On The Use Of Lacunarity Analysis Of Cgr Images For Metagenomic Classification

Aiswarya Prasad
New Horizon College of Engineering, Bangalore
aiswaryaprasad13@gmail.com

Abstract: Metagenomic projects collect DNA from uncharacterized environments that may contain thousands of species per sample. Because the output from metagenomic sequencing is a large set of reads of unknown origin, clustering reads together that were sequenced from the same species is a crucial analysis step . This paper compares the the use of lacunarity analysis of the Chaos Game Representation of metagenomic sequences of various lengths.. It demonstrates the efficiency of lacunarity analysis for shorter nucleotide sequences.

Keywords: Metagenomes, Chaos Game Representation, Lacunarity, Gliding Box Algorithm

I. Introduction

Sequencing of environmental DNA(often called metagenomes) [i] has shown tremendous potential to uncover the vast number of unknown microbes that cannot be cultured and sequenced by traditional methods. Two main challenges of metagenomics are to determine the species present in the mixture and their proportion. But we have only a fragmented assembly of short sequences attaining these goals difficult. Advances in computational analysis techniques are essential to move the field forward.

Phymm, a classification method based on interpolated markov model, was proposed to characterize taxonomic groups [ii]. PhyloPythia, a composition based classifier could accurately classify genomic fragments of 13 kb for all taxonomic ranks considered; but for fragments shorter than 5 kb, and in particular for those shorter than 3 kb, the sensitivity decreases markedly[iii]. An unsupervised sequence clustering method called SCIMM(Sequence Clustering with Interpolated Makov Models) used the Classification Expectation Maximization(CEM) algorithm.

Mandelbrot has come up with the idea of using lacunarity as a measure to classify fractals. Lacunarity is a measure of how patterns fill the space. At a given scale, low lacunarity indicates being homogeneous and high lacunarity indicates being heterogeneous. Hence lacunarity is considered to be a scale dependent measure of heterogeneity or texture. It is possible to utilize these properties of lacunarity to make it useful in the field of metagenomic sequence analysis[iv].

In this work, the Chaos Game Representation(CGR)[v] was used to represent the nucleotide sequences followed by extraction of feature vector by lacunarity analysis. The features were compared for metagenomic sequences of various lengths. Even for shorter nucleotide sequences , the lacunarity values are comparable to the longer nucleotide sequences which makes it possible to apply this idea into classification problem atleast at the phylum level.

II. Material and Methodology

A.Chaos Game Representation

The chaos game representation(CGR) is a scatter plot derived from any sequence, with each point of the plot corresponding to one element of the sequence. It is a new tool for investigating gene structure. The CGR can recognize patterns in the nucleotide sequences, of a class of genes using the techniques of fractal structures .Since the patterns are unique to genomes this can be used to identify genome fragments. DNA sequence is having some kind of non-randomness. The Chaos Game can be also be used to display certain kinds of non-randomness present in the DNA sequence visually. By considering DNA sequences as strings composed of four units, G, A, T and C, the chaos game algorithm to plot the CGR image for a short DNA sequence ’gaattc ’is ’ as follows and shown in figure 1[v][ix].

1) Take a square whose corners are marked with the alphabets

a, g, c and t which represents the four nucleotides.

2) The first ’g’ is plotted half way between the center of the square and the ’g’ corner.

3) The next base, ’a’, is plotted half way between the point just plotted and the ’a’ corner.

4) The base ’a’ is plotted half way between the previous point and the ’a’ corner.

5) Next, ’t’ is plotted half way between the previous point and the ’t’ corner.

6) Repeat for all the nucleotides in the sequence.

