Communications in Information and Systems
Volume 20, Number 1, 61–76, 2020

Whole-genome phylogeny of giant viruses by Fourier transform

Changchuan Yin and Stephen S.-T. Yau

Dozens of giant viruses have been surprisingly discovered since 2003. Giant viruses are notably larger than typical bacteria and have extremely large genomes that encode thousands of genes. Because the giant viruses have super-sized shapes and genome sizes, they are distinguished from classical viruses and bacteria. The evolutionary origin of the giant virus is still an open question. The current phylogenetic studies on giant virus use specific genes or proteins, and can not elucidate the origins of the giant viruses from the global genome perspective. In this study, we perform the whole-genome phylogenetic analysis of the giant viruses using the Fourier transform-based alignment-free method. The phylogenetic analysis shows that the typical giant virus *Pandoravirus* and tailed giant virus *Tupanvirus* are closely related to archaea. This new finding suggests that giant viruses may origin from archaea and supports the reductive model on the giant virus evolution.

1. Introduction

Viruses are common infectious agents of small size and have a few essential genes for DNA replication and production of viral capsids. Surprisingly, giant viruses discovered since 2003 are larger than some bacteria. The genomes of giant viruses can be more than 2M bp and encode thousands of genes. The giant virus genomes may also encode the components for protein translation [17], which are the signatures of cellular organisms. Therefore, giant viruses are complex systems and different from regular viruses and bacteria. The controversial hypothesis was proposed that so-called giant viruses are descendants of a vanished group of cellular organisms, the fourth domain of life [4].

From the genome characters, giant viruses belong to the nucleocytoplasmic large DNA viruses (NCLDVs) family [5]. The traditional NCLDVs include five families *Poxviridae, Asfarviridae, Iridoviridae, Ascoviridae*, and *Phycodnaviridae*. The giant virus families *Mimiviradae, Pandoravirus*, and *Marseillevirus* were recently included in the NCLDVs families. However, the disparities in virion shapes and replication modes among NCLDV suggest that giant virus families are not necessarily a taxonomic group in NCLDV.

Therefore, the evolutionary origin and ancestry of giant viruses are controversial questions.

Two evolutionary models, a reductive model and an expansion model, have been presented for the origin of giant viruses [5]. In the reductive model, the giant viruses were from an ancestral cellular organism, which became reduced in size, leading to the dependence of the resulting genome on host cells. The presence of genes carrying cellular functions in almost any giant virus, for example, translation components, is consistent with this model. In the expansion model, current giant viruses originated from smaller ancestral viruses which carried only a few dozens of genes, could expand the genome through gene duplications and horizontal gene transfer. This model is supported by the discovery of massive gene exchange between giant viruses and a variety of organisms. However, the actual origin of giant viruses is a continuum.

Whole-genome sequence phylogenetic analysis provides a global view of the evolutionary aspect of the genomes. Due to the oversize genomes of giant viruses, the alignment-based method cannot be used in the phylogenetic analysis. This challenge hinders the whole-genome analysis of giant viruses by multiple sequence alignments. The alignment-free k-mer method has been applied to giant virus analysis [22], but the k-mer method highly depends on the selection of k-mer sizes, various k-mer size may create a different phylogenetic tree, therefore, the k-mer based method may not produce consistent phylogenetic trees. Besides, the previous whole proteome phylogenetic analysis only studied protein-coding regions of giant viruses and did not compare with bacteria and archaea, therefore, global phylogeny of microorganisms is not achieved.

Here, we analyze giant virus whole-genomes and compare the giant virus genomes with bacterial and archaeal genomes using the alignment-free method by Fourier transform [23, 24]. The results of whole-genome phylogeny show that giant viruses have a close relationship with archaea. This study suggests that giant virus may originate from archaea and supports the reduction model about the giant virus evolution.

2. Methods and algorithms

Previously we presented an alignment-free method for comparing whole genomes [23, 24]. The method uses Fourier transform features of genomes, which are first numerically represented by binary indicator sequences. To compare genomes of different lengths in the same Euclidean space, we proposed an even scaling algorithm to extend all sequences to the same length. The detailed alignment-free method for constructing phylogenetic trees of

the genomes of the giant viruses in this study is described in the following sections.

