Physica A 482 (2017) 87–94

Contents lists available at ScienceDirect

Physica A

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

A novel genome signature based on inter-nucleotide distances profiles for visualization of metagenomic data

Xian-Hua Xie [a,b], Zu-Guo Yu [a,c,*], Yuan-Lin Ma [a], Guo-Sheng Han [a], Vo Anh [c]

[a] Hunan Key Laboratory for Computation and Simulation in Science and Engineering and Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan, Hunan, 411105, China
[b] Key Laboratory of Jiangxi Province for Numerical Simulation and Emulation Techniques, Gannan Normal University, Jiangxi, 341000, China
[c] School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Q4001, Australia

HIGHLIGHTS

- We visualize metagenomic data using inter-nucleotide distances profile.
- We analyze these profiles by principal component analysis.
- Three benchmark data sets are evaluated by the novel method.
- Our method is efficient and powerful for visualization of metagenomic data.

ARTICLE INFO

Article history:
Received 6 December 2016
Received in revised form 12 March 2017
Available online 20 April 2017

Keywords:
Alignment-free method
Inter-nucleotide distance
Metagenome
Visualization

ABSTRACT

There has been a growing interest in visualization of metagenomic data. The present study focuses on the visualization of metagenomic data using inter-nucleotide distances profile. We first convert the fragment sequences into inter-nucleotide distances profiles. Then we analyze these profiles by principal component analysis. Finally the principal components are used to obtain the 2-D scattered plot according to their source of species. We name our method as *inter-nucleotide distances profiles (INP)* method. Our method is evaluated on three benchmark data sets used in previous published papers. Our results demonstrate that the INP method is good, alternative and efficient for visualization of metagenomic data.

1. Introduction

Ever since Handelsman et al. [1] first proposed the concept of metagenome, great progress has been made in the research of metagenomics [2,3]. High-throughput sequencing with massively parallel 454 pyrosequencing was used in the first metagenomic studies [4]. Assuming an average genome size of 4 Mb, only about 5%–10% of the clones within a fosmid library (40 kb insert size) harbor phylogenetic marker genes like 16S rDNA, rpoA, recA, etc. and can therefore be assigned to a certain species or taxonomic group [5]. Most of the clones do not contain any marker genes. A key step in genome recovery from metagenomic sequence data is the classification of sequences assembled from metagenomic reads into bins. These bins represent composite genomes of individual populations that comprise the microbial community. Several approaches have been developed to bin assembled sequences from metagenomic data [6–15]. Among these techniques, one of the most widely used is emergent self-organizing maps (ESOMs), which have been used to bin assembled sequences based on tetranucleotide

* Correspondence to: School of Mathematics and Computational Science, Xiangtan University, Hunan, 411105, China.
E-mail address: yuzuguo@aliyun.com (Z. Yu).

http://dx.doi.org/10.1016/j.physa.2017.04.064
0378-4371/

frequencies [6,12]. ESOMs based on tetranucleotide frequencies can be applied to individual metagenomic data sets. Based on tetranucleotide frequencies of fragments of genome, Laczny et al. [16] reported a scalable approach for the visualization of metagenomic data via nonlinear dimension reduction.

Alignment-free methods has been increasingly used for genome comparison, in particular for genome-based phylogeny reconstruction, including information-based analysis [17,18], singular value decomposition (SVD) [19,20], principal component analysis [21], the ZL compression distance [22] which is based on the algorithm of Ziv and Lempel [23], fractal methods [24–27], Markov model [28–32], dynamical language model [33–35], the average common substring (ACS) distance [36], the maximum unique matches (MUM) distance [37,38], feature frequency profiles (FFP) [39–41], distance measure based on return time distribution [42], method using spaced-word frequencies [43,44]. The inter-nucleotide distance [45] is a novel DNA numerical profile to explore the correlation structure of DNA. Afreixo et al. [46] explored this method to construct phylogenetic tree using genomic data. Then Afreixo et al. [47] introduced the distance to the nearest dissimilar nucleotide to construct the phylogenetic tree. Based on the inter-nucleotide distance sequence, Chang and Wang [48] proposed the conditional multinomial distribution profile for phylogeny on genomic level. These profiles can be used to define a distance that reflects the evolutionary relationships between genomic sequences [48]. Inspired by Afreixo et al. [46], Xie et al. [49] proposed the inter-amino-acid distance profile for phylogeny on whole-proteome level.

In the present study, we propose a new method as characteristic of fragments of genome to plot the visibility graph (the difference between our new method and those in Refs. [46–48] is given in Remark 1 in Section 2). We evaluate our new method on three benchmark data sets used in previous published papers.

