Computational and Structural Biotechnology Journal xxx (2017) xxx–xxx

ELSEVIER

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

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

Assessing Species Diversity Using Metavirome Data: Methods and Challenges

Damayanthi Herath[a,b,*], Duleepa Jayasundara[c], David Ackland[a], Isaam Saeed[d], Sen-Lin Tang[e], Saman Halgamuge[f]

[a] Department of Mechanical Engineering, University of Melbourne, Parkville, 3010 Melbourne, Australia
[b] Department of Computer Engineering, University of Peradeniya, Prof. E. O. E. Pereira Mawatha, Peradeniya, 20400, Sri Lanka
[c] School of Public Health and Community Medicine, University of New South Wales, Randwick, NSW 2052, Australia
[d] PERSUIT, 134-146 Cambridge Street, Collingwood, Victoria 3066, Australia
[e] Biodiversity Research Center, Academia Sinica, Nan-Kang, Taipei 11529, Taiwan
[f] Research School of Engineering, College of Engineering and Computer Science, The Australian National University, Canberra 2601, ACT, Australia

ARTICLE INFO

Article history:
Received 13 March 2017
Received in revised form 1 September 2017
Accepted 11 September 2017
Available online xxxx

Keywords:
Metagenomics
Phage studies
Biodiversity
Species diversity
Metavirome data
Bioinformatics

ABSTRACT

Assessing biodiversity is an important step in the study of microbial ecology associated with a given environment. Multiple indices have been used to quantify species diversity which is a key biodiversity measure. Measuring species diversity of viruses in different environments remains a challenge relative to measuring the diversity of other microbial communities. Metagenomics has played an important role in elucidating viral diversity by conducting metavirome studies. However, the metavirome data are of high complexity requiring robust data preprocessing and analysis methods. In this review existing bioinformatics methods for measuring species diversity using metavirome data are categorised broadly as either sequence similarity-dependent methods or sequence similarity-independent methods. The former includes a comparison of DNA fragments or assemblies generated in the experiment against reference databases for quantifying species diversity, whereas estimates from the latter are independent of the knowledge of existing sequence data. Furthermore, current methods and tools are discussed in detail with examples of their applications and their limitations. Drawbacks of the state-of-the-art method are demonstrated through results from a simulation. In addition, alternative approaches are proposed to overcome the challenges in estimating species diversity measures using metavirome data.

1. Introduction

Most viruses in the environment exist in the form of parasites that infect prokaryotes and hence are frequently termed phages or bacteriophages. Recent studies [1,2] have shown that despite being identified as parasites, viruses may have symbiotic relationships that are beneficial to the host as well. Viruses represent the most abundant biological entity in the biosphere with an estimated phage population of ~10^{31} [3]. Many microbiological experiments conducted in the past highlight the effect that viruses have on different processes in our biosphere. Examples are their effects on food web and organic carbon flow in the oceans [4], and population structure of bacterial communities in the human gut [5,6]. The influence of viruses on driving ecological functionalities and evolutionary changes of prokaryotes has been previously highlighted, as well as the effect of viruses on the gene transfer across species [7]. One study [8] has illustrated the connection between the diversity of viruses and climate change with eight case studies concluding that viruses are significantly influenced by climate change and in turn, are affecting biological processes contributing to climate changes. These studies stress the importance of studying viral ecology in different environments.

The conventional method of analysing the behaviour of viruses involves infecting them into cultured prokaryotic hosts. Such culture-dependent approaches are limited in applicability because a large number of microbial hosts have not been cultured [9]. One way of studying microbes in a culture-independent manner is the use of

* Corresponding author at: Department of Mechanical Engineering, University of Melbourne, Parkville, Melbourne 3010, Australia.
E-mail address: damayanthi@ce.pdn.ac.lk (D. Herath).

https://doi.org/10.1016/j.csbj.2017.09.001

taxonomic marker genes like 16S ribosomal RNA gene (16S rRNA) that are conserved in genomes of all the species being studied [10]. However, due to the absence of such a conserved genomic region, the traditional marker genes based methods such as Polymerase Chain Reaction (PCR) and Fluorescence in situ hybridization (FISH) cannot be used to study viruses [9].

The emergence of Metagenomics helped in overcoming these challenges in studying the dynamics of viruses in different environments. Metagenomics refers to the biotechnological and bioinformatics methods involved in culture-independent analysis of genetic material of all microbial organisms in an environmental sample. A metagenome is the collection of genomic sequences of all the organisms in a given environment [9]. Advancements in high-throughput DNA sequencing and assembling techniques [11–13] have made metagenomics a popular approach for studying microbial ecology. The major steps involved in a metagenomics study have been previously reviewed [14] and include sample collection; extraction of DNA and removal of unwanted genetic material such as proteins, organelles and membranes; fragmentation of DNA using enzymes or mechanical techniques; sequencing of DNA; and bioinformatic analysis [14]. Metagenomics have a range of applications such as production of novel enzymes, discovery of new antibiotics and production of biosurfactants [15] and metagenomics related researches are being conducted around the world [16]. Moreover, metagenomics is expected to be highly effective in enteric disease diagnostics [17]. Bioinformatic analyses conducted on metagenomic data helps in expanding our knowledge on microbes in terms of taxonomic profiles, metabolic pathways and inter-species interactions etc. [18].

