Molecular Phylogenetics and Evolution 99 (2016) 53–62

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Virus classification in 60-dimensional protein space

Yongkun Li [a], Kun Tian [a], Changchuan Yin [b], Rong Lucy He [c], Stephen S.-T. Yau [a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China
[b] Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607-7045, USA
[c] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA

ARTICLE INFO

Article history:
Received 14 June 2015
Revised 24 January 2016
Accepted 10 March 2016
Available online 15 March 2016

Keywords:
Hausdorff distance
Natural vector
Natural graphical representation
Virus classification

ABSTRACT

Due to vast sequence divergence among different viral groups, sequence alignment is not directly applicable to genome-wide comparative analysis of viruses. More and more attention has been paid to alignment-free methods for whole genome comparison and phylogenetic tree reconstruction. Among alignment-free methods, the recently proposed "Natural Vector (NV) representation" has successfully been used to study the phylogeny of multi-segmented viruses based on a 12-dimensional genome space derived from the nucleotide sequence structure. But the preference of proteomes over genomes for the determination of viral phylogeny was not deeply investigated. As the translated products of genes, proteins directly form the shape of viral structure and are vital for all metabolic pathways. In this study, using the NV representation of a protein sequence along with the Hausdorff distance suitable to compare point sets, we construct a 60-dimensional protein space to analyze the evolutionary relationships of 4021 viruses by whole-proteomes in the current NCBI Reference Sequence Database (RefSeq). We also take advantage of the previously developed natural graphical representation to recover viral phylogeny. Our results demonstrate that the proposed method is efficient and accurate for classifying viruses. The accuracy rates of our predictions such as for Baltimore II viruses are as high as 95.9% for family labels, 95.7% for subfamily labels and 96.5% for genus labels. Finally, we discover that proteomes lead to better viral classification when reliable protein sequences are abundant. In other cases, the accuracy rates using proteomes are still comparable to that of genomes.

Published by Elsevier Inc.

1. Introduction

With fast development of sequencing technology, an increasing number of viral sequences have been available. Phylogenetic and taxonomic studies on viral sequences become increasingly important for understanding diversities and origins of viruses (Holmes, 2010). Traditional approaches mostly are based on pairwise and multiple sequence alignment. There is high rate of divergence between different virus sequences due to gene mutation, horizontal gene transfer, gene duplication, gene insertion and deletion (Duffy et al., 2008). These features pose a challenge to phylogenetic investigation of viruses. Furthermore, whole genome sequences generally supply more comprehensive information for inferring the phylogeny of viruses than a few orthologous genes (Wong et al., 2008). Since genomes or proteomes include a lot of genes or proteins, the existing methods relating to multiple sequence comparison are computationally intensive (Vinga and Almeida, 2003). Thus they are not suitable for genome-wide phylogeny analysis.

In the past ten years, there has been a growing interest in genome based alignment-free methods for evolutionary studies. Among them, the *k*-mer related methods are all based on word frequencies, which ignore the position of nucleotides (Dai et al., 2008; Wu et al., 2009). In comparison, the natural vector method characterizes both the count and position information of nucleic acids (Deng et al., 2011). The NV method has succeeded in classifying viruses and reconstructing phylogenetic trees (Yu et al., 2013). The NV representation builds a one-to-one correspondence between a DNA sequence and a 12-dimensional numerical vector. Thus we establish a 12-dimensional genome space. Since the Euclidean distance between points in this space can represent their biological similarity to some extent, it allows comparing viruses simultaneously at family level, subfamily and genus levels. As some viral genomes are in the form of several segments, each segment corresponds to a point in R^{12} by the NV method, and then each virus corresponds to a set of points in R^{12}. Recall the general definition of Euclidean distance:

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).

http://dx.doi.org/10.1016/j.ympev.2016.03.009
1055-7903/Published by Elsevier Inc.

$d(a,b)=\sqrt{(a_1-b_1)^2+\cdots+(a_d-b_d)^2}$, where $a=(a_1,\ldots,a_d)$, $b=(b_1,\ldots,b_d)$, d is the dimension of vectors a and b. So it is only used to measure the distance between two points. To solve this, the Hausdorff distance is used to measure the distance between point sets, which results in the global comparison of multiple segmented viruses, including single-segmented viruses as well (Huang et al., 2014).

It is of importance to determine whether virus classification using whole-proteomes is indeed better than classification using whole genome. Although one gets nucleotide sequences first, it is increasingly feasible to get corresponding protein sequences as many of gene annotations have been done automatically or manually. Moreover, proteome sequences may be directly involved in determining the variety of functions and the structure of viruses. Mutation in a protein may directly affect its functions which likely result in phenotype changes in evolution. However, changes to nucleic acids may not lead to a protein mutation, because of degeneracy of genetic codons and presence of introns. Even though it was suggested that using proteome sequences was better than using whole genome DNA sequences for genome-based phylogeny reconstruction (Xu and Hao, 2009; Yu et al., 2010a,b; Xie et al., 2015), these studies were only based on a specific Baltimore class or certain families of viruses in the National Center for Biotechnology Information database (NCBI). Virus classification by proteomes has not been systematically characterized. Therefore, we have done a large scale test using almost all viruses in RefSeq database, which is a reliable, non-redundant, and annotated reference subset of NCBI. We compare the results obtained through whole proteome sequences with those by whole DNA genomes.