2. Inter-nucleotide distance based model

2.1. Inter-nucleotide distance in a DNA sequence

Consider the alphabet $\Omega = \{A, C, G, T\}$ and let $s = s_1 s_2 \cdots s_M$ be a DNA sequence, where each s_i is a symbol in Ω. The sequence distance from s_i to s_{i+k} is k for $k \geq 1$. For a pair of nucleotides $(x, y) \in \Omega \times \Omega$, the inter distance between x and y (we call it *inter-nucleotide distance*) is the sequence distance of nucleotide x and the first y after the x, here we require x is the nearest one to y. For each pair of x and y in a DNA sequence s, we denote by $d^{(x,y)}$ the vector of all these distances presents in s. As an example, all vectors of distances for the short DNA fragment ACTCTCCATA are:

$$\begin{aligned} &d^{(A,A)} = (7, 2)\ d^{(A,C)} = (1),\ \ d^{(A,T)} = (2, 1),\\ &d^{(C,A)} = (1),\ \ d^{(C,C)} = (2, 2, 1),\ \ d^{(C,T)} = (1, 1, 2),\\ &d^{(T,A)} = (3, 1),\ \ d^{(T,C)} = (1, 1),\ \ d^{(T,T)} = (2, 4), \end{aligned} \tag{1}$$

here, we deal with the DNA sequence as acyclic. From $d^{(x,y)}$, we calculate the observed probability $f_0^{(x,y)}(k)$ for each inter-amino-acid distance k. We have $f_0^{(C,C)}(1) = \frac{1}{3}$ and $f_0^{(C,C)}(2) = \frac{2}{3}$ in the above example.

2.2. The nearest dissimilar distance in a DNA sequence

Consider the alphabet $\Omega = \{A, C, G, T\}$ and let $s = s_1 s_2 \cdots s_M$ be a DNA sequence, where each s_i is a symbol in Ω. According to Ref. [47], consider a numerical vector, w^x, that represents the distance to the nearest dissimilar of symbol $x \in \Omega$. For the last symbol s_M, if $s_M \neq s_{M-1}$, the nearest dissimilar distance of s_M is 1, otherwise, the nearest dissimilar distance of s_M is the repeat length of s_M. As an example, all vectors of distances for the short DNA fragment ACCCTTCTCTCCATGGAAGGACCT are:

$$w^A = (1, 1, 2, 1),\ \ w^C = (3, 1, 1, 2, 2),\ \ w^G = (2, 2),\ \ w^T = (2, 1, 1, 1, 1), \tag{2}$$

here, we deal with the nucleotide sequence as acyclic. From w^x, we calculate the observed probability $f_1^x(k)$ for each nearest dissimilar distance k. We have $f_1^A(1) = \frac{3}{4}, f_1^A(2) = \frac{1}{4}, f_1^C(1) = \frac{2}{5}, f_1^C(2) = \frac{2}{5}, f_1^C(3) = \frac{1}{5}, f_1^G(2) = 1, f_1^T(1) = \frac{4}{5}$, and $f_1^G(2) = \frac{1}{5}$ in the above example.

2.3. Characteristics extraction of fragments in genome

For a fixed K, we consider $f_0^{(x,y)}(k)$ and $f_1^x(k)$, $k = 1, 2, \ldots, K$, and x and y are symbols in Ω. There are a total of $N = 4 \times K + 4 \times 4 \times K$ numbers (we set $f_0^{(x,y)}(k) = 0$ and $f_1^x(k) = 0$ if vectors $d^{(x,y)}$ and w^x do not have number k as a component, respectively) for each fragment of a genome. We can arrange components of $f_0^{(x,y)}(k)$ and $f_1^x(k)$ according to a fixed alphabetical order of (x, y) and x to form a new vector:

$$X(s) = (f_1^A(1), f_1^A(2), \ldots, f_1^A(K), f_1^C(1), \ldots, f_1^T(1), f_1^T(2), \ldots, f_1^T(K), f_0^{AA}(1), \ldots, f_0^{TT}(K)) \tag{3}$$

for fragment s. We name the vector $X(s)$ as the *Inter-nucleotide distance profile (INP)* of fragment s.

Remark 1. The difference between our INP method and those in previous works [46–48] is the following. Our INP uses the nearest dissimilar distance and the inter-nucleotide distances of all possible pairs $(x, y) \in \Omega \times \Omega$. The methods in Refs. [46,48] used the inter-nucleotide distances of pairs $(x, x) \in \Omega \times \Omega$ with $x = y$. The method in Ref. [47] just replaced "the inter-nucleotide distances of pairs $(x, x) \in \Omega \times \Omega$ with $x = y$" by "the distance to the nearest dissimilar nucleotide".