Accepted Manuscript

A new method to cluster genomes based on cumulative Fourier power spectrum

Rui Dong, Ziyue Zhu, Changchuan Yin, Rong L. He, Stephen S.-T. Yau

PII:	S0378-1119(18)30698-X
DOI:	doi:10.1016/j.gene.2018.06.042
Reference:	GENE 42975
To appear in:	*Gene*
Received date:	7 March 2018
Revised date:	12 June 2018
Accepted date:	14 June 2018

Please cite this article as: Rui Dong, Ziyue Zhu, Changchuan Yin, Rong L. He, Stephen S.-T. Yau , A new method to cluster genomes based on cumulative Fourier power spectrum. Gene (2018), doi:10.1016/j.gene.2018.06.042

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

A new method to cluster genomes based on cumulative Fourier power spectrum

Rui Dong[a,1], Ziyue Zhu[a,1], Changchuan Yin[b], Rong L. He[c], Stephen S.-T. Yau[a,*]
[a]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
[b]Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, IL 60607, USA
[c]Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA

[*]Corresponding author. Email: yau@uic.edu, Address: Department of Mathematics, Tsinghua University, Beijing 100084, China
[1]These authors contributed equally to this work

Abstract

Analyzing phylogenetic relationships using mathematical methods has always been of importance in bioinformatics. Quantitative research may interpret the raw biological data in a precise way. Multiple Sequence Alignment (MSA) is used frequently to analyze biological evolutions, but is very time-consuming. When the scale of data is large, alignment methods cannot finish calculation in reasonable time. Therefore, we present a new method using moments of cumulative Fourier power spectrum in clustering the DNA sequences. Each sequence is translated into a vector in Euclidean space. Distances between the vectors can reflect the relationships between sequences. The mapping between the spectra and moment vector is one-to-one, which means that no information is lost in the power spectra during the calculation. We cluster and classify several datasets including Influenza A, primates, and human rhinovirus (HRV) datasets to build up the phylogenetic trees. Results show that the new proposed cumulative Fourier power spectrum is much faster and more accurately than MSA and another alignment-free method known as k-mer. The research provides us new insights in the study of phylogeny, evolution, and efficient DNA comparison algorithms for large genomes. The computer programs of the cumulative Fourier power spectrum are available at GitHub (https://github.com/YaulabTsinghua/cumulative-Fourier-power-spectrum).

Keywords:

cumulative Fourier power spectrum, moment vectors, DNA sequences, phylogenetic trees

1.Introduction

In molecular biology, mathematical methods are often used to interpret biological sequence information. Mathematics can transform biological sequences into numerical representations to analyze them quantitatively. Genetic recombination and, in particular, genetic shuffling are at odds with sequence comparison by alignment, which assumes conservation of contiguity

Download English Version:

https://daneshyari.com/en/article/8644611

Download Persian Version:

https://daneshyari.com/article/8644611

Daneshyari.com