Fast Comparison of Microbial Genomes Using the Chaos Games Representation for Metagenomic Applications

Swain, Martin

Published in:
Procedia Computer Science

DOI:
10.1016/j.procs.2013.05.304

Publication date:
2013

Citation for published version (APA):
Swain, M. (2013). Fast Comparison of Microbial Genomes Using the Chaos Games Representation for Metagenomic Applications. *Procedia Computer Science*, *18*, 1372-1381. https://doi.org/10.1016/j.procs.2013.05.304

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 18 (2013) 1372 – 1381

Procedia Computer Science

International Conference on Computational Science, ICCS 2013

Fast comparison of microbial genomes using the Chaos Games Representation for metagenomic applications

Martin T. Swain[1,*]

Institute of Biological, Environmental and Rural Sciences (IBERS), Aberystwyth University, Penglais, Aberystwyth, Ceredigion, SY23 3DA

Abstract

Genome sequencing technology is generating large databases of sequence at such a rate that advances in computer hardware alone are not adequate to handle them: more efficient algorithms are needed. Here an alignment-free method of sequence comparison and visualisation based on the Chaos Games Representation (CGR) and multifractal analysis is explored as an approach to search and filter through a data set of over 1500 microbial genomes. Whereas BLAST takes 25 hours to search this data set with large sequence fragments (e.g. 100 Kb), the method introduced here can reduce this data set by 95% (from 1550 target species to just 50) in about 15 minutes, and it is able to predict the exact species correctly in 67% of cases. The results presented here demonstrate that CGR is worth further investigation as a fast method to perform genome sequence comparison on large data sets, and various ways to further develop the method are discussed.

Keywords: Genomics, Visualisation, Biological sequence analysis, Large data sets, Data mining

1. Introduction

DNA sequences are currently being generated at a rate that outstrips Moore's Law [1]. In addition new sequencing technologies such as nanopore sequencing have recently claimed to be able to generate very long read lengths of 100,000 base-pairs in length (100 Kb) [2]. If this is realised, then there will be a significant acceleration in the size of genome databases. In order to analyse these sequences it has become clear that simply buying more advanced computational hardware is no longer possible. Instead it is essential to develop more efficient computational methods able to quickly search these vast collections. In this article preliminary results are given for a fast approach to comparing and identifying microbial genome sequences.

The most widely used methods of comparing fragments of DNA are based on sequence alignment, using approaches based on dynamic programming [3] and implemented in tools like BLAST [4]. However, methods have been proposed for alignment-free sequence comparison. In a review of such methods [5], two basic approaches were identified. The first approach considers the statistical analysis of word frequency, where a word is a small oligomer or k-mer. The second approach includes the use of Kolmogorov complexity and Chaos Theory. In this paper the

* *Email address:* mts11@aber.ac.uk (Martin T. Swain)
[1]Corresponding author

Selection and peer review under responsibility of the organizers of the 2013 International Conference on Computational Science
doi:10.1016/j.procs.2013.05.304

alignment-free sequence comparison method we are investigating belongs to the second approach, and is known as the *Chaos Games Representation* or *CGR* [6, 7]. It uses a DNA walk model to highlight how the sequence of nucleotides differs from that of a random sequence and represents it in a scale-independent manner that has fractal properties. In addition, the CGR can be used to visualise the sequence: it may display the set of word frequencies in a single square image by using pixels to represent single words, coloured according to word frequency. See Figure 1 for examples.

Multifractal analysis methods can be applied to CGRs. One such method proposed to classify bacterial genomes by applying multifractal analysis, along with spectral analysis, to the CGRs of just over 33 bacterial genomes [8]. The authors showed that phylogenetically similar bacteria clustered together in 2 and 3 dimensional spaces that were derived from the CGR fractal dimension. By deriving a property from a CGR analogous to the concept of specific heat capacity in physics, Yu et al. [9] were able to distinguish between coding and noncoding exons in bacterial genomes. They found the fractal dimension was higher in noncoding sequences than in coding sequences. Similar approaches were used recently to analyse the human genome [10] and to distinguish between HIV-1 genomes [11]. However, it is not clear how well these approaches scale, and if they can be applied to the thousands of genomes that now reside in public repositories.