2.1. Numerical representations of DNA sequences by 4-D binary indicators

DNA sequences or genomes are composed of four nucleotides, adenine (A), thymine (T), cytosine (C), and guanine (G). To apply the digital signal processing (DSP) techniques or mathematical analysis methods on DNA sequences, the symbolic DNA sequences should be represented by numerical values. The numerical representation of a DNA sequence is to use 4-dimensional (4-D) binary indicator sequences [20]. In the 4-D binary indicator representation, a DNA sequence, denoted as, $s(0), s(1), \ldots, s(N-1)$, is decomposed into four binary indicator sequences, $u_A(n), u_T(n), u_C(n)$, and $u_G(n)$, which indicate the presence or absence of four nucleotides, A, T, C, and G at the n-th position, respectively. The 4-D binary indicator mapping of DNA sequences is defined as follows:

$$u_\alpha(n) = \begin{cases} 1, s(n) = \alpha \\ 0, \text{otherwise} \end{cases} \tag{1}$$

where $\alpha \in \{A, T, C, G\}, n = 0, 1, 2, \ldots, N-1$. The four indicator vectors correspond to the appearance of the four nucleotides at each position of the DNA sequence. For example, the indicator sequence, $u_G(n) = 0001010111\ldots$, represents that the nucleotide G occurs in the positions of 4, 6, 8, 9, and 10 of the DNA sequence.

2.2. Discrete Fourier transform

Discrete Fourier Transform (DFT) is the primary digital signal processing technique that transforms N observation data (time domain) to N new values in the frequency domain [18]. In the frequency domain, DFT may reveal the hidden periodic features buried in the time domain. DFT spectral analysis of DNA sequences may detect any latent or hidden periodical signal in the original sequences. The DFT method has been often used for finding the approximate repeats that are often difficult to detect by using a tandem repeat search.

Let U_A, U_T, U_C, and U_G be the DFT of the 4-D binary sequences u_A, u_T, u_C, and u_G, the DFT of the numerical series u_x of length N is defined as

$$U_x(k) = \sum_{n=0}^{N-1} u_x(n) e^{-i\frac{2\pi}{N}kn} \tag{2}$$

where $i = \sqrt{-1}$. The DFT power spectrum of the signal u_x at the frequency k is defined as

$$PS(k) = \sum_{x \in \{A,T,C,G\}} |U_x(k)|^2, k = 0, 1, 2, \cdots, N-1 \tag{3}$$

where $U[k]$ is the k-th DFT coefficient.

Fourier Transform gives a unique feature representation of the original underlying signal in frequency domain. The frequency domain vector $U_x(k)$ contains all the information about $u_x(n)$. The Parseval's Theorem for Fourier Transform signifies equivalence in the energy levels of signal in time and frequency domains. Therefore, we may measure the distance of two DNA sequences after DFT of the sequences.

2.3. Even scaling of Fourier power spectrum of different lengths

From the definition of the Fourier power spectrum, DNA sequences of various lengths have power spectra of different lengths and thus the power spectra cannot be used as a direct comparison of DNA sequences. To overcome this problem, we proposed an even scaling algorithm for extending the DFT power spectra of different lengths into the same length [24]. In details, let T_n denote the original power spectrum of length n and T_m denote the extended power spectrum of length m from even scaling of T_n, and $m > n$. The symbol $\lfloor \cdots \rfloor$ denotes the floor function on non-integers. The even scaling operation on the original power spectrum T_n to T_m is defined as follows and in Algorithm 1.

$$T_m(k) = \begin{cases} T_n(Q), \ \text{if}\, Q \in Z^+ \\ T_n(R) + (Q - R)\left(T_n(R+1) - T_n(R)\right), \ \text{if}\, Q \notin Z^+ \\ \quad \text{where}\, Q = \dfrac{kn}{m}, \ R = \left\lfloor \dfrac{kn}{m} \right\rfloor \end{cases} \tag{4}$$

2.4. Algorithm for computing the pairwise Euclidean distances of DNA sequences

The most common distance measure for time series is the Euclidean distance, which is the optimal distance measure for estimation if signals corrupted by additive Gaussian noise [2, 26]. The *Euclidean metric* on $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is

Input: T_n of length n, new length m, $m > n$
Output: T_m of length m
$T_m(1) = T_n(1)$
for $k \leftarrow 2$ ***to*** m **do**
  $Q = \frac{kn}{m}$
  $R = \lfloor \frac{kn}{m} \rfloor$
  if $R == 0$ **then**
    $R \leftarrow 1$
  end
  if $Q \in Z^+$ **then**
    $T_m(k) = T_n(Q)$
  else
    $T_m(k) = T_n(R) + (Q - R) * (T_n(R + 1) - T_n(R))$
  end
end
return T_m

Algorithm 1: Even scaling a time series T_n.

