Article

Microbial Diversity Based on Multifractal Analysis of Metagenomes

Xianhua Xie [1,2,*], Yuanlin Ma [2], Zuguo Yu [2] and Guosheng Han [2]

[1] Key Laboratory of Jiangxi Province for Numerical Simulation and Emulation Techniques, Gannan Normal University, Jiangxi, 341000, P.R. China
[2] 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, Hunan 411105, P.R. China; 252928786@qq.com(Y.M); yuzg@xtu.edu.cn(Z.Y.);hangs@xtu.edu.cn(G.H.)
* Correspondence: xxianhua@sina.com.cn

Abstract: Species diversity in microbiome is a cutting-edge concept in metagenomic research. In this study, we propose a multifractal analysis for metagenomic research. From the chaos game representation (CGR) visualization of simulated and real metagenomes, we find that there exists self-similarity in the visualization of metagenomes. Then we compute the multifractal dimensions for simulated and real metagenomes. For simulated metagenomes, we also compute their diversity indices, such as species richness indices, Shannon's diversity indices and Simpson's diversity indices respectively for varying value of q. Fom the Pearson correlation coefficients between their multifractal dimensions and traditional species diversity indices, we find that the correlation coefficients between the multifractal dimensions and species richness indices and Shannon diversity indices reach their maximums at $q = 0$， 1respectively. The correlation coefficients between the multifractal dimensions and Simpson's diversity indices reach their maximums at $q = 2$ nearly. So the traditional diversity indices can be unified by the frame of multifractal analysis. These results coincided with the similar results in macrobial ecology. Finally, we apply our methods to real metagenomes of 100 infants' gut microbiomes when they are newborn, 4 months and 12 months. Our results show that multifractal dimensions of infants' gut microbiomes can discriminate the age difference.

Keywords: Diversity index; multifractal; metagenome; gut metagenome

1. Introduction

Species diversity in ecology has been long studied [1,2]. Generally, diversity indices can be divided into two classes (α diversity indices and β diversity indices). All diversity indices referred in this report are α diversity indices. In macrobial (plants/animals), α diversity can be characterized by species richness, Shannon diversity index and Simpson diversity index. Usually, in the field of macrobial ecology, with the increasing of ecology area, species richness is increasing. Generally, species-area relationship (SAR) can be formulated as $S(A)=cA^{z}$, where A is area, $S(A)$ is the number of species in A, c and z are constant. SAR is a famous formula in ecological study [3]. On the basis of SAR, Harte and Kinzig pointed out that the formula indicates the self-similarity of species number and area [4]. As main feature of fractals, self-similarity can be described by

$$z_q = \lim_{A\to+\infty} \frac{1}{1-q} \cdot \frac{\ln \sum_{i=1}^{S(A)} p_i^q}{\ln(A)}$$

and

$$z_1 = \lim_{A\to+\infty} \frac{-\ln \sum_{i=1}^{S(A)} p_i \ln(p_i)}{\ln(A)} .$$

Generally, for $q<0$, z_q emphasis the character of rare species, for $q>0$, z_q emphasis the common species. Particularly, z_0 implies the relationship of the logarithm of species richness ($\ln(S(A))$) and the logarithm of the area ($\ln(A)$). z_1 implies the relationship of the logarithm of Shannon diversity (SHD) index and the logarithm of the area. z_2 implies the relationship of the logarithm of Simpson diversity (SID) index and the logarithm of the area.

In microbial diversity study, identifying bacterial strains in metagenome and microbiome samples using computational analyses of short-read sequences remains a difficult problem [5], so that the main difference of diversity indices between macrobial and microbial is that the concept of "species" had been substituted by "OTUs". The number of operation taxonomic units (OTUs) within a community is akin to species richness within macrobial systems [6]. Similar to macrobial ecology, species richness, Shannon diversity index and Simpson diversity index were used to describe the species diversity of microbial community [7]. Up to now, there is no report to unify these diversity indices into one frame.

Fractal analysis has been applied in DNA sequence analysis more than 30 years [8,9]. For example, Chaos Game Representation (CGR) is a classical method [10]. CGR map DNA sequence into unit square by

$$CGR_{\rm i} = CGR_{\rm i\text{-}1} + 0.5*(P_i - CGR_{i-1}),\quad P_i{=}P_A,\ P_C,\ P_G \text{ or } P_T$$

where $P_A=(0,0)$, $P_C=(0,1)$ $P_G=(1,0)$ and $P_T=(1,1)$ is corresponding to four nucleotides *A*, *C*, *G* and *T* respectively, $CGR_0=(0.5,0.5)$.