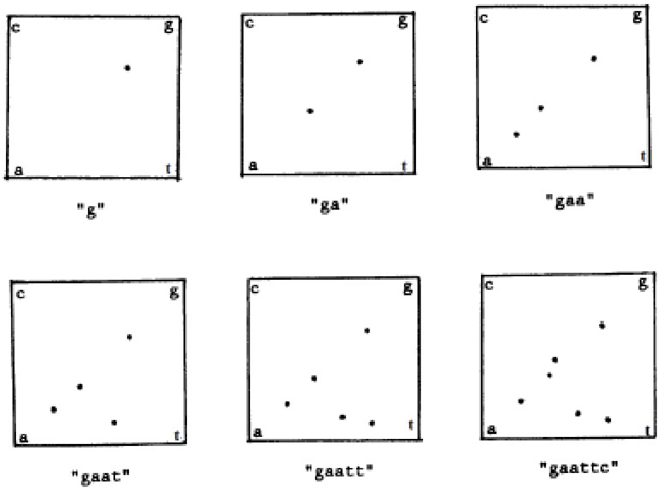


Fig.1. Plotting CGR image of a short length DNA sequence ‘GATTC’

Corresponding to this final CGR image, a binary matrix can be derived by giving a value of 1 for the presence of a dot and 0 for white.

B. Lacunarity using Gliding Box Algorithm

A simple way to calculate lacunarity on a binary image is the gliding box method [viii]. The algorithm is as follows.Take an MXM matrix, with a value 1 for the presence of an object and 0 for the absence. Place rxr (r can vary from 1 to M) box over the top left corner of the matrix. Box mass is taken to be equal to the number of occupied sites.Move the box one column to the right and calculate all the new box masses .Over all the columns and rows, this process is repeated.

The total number of boxes containing P occupied sites is taken as n(P,r) and the total number of boxes of size r as N(r). The map size is M.

$$N(r)=(M-r+1)^2$$

The frequency distribution n(P,r) is converted into probability distribution Q(P,r) as

$$Q(P,r)=n(P,r)/N(r).$$

$$Z^{(1)} = \quad \sum PQ(P$$

$$Z^{(2)} = \quad \sum F(P,r)$$

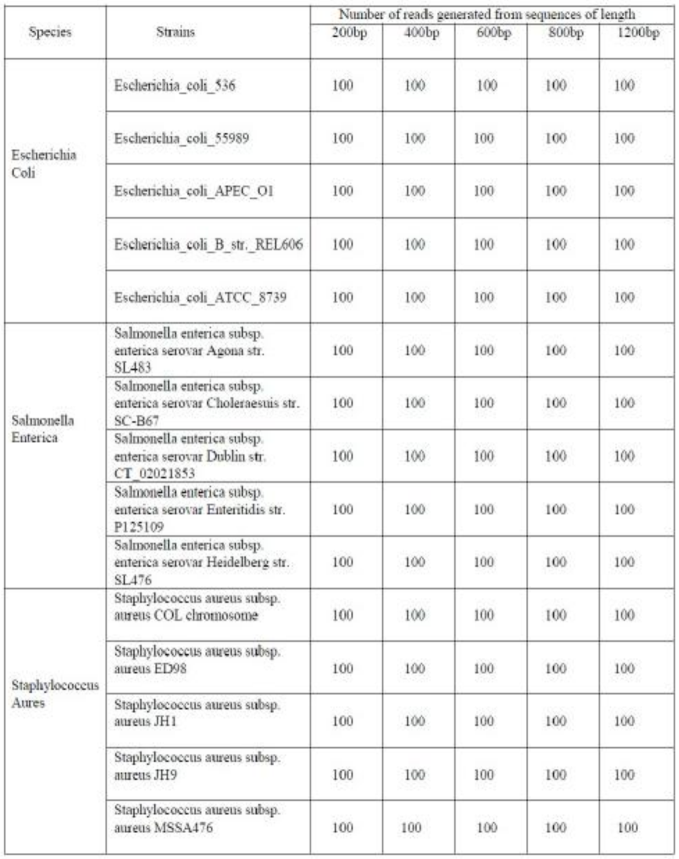
where Z $^{(1)}$ and Z $^{(2)}$ are the first and second moments of the frequency distribution.

Lacunarity is given by,

$$\lambda(r)=Z^{(2)}/(Z^{(1)})^2$$

C. Datasets used

The chromosomal DNA sequences of 5 different strains of 3 microbial species were obtained from NCBI Organelle database (www.ncbi.nlm.nih.gov)[x]. From these 15 strains, 100 reads of lengths 1200bp, 800bp, 600bp, 400bp, 200bp were generated.