Data: DNA SEQ1(length N1), SEQ2(length N2), SEQ3(length M), with M>N1, M>N2
Result: Pairwise distance of SEQ1, SEQ2 and SEQ3
Steps

1. Convert SEQ1, SEQ2, SEQ3 to binary indicator sequence BS1, BS2, BS3
2. Compute Fourier power spectrum PS1, PS2, and PS3M from BS1, BS2, BS3
3. Even scale PS1 as PS1M from length N1 to length M
4. Even scale PS2 as PS2M from length N2 to length M
5. Compute the Euclidean distance d(PS1M,PS2M), d(PS2M,PS3M), d(PS1M,PS3M) in an M-dimensional space

Algorithm 2: Algorithm for computing pairwise distances of two genomes.

defined by the function d,

$$d\left((x_1, \ldots, x_n), (y_1, \ldots, y_n)\right) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{5}$$

After even scaling the DFT spectrum, we measured the Euclidean distances of DNA sequences by the full DFT power spectra of the sequences. The distance matrix of the DNA sequences is then used for producing a phylogenetic tree of the sequences.

The algorithm for computing the pairwise Euclidean distances of DNA sequences SEQ1, SEQ2, and SEQ3 is described in Algorithm 2.

2.5. Whole-genome phylogenetic analysis

In this study, the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) phylogenetic trees [19], a bottom-up hierarchical clustering method, are built based on the Euclidean distance matrices of the genomes. The phylogenetic trees are built using MATLAB R2015a [15].

2.6. Giant virus dataset

The definition of giant viruses can be based on shape sizes or genomes. In this study, we adopt the definition of giant viruses as eukaryote-infecting viruses with at least 500 protein-coding genes [5]. All sequence data were obtained from GenBank on the NCBI and are listed in Table A1 in the supplementary materials.

3. Results

Giant virus genomes are large in size and diverse in sequences. The commonly used multiple sequence alignment (MSA) methods cannot be used for the phylogenetic analysis of large and diverse genomes because of the computational complexity of the MSA methods. To perform the whole-genome phylogeny of giant viruses, we apply the previous alignment-free method that is based on Fourier transform features of DNA sequences [23, 24]. Specially, we combine the DFT of 4-D numerical representations of DNA sequences [23] and the even-scaling method [24]. Using the Fourier transform alignment-free method, we first analyze *Influenza* virus and bacterial genomes, the results are to validate the effectiveness of the method. Then we compare the giant virus genomes and bacterial genomes to identify the evolutionary relationship among the giant viruses and bacteria.

3.1. Alignment-free phylogenetic analysis of viruses and bacteria

We demonstrate the effectiveness of the DFT-based phylogenetic analysis. We use the genome DFT method to classify *Influenza* A viruses. *Influenza* A virus is a pathogenic agent causing influenza in birds and humans. The viruses are single-stranded RNA viruses. It is the only species of the *Alphainfluenzavirus* genus of the *Orthomyxoviridae* family of viruses. The subtypes are classified according to the viral surface proteins hemagglutinin (H) and neuraminidase (N). Eighteen H subtypes of antigens (H1-H18) and eleven N subtypes of antigens (N1-N11) have been identified [10]. In the classification test of *Influenza* A viruses, we used neuraminidase (NA) gene of