The objective of this paper is to investigate the utility of CGRs for comparative and metagenomic studies, where fragments of genomes from hundreds or thousands of microbial species need to be classified taxonomically and identified [12]. The outline of this paper is as follows. First, the methods used to generate, analyse and compare CGRs are described, followed by the test data sets used, and then a number of results are given that are important for calibrating the method. The method is compared to BLAST, and before the conclusion, a discussion that includes ideas for future work is given.

2. Methodology

2.1. The Chaos Games Representation and multifractal analysis

A detailed explanation of the CGR is given by Almeida *et al.* [13]. Due to space limitations we can only give a brief summary of the method here.

A CGR plot is based upon a 2-dimensional graph [6, 7]. It is square in shape, and the vertices of the square represent the four nucleotide bases g, such that $A = (0, 0)$, $T = (1, 0)$, $C = (0, 1)$ and $G = (1, 1)$. A sequence can then be plotted within the area of the square, such that the coordinates of the position r_i, corresponding to the nucleotide i, are given by:

$$r_i = 0.5 \times (r_{i-1} + g_i) \tag{1}$$

where the first nucleotide of the sequence is plotted at the centre of the square, in the position $(0.5, 0.5)$. Using an iterative approach, the genome sequence is then transformed into a series of points comprising a pseudo-random walk around the area of the square graph. The location of each point is representative of the base-pairs in the sequence immediately preceding that point. Hence it is possible to transform the CGR plot into a representation of the words or k-mer frequencies in the DNA sequence.

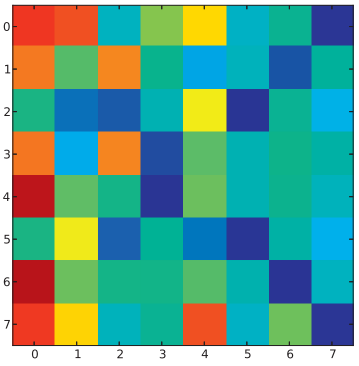
Following an approach previously used for multifractal analysis of CGRs [14, 9, 11], boxes of size r were defined; for instance 64 equal sized boxes will each have a side length of $1/8$ given that the dimensions of the graph are of unitary length. Once the boxes are defined the density of points in each box may be calculated. Any box size is possible, and the smaller the box size, the higher the resolution of the representation. However, this study has divided the CGR square such that the number of boxes is always a multiple of 4 (due to the 4 DNA bases). Thus a box size of $1/8$ corresponds to 4^3 boxes, i.e. all possible nucleotide trimers or 3-mers.

Given the density of the boxes, the partition sum is defined for all non-empty boxes i as:

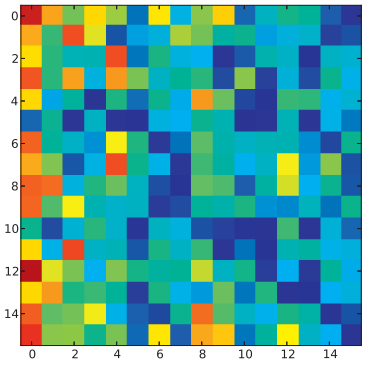
$$Z_r(q) = \sum_i P_i^q \tag{2}$$

here q is any real number, and acts to emphasis relatively sparse and dense regions. In this study it is given values between -15 to 15. The spectrum of scaling exponents τ_q may then be given by:

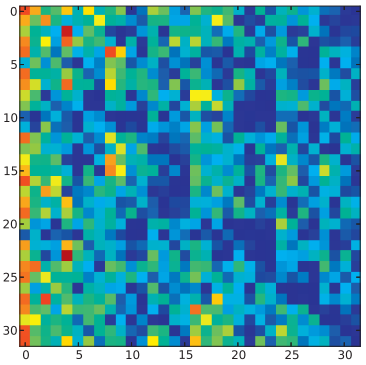
$$\tau_q = \lim_{r \to 0} \frac{\log(Z_r(q))}{\log(r)} \tag{3}$$