Species	Strains	Number of reads generated from sequences of length				
		200bp	400bp	600bp	800bp	1200bp
Escherichia Coli	Escherichia_coli_536	100	100	100	100	100
	Escherichia_coli_55989	100	100	100	100	100
	Escherichia_coli_APEC_O1	100	100	100	100	100
	Escherichia_coli_B_str._REL606	100	100	100	100	100
	Escherichia_coli_ATCC_8739	100	100	100	100	100
Salmonella Enterica	Salmonella enterica subsp. enterica serovar Agona str. SL483	100	100	100	100	100
	Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	100	100	100	100	100
	Salmonella enterica subsp. enterica serovar Dublin str. CT_02021853	100	100	100	100	100
	Salmonella enterica subsp. enterica serovar Enteritidis str. P125109	100	100	100	100	100
	Salmonella enterica subsp. enterica serovar Heidelberg str. SL476	100	100	100	100	100
Staphylococcus Aures	Staphylococcus aureus subsp. aureus COL chromosome	100	100	100	100	100
	Staphylococcus aureus subsp. aureus ED98	100	100	100	100	100
	Staphylococcus aureus subsp. aureus JH1	100	100	100	100	100
	Staphylococcus aureus subsp. aureus JH9	100	100	100	100	100
	Staphylococcus aureus subsp. aureus MSSA476	100	100	100	100	100

Table 1. Datasets used for classification

D. Mapping of sequences into CGR

The Chaos Game Representation(CGR) of 1500 reads were plotted for nucleotide sequences of length 200bp, 400 bp, 600bp, 800bp and 1200bp by the method described in section II.A. To extract features, lacunarity of these CGR images was found for various boxes of size rxr where r varies from 1 to in multiples of 2

III. Results and Tables

For each test, 15 strains of 3 microbial species(5 from each) were chosen. 100 reads of length 200bp, 400bp, 600bp, 800bp and 1200bp from each of these 15 strains(1500 in total) were generated.

Fig. 2 shows the CGR image of Escherichia coli-536 for nucleotide sequence of length 1200bp.

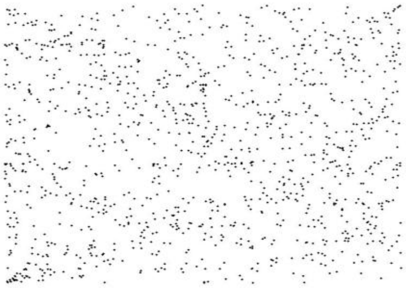


Fig.2. CGR image of Escherichia Coli-536

Lacunarity plots of CGR images of 3 microbial species are shown the figure 3 for nucleotide sequences of length 600bp, 800 bp, 1200 bp respectively.