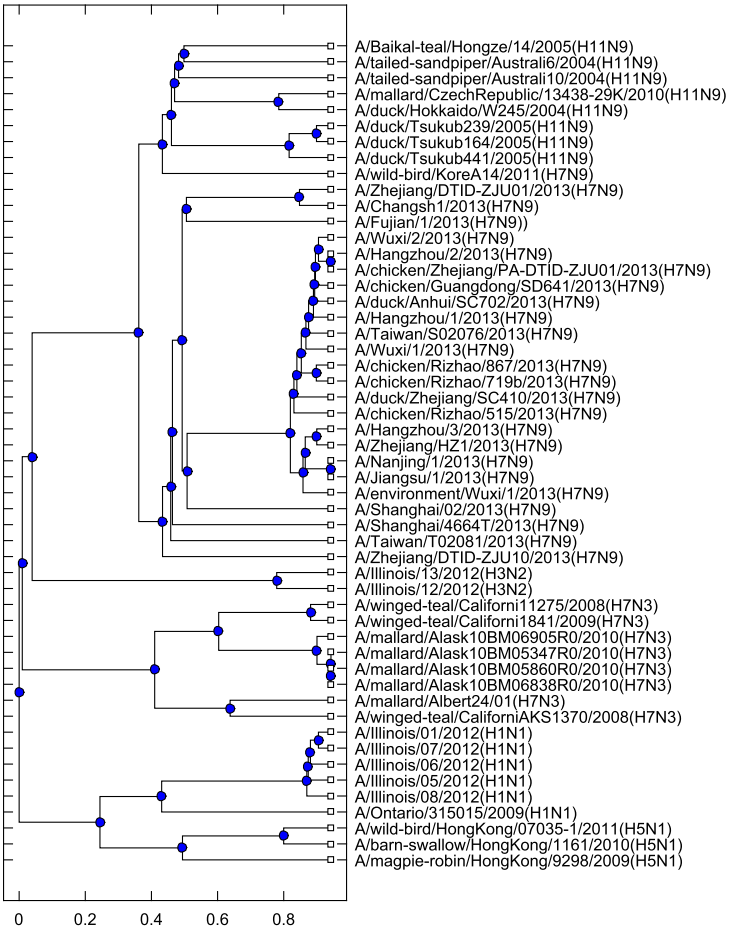

Figure 1: The phylogenetic analysis of *Influenza* A viruses. The phylogeny is built upon the DFT distances of neuraminidase (NA) genes in the virus.

Influenza A virus because this gene is associated with pandemic influenza and a wide range of natural hosts, including man, birds, and other animals. The classification result of *Influenza* A virus shows the accurate phylogenetic relationship of the subtypes of the viruses (Fig. 1). For instance, the virus subtypes N7N9 are clustered in one branch, the subtype variants are classified correctly according to the geographical locations and times of the isolated strains.

We also apply the DFT methods to classify the bacteria using whole-genomes. The bacterial genomes have the sizes in the range of about 1 Mb to

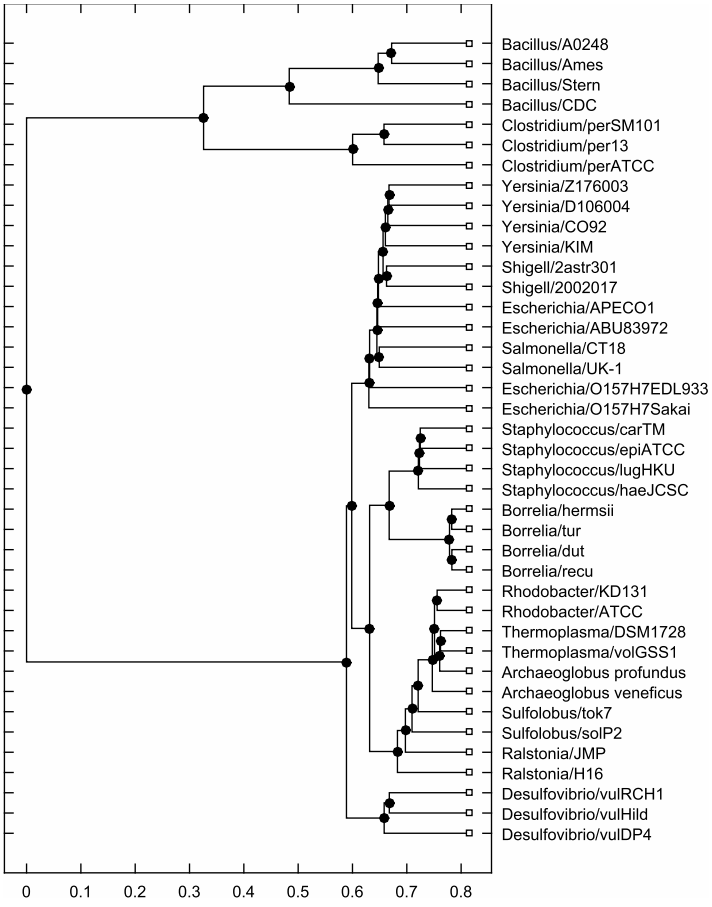

Figure 2: Whole-genome phylogenetic analysis of bacteria.

5.5M bp. The classification result of bacteria (Fig. 2.) by whole-genomes is in agreement with the taxonomy by the morphologic characteristics and biological properties. These results validate that the Fourier transform alignment-free method of whole-genome comparison is accurate and effective.

3.2. Whole-genome phylogenetic analysis of giant virus

The phylogenetic analysis of the whole-genomes of giant viruses shows that similar giant viruses are clustered as a taxonomic group (Fig. 3). Especially, in the phylogenetic tree, *Pandoravirus* is separated from the other giant viruses due to the large genome of *Pandoravirus*.