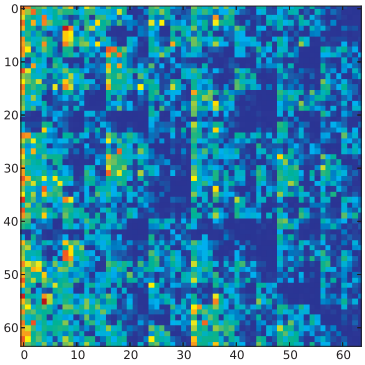
Box size 1/8, 3-mers



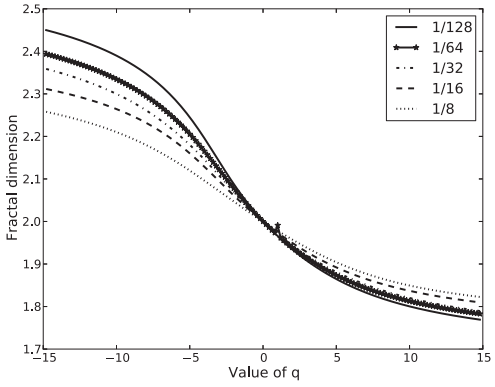
Box size 1/16, 4-mers



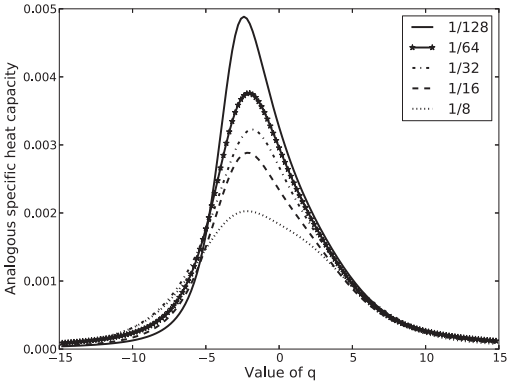
Box size 1/32, 5-mers



Box size 1/64, 6-mers



Fractal dimension



Anal. sp. heat capacity

Figure 1: CGR box density representations of the whole *Lactobacillus plantarum* WCF218 genome using different box sizes. In the top four figures, each coloured box represents a particular k-mer frequency. The lowest two figures show how the fractal dimension and analogous specific heat capacity depend on q for box sizes 1/8, 1/16, 1/32, 1/64, 1/128.

The generalised fractal dimension spectrum is defined for $q \neq 1$ as:

$$D_q = \frac{\tau_q}{q-1} \tag{4}$$

and for $q = 1$ as:

$$D_q = \lim_{r \to 0} \frac{\sum p_i \log(p_i)}{\log(r)} \tag{5}$$

A quantity analogous to the concept of specific heat spectra, C_q, can also be calculated by taking the second differential of τ with respect to q. The calculation is typically performed [9, 11] using the following approximation:

$$C_q = \frac{\delta^2 \tau_q}{\delta q^2} \approx 2\tau_q - \tau_{q-1} - \tau_{q+1} \tag{6}$$

See Figure 1 for some example CGR box density, fractal dimension and analogous specific heat capacity plots.

2.2. The microbial genome data sets used

The Genometa software package includes a large collection of curated microbial reference sequences from five sequencing initiatives [15]. The full data set (as of April 2012) consists of 2550 genomes, while the "one genome per species" subset consists of 1551 genomes, each from a different species. By subtracting this "one genome per species" data set from the full data set, a data set of 999 genomes is created, each genome representing a different strain of the various species present in the "one genome per species" data set.

In the tests performed here, the "one genome per species" data set is referred to as the *species-target* data set, while the data set of 999 different strains will be referred to as the *strain-query* data set. Due to issues when pre-processing the data, including the automatic parsing of ambiguous genome names, the strain-query data set was reduced to 977 genomes.