According to [11], CGRs have also been subjected to multifractal analysis (which measures the degree of self-similarity within the image).On the basis of visualization for DNA sequence, one can define its multifractal spectrum by

$$D_q(\varepsilon)=\begin{cases} \dfrac{\sum_i \left(\dfrac{M_{\rm i}}{M_0}\right)\ln\left(\dfrac{M_{\rm i}}{M_0}\right)}{\ln(\varepsilon)} & q=1, \\[2ex] \dfrac{1}{1-q}\dfrac{\ln\left(\sum_i \left(\dfrac{M_{\rm i}}{M_0}\right)^q\right)}{\ln(\varepsilon)} & q\neq 1. \end{cases} \qquad (1)$$

where ε is the side length of grid, $M_{\rm i}$ is the count of point in the *i*-th grid, M_0 is the summation of all $M_{\rm i}$. Furthermore, one can define multifractal dimensions by $D(q)=\lim_{\varepsilon\to 0} D_q(\varepsilon)$. In practical computation, one can rewrite the above-mentioned formula to

$$\ln\left(\sum_{\rm i} M_i^q\right) = D_q(\varepsilon)(q-1)\ln(\varepsilon) + (q-1)\ln\left(M_0^q\right). \qquad (2)$$

Then one can compute $D(q)$ by linear fitting between $M_{\rm i}^q$ and $\ln(\varepsilon)$.

Inspired by [10], research group of Vélez studied the *Caenorhabditis elegans* genome [12] and the human genome [13] by multifractal formalism. Their results showed that human (*Homo sapiens*) genome has stronger multifractality than that of *Caenorhabditis elegans* at chromosome level. Similarly, Zhou et al. studied the discrimination problem of

doi:10.20944/preprints202107.0560.v1

coding and non-coding DNA sequence [14]. Their results suggest that coding and non-coding DNA sequence have different multifractal characteristic in the same genome.Pandit et al. Studied the classification of HIV-1 by use of multifractal dimensions of their genome [15]. These results suggested that multifractal characteristic can measure the complexity of gene and genome. Recently. Olyaee et al. Used CGR method to extract several valuable features from genomic sequences of SARS-CoV-2 [16]. In [17], Kania and Sarapata studied the robustness of the chaos game representation to mutations and its application in free-alignment methods.

In [18], Ge et al. generalized CGR to higher dimensional spaces while maintaining its bijection, keeping such method sufficiently representative and mathematically rigorous compare to previous attempts. In this frame, Dick and Green studied Proteome-Wide Protein Prediction problem by chaos game representations and deep learning [19]. Ni et al. studied gene sequence phylogenetic problem by frequency chaos game representation with perceptual image hashing [20] also.

For additive methods for genomic signatures of CGR, Karamichalis et al. studied this problem in [16]. They proposed the general concept of additive DNA signature of a set (collection) of DNA sequences. For example, the composite DNA signature (combines information from *n* DNA fragments and organellar), the assembled DNA signature (combines information from many short DNA subfragments (e.g., 100 basepairs) of a given DNA fragment). They concluded that such additive signatures could be used with raw unassembled next-generation sequencing (NGS) read data when high-quality sequencing data is not available.

Motivated by [21], in this study, we apply the fractal and multifractal method to species diversity analysis of microbiome. First, we visualize the simulated metagenomes and real metagenomes. Then we compute the multifractal dimensions of simulated metagenomes and study the relationship between its multifractal dimensions and species diversity indices. Last, we compute multifractal dimensions of real metagenomes of 100 infants' gut microbiomes when they are newborn, 4 months and 12 months.

2. Materials , Methods and Results

2.1 Metagenome Datasets

The whole genomic sequences (WGS) (.fasta files) were downloaded from the NCBI database (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The WGS for real metagenomes (.gz files) were downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra).

Data set 1: Simulated high-diversity metagenome set generated from the genomes of ten distantly related major bacterial species used in [22]. The high-diversity set include 100 metagenomes generated from the genomes of ten distantly related major bacterial species accounting for more than 90 % of all reads in Chinese group: The species used in data set 1 are listed in Table 1. The abundances in data set 1 are listed in Table S1 of Supplementary Materials.

Table 1. Species and accession numbers used in Data set 1.