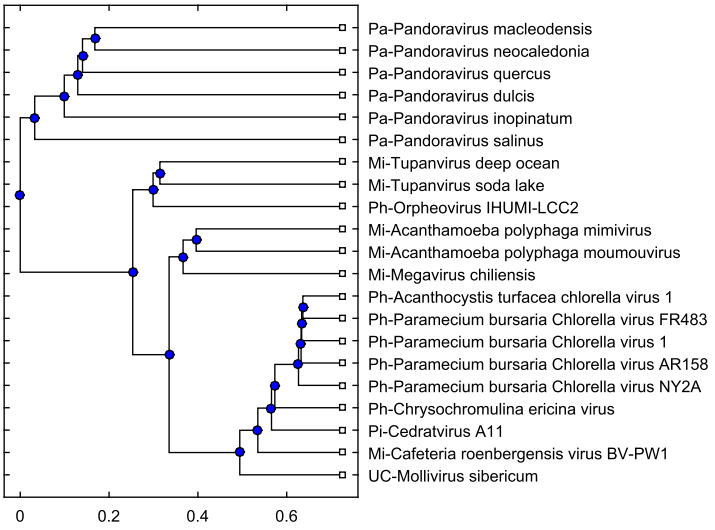

Figure 3: Whole-genome phylogenetic analysis of the giant viruses.

The phylogenetic analysis of the whole-genomes of giant viruses and bacteria shows that giant virus *Pandoravirus* is close to archaea *Thermococcus*, *Methanobrevibacter* and tailed giant viruses *Tupanvirus* is most closely related to *Thermoplasma* (Fig. 4). The thermophile *Aquifex aeolicus* of 1.5M bp size is known as the bacterium of the smallest genome. The archaeon *Thermoplasma* lives in an environment of high temperature or salinity. *Pandoraviruses* is a genus of the giant virus, and have the largest genome size of any known viral genus. The *Pandoraviruses* genome size can be 2.5M bp that of parasitic eukaryotes [16]. These two-tailed giant viruses have the most complete translational apparatus in all viruses [1], only the ribosome is lacking. The shapes, genomes, and environments of both giant viruses and archaea suggest a theory that giant viruses may originate from archaea.

The whole-genome phylogeny in this study shows that the giant viruses are close to some archaea, suggesting that giant viruses may originate from archaea. This study provides solid evidence to support the reductive model of the evolutionary origin of giant viruses.

4. Discussion

The discovery of *Mimivirus*, a virus infecting *Acanthamoeba*, initiated a reappraisal of the upper limits of the viral world, both in terms of particle size and genome complexity, dimensions typical of parasitic bacteria. The diversity of these giant viruses *Megaviridae* was assessed by sampling a variety of