2.3. Using CGRs for genome comparison

This study has investigated the utility of CGRs for searching through the species-target data set with the strain-query data set using different box sizes r and three measures or properties of the CGR:

1. The box densities P_i.
2. The fractal dimension D_q, given by Equations 4 and 5.
3. The analogous specific heat capacity C_q, given by Equation 6.

An RMSD measure has been used to compare a pair of CGR properties. For example, if each CGR box density is calculated with N_B boxes of the same size, then the RMSD score is given by:

$$\sqrt{\sum_{i,j}^{N_B} \frac{(d_i - d_j)^2}{N_B}} \tag{7}$$

where d_i and d_j are the density of boxes i and j in the two CGRs.

A similar approach is used when comparing the other two properties derived from the CGR. In these cases the d_i and d_j in Equation 7 are the value of either the fractal dimension or the analogous specific heat capacity for a particular value of q, and the sum is over all q values.

Note that because the strains in the strain-query data set are not included in the species-target data set, exact matches between the two data sets are not possible – at best the lowest RMSD score as given by Equation 7 should identify a genome from the species to which the strain belongs. Therefore, to test the accuracy of this identification process, the taxonomic information of each genome are compared. This is performed using taxonomic information downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/pub/taxonomy/) and an in-house script to build the taxonomic tree. Each genome in both data sets was allocated the following 6 attributes of taxonomic information, in order of the lowest member of the hierarchy to the highest: species, genus, family, order, class, and phylum.

3. Results

To calibrate and evaluate the method, the following experiments were performed.

3.1. Size of sequence fragment

A CGR generated from a fraction of a genome sequence is similar to the whole sequence. However, it is known that the fractal properties differ significantly for CGRs generated from coding and noncoding regions in a genome [9]. Hence it is important to investigate the variability of the genome sequence in order to understand if there is a minimum fragment size that would be suitable for CGR-based analysis.

Figure 2 shows the average and maximum RMSD scores for fragments of different size (from 3.2 Kb to just over 204 Kb) taken from an *E. coli* genome. Fragments of each size were taken in regular intervals along the entire genome sequence, each fragment overlapping 50% of its sequence with the previous fragment.

For the CGR box density measure, the figure shows that the average RMSD is much the same for all box sizes. However, it does change according to the fragment size, leveling off after about 50 Kb. There is also quite a lot of variability in the measure, as shown by the maximum RMSD scores: these are much greater than the average scores. The maximum RMSD plots indicates the maximum variability of CGR box density approach to comparing sequences. A large maximum RMSD indicates that the CGR box density may not be able to uniquely identify the species from which the fragment is derived when searching through large databases of genome sequences. For small fragment sizes there is big difference between the maximum and the average RMSD values, especially for CGR densities constructed with large box sizes (e.g. 1/8). However, for fragments of size 100 Kb or more, the RMSD plots for all box sizes tend to converge to similar maximum RMSD values (these are about twice the value of the average RMSD). Therefore, for fragments of about 100 Kb or greater size, relatively large box sizes can be used – which will reduce the time required to calculate Equation 7. Note also that for small box sizes (e.g. 1/128) the number of boxes may be greater than the number of base-pairs (i.e. points) in the CGR plot, and hence there will be many zero values included in Equation 7, which may lower the similarity score but not increase its ability to distinguish between fragments between different genomes: this is especially relevant for smaller sequence fragments.

For the fractal dimension measure there is slightly more variability, as shown by the maximum RMSD score. The plots also show that the largest box size (1/8) has the highest accuracy. While the actual RMSD scores are much smaller for the analogous specific heat capacity measure, the general behaviour is similar to that of the fractal dimension measure. These RMSD plots are not so well behaved as for the CGR box densities: there is no overall reduction in the variability of the RMSD scores as the fragment sizes become larger, instead the RMSD scores tend to jump about for different fragment sizes. This may indicate the presence of noise which may weaken the discriminatory power of these two measures.