Organism	Accession number
Akkermansia muciniphila ATCC BAA-835	NC_010655.1
Alistipes shahii WAL 8301	NC_021030.1
Bifidobacterium adolescentis ATCC 15703	NC_008618.1
Bacteroides vulgatus ATCC 8482	NC_009614.1
Coprococcus sp. ART55/1	FP929039.1
[Eubacterium] eligens ATCC 27750	NC_012778.1
Faecalibacterium prausnitzii A2-165	ACOP02000001.1
Lachnospiraceae bacterium 1_4_56FAA	NZ_GL945163.1
Prevotella copri DSM 18205	NZ_GG703878.1
Ruminococcus champanellensis type strain 18P13T	NC_021039.1

Data set 2: Simulated low-diversity metagenome set generated from the genomes of ten closely related major bacterial species used in [22]. The species used in data set 2 are listed in Table 2. The abundances in data set 2 are listed in Table S2 of Supplementary Materials.

Table 2. Species and accession numbers used in Data set 2.

Organism	Accession number
Bacteroides caccae strain ATCC 43185	NZ_CP022412.2
Bacteroides dorei CL03T12C01	NZ_CP011531. 1
Bacteroides ovatus strain ATCC 8483	NZ_CP012938. 1
Bacteroides ovatus V975	NZ_LT622246. 1
Bacteroides ovatus SD CMC 3f	NZ_ADM001000156. 1
Bacteroides stercoris ATCC 43183	NZ_DS499677. 1
Bacteroides thetaiotaomicron VPI-5482	NC_004663. 1
Bacteroides uniformis ATCC 8492	NZ_DS362249. 1
Bacteroides vulgatus ATCC 8482	NC_009614. 1
Bacteroides xylanisolvens CL03T12C04	NZ_JH724294. 1

Data set 3: 400 WGS for real metagenomes of 100 infants' and their mother's gut microbiota. It includes 300 infant's fecal metagenomes when they are new born, 4 month and 12 month; and 100 fecal metagenomes of their mothers. This data set was used in [23] and the accession number is PRJEB6456.

2.2 Visualization of metagenomes.

Consider the alphabet Ω ={*A*, *C*, *G*, *T*} and let $S=\{s_1,s_2,\cdots,s_m\}$ be a WGS metagenome dataset, $s_i = s_{i1}s_{i2}\cdots s_{i,n_i}$ be the *i*-th reads in S, $s_{ik}\in\Omega$ is the *k*-th nucleotide of reads s_i. To represent a WGS dataset of metagenome in the form of a CGR plot, a unit square was used, whose 4 vertices were labeled as $A=(0,0)$, $C=(0,1)$, $G=(1,0)$, $T=(1,1)$. For a given metagenome dataset $S=\{s_1,s_2,\cdots,s_m\}$ which includes *m* reads, the *k*-th nucleotide s_{ik} of reads s_i correspondes to

$$CGR_{\mathrm{ik}} = CGR_{\mathrm{i,k\text{-}1}} + 0.5 * \left(P_{ik} - CGR_{i,k-1}\right), \qquad P_i{=}P_A,\ P_C,\ P_G \text{ or } P_T,\ i{=}1,2,\ldots,m$$

where $P_A = (0,0)$, $P_C = (0,1)$ $P_G = (1,0)$and $P_T = (1,1)$ is corresponding to four nucleotides *A*, *C*, *G* and *T* respectively, $CGR_{i0} = (0.5,0.5)$.

In order to avoid "large number annihilating small number", we disgarded the first 10 points of each reads. The visualization of a simulated metagenome in data set 1 is demonstrated by Figure 1 as an example.