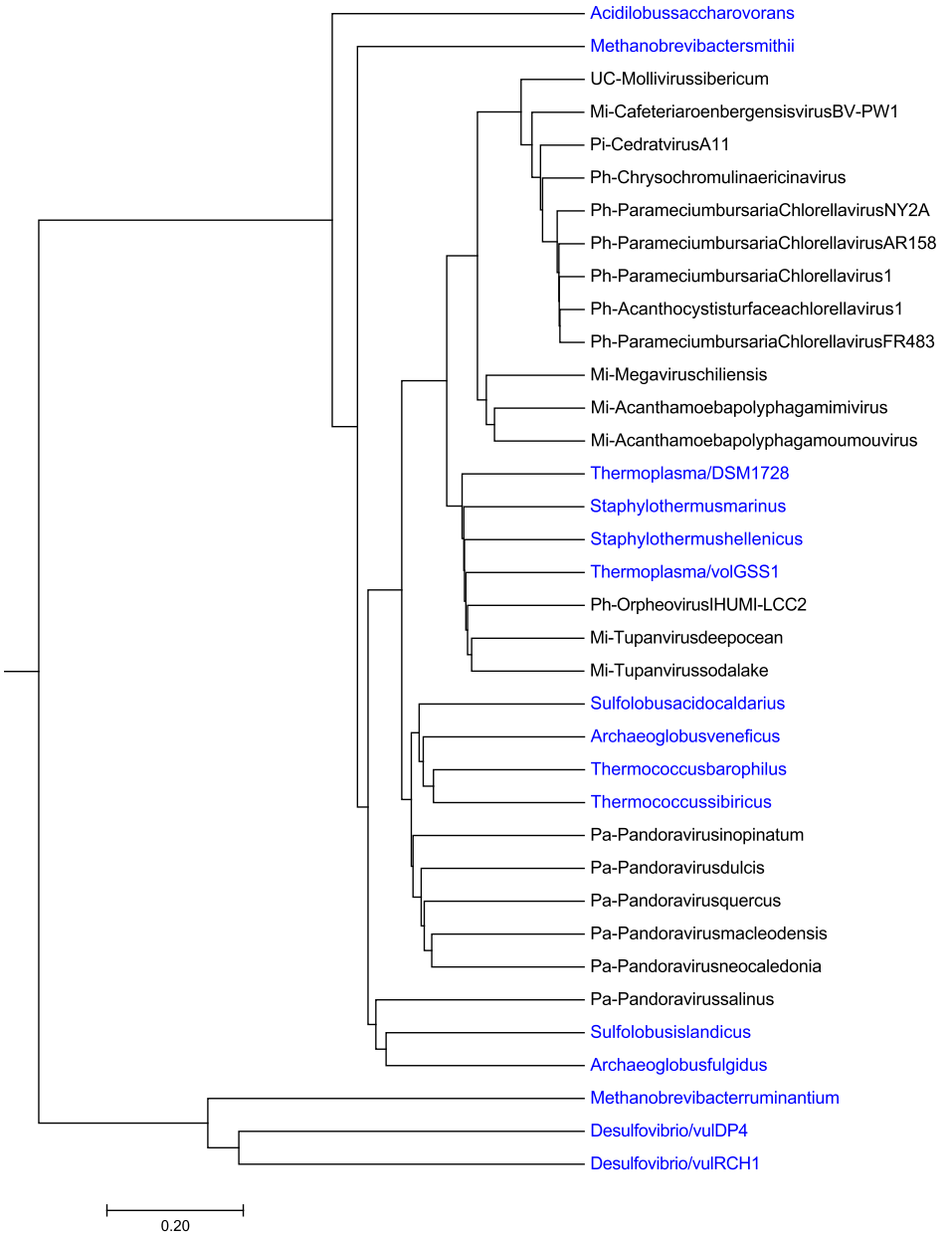

Figure 4: Whole-genome phylogenetic analysis of the giant virus with archaea.

aquatic environments and their associated sediments worldwide. We report the isolation of two giant viruses, one off the coast of central Chile, the other from a freshwater pond near *Melbourne*, without morphological or genomic resemblance to any previously defined virus families. Their micrometer-sized particles contain DNA genomes of at least 2.5M and 1.9M bp, respectively. These viruses are the first members of the proposed *Pandoravirus* genus, a term reflecting their lack of similarity with previously described microorganisms and the surprises expected from their future study.

Inferring the phylogeny of giant viruses from the whole-genome comparison opens a new channel in evolution study especially where there are no common genes that can be used for phylogenetic analysis. Cautions should be taken in whole-genome analysis when large arrangements occur in genomes, the problem can be mitigated by analyzing the genome signature characters and protein-coding regions of the genomes. Therefore, to examine the detailed evolutionary relationships of the giant viruses and archaea, we will further compare the whole-genome signatures and proteomes from two giant viruses *Pandoravirus* and *Tupanvirus* with corresponding archaea.

For whole-genome comparison, when the genomes have different lengths, the proposed even scaling method may effectively extend or shrink the DFT spectra to the same length. The even scaling method can be used in comparing any time series of different lengths. The other commonly used method for dealing with different lengths is zero paddings in time domain before DFT. The zero paddings method refers to adding zeros to the end of a time-domain signal to increase its length [18]. However, in the numerical mapping, DNA sequences are mapped to binary indicator vector, the padding zeros approach may add additional information to the vector, then the frequency domain length is increased and has an increased resolution. Therefore, zero paddings in time series may not work well in the genome comparison by Fourier transform.

This study compares and classifies giant viruses and bacteria using the whole-genome sequences, however, some whole genomes are not a single component, but consist of multiple segments. The DFT based alignment-free method may not process the multiple-segment genomes. To mitigate this problem in comparing multiple-segment genomes, existing studies on the classification of multiple-segment genomes employ the Hausdorff distance of the feature vectors of genomes [7, 27], because Hausdorff distance is to measure the distance two subsets of a metric space, for example, the dissimilarity of the multiple segments of two genomes. We will apply the Hausdorff distance and DFT transform to investigate the evolutionary origins of giant viruses from multiple coding components of the whole genomes or multiple segments of the whole genomes.

The phylogeny of the giant virus genomes shall be supported by the neutral theory of molecular evolution, which holds that the molecular evolution may be caused by random genetic drifts, rather than natural selections. Since genomes may evolve adaptively and quickly, a phylogenetic inference by genomes must use only neutral variations and should select a set of slow-evolving sequences for constructing the phylogeny of genomes [13, 21]. As proteins usually evolve slowly, the protein-coding sequence sets in a genome can be only considered as neutral variations in phylogenetic analysis. The phylogenetic trees constructed by protein-coding sequences from genomes may reflect the evolutionary origin of the virus species. The challenge in using proteomes of giant virus genomes for phylogenetic analysis is that most protein-coding sequences found in the giant viruses only exist in giant viruses. The protein-coding sequences do not have homologous counterparts in any organism other than the giant viruses [11].