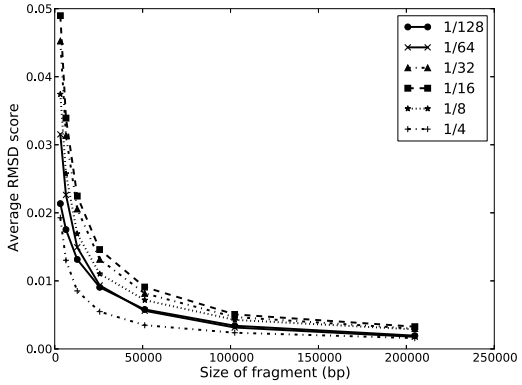
3.2. Comparison with BLAST

Here the utility of the CGR properties for searching and filtering large genomic data sets is investigated by comparing the results of the CGR approaches to those from BLAST. Four sequence fragments with sizes of 100 bp, 1000 bp, 10 Kb, and 100 Kb were taken from each of the 977 genomes in the strain-query data set, and compared against the 1550 genomes in the species-target data set.

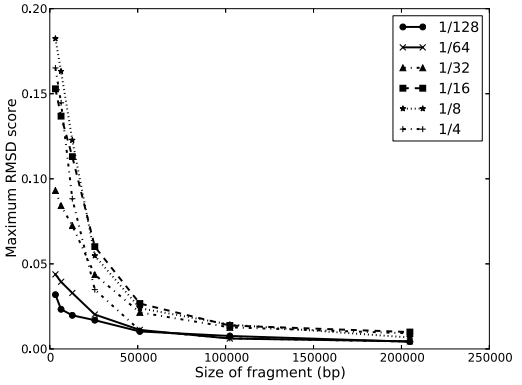
Using the BLAST programs, a database of the 1550 species-target data set was created. Each sequence fragment from the strain-query data set was queried against this database, the top scoring BLAST hit was recorded, and the correct taxonomic classifications of this BLAST hit were calculated as described in Section 2.3.

For each sequence fragment, CGR properties were created with different box side lengths. Using Equation 7 the CGR properties of the fragment were compared to those of the 1550 whole target genomes, created using the same box side lengths. In the left-hand side of Figure 3 we consider how well the CGR properties perform at predicting the species from which the fragments are derived. For these results, only the predictions made by the top ranking hit were considered.

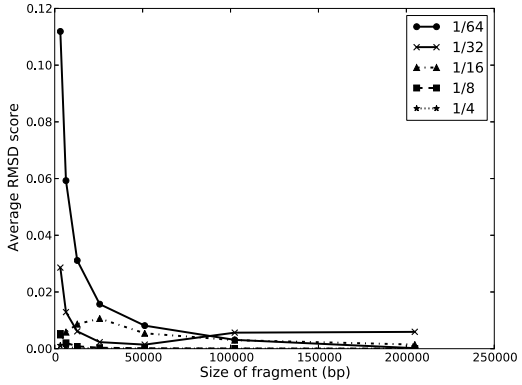
The left-hand side of Figure 3 indicates that the CGR box density property is the most useful for identifying the strain's correct species, especially for larger fragment sizes. Out of the 1550 species in the species-target data set, the exact species was predicted for just over 67% of the 977 strains when the fragment size was 100 Kb, and the box size was 1/16 or 1/32. For a fragment size of 10 Kb, 46% of species predictions were correct for the 1/16 box size. For