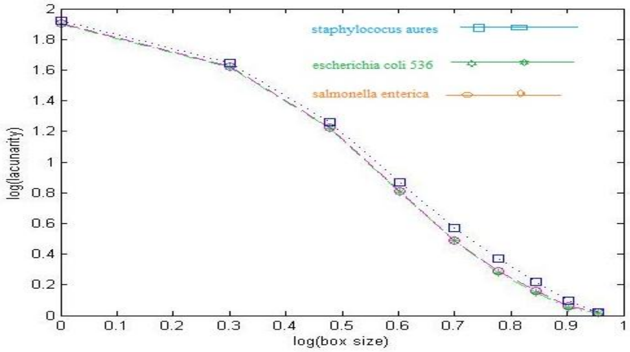


Fig. 3. Lacunarity plots of CGR images for length 800bp

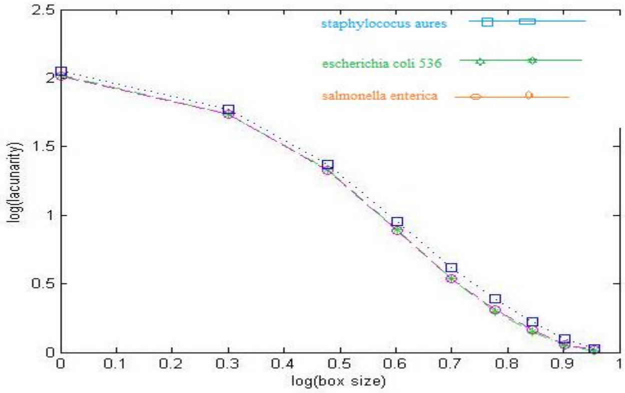


Fig. 4. Lacunarity plots of CGR images for length 800bp

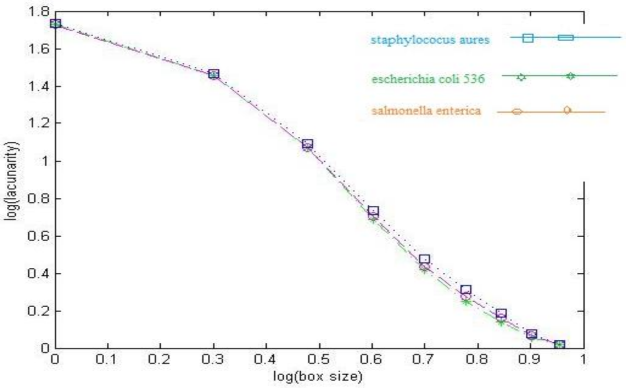


Fig. 5. Lacunarity plots of CGR images for length 1200bp

The curves shown in figures 3,4 and 5 are similar to the plots obtained that for the lacunarity analysis of fractal levy dust by Plotnick[vii] . It shows that nucleotides in the genomes are also hierarchically clumped. Eventhough for shorter nucleotide sequences also it provides better classification accuracy. From these graphs, it is clear that lacunarity of CGR images of metagenomic sequences can be used for classification. Both E.coli and Salmonella enteric belongs to the same family at phylum level. So their lacunarity plots are overlapping.

IV. Conclusions

In this paper work, the use of lacunarity analysis of CGRs of metagenomic sequences as a promising sequence analysis method have been demonstrated. This study can be extended to metagenomic classification using supervised and unsupervised methods which are expected to give better accuracies than the existing methods.

Acknowledgement

I would like to thank researchers and faculties at New horizon College of Engineering, Bangalore, for their contribution towards this work.

References

i. *Gail L. Rosen, Bahrad A. Sokhansanj, “Signal Processing for Metagenomics ”, Current Genomics, Vol.10, No.7, 2009.*

ii. *Arthur Brady, Steven L Salzberg, “Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models ”,Nature Methods, Vol.6 No.9, September 2009.*

iii. *Alice Carolyn McHardy, Hector Gracia Martin, Aristotelis Tsirigos, Philip Hugenholtz, Isidore Rigoutsos, “Accurate phylogenetic classification of variable-length DNA fragments”, Nature Methods, Vol.4 No.1, January 2007.*

iv. *Gopakumar G, Achuthshankar S. Nair ”Lacunarity Analysis of genomic sequences: A potential bio-sequence analysis method”,2011, IEEE*

v. *H. Joel Jefrey, “Chaos game representation of gene structure ”, Nucleic Acids Research, Vol.18, No.8, 1990.*

vi. *B.Mandelbrot, the fractal geometry of nature. Freeman, Newyork, 1983*

vii. *R.E. Plotnick, R.H.Gardner, W.W Hargrove, K.Prestegaard and M . Perlmutter, “Lacunarity analysis: A general technique for the analysis of spatial patterns”, Phys. Rev.E.vol.53,pp.5461-5468,1996*

viii. *C.Allain and M.Cloitre, “Characterising Lacunarity of random and deterministic fractal sets”, Phys. Rev. A. vol.44, pp.3552-3558,1991.*

ix. *Patrick J. Deschavanne, Giron A., Vilain J., Fagot G. and Fertil “Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences”,Mol. Biol. Evol. 16(10):13911399, 1999.*

x. *http:www.ncbi.nlm.nih.gov*