Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

ViPTree: the viral proteomic tree server

Yosuke Nishimura[1,2], Takashi Yoshida[2], Megumi Kuronishi[1], Hideya Uehara[3], Hiroyuki Ogata[1] and Susumu Goto[1,*]

[1]Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan, [2]Graduate School of Agriculture, Kyoto University, Kitashirakawa-Oiwake, Sakyo-ku, Kyoto, 606-8502, Japan, [3]SGI Japan, Ltd., Yebisu Garden Place Tower 31F, 4-20-3 Ebisu, Shibuya-ku, Tokyo 150-6031, Japan

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: ViPTree is a web server provided through GenomeNet to generate viral proteomic trees for classification of viruses based on genome-wide similarities. Users can upload viral genomes sequenced either by genomics or metagenomics. ViPTree generates proteomic trees for the uploaded genomes together with flexibly selected reference viral genomes. ViPTree also serves as a platform to visually investigate genomic alignments and automatically annotated gene functions for the uploaded viral genomes, thus providing virus researchers the first choice for classifying and understanding newly sequenced viral genomes.
Availability: ViPTree is freely available at: http://www.genome.jp/viptree
Contact: goto@kuicr.kyoto-u.ac.jp
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Viruses are the most abundant biological entities and a reservoir of the greatest genetic diversity on Earth (Edwards and Rohwer, 2005; Suttle, 2005). Viruses are found in various habitats including aquatic, terrestrial, animal-/plant-associated, and engineered environments (Paez-Espino, et al., 2016), where they are considered to infect all types of cellular organisms (Fuhrman, 1999). Furthermore, viral pathogens of humans, crops and livestock are highly diverse. Thus, viral studies and their classification are crucial in various research fields including epidemiology, clinical microbiology, ecology, and evolutionary biology.

Recent advances in sequencing technologies have lead to an accelerated accumulation of sequenced viral genomes including those from environmental samples (Paez-Espino, et al., 2016; Roux, et al., 2016). However, viral genomes do not contain universally conserved genes like rRNAs, making it difficult to classify them using gene-based approaches (Rohwer and Edwards, 2002). To overcome this, several methodologies have been proposed to classify viruses based on a whole gene set encoded in viral genomes, including the viral (phage) proteomic tree (Adriaenssens, et al., 2015; Bellas, et al., 2015; Bhunchoth, et al., 2016; Mizuno, et al., 2013; Rohwer and Edwards, 2002) and various others (Glazko, et al., 2007; Iranzo, et al., 2016; Iranzo, et al., 2016; Lavigne, et al., 2009; Lavigne, et al., 2008; Lima-Mendez, et al., 2008; Roux, et al., 2016; Roux, et al., 2015; Wu, et al., 2009). These methods were used to classify members of the three families in the order *Caudovirales*, namely *Podoviridae* (Lavigne, et al., 2008), *Myoviridae* (Lavigne, et al., 2009), and *Siphoviridae* (Adriaenssens, et al., 2015), which resulted in the proposal of new subfamilies and genera that were later ratified by the International Committee on Taxonomy of Viruses. These methods have also been applied to non-isolated viral genomes sequenced from environmental DNA (Bellas, et al., 2015; Mizuno, et al., 2013; Roux, et al., 2016).

These genome-based methods have advantages over gene phylogenic analyses (Rohwer and Edwards, 2002; Wu, et al., 2009). Namely, (*i*) highly diverse viral genomes can be analyzed together since a sequence alignment is not needed, (*ii*) no conserved gene among analyzed genomes is needed, and (*iii*) it is likely less sensitive to genome rearrangement including horizontal gene transfer. However, to our knowledge, no ready-to-use software or web server is available to perform proteomic tree reconstructions for analyzing newly sequenced viral genomes.

We have thus developed the viral proteomic tree server GenomeNet/ViPTree (http://www.genome.jp/viptree), providing virus

researchers a convenient way of generating proteomic trees for newly sequenced viral genomes together with reference viral genomes flexibly chosen by the users, and to help them gain quick insights into the classification of target viral genomes. ViPTree also provides automatically generated gene annotations as well as user-friendly views of genomic alignments. Various parameters are available for visualization of proteomic trees and genomic alignments, and all resultant images can be downloaded in a scalable vector format (SVG), serving as ready-to-use figures for publication. Thus, ViPTree provides the first choice for classifying and understanding newly sequenced viral genomes.