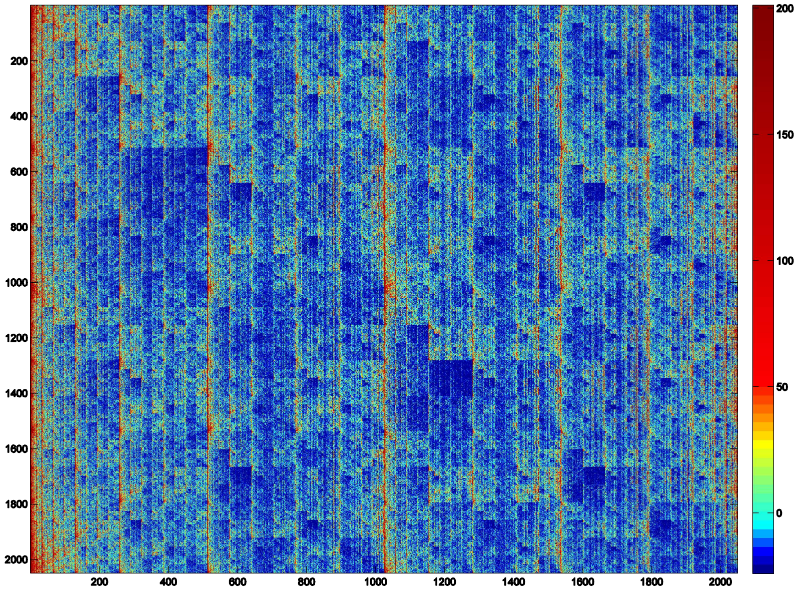

Figure 1. Heat map of simulated metagenome

2.3 *Fractal and multifractal spectrum of metagenome*

We found all CGRs (e.g. Figure 1) seem to be self-similar. So we intend to study their fractal and multifractal properties. On the basis of visualization of metagenome sequence, one can define its multifractal spectrum by (1).

Furthermore, one can define multifractal dimension by $D(q) = \lim_{\varepsilon \to 0} D_q(\varepsilon)$. Figure. 2 shows the linear fit between $\ln\left(\sum_{\mathrm{i}} M_{\mathrm{i}}^{q}\right)$ (i.e. $\ln(\mathrm{M}(\varepsilon,\mathrm{q}))$) and $\ln(\varepsilon)$ of simulated metagenome.

doi:10.20944/preprints202107.0560.v1

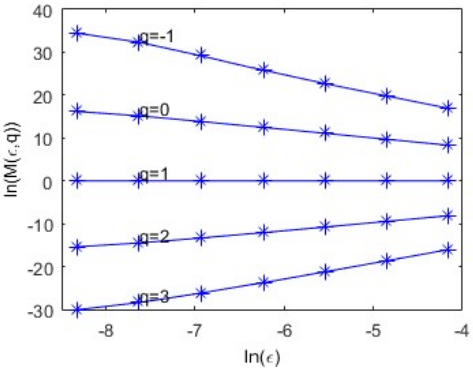

Figure 2. Linear fit of $\ln(\mathrm{M}(\varepsilon,\mathrm{q}))$ and $\ln(\varepsilon)$.

In practical computation, one can compute $D(q)$ by linear fitting between $\ln(\mathrm{M}(\varepsilon,\mathrm{q}))$ and $\ln(\varepsilon)$ according to (2). In metagenomic research, for a given community (i.e. given abundance values of bacteria), a WGS dataset of metagenome is actually a collection of sampling reads from the give community. Here, we simulate 100 metagenomes from a given abundance of ten bacteria. Figure 3 demonstrates the multifractal dimensions of 100 simulated metagenoms from data set 1 and 100 simulated metagenoms from data set 2.

From Figure 3, we can find that multifractal dimension curves of different simulated metagenomes from the same abundance are unstable when $\mathrm{q}<0$, they are stable when $\mathrm{q}\geq 0$. So we only consider $D(q)$ for $\mathrm{q}\geq 0$ in multifractal spectrum of metagenome.

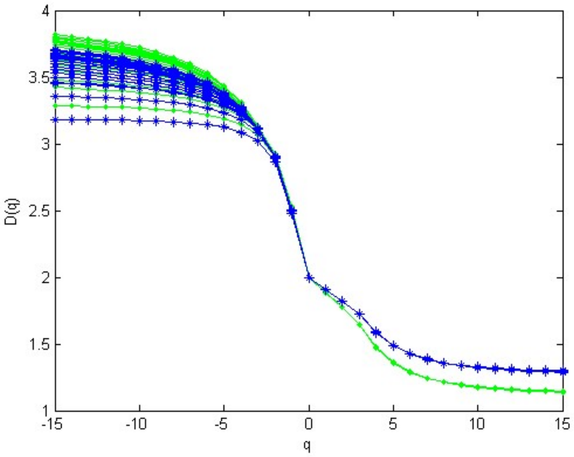

Figure 3. Multifractal dimensions of simulated genome. The number of reads is 10M, read length is 1000bp, green asterisk represented the D(q) of samples simulated from high-diversity communities, blue dot represented the D(q) of samples simulated from low-diversity communities. For each sample from the same community, the abundances are given in appendix 1.(the last line in the table)

2.4 The relationship between multifractal spectrum and microbial diversity index of metagenomes

In order to study the relationship between multifractal spectrum and diversity indices of metagenomes, we simulated 100 metagenomes whose abundance are known, then their species richness index, Shannon diversity index, Simpson diversity index, and multifractal dimensions are computed. Then the Pearson correlation coefficients are computed according to varying q.

