Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Trends in Bioinformatics

ISSN 1994-7941
DOI: 10.3923/tb.2018.1.6

Research Article
ArchaeVir: A Comprehensive Genometrics Database of Archaeal Viruses

[1]Krishna Kumar Ojha, [2]Swapnil Mishra and [3]Lalit Kumar Pandey

[1]Centre for Biological Sciences (Bioinformatics), Central University of South Bihar, Patna, India
[2]Centre for Bioinformatics, IIDS, University of Allahabad, Allahabad, India
[3]Department of Chemistry, Indian Institute of Technology (BHU) Varanasi, Varanasi 221005, India

Abstract

Background and Objective: The era of high throughput sequencing technology has poured thousands of full genomes sequences in public databases. A remarkable portion of these genomic sequences belongs to the viruses of which viruses infecting archaea are least studied group. Archaeal viruses are morphologically more diverse than bacterial ones and display unique morphotypes. All sequenced archaeal viruses have DNA as their genetic material, which is double stranded in most of the cases. In this study, applied biometric method to understand the local and global feature of nucleotide distribution in all available archaeal genomes on Genbank. ArchaeVir database is dedicated to characterization of archaeal virus genomes with standardized genometrics of Z-curve and cumulative GC and AT skews. **Material and Methods:** All archaea viruses genomes were downloaded from NCBI. A script was written in MATLAB to calculate and plot the cumulative Z-curve and nucleotide skews. Database was created using MySQL and PHP, AJAX was used to create front end of the database. **Results:** Genometrics is a biometric analysis of chromosomes which is capable of identifying, at the level of whole genomes, features inherent to chromosome organization and functioning. This database also hosts the basic genomic records like genome length, GC%, total genes and proteins all sequenced viral genomes of archaea. In addition to this, two tools, one for finding the repeats and nucleotide distribution and second for plotting the geometric skew of nucleotide sequence has also been integrated with this database. **Conclusion:** Genometrics data of archaeal viruses available on the database could be readily accessed by the scientific community for further analysis of local and global features of viral genomes as well as for tutorial purpose.

Key words: Archaeal viruses, Z-curve, nucleotide skew, BLOB, DNA walk

Citation: Krishna Kumar Ojha, Swapnil Mishra and Lalit Kumar Pandey, 2018. ArchaeVir: A comprehensive genometrics database of archaeal viruses. Trends Bioinform., 11: 1-6.

Corresponding Author: Krishna Kumar Ojha, Centre for Biological Sciences (Bioinformatics), Central University of South Bihar, Patna, India

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Over the past few years, the viruses of prokaryotes have been transformed in the view of microbiologists from simply being convenient experimental model systems into being a major component of the biosphere. They are the global champions of diversity and constitute a majority of organisms on the planet, they have large roles in the planet's ecosystems, who exert a significant force on the evolution of their bacterial and archaeal hosts and they have been doing this for billions of years, possibly for as long as there have been cells[1]. The viruses infecting the archaea constitute a small portion of total viruses whose full genomes is available on public database. Many of these viruses, specifically the species that infect hyperthermophilic hosts, display several morphotypes for example, bottle shaped, spindle shaped, droplet shaped, coil shaped and bacilli form[2]. Moreover, the ways in which these viruses interact with their archaeal hosts are also unique, as indicated by a unique virion progress mechanism, which involves formation of pyramidal portion on the cell surface[3]. All sequenced archaeal viruses genomes exclusively carry double stranded deoxyribonucleic acid (DNA) and only a few species have single stranded DNA. Remarkably, the virosphere has now been shown to extend to almost every known environment on earth, including the extreme acidic, thermal and saline environments where archaeal organisms can be dominant. Thus, because of their abundance and variety, viruses were now thought to represent the greatest reservoir of genetic diversity on the planet[4].

Archaeal viruses were growing appreciation for their role in organic evolution. Remarkably with >500 cellular genomes sequenced to date, most show a significant amount of viral or virus-like sequence within their genome, further evidence that viruses play a central role in horizontal gene transfer and help drive the evolution of their hosts[4]. Current hypotheses contend that viruses have catalyzed several major evolutionary transitions, including the invention of DNA and DNA replication mechanisms[5] and thus a role in the formation of the three domains of life. The genome of archaeal viruses were point of interest due to their exceptional habitat and higher rate of evolution[6]. Out of several techniques available to analyses the genomic sequences *in silico* analysis have surpassed all other due to ease of availability and short time requirement. Genometrics is one of the important *in silico* method to analyses the DNA sequences of any organism. Although several databases available for the study of the genometrics of archaea genomes but there is no dedicated database to host the comparative genomics of all archaeal viruses[7-10]. The development of ArchaeVir database is an effort to fill this lacuna and provide a dedicated platform to analyze and compare the Z-curve and cumulative amino and keto skews of all sequenced archaeal viruses.

