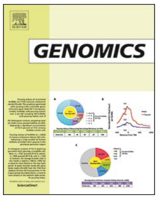
Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

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



Applying MSSIM combined chaos game representation to genome sequences analysis

Hai ming Ni[a], Da wei Qi[b,*], Hongbo Mu[b]

[a] College of Mechanical and Electrical engineering, Northeast Forestry University, Hexing Road 26, Harbin, Heilongjiang Province 150040, PR China
[b] College of Science, Northeast Forestry University, Hexing Road 26, Harbin, Heilongjiang Province 150040, PR China

ARTICLE INFO

Keywords:
Chaos game representation
Structural similarity
Hierarchical clustering analysis
Genome sequences

ABSTRACT

Converting DNA sequence to image by using chaos game representation (CGR) is an effective genome sequence pretreatment technology, which provides the basis for further analysis between the different genes. In this paper, we have constructed 10 mammal species, 48 hepatitis E virus (HEV), and 10 kinds of bacteria genetic CGR images, respectively, to calculate the mean structural similarity (MSSIM) coefficient between every two CGR images. From our analysis, the MSSIM coefficient of gene CGR images can accurately reflect the similarity degrees between different genomes. Hierarchical clustering analysis was used to calculate the class affiliation and construct a dendrogram. Large numbers of experiments showed that this method gives comparable results to the traditional Clustal X phylogenetic tree construction method, and is significantly faster in the clustering analysis process. Meanwhile MSSIM combined CGR method was also able to efficiently clustering of large genome sequences, which the traditional multiple sequence alignment methods (e.g. Clustal X, Clustal Omega, Clustal W, et al.) cannot classify.

1. Introduction

Genomic sequences analysis, as one of the most important part of bioinformatics, which was considered to reveal the essence of all life phenomenons, has been developing rapidly in recent years [1–5]. Genomic sequences analysis is mainly contributed to similarities and differences in the intergenic arrangement based on the four bases oligonucleotides [6]. The difference orders of the four basic oligonucleotides determined the different genome's functions. From the early sequence alignment to the recent global DNA sequences prediction, the genomic sequences analysis technology has been developing toward the direction of intelligence, high efficiency and automation [7–9].

Chaos game representation (CGR) is an iterative mapping technique that divides the genomic sequences into certain units, and further research for finding the correlation of their position in a continuous gene sequence [10]. The CGR method is a technology of constructing pictures to reveal the patterns of gene sequence [6]. Firstly, the gene sequence can be regarded as an alphabet set, which are composed of four basic characters A, T, G, and C. Then, the whole alphabet set of frequencies of the words found in a certain gene sequence can be expressed as a single gray level image that have a one to one correspondence between each pixel and each specific word. Since the CGR method was invented by Jeffrey in 1990 [11], this novel genome sequences analysis method has been widely used in genome sequences analysis and protein analysis [12–16].

The image structural similarity (SSIM) index is a comprehensive evaluation indicator that is used in assessment the image quality. The SSIM index is mainly composed of luminance, contrast ratio and structure factor. The mean structural similarity (MSSIM) index has been put forward to compare the overall similarity between a pair of images [17].

In this paper, we proposed a new dendrogram construction method, which combine CGR and structural similarity method to construct a MSSIM coefficient diagonal matrix, and then establish dendrogram using hierarchical clustering analysis. This experiment indicates that our dendrogram establishment method can accurately identify the genetic relationship of different biology, and this method was generally applicable to various organisms.

2. Methods

In this study, the novel method combining the CGR and the MSSIM was applied to genome sequence data. Firstly, the genome sequence was converted into an image through the iterative mapping technique CGR, which was proposed by Jeffrey [11]. For the fixed word length, the size of CGR images was also determined for different genomes, even

[*] Corresponding author.
E-mail address: qidw9806@126.com (D.w. Qi).

http://dx.doi.org/10.1016/j.ygeno.2017.09.010
Received 18 March 2017; Received in revised form 11 September 2017; Accepted 19 September 2017
0888-7543/ © 2017 Elsevier Inc. All rights reserved.

Table 1
The list of 10 different species of mammalian chromosomes DNA sequences information considered in this work.

Species	GI number	Strain name	Nucleotide length
Cat	1,098,523	CA	17,009
Common chimpanzee	643,684	CC	16,563
Cow	12,800	CO	16,338
Dog	7,534,303	DO	16,727
Gorilla	643,689	GO	16,364
Mouse	13,838	MO	16,295
Opossum	452,251	OP	17,084
Pigmy chimpanzee	643,679	PC	16,554
Rat	854,269	RA	16,300
Sheep	3,445,513	SH	16,616

if the length of genome sequences were unequal. The size of CGR image can be described as 4^n (the total pixels of image), where n is the size of words and 4^n is the total number of words. Each pixel gray value in the CGR image represents the occurrence frequency of the corresponding words in the gene sequences. The pixels of natural image signals have strongly dependencies, especially when they are spatial similarity. Therefore, we can explore the relationship between two gene sequences through comparing the similarity of their CGR image. The MSSIM coefficient can be calculated by using the method of literature [17]. The specific calculation process was shown as follows:

First of all, for discrete signal, we use average gray level as a luminance measurement estimates:

$$\alpha_x = \frac{1}{N}\sum_{1}^{N} x_i \tag{1}$$

where x_i is the gray value of image, N is the size of image.