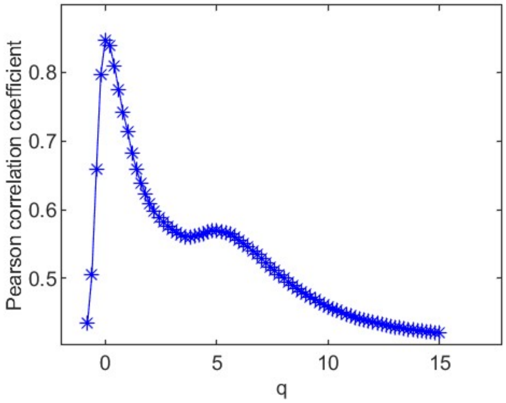

Figure 4. Pearson correlation coefficient of species richness and multifractal dimension $D(q)$

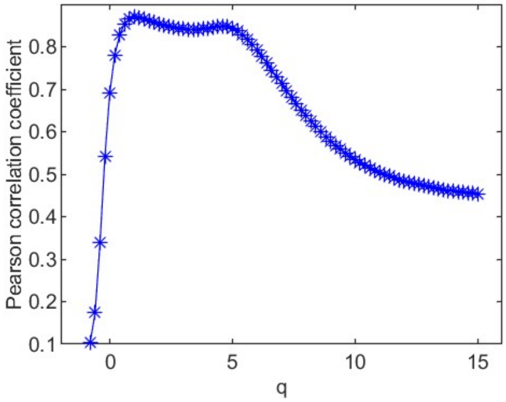

Figure 5. Pearson correlation coefficient of species shannon diversity index and multifractal dimension $D(q)$

The Pearson correlation coefficients between species richness diversity index and multifractal dimension are plotted in Figure 4. The plot suggests that Pearson correlation coefficient between species richness indices and multifractal dimensions reach its maximum (0.85) at $q = 0$. Similarly, the Pearson correlation coefficients between species Shannon diversity indices and multifractal dimensions reach its maximum (0.88) at $q = 1$. The Pearson correlation coefficient between species Simpson diversity indices and multifractal dimensions reach 0.87 at $q = 2$.

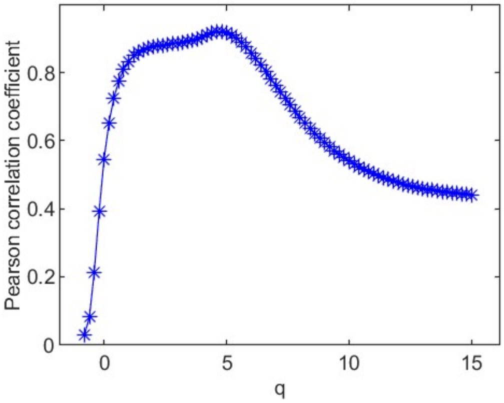

Figure 6. Pearson correlation coefficient of species Simpson diversity index and multifractal dimension $D(q)$

Based on the Pearson correlation coefficient plot, in order to study the relationship between species diversity indices and multifractal dimensions of metagenomes, we plot the scatter plot of Shannon diversity indices and $D(1)$s in Figure 7 and that of Simpson diversity index and $D(2)$ in Figure 8 respectively .