Experiments have shown that the Z-curve can be used to identify the replication origin in various prokaryotes[11] and specially in several archaea[12-13]. One study analyzed the Z-curve for multiple species of archaea and found that the OriC is located at a sharp peak on the curve followed by a broad base. This region was rich in AT bases and had multiple repeats, which was expected for replication origin sites. This and other similar studies were used to generate a program that could predict the origins of replication using the Z-curve.

The Z-curve has also been experimentally used to determine phylogenetic relationships of *Coronaviruses*. It was determined that similarities and differences in related species can be quickly determined by visually examining their Z-curves[14]. An algorithm was created to identify the geometric center and other trends in the Z-curve of 24 species of *Coronaviruses*. The data were used to create a phylogenetic tree. The results matched the tree that was generated using sequence analysis. The ArchaeVir database hosts the Z-curve of all sequenced archaeal viruses and thus will help the biologists and scientist to investigate and infer the local and global feature of archaeal viruses' genomes.

MATERIALS AND METHODS

All viral genomes were downloaded from the viral genome section of NCBI. A MATLAB script was written to analyses genome and plot Z-curve and skew of the all genomes. A relational database MySQL is used as the primary DBMS for storing the genomic and graphical data for rapid retrieval and easily maintainable. PHP, HTML along with AJAX is used to create the front-end GUI for accessing database by the user. To generate real-time nucleotide skew from the database BioPHP code has been incorporated. To find the direct repeats and nucleotide distribution percentage PHP library has been used. The database front end is dynamic and adaptive to all size screen devices (Fig. 1).

Fig. 1: Schematic diagram of archaeVir database

RESULTS AND DISCUSSION

User interface: The ArchaeVir database searching is very interactive and simple for the user. There were several independent menus to directly go the particular section of the database like tools, method, archaeal viruses and contact pages. There were two search boxes on home page for database search with AJAX auto-suggest functionality, it suggests up to 10 most similar organism names filled by the user my matching the similar word in the database (Fig. 2). This auto suggest feature nullifies the need to memories complex full name and correct spelling of archaeal viruses. There was another search box through which database can be searched using the NCBI-ID of the viral genome. User can use search database using either by virus name or corresponding NCBI-ID of the archaeal virus.

The landing page of the query search gives two diagrams one was the Z-curve in 3-D shape and the other was the X and Y component of the Z-curve plotted on the same graph in 2-D plane. The X and Y components are plotted in two different color i.e., Red and blue colors, respectively (Fig. 3).

All data has been stored in a single table in MySQL database for fast access. There was a separate table for the storage of genometrics image in BLOB data type. PHP is used to link the web front with the database backend. BLOB data type is a collection of binary data stored as a single entity in a database management system. Blobs are typically images, audio or other multimedia objects. There are two tools which have been integrated with database to enhance usability of ArchaeVir database. One for finding repeats and other for generating the genometrics graph of the given nucleotide sequences.

Repeat finder tool written in PHP to find direct repeat of the given signature and it's accurate place in the genomic sequence, in addition to that this tool was also equipped to calculate the nucleotide frequency on user choice.

Skew plot generator code has been incorporated form BioPHP website which is an open source consortium for developing biological tools written in PHP. This tool plots the different skew plot like MK, RY and GC skew specified by users by selecting respective check boxes. All these plots can be generated either for one strain or both strain of the nucleotide sequence. This tool has also flexibility to generate oligo-nucleotide skew plot ranging from di- to hexa-nucleotide range.