The luminance comparison function $l(x,y)$ is a function of α_x and α_y.

Then, remove the average gray value from the signal. For discrete signal $x - \alpha_x$, the standard deviation can be used as an estimate of the contrast ratio.

$$\beta_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \alpha_x)^2\right)^{1/2} \tag{2}$$

The contrast function $c(x,y)$ is then a function of β_x and β_y.

Next, the signal is divided by its own standard deviation. The structure comparison function $s(x,y)$ is defined as a function of $\frac{x-\alpha_x}{\beta_x}$ and $\frac{y-\alpha_y}{\beta_y}$.

Finally, the three elements are combined into a complete similarity measurement function:

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{3}$$

where the similarity measurement function $S(x,y)$ satisfy the following three conditions:

Symmetry:$S(x,y) = S(y,x)$;

Boundedness:$S(x,y) \le 1$;

Unique maximum:If and only if $x = y$, $S(x,y) = 1$.

Next, we need to defined the three functions (x,y), $c(x,y)$ and $s(x,y)$ as follows:

The luminance contrast function:

$$l(x,y) = \frac{2\alpha_x\alpha_y + C_1}{\alpha_x^2 + \alpha_y^2 + C_1} \tag{4}$$

where the constant C_1 is defined to avoid the instability when ${\alpha_x}^2 + {\alpha_y}^2$ is close to 0.

In particular, we select $C_1 = (K_1L)^2$, where L is the gray level of image, K_1 is a constant.

The contrast discrimination function:

$$c(x,y) = \frac{2\beta_x\beta_y + C_2}{\beta_x^2 + \beta_y^2 + C_2} \tag{5}$$

where $C_2 = (K_2L)^2$ and $K_2 \ll 1$ are constants.

The structure comparison function:

$$s(x,y) = \frac{\beta_{xy} + C_3}{\beta_y\beta_y + C_3} \tag{6}$$

where $\beta_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \alpha_x)(y_i - \alpha_y)$.

Finally, we combine the three comparison functions Eqs. (4)–(6), and the complete similarity measurement function SSIM is obtained:

$$SSIM(x,y) = [l(x,y)]^{\rho}[s(x,y)]^{\varepsilon}[s(x,y)]^{\delta} \tag{7}$$

where $\rho > 0, \varepsilon > 0, \delta > 0$ are used to adjust the relative importance of the three functions.

In order to obtained the simplify form, we set $\rho = \varepsilon = \delta = 1$, and $C_3 = C_2/2$. The simplified SSIM index is as follows:

$$SSIM(x,y) = \frac{(2\alpha_x\alpha_y + C_1)(2\beta_x\beta_y + C_2)}{(\alpha_x^2 + \alpha_y^2 + C_1)(\beta_x^2 + \beta_y^2 + C_2)} \tag{8}$$

By comparing the SSIM index of the CGR image between every pair of genome sequences using hierarchical clustering analysis method, the phylogenetic tree can be obtained.

3. Results and discussion

To verify the validity of this method, we have chosen the widely used Clustal X method as a comparison. The distance matrix of our MSSIM combined CGR method is calculated by using Matlab 2012a. All the dendrograms in this research are drawn using MEGA 6.0 [18]. Computations in this research are performed by using a server with the configuration of dual Intel Core i7 processors running at 2.4 GHz with 16 RAM.

3.1. Eutherian mammals

In this study, we have researched the phylogeny of 10 different species of mammalian chromosomes DNA sequences by using a federation approach combining CGR and MSSIM index. Jie Tang [19] and Chun Li [20] et al. have been correctly classified them in previous studies. The 10 different species of chromosomes DNA sequences data, which calculated in this work were obtained from the National center for biotechnology information (NCBI) genome database (http://www.ncbi.nlm.nih.gov/). The details information (including bacteria species, GI number, Strain name and Nucleotide length) of these genome sequences were listed in Table 1.

In the first step, we considered the words' length $n = 8$, and obtained a 256 × 256 pixel CGR grayscale image. Save the CGR grayscale image as JPEG format. We have chosen 10 mammalian chromosomes DNA sequences as research object which have close kinship. They are Cat, Common chimpanzee, Cow, Dog, Gorilla, Mouse, Opossum, Pigmy chimpanzee, Rat, Sheep. The CGR images for these 10 different species of mammalian chromosomes DNA sequences at the words length $n = 8$ were shown in Fig. 1.

In the further step, according to the related literature [17], we chose an 11 × 11 symmetry Gaussian weighting function $W = \{\omega_i \mid i = 1, 2, \ldots, N\}$ as the weighted window. With such a windowing approach, the compared images were divided into many contiguous local windows. Meanwhile, we set the standard deviation of 1.5 and the SSIM measure parameter $K_1 = 0.01$, $K_2 = 0.03$. According to the normalized law, we defined the unit sum $\sum_{i=1}^{N}\omega_i = 1$. Then the estimate value of α_x, β_x and β_{xy} can be expressed as follows:

Download English Version:

https://daneshyari.com/en/article/8646325

Download Persian Version:

https://daneshyari.com/article/8646325

Daneshyari.com