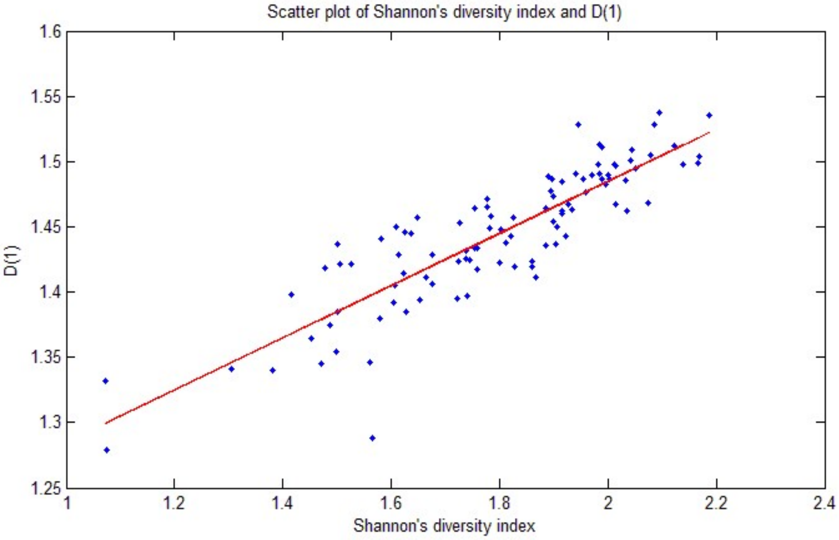

Figure 7. Scatter plot of Shannon diversity index and $D(1)$

doi:10.20944/preprints202107.0560.v1

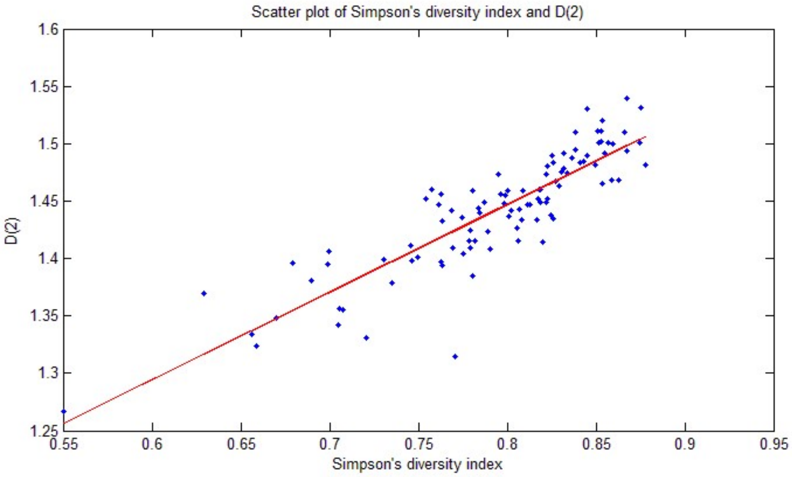

Figure 8. Scatter plot of Simpson diversity index and $D(2)$

2.5 Application of multifractal dimension in metagenomes to infant's gut microbiome

In order to apply the multifractal analysis to real metagenomes, we selected 100 infants' fecal WGS datasets of 300 metagenomes (There are 3 samples, including 12 Month (12 M), 4 Month (4 M) and new born (baby) for each infant) and 100 corresponding gut metagenomes of their mothers to mine potential information of its multifractal dimensions.

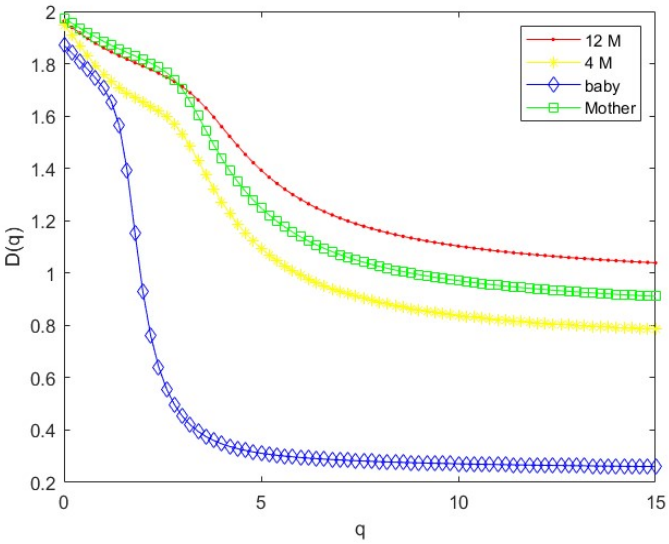

Figure 9. Multifractal dimension of gut microbiome of 12 month(12 M), 4 month(4 M), newborn baby (baby), and her mother (M)

As an example, we plot multifractal dimensions of a selected gut microbiome of a baby in Figure 9. The plot demonstrates the multifractal dimensions of gut microbiomes of an infant and its mother when he/she is a newborn (baby), 4 month, 12 month. Figure 9 suggests that the $D(0)$ (fractal dimension), $D(1)$ (information dimension) and $D(2)$

doi:10.20944/preprints202107.0560.v1

(correlation dimension) are increasing with growing. In other words, their gut microbial diversity is developing with growing. In order to study the generality of this property, we computed the mean value of 100 multifractal dimensions of 12 Month, 4 Month, new born and their mothers, respectively. From Figure 10, we can find that the similar results in average.