The idea of Z-curve was mooted first by Lobry[15] while creating a 3-D DNA walk of DNA sequences. Cumulative Z-curve is zig-zag, three-dimensional curve which constitutes a unique representation of a DNA sequence, by means of cumulative counting of nucleotide matrices in through a counting window. Due to discrete and vector nature of all metrics Z-curve and the given DNA sequence each can be

Fig. 2: Landing page of ArchaeVir database

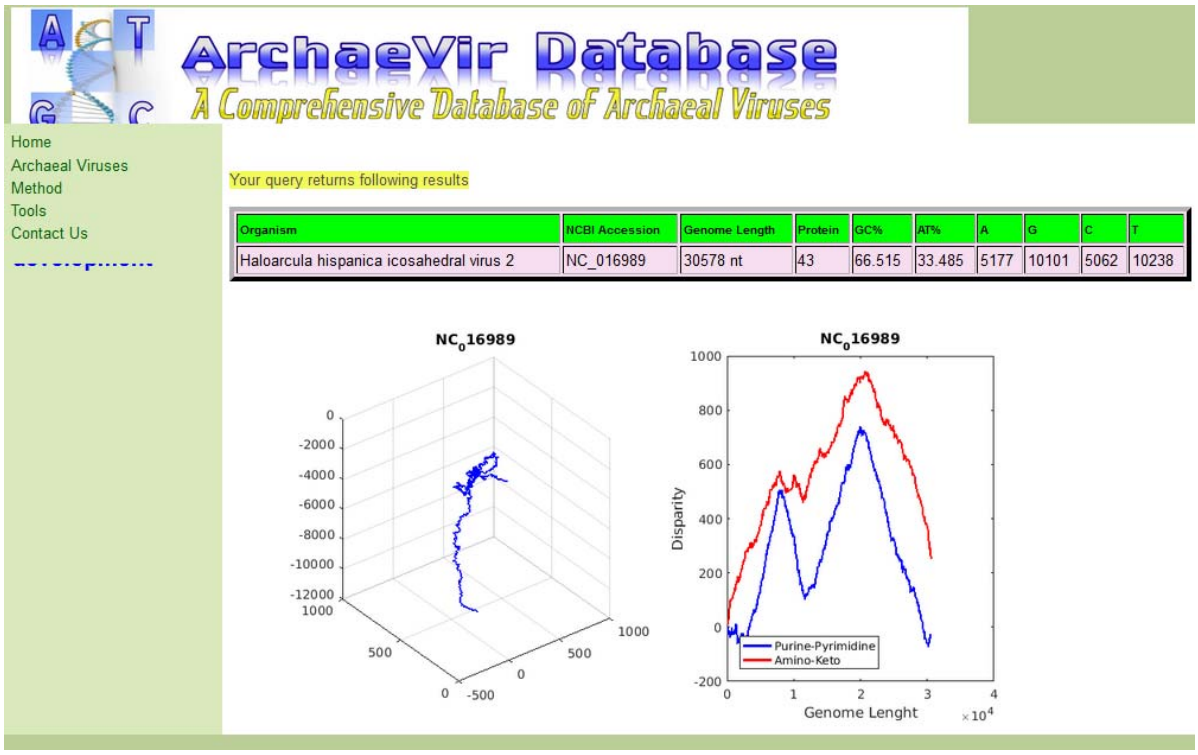

Fig. 3: Search page of ArchaeVir database

uniquely reconstructed from the other[10]. Different properties of the Z-curve, such as its symmetry and periodicity can give unique information on the DNA sequence[16]. The Z-curve was generated from a series of nodes, P0, P1,⋯ PN, with the coordinates Xn, Yn and Zn (n = 0, 1, 2⋯ N, with N being the length of the DNA sequence). The Z-curve was created by connecting each of the nodes sequentially[17] using vector components of X, Y, Z over a sliding windows:

$$\text{Components of Z} - \text{curve}\begin{cases}\text{Xn} = (\text{An} + \text{Gn}) - (\text{Cn} + \text{Tn})\\ \text{Yn} = (\text{An} + \text{Cn}) - (\text{Gn} + \text{Tn})\\ \text{Zn} = (\text{An} + \text{Tn}) - (\text{Gn} + \text{Cn})\end{cases}$$

These interconnected nodes are created using three chemical-physico property of nucleotide in three axis which are as follows:

- Chemical structure of having single or double rings nucleotides:

$$\text{Bases}\begin{cases}\text{Purine R} = \text{A G}\\ \text{Pyrimidine Y} = \text{C T}\end{cases}$$

- Chemical structure of having an amino or keto group:

$$\text{Bases}\begin{cases}\text{Amino M} = \text{AC}\\ \text{Keto K} = \text{G T}\end{cases}$$

- Structure of the double helix forming 2 or 3 hydrogen bonds in the Watson-Crick pair:

$$\text{Bases}\begin{cases}\text{Weak W} = \text{A T}\\ \text{Strong S} = \text{G C}\end{cases}$$

There are two figures generated using this approach one is the 3D Z-curve itself and second consist two components X (purine-pyrimidine) and Y (amino-keto) of the Z-curve.