2 Materials and Methods

Reference viral sequences and taxonomies are based on the GenomeNet/Virus-Host DB (Mihara, et al., 2016). ViPTree performs proteomic tree construction as reported (Bellas, et al., 2015; Bhunchoth, et al., 2016; Mizuno, et al., 2013). Specifically, normalized tBLASTx scores (S_G; $0 \leq S_G \leq 1$) between viral genomes are calculated (Bhunchoth, et al., 2016). A proteomic tree is generated by BIONJ based on the genomic distances (i.e., $1-S_G$). Gene finding and automated gene functional annotation are also performed. Further information of materials and methods is written in Supplementary Text S1.

3 Features and Implementation

Pre-calculated proteomic trees for reference viral genomes classified based on their nucleic acid types (e.g., dsDNA/ssDNA/dsRNA/ssRNA viruses) are viewable. Users can upload their viral genome sequences and choose reference viruses by the nucleic acid types (and, optionally, host categories (i.e., prokaryotes/eukaryotes)) to generate a proteomic tree. The maximum number of user sequences that can be uploaded in a session is five. Computation for a session is normally completed within a few hours. The main ViPTree web server interfaces are listed below.

- Proteomic tree view (Supplementary Fig. S1, S2): circular (comprehensive view) and rectangular (detailed view) representations are provided. Selected genomes can be highlighted. Where inner nodes of a tree are shown as filled circles, each of them links to a genomic alignment of sequences included in its subtree. Various visualization parameters (e.g., sizing images) are available.
- Genomic alignment view (Supplementary Fig. S3): Genomic alignments and dot plots based on tBLASTx results are shown. The order of the genomes in an alignment, an orientation and a start position of each genome can be automatically/manually configured. Annotated gene functions can be indicated on the genomes. In addition, various visualization parameters are available.

In addition, for each uploaded genome, tables of S_G scores to reference viral genomes and gene annotations can be browsed. The selection of viral genomes can be altered to generate proteomic trees of publication quality. All visualizations, tables, reference virus data are downloadable. Standalone software for proteomic tree generation is also available at the ViPTree website. The ViPTree web server is implemented mainly by Ruby. Visualization of trees and alignments uses the D3.js library (Bostock, et al., 2011).

Acknowledgements

Computation time was provided by the SuperComputer System, Institute for Chemical Research, Kyoto University.

Funding

This work was supported by The Canon Foundation (grant number 203143100025), Japan Society for the Promotion of Science (JSPS)/KAKENHI (grant numbers 26430184, 16H06429, 16K21723 and 16H06437), and the Collaborative Research Program of the Institute for Chemical Research, Kyoto University (grant number 2016-28).

Conflict of Interest: none declared.

References

Adriaenssens, E.M., *et al.* Integration of genomic and proteomic analyses in the classification of the Siphoviridae family. *Virology* 2015;477:144-154.

Bellas, C.M., Anesio, A.M. and Barker, G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. *Front Microbiol* 2015;6:656.

Bhunchoth, A., *et al.* Two asian jumbo phages, ΦRSL2 and ΦRSF1, infect Ralstonia solanacearum and show common features of ΦKZ-related phages. *Virology* 2016;494:56-66.

Bostock, M., Ogievetsky, V. and Heer, J. D(3): Data-Driven Documents. *IEEE Trans Vis Comput Graph* 2011;17(12):2301-2309.

Edwards, R.A. and Rohwer, F. Viral metagenomics. *Nature reviews* 2005;3(6):504-510.

Fuhrman, J.A. Marine viruses and their biogeochemical and ecological effects. *Nature* 1999;399(6736):541-548.

Glazko, G., *et al.* Evolutionary history of bacteriophages with double-stranded DNA genomes. *Biol Direct* 2007;2:36.

Iranzo, J., *et al.* Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *J Virol* 2016;90(24):11043-11055.

Iranzo, J., Krupovic, M. and Koonin, E.V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *MBio* 2016;7(4):e00978-00916.

Lavigne, R., *et al.* Classification of Myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol* 2009;9:224.

Lavigne, R., *et al.* Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res Microbiol* 2008;159(5):406-414.

Lima-Mendez, G., *et al.* Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* 2008;25(4):762-777.

Mihara, T., *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* 2016;8(3):66.

Mizuno, C.M., *et al.* Expanding the marine virosphere using metagenomics. *PLoS Genet* 2013;9(12):e1003987.

Paez-Espino, D., *et al.* Uncovering Earth's virome. *Nature* 2016;536(7617):425-430.

Rohwer, F. and Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* 2002;184(16):4529-4535.

Roux, S., *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 2016;537(7622):689-693.

Roux, S., *et al.* Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 2015;4:e08490.

Suttle, C.A. Viruses in the sea. *Nature* 2005;437(7057):356-361.

Wu, G.A., *et al.* Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc Natl Acad Sci U S A* 2009;106(31):12826-12831.