The alignment-free methods for phylogenetic analysis explicitly rely on a measure of genetic distance between the sequences. The mismatches in DNA or protein sequences are related to phylogeny. The approximately linear correlation of the genetic distances and the DFT-based Euclidean distances was validated previously [23, 24]. The genetic equidistance result directly provided the formulation of the molecular clock [14, 31]. However, the molecular clock assumes the linear changing rate of genetic variants and does not reflect the saturated maximum of genetic distances (MGD) [3, 6, 8, 28]. Therefore, the neutral variations in genomes are explained by the neutral theory whereas the functional variations, which are correlated with physiology, can be explained by the MGD theory. The evolution principles of molecular and the neutral theory has also been tested using ancient proteomes [12]. Consequently, to confirm the evolutionary origin of giant viruses, the protein-coding regions of giant virus genomes will be examined for phylogeny in the future.

For accurately inferring the phylogeny of the giant viruses, only a small set of slowest evolving protein genes will be considered because most proteins of unimportance for survival are fast evolving or adaptive, reaching the saturation MGD levels in their distances among orthologs. For example, the phylogenetic inference of humans only requires about 500 proteins [9, 29]. To this end, we will examine the SNPs (Single-Nucleotide Polymorphism) on the protein-coding regions of the virus genomes to identify the slow-evolving proteins [25, 29], which are mostly the essential proteins in the genomes [30]. These slow-evolving protein sequences in the giant viruses will be concatenated as proteomes, and then the alignment-free DFT method will be applied to the proteomes for phylogenetic analysis.

Appendix A. Supplementary materials

Table 1: Genbank accession numbers of the genomes of giant virus and bacteria

Genbank ID	Name	Length (bp)
NC_014649	Mi-Acanthamoeba polyphaga mimivirus	1181548
NC_020104	Mi-Acanthamoeba polyphaga moumouvirus	1021348
NC_008724	Ph-Acanthocystis turfacea chlorella virus 1	288046
NC_014637	Mi-Cafeteria roenbergensis virus BV-PW1	617453
NC_032108	Pi-Cedratvirus A11	573918
NC_028094	Ph-Chrysochromulina ericina virus	473558
NC_016072	Mi-Megavirus chiliensis	1259013
NC_027867	UC-Mollivirus sibericum	651523
NC_036594	Ph-Orpheovirus IHUMI-LCC2	1473473
NC_021858	Pa-Pandoravirus dulcis	1908524
NC_026440	Pa-Pandoravirus inopinatum	2243076
NC_037665	Pa-Pandoravirus macleodensis	1838258
NC_037666	Pa-Pandoravirus neocaledonia	2003191
NC_037667	Pa-Pandoravirus quercus	2077288
NC_022098	Pa-Pandoravirus salinus	2473870
NC_000852	Ph-Paramecium bursaria Chlorella virus 1	330611
NC_009899	Ph-Paramecium bursaria Chlorella virus AR158	344686
NC_008603	Ph-Paramecium bursaria Chlorella virus FR483	321240
NC_009898	Ph-Paramecium bursaria Chlorella virus NY2A	368683
MF405918	Mi-Tupanvirus deep ocean	1515867
KY523104	Mi-Tupanvirus soda lake	1439485
CP000527	Desulfovibrio/vulDP4	3462887
CP002297	Desulfovibrio/vulRCH1	3532052
NC_002578	Thermoplasma/DSM1728	11564906
NC_002689	Thermoplasma/volGSS1	1584804
CP001742	Acidilobus saccharovorans	1496453
CP002426	Sulfolobus islandicus	2655198
CP000077	Sulfolobus acidocaldarius	2225959
AE000782	Archaeoglobus fulgidus	2178400
CP002588	Archaeoglobus veneficus	1901943
CP001719	Methanobrevibacter ruminantium	2937203
CP000678	Methanobrevibacter smithii	1853160
CP002372	Thermococcus barophilus	2010078
CP001463	Thermococcus sibiricus	1845800
CP002051	Staphylothermus hellenicus	11580347
CP000575	Staphylothermus marinus	1570485

Acknowledgement

This research was partially supported by National Natural Sciences Foundation of China (31271408), Tsinghua University Education Foundation fund (042202008), and Tsinghua University Start-up Research Fund (to S.S.-T. Yau).

References

[1] Abrahão, J., Silva, L., Silva, L. S., Khalil, J. Y. B., Rodrigues, R., Arantes, T., Assis, F., Boratto, P., Andrade, M., Kroon, E. G., et al. (2018). Tailed giant tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nature Communications*, 9(1):749.

[2] Agrawal, R., Faloutsos, C., and Swami, A. (1993). *Efficient similarity search in sequence databases*. Springer.

[3] Ayala, F. J. (2000). Neutralism and selectionism: the molecular clock. *Gene*, 261(1):27–33.

[4] Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., and Raoult, D. (2010). Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. *PLoS One*, 5(12):e15530.