A metagenome of a viral population is termed a 'metavirome' [19]. The first metavirome study was an experiment carried out to study the ecology of viruses in marine environments using samples extracted from the two oceans Scripps Pier, CA and Mission Bay, San Diego. [20,21]. Thereafter, many studies have been conducted to analyse metaviromes of samples collected from different environments such as sea water [20,22], marine sediments [23], soil [24], human faeces [25,26] and the human gut [27–29].

Biodiversity is an important ecological parameter in understanding the dynamics of a given environment as there is a strong relationship between biodiversity and the stability of an ecosystem [30]. It can be quantified in three ways: α-diversity referring to the diversity of a given sample or environment, γ-diversity quantifying the collective diversity of multiple environments and β-diversity capturing the difference in diversity among environments [31]. Implications of α, β and γ diversities have been reviewed comprehensively [32,33]. One aspect often considered in a metagenomics study is α-diversity which is also termed 'species diversity'.

The definition of a *virus species* has been debated about [34,35], and is being updated [36]. Generally, the term *species* is used to refer to the lowest category in biological classification. However, whether the term species should be referred to an individual entity or an abstract class or category remains a debate [35]. Initially, the concept of species was considered to be not applicable for viruses because the early definition of species as *groups of interbreeding natural populations which are reproductively isolated from other such groups*, may not be related to viruses [34]. The International Committee on Taxonomy of Viruses (ICTV) which acts as the body responsible for maintaining the virus taxonomy [37], has accepted the formal definition of a virus species as "a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological niche" [34,38]. A *polythetic* class consists of members having multiple properties in common, but may not be defined by a single property [39]. Metagenomics can help in obtaining the assemblies of complete genome sequences of new viruses, however the obtained assemblies may lack information of their biological properties raising the concern how to define a virus species based on metagenomics data [36]. The term viral *genotype* has been used in the first metagenomic experiment of viruses [20] referring to in silico conditions assuring that sequences of different phage genomes may not assemble together [20,40]. The complexities in defining taxonomy of viruses as mentioned have been reviewed comprehensively [35] and implications of metagenomics in defining taxonomy of viruses have been discussed [36]. In 2016, ICTV endorsed a proposal made to classify viruses solely based on metagenomics sequence data. This proposal recommends retaining the ICTV definition of a virus species and using biological characteristics that may be inferred from sequence data such as genome organization, replication strategy, presence of homologous genes and host range or type of vector [36].

Alternative approaches to quantify biodiversity instead of measures of species diversity have been proposed [41,42]. An example is the suggestion to use statistical properties of communities with straightforward biological interpretations [41]. However, as far as metavirome studies are considered, estimation of species diversity is a key step in the bioinformatics analysis pipeline [43]. As far as viral communities are considered species diversity indices estimates are used to answer multiple questions. Examples are: use of species diversity estimates to learn the relationship between species richness and range size distributions in plants [44,45], demonstration of factors leading to the differences between the ambient and induced viral communities [46] considering species diversity of viruses, and prediction of zoonotic potential of mammalian viruses [47], modelling predator-prey dynamics based on rank-abundance distributions [48], use of evenness indices to determine factors affecting horizontal gene transfer and functional microbiome evolution in chicken cecum microbiome [49].

This review summarises the existing bioinformatics methods and tools for quantifying viral diversity from metavirome data. The widely considered species diversity measures in metavirome studies are described in brief with their definitions. The existing methods for estimating viral diversity measures are reviewed comparatively and their limitations are identified. Furthermore, possible alternative approaches are proposed to address the limitations in existing methods. Previous reviews have summarised various bioinformatics strategies used in existing methods for studying viruses [50,51]. This review discusses further methods for measuring species diversity from metavirome data with comparisons between them.

2. Common Measures of Viral Diversity

Three commonly considered species diversity measures in previous metavirome studies are species richness, Shannon-Wiener index and evenness. They represent the key quantitative species diversity measures: species richness, heterogeneity and equability [52]. The rank-abundance distribution and the relative abundance of genomes have been considered in addition (e.g.: [20,53-55]).

Species richness is the total number of species in a population and is estimated from a sample, a representative subset of the population. While two environments may have equal species richness, if some species are dominant in number in one environment (i.e. less diverse) these two environments should be considered as different in diversity. Evenness captures how uniformly the species are distributed in number in an environment and is related with the relative abundance of species. If there are n_i number of individuals from ith species, its relative abundance, $f_i = n_i \Big/ \sum_{i=1}^{M} n_i$ where M is species richness. Heterogeneity measures combine species richness with evenness [52]. A commonly used heterogeneity measure is the *Shannon - Wiener* index. Shannon - Wiener index [56]considers both species richness and relative abundance and is defined as $H' = -\sum_{i=1}^{M} f_i \quad ln \quad f_i$.

Please cite this article as: D. Herath et al., Assessing Species Diversity Using Metavirome Data: Methods and Challenges, Computational and Structural Biotechnology Journal (2017), https://doi.org/10.1016/j.csbj.2017.09.001