The phylogenetic tree is a useful tool for classifying and inferring the origin of organisms. Traditionally, this tree has been constructed on the basis of a distance or dissimilarity matrix of species. Many algorithms such as the neighbor-joining algorithm (Saitou and Nei, 1987), have been designed to recover this tree from this matrix. But there are some disadvantages to the resulting phylogeny tree. For instance, the tree may not be unique if the dissimilarity matrix doesn't obey the triangle inequality (Buneman, 1974). To overcome these limitations, the natural graphical representation was proposed (Yu et al., 2013). It has been shown to perform well and can be computed efficiently.

In this paper, we determine the classification of 4021 viruses in seven Baltimore classes based on the NV representation of proteomes and Hausdorff distance. Additionally we also apply the natural graphical representation to show the viral phylogeny. To validate the advantages of proteomes in virus classification, we further process the single-segmented viruses in Baltimore class IV with k-mer method and NV approach based on genomes and Euclidean distance as comparison.

2. Materials and methods

2.1. Overview of the viral data sets

Viruses exhibit more biological diversity than the rest of bacterial, plant, and animal kingdoms. Genomes of viruses may be single-stranded or double-stranded, linear or circular, and in a single-segmented or multi-segmented configuration. In this work, we first downloaded all the referenced protein sequences corresponding to the 4021 viruses as well as their referenced genome sequences from RefSeq database release 69 (January 7, 2015) from NCBI. Traditionally, viruses are classified into seven Baltimore classes. The information for each class in the proteome data set is summarized in Table 1. The 4021 viruses consist of 91 families, 22 subfamilies and 523 genera in total. The viruses in Baltimore VII class have no subfamily labels, thus we use zero to denote the number of their subfamilies in this table. To be convenient, we number the viruses by integers.

2.2. Natural vector and protein space

Let $\mathscr{L}$ be the set of 20 types of amino acids, i.e., $\mathscr{L}=\{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$, and $S=(s_1,s_2,\ldots,s_n)$ be a protein sequence of length n, that is, $s_i\in\mathscr{L}$, $i=1,2,\ldots,n$. For $k\in\mathscr{L}$, define $w_k(\cdot):\mathscr{L}\to\{0,1\}$ such that $w_k(s_i)=1$ if $s_i=k$ and 0 otherwise.

(1) Let $n_k=\sum_{i=1}^{n}w_k(s_i)$ denote the number of letter k in S.
(2) Let $\mu_k=\sum_{i=1}^{n}i\cdot\frac{w_k(s_i)}{n_k}$ be the mean position of letter k.
(3) Let $D_2^k=\sum_{i=1}^{n}\frac{(i-\mu_k)^2w_k(s_i)}{n_kn}$ be the normalized 2-nd central moment of positions of letter k.

For ambiguous amino acids, 1-letter B represents N or D; Z for E or Q; J for I or L; and X for all possible 20 types of amino acids. Thus for $k\in\mathscr{L}$ we define the weight $w_k(s_i)$ as the expected count of letter k in position i. For instance,

$$w_N(s_i)=\begin{cases}1, & s_i=N\\ 0.5, & s_i=B\\ 0.05, & s_i=X\\ 0, & \text{otherwise.}\end{cases}$$

The 60-dimensional NV of a protein sequence S is defined by $(n_A,n_R,\ldots,n_V,\mu_A,\ldots,\mu_V,D_2^A,\ldots,D_2^V)$. For nucleotide sequences, we have similarly defined NV, see Yu et al. (2013).

2.3. Hausdorff distance

Once each protein sequence is mapped to a unique point in the 60-dimensional NV space, each virus then corresponds to a set of points. But in our dataset, three viruses (#121, #297 and #700 in Baltimore class II) share the same set of proteins with other viruses. To ensure one-to-one correspondence between viruses and set of NVs, these three viruses were excluded from the subsequent study. The Hausdorff distance is utilized to measure the pairwise distance between point sets (Huttenlocher et al., 1993). This distance has been suitable to reconstruct the phylogenetic tree for multi-segmented viral genomes from different families when combined with Lempel–Ziv complexity or NV representation of nucleotide sequences (Yu et al., 2014; Huang et al., 2014). The extended version of it, Yau–Hausdorff distance, has achieved success in matching graphical curves of DNA or protein sequences (Tian et al., 2015).

To be precise, suppose A and B are two finite point sets in R^n. Their Hausdorff distance is defined by

$$h(A,B)=\max\left\{\max_{a\in A}\min_{b\in B}\ d(a,b),\max_{b\in B}\min_{a\in A}\ d(a,b)\right\},$$

where $d(a,b)$ is the Euclidean distance between two numeric vectors a and b. Note that when both the set A and B have only one member, the Hausdorff distance is reduced to the Euclidean distance. Additionally, unlike many similarity/distance measures in genomics, the Hausdorff distance is a true distance in the sense of mathematics, i.e. it is nonnegative, symmetric and satisfies triangle inequality. When comparing two viruses, this distance is free from the order of viral protein sequences in the form of NVs. The viral classification and phylogenetic tree can be built efficiently using the Hausdorff distance.

Download English Version:

https://daneshyari.com/en/article/5918517

Download Persian Version:

https://daneshyari.com/article/5918517

Daneshyari.com