
- As: Asian
- Am: American
- Cos: Cosmopolitan
- AsAm: Asian-American

- RTD statistics at k = 5

Total number of RTDs = 4^5 i.e. 1024

Size of numeric vector for each genome = 2*1024 (2 parameters of each RTD, μ and σ)

Time required for computations using PERL script:

Reference dataset: ~3 seconds.

Test dataset: ~40 minutes

(System configuration: 32-bit OS, 2.80GHz processor, 4 GB RAM)

- Advantages of the proposed method

No need of sequence alignment

Computationally efficient

Equally accurate

Faster than alignment-based methods

Conclusions

- To the best of our knowledge, this is the first application of the concept of RTD from stochastic process theory for genotyping.
- RTD is capable of catching the pattern and provides unique representation for genomic sequences.
- It is observed that RTD based alignment-free method successfully genotypes DENV. This robust approach can be extended for the genotyping of other viruses.
- The initial results are encouraging. Work is under progress for the genotyping of other viruses using RTD.

References

1. Balmaseda et al. (2006) Serotype-specific differences in clinical manifestations of Dengue, Am J Trop Med Hyg, 74, 449-456.
2. Voelkerding , K.V., Dames, S.A. and Durtschi, J.D. (2009) Next-Generation Sequencing: From Basic Research to Diagnostics, Clin Chem, 55, 641-658.
3. Kolekar PS, Kale MM, Kulkarni_Kale U (2010) `Inter_Arrival Time' Inspired Algorithm and its Application in Clustering and Molecular Phylogeny.
4. *AIP Conference Proceedings*, 1298(1):307-312.
5. Viral Bioinformatics Resource Center: [http://vbrc.org/index.asp]
6. Resch W, Zaslavsky L, Kiryutin B, Rozanov M, Bao Y, Tatusova T (2009) Virus variation resources at the National Center for
7. Biotechnology Information: dengue virus. *BMC Microbiology*, 9(1):65.
8. PHYLIP: [http://www.phylip.com/]
9. FigTree: [http://tree.bio.ed.ac.uk/software/figtree/]

Acknowledgments

PSK acknowledges the BINC (BioInformatics National Certification) fellowship awarded by Department of Biotechnology, Government of India. This work is financially supported by Department of Biotechnology, Government of India under Centre Of Excellence (COE) grant to Bioinformatics Centre, University of Pune, Pune.

Contact:
PSK: pandurang@bioinfo.ernet.in
MMK: mmkale@stats.unipune.ac.in
UKK: urmila@bioinfo.ernet.in

Homepage: http://bioinfo.ernet.in