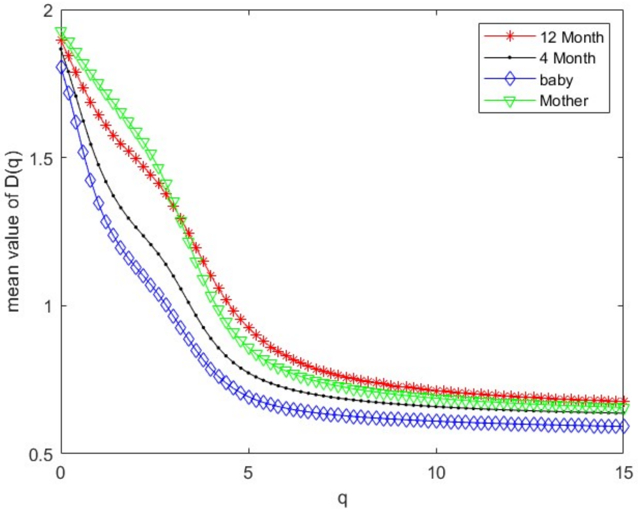

Figure 10. Mean values of multifractal dimension of 100 infants' gut microbiomes of 12 month(12 M), 4 month(4 M), newborn baby(baby), and their mother(M)

For data set 3, there are 100 infants' gut microbiomes. If we consider each infant gut micobiomes as one group, there are 100 groups gut microbiomes. For each group, we compute the difference between 12 M and 4 M, 12 M and new born, 4M and new born respectively. In order to observe the overall characteristic of these multifractal dimensions, we plotted the mean value of 100 multifractal dimensions of gut microbiomes in Figure 10.

In order to evaluate the discriminating power of gut microbiomes' multifractal dimensions in ages of infants, we use multifractal dimensions of 12M,4M, baby and Mother gut microbiomes to discriminate by Support Vector Machine (SVM) [24]. Table 3 demonstrates accurate rates of discriminating metagenomes of 12M, 4M, baby and Mothers by SVM. Within infants' gut microbiomes, the accurate rate of 12M and baby, 12M and 4M, baby and 4M is decreasing.

Table 3. Accurate rates of discriminating metagenomes of 12M, 4M, baby and Mothers by SVM.

Accurate rate	12 Month	4 Month	baby
4 Month	84.0%	-	
baby	91.5%	70.5%	-
Mother	91.5%	95.5%	94.5%

3. Discussions and conclusions

In this study, we studied metagenomes by multifractal analysis. From the results above, we can draw the following conclusions.

(i) From the CGR visualization of metagenomes by, we can see there exists statistical self-similarity in these plots. Figure 3 demonstrates 100 simulated WGS metagenomes sampling from a given abundance, it suggests that $D(\mathrm{q})$ of metagenomes is stable when $q \geq 0$ and unstable when $q < 0$. These results guide us to study multifractal dimensions of metagenomes only for $q \geq 0$ in the following study. These results show that there is multifractal character in CGRs of metagenomes.

(ii) From Figure 4, we can see that the Pearson correlation coefficients of species richness indices and $D(\mathrm{q})$ reach their maximums when $q = 0$. Similarly, we can find that the Pearson correlation coefficients of Shannon diversity indices and $D(\mathrm{q})$ reach their maximums when $q = 1$ from Figure 5, the Pearson correlation coefficients of Simpson diversity indices and $D(\mathrm{q})$ approach their maximums when $q = 2$ from Figure 6. These results coincide with the results of macrobial ecology in [4]. On the whole, the scatter plot of Shannon diversity indices and corresponding $D(1)$ in Figure 7 shows that $D(1)$ is increasing with the increasing of Shannon diversity indices of metagenome. Figure 8 shows that $D(2)$ is increasing with the increasing of Simpson diversity indices of metagenome. These results suggest that multifractal dimensions can reflect the microbial diversity in metagenomic research and the traditional diversities can be unified by the frame of multifractal analysis.

(iii) In research on real metagenomes, multifractal dimensions of gut mirobiome of one mother and her baby is demonstrated in Figure 9, this plot shows that the multifractal dimensions of gut microbiome of baby is increasing with aging (new born, 4 M and 12M). Figure 10 shows this law holds on the whole for baby in average.The discriminated power of multifractal dimensions of gut microbiomes of infants demonstrated in Table 3 shows that the infants' age can be discriminated by their multifractal spectrum of CGR visualization of gut microbiomes.This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

Supplementary Materials: Table S1: The abundances in data set 1; Table S2: The abundances in data set 2.