[5] Brandes, N. and Linial, M. (2019). Giant viruses-big surprises. *Viruses*, 11(5):404.

[6] Hu, T., Long, M., Yuan, D., Zhu, Z., Huang, Y., and Huang, S. (2013). The genetic equidistance result: misreading by the molecular clock and neutral theory and reinterpretation nearly half of a century later. *Science China Life Sciences*, 56(3):254–261.

[7] Huang, H.-H., Yu, C., Zheng, H., Hernandez, T., Yau, S.-C., He, R. L., Yang, J., and Yau, S. S.-T. (2014). Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Molecular Phylogenetics and Evolution*, 81:29–36.

[8] Huang, S. (2008). The genetic equidistance result of molecular evolution is independent of mutation rates. *Journal of Computer Science and Systems Biology*, 1:92.

[9] Huang, S. (2012). Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. *Science China Life Sciences*, 8(55):709–725.

[10] Hutchinson, E. C. (2018). Influenza virus. *Trends in Microbiology*.

[11] Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J.-M., Beucher, L., Philippe, N., Bertaux, L., Christo-Foroux, E., et al. (2018). Diversity and evolution of the emerging Pandoraviridae family. *Nature communications*, 9(1):2285.

[12] Liu, T. and Huang, S. (2019). Testing the basic tenet of the molecular clock and neutral theory by using ancient proteomes. *bioRxiv*, page 821736.

[13] Luo, D. and Huang, S. (2016). The genetic equidistance phenomenon at the proteomic level. *Genomics*, 108(1):25–30.

[14] Margoliash, E. (1963). Primary structure and evolution of cytochrome c. *Proceedings of the National Academy of Sciences of the United States of America*, 50(4):672.

[15] MATLAB (2015). *version 8.5.0 (R2015a)*. The MathWorks Inc., Natick, Massachusetts.

[16] Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., et al. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286.

[17] Schulz, F., Yutin, N., Ivanova, N. N., Ortega, D. R., Lee, T. K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G. J., et al. (2017). Giant viruses with an expanded complement of translation system components. *Science*, 356(6333):82–85.

[18] Smith, J. O. (2007). Mathematics of the discrete fourier transform (DFT). *W3K: Charleston, SC, USA*, pages 7–9.

[19] Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, 38:1409–1438.

[20] Voss, R. (1992). Evolution of long-range fractal correlation and 1/f noise in DNA base sequences. *Physical Review Letters*, 68:3805–3808.

[21] Wang, D., Liu, F., Wang, L., Huang, S., and Yu, J. (2011). Nonsynonymous substitution rate (ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biology Direct*, 6(1):13.

[22] Wu, G. A., Jun, S.-R., Sims, G. E., and Kim, S.-H. (2009). Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proceedings of the National Academy of Sciences*, 106(31):12826–12831.

[23] Yin, C., Chen, Y., and Yau, S. S.-T. (2014). A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. *Journal of Theoretical Biology*, 359:18–28. MR3248415

[24] Yin, C. and Yau, S. S.-T. (2015). An improved model for whole genome phylogenetic analysis by Fourier transform. *Journal of Theoretical Biology*, 359(21):18–28. MR3385919

[25] Yin, C. and Yau, S. S.-T. (2019). Whole genome single nucleotide polymorphism genotyping of Staphylococcus aureus. *Communications in Information and Systems*, 19(1):57–80. MR3946079

[26] Yu, C., Cheng, S.-Y., He, R. L., and Yau, S. S.-T. (2011). Protein map: An alignment-free sequence comparison method based on various properties of amino acids. *Gene*, 486(1):110–118.

[27] Yu, C., He, R. L., and Yau, S. S.-T. (2014). Viral genome phylogeny based on lempel–ziv complexity and hausdorff distance. *Journal of Theoretical Biology*, 348:12–20. MR3178820

[28] Yuan, D. and Huang, S. (2017). Genetic equidistance at nucleotide level. *Genomics*, 109(3-4):192–195.

[29] Yuan, D., Lei, X., Gui, Y., Zhu, Z., Wang, D., Yu, J., and Huang, S. (2017). Modern human origins: multiregional evolution of autosomes and east asia origin of Y and mtDNA. *bioRxiv*. doi: https://doi.org/10.1101/101410.

[30] Zhang, J. and Yang, J.-R. (2015). Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, 16(7):409–420.

[31] Zuckerkandl, E. (1962). Molecular disease, evolution, and genetic heterogeneity. *Horizons in Biochemistry*, pages 189–225.

Changchuan Yin
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
USA
E-mail address: cyin1@uic.edu

Stephen S.-T. Yau
Department of Mathematical Sciences
Tsinghua University
Beijing 100084
China
E-mail address: yau@uic.edu

Received October 20, 2019