CONCLUSION

There are several databases which describe genometrics of prokaryotic genomes using various methods like GC skew Z-curve and amino-keto skews. Most of them are for prediction of origin of replication in prokaryote genomes but none of them covers the all archaeal viruses genometrics analysis. Archaevir is the first database which has completed and archived all archaeal virus genomes for Z-curve along with amino-keto and purine pyrimidine skews with their basic genomics record like between genome length and protein, total gene, nucleotide distribution etc. Data available on this database may be very useful for scientists and researchers for understanding the local and global genometric feature of archaeal virus genomes. The data base is dynamic and will be updated time to time to add new feature and at the same time new genomes analysis will be archived as soon as it will be available in the NCBI.

SIGNIFICANCE STATEMENTS

This study discovered the genometrics of archaeal viruses based on the Z-curve. Two components of Z-curve i.e., amino-keto and purine-pyrimidine shows high variation in distribution of the nucleotide in a given cumulative windows, this study reveals the local and global genometric feature of a DNA chain. This database also hosts some basic genomic records of archaeal viruses which may also be helpful for the users.

REFERENCES

1. Krupovic, M., D. Prangishvili, R.W. Hendrix and D.H. Bamford, 2011. Genomics of bacterial and archaeal viruses: Dynamics within the prokaryotic virosphere. Microbiol. Mol. Biol. Rev., 75: 610-635.
2. Hendrix, R.W., 1999. Evolution: The long evolutionary reach of viruses. Curr. Biol., 9: R914-R917.
3. Norrby, E., 2008. Nobel Prizes and the emerging virus concept. Arch. Virol., 153: 1109-1123.
4. Sorek, R., V. Kunin and P. Hugenholtz, 2008. CRISPR-A widespread system that provides acquired resistance against phages in bacteria and archaea. Nat. Rev. Microbiol., 6: 181-186.
5. Bell, P.J.L., 2001. Viral eukaryogenesis: Was the ancestor of the nucleus a complex DNA virus? J. Mol. Evol., 53: 251-256.
6. Zillig, W., D. Prangishvili, C. Schleper, M. Elferink and I. Holz *et al*., 1996. Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. FEMS Microbiol. Rev., 18: 225-236.
7. Ojha, K.K. and D. Swati, 2015. ArchaeProfile: A database of Archaea and their origins of replication. J. Comput. Sci. Syst. Biol., 8: 96-98.
8. Roten, C.A.H., P. Gamba, J.L. Barblan and D. Karamata, 2002. Comparative Genometrics (CG): A database dedicated to biometric comparisons of whole genomes. Nucl. Acids Res., 30: 142-144.
9. Frank, A.C. and J.R. Lobry, 2000. Oriloc: Prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics, 16: 560-561.

10. Zhang, C.T., R. Zhang and H.Y. Ou, 2003. The Z curve database: A graphic representation of genome sequences. Bioinformatics, 19: 593-599.
11. Ojha, K.K. and D. Swati, 2010. *In silico* detection of origins of replication in bacteria and Archaea. J. Int. Acad. Phys. Sci., 14: 531-541.
12. Zhang, R. and C.T. Zhang, 2005. Identification of replication origins in archaeal genomes based on the Z-curve method. Archaea, 1: 335-346.
13. Ojha, K.K. and D. Swati, 2010. Mapping of origin of replication in themococcales. Bioinformation, 5: 213-218.
14. Zheng, W.X., L.L. Chen, H.Y. Ou, F. Gao and C.T. Zhang, 2005. Coronavirus phylogeny based on a geometric approach. Mol. Phylogenet. Evol., 36: 224-232.
15. Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol., 13: 660-665.
16. Zhang, R. and C.T. Zhang, 1994. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. J. Biomol. Struct. Dyn., 11: 767-782.
17. Zhang, C.T., 1997. A symmetrical theory of DNA sequences and its applications. J. Theoret. Biol., 187: 297-306.