Author Contributions: Conceptualization, X.X. and Y.M.; methodology, X.X. and Z.Y.; software, X.X. and Y.M; validation, X.X., Y.M., Z.Y. and G.H.; formal analysis, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X. and Z.Y.; writing—review and editing, X.X. and Z.Y.; visualization, X.X.; supervision, Z.Y.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 11871061; open project of Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education (Xiangtan University), grant number 2018ICIP04; the Science and Technology Project of Jiangxi Provincial Education Department, grant number GJJ170820.

Data Availability Statement: The source codes for our algorithm and datasets used in this report can be provided via email.

Conflicts of Interest: The authors declare no conflict of interest.

doi:10.20944/preprints202107.0560.v1

References

1. Kempton R. A.; Taylor L R. Models and statistics for species diversity. *Nature*, 1976, 262,818-20.
2. Hubalek Z. Measures of species diversity in ecology: an evaluation. *FOLIA ZOOL*, 2000, 49,241-260.
3. Borda-de-Água L.; Hubbell SP; McAllister M. Species-Area Curves, Diversity Indices, and Species Abundance Distributions: A Multifractal Analysis. *The American Naturalist*, 2002, 159,138-155.
4. Harte J .; Kinzig A P. On the Implications of Species-Area Relationships for Endemism, Spatial Turnover, and Food Web Patterns. *Oikos*, 1997, 80,417.
5. Kuleshov V.; Jiang C.; Zhou W.; et al. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature Biotechnology*, 2016, 34,64–69.
6. Stegen J C.; Hurlbert A H.; Bond-Lamberty B.; et al. Aligning the Measurement of Microbial Diversity with Macroecological Theory. *Frontiers in Microbiology*, 2016, 7,1487.
7. Leinster T.; Cobbold C A. Measuring diversity: the importance of species similarity. *Ecology*, 2012, 93,477-489.
8. Joel J. H. Chaos game representation of gene structure. *Nucleic Acids Research*, 1990,8,2163-2170.
9. Berthelsen C. L.; Glazier J. A.; Skolnick M. H. Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Physical Review A*, 1992, 45, 8902.
10. Berthelsen C. L.; Glazier J. A.; Skolnick M. H. Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Physical Review A*, 1992, 45, 8902.
11. Joseph J.; Sasikumar R. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, 2006, 7,243.
12. Karamichalis R. Molecular Distance Maps: An alignment-free computational tool for analyzing and visualizing DNA sequences,Doctor of Philosophy. The University of Western Ontario. Ontario, Canada, 2016.
13. Vélez P. E.; Garreta L. E.; Martínez E.; et al. The Caenorhabditis elegans genome: a multifractal analysis. *Genetics & Molecular Research Gmr*, 2010, 9,949.
14. Moreno P. A.; Patricia E.; Vélez, E.; Martínez; et al. The human genome: A multifractal analysis. *BMC Genomics*, 2011, 12,506.
15. Zhou L. Q.; Yu Z G.; Deng J Q.; et al. A fractal method to distinguish coding and noncoding sequences in a complete genome based on a number sequence representation, *J. Theor. Biol.*, 2005, 232, 559-567.
16. Pandit A.; Dasanna A K.; Sinha S. Multifractal analysis of HIV-1 genomes. *Molecular Phylogenetics & Evolution*, 2012, 62,756-763.
17. Olyaee M. H.; Pirgazi J.; Khalifeh K.; et al. RCOVID19: Recurrence-based SARS-CoV-2 features using chaos game representation. *Data in Brief*, 2020,32,106144.
18. Kania A.; Sarapata K. The robustness of the chaos game representation to mutations and its application in free-alignment methods. *Genomics*, 2021, 1428-1437.
19. Ge L.; Liu, J.; Zhang Y. et al. Identifying anticancer peptides by using a generalized chaos game representation. *J. Math. Biol.* 2019,78, 441–463.
20. Dick K.; Green J. R. Chaos Game Representations & Deep Learning for Proteome-Wide Protein Prediction. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2020.
21. Ni H.; Mu H.; Qi D. Applying frequency chaos game representation with perceptual image hashing to gene sequence phylogenetic analyses. *Journal of Molecular Graphics and Modelling*, 2021,107,107942.
22. Karamichalis R .;Kari L.; Konstantinidis S.; et al. Additive methods for genomic signatures. *BMC Bioinformatics*, 2016, 17,313.
23. Dubinkina V.; Ischenko D.; Ulyantsev V.;Tyakht A.; AlexeevD. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, 2016, 17,38.
24. Bäckhed F.; Roswall J.; Peng Y.; et al. Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe*, 2015, 17, 690-703.