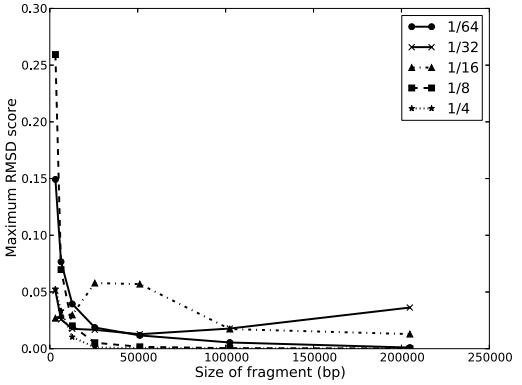
CGR box density, average RMSD



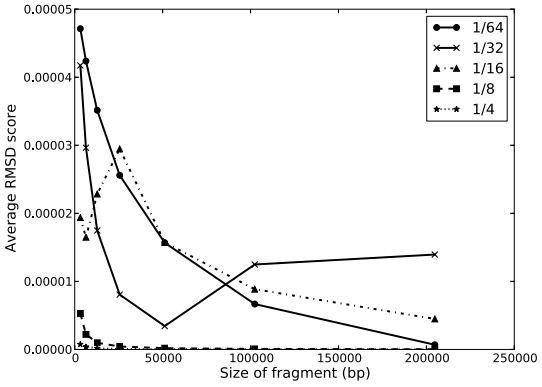
CGR box density, maximum RMSD



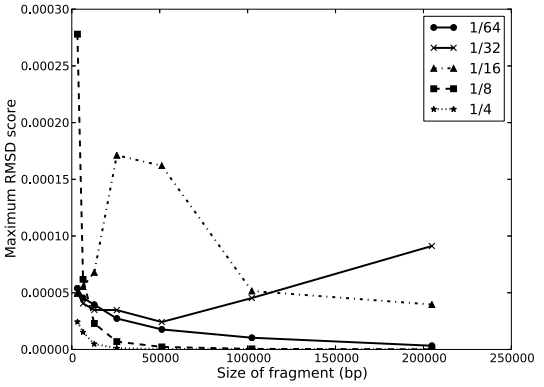
Fractal dimension, average RMSD



Fractal dimension, maximum RMSD

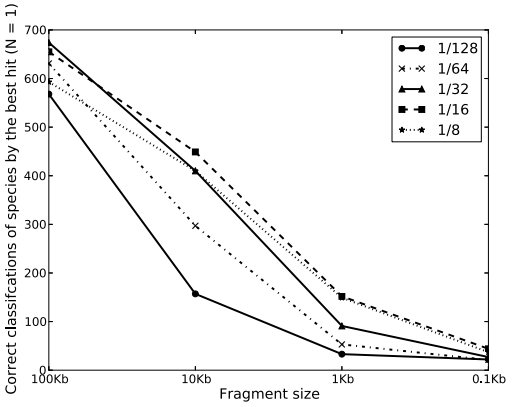


Anal. sp. heat capacity, average RMSD

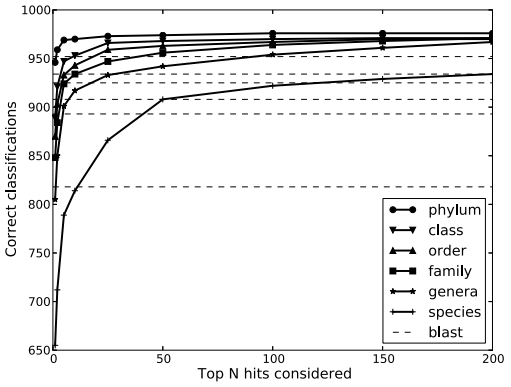


Anal. sp. heat capacity, maximum RMSD

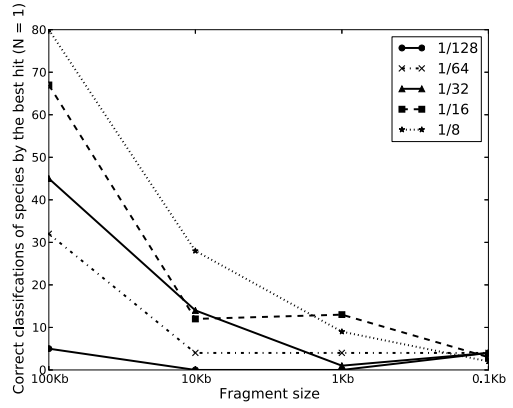
Figure 2: Average and maximum RMSD score for CGR densities, fractal dimensions, and analogous specific heat capacities, generated from fragments of an *E. coli* genome that have been compared to the corresponding measure generated for the whole genome.



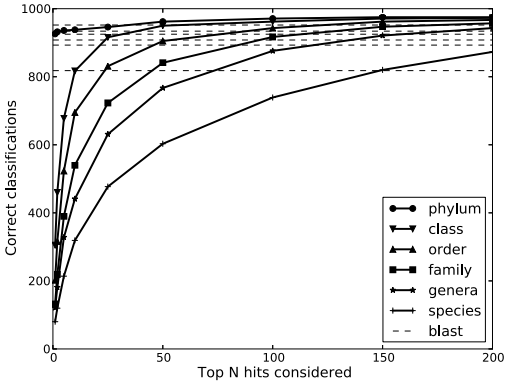
CGR box density



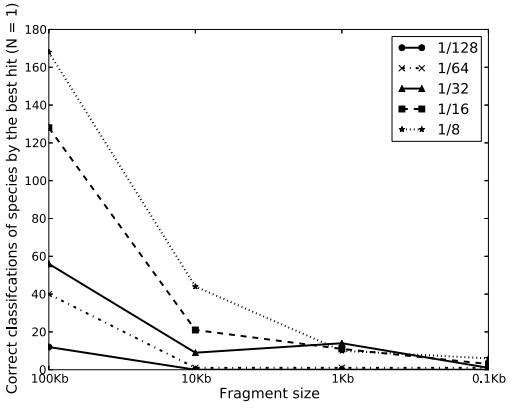
CGR box density for box size 1/16



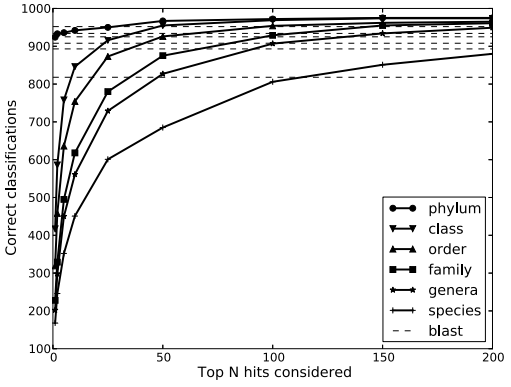
Fractal dimension



Fractal dimension for box size 1/8



Anal. sp. heat capacity



Anal. sp. heat capacity for box size 1/8

Figure 3: The left-hand column is the number of correct taxonomic classifications of species as predicted by the number 1 top ranked hit (i.e. the lowest RMSD score) for different fragment sizes. The right-hand column indicates how large (i.e. the number N) the set of top ranked hits should be in order to contain correct classifications at different taxonomic levels. The lowest BLAST line corresponds to the number of correct species identified by the top BLAST hit, with higher lines corresponding to higher taxonomic classifications. The fragment size is 100 Kb and the box size used gives the highest number of correct species classifications according to the left-hand side plots.

Table 1: Comparison of timings using CGR box densities and BLAST for the experiment described in Section 3.2. The results are combined timings based on all four fragment sizes (100 bp, 1 Kb, 10 KB, 100 Kb).

Box side length	Time
1/8	23 mins
1/16	39 mins
1/32	1 hour 40 mins
1/64	5 hours and 12 mins
1/128	18 hours 24 mins
BLAST	31 hours and 44 minutes

Table 2: Breakdown of timings using CGR box density of 1/16 and BLAST for the four fragment sizes. The number of correct species predictions made by the top BLAST hit are also given (out of 977).

Fragment size	CGR time	BLAST time	BLAST no. correct
100 bp	8 mins	54 mins	667
1 Kb	8 mins	1 hour and 21 mins	737
10 Kb	9 mins	4 hours and 21 mins	788
100 Kb	15 mins	25 hours and 8 mins	818

smaller fragment sizes, and for the fractal dimension and analogous specific heat capacity measures, the prediction rate drops quite considerably.

The results for the CGR box densities demonstrate that this measure may be used to pre-filter assembled metagenomic reads. In this scenario, the correct species would be contained with a data set corresponding to the set of species predicted by a number N of the highest ranked hits. The question then is how many of the top hits should be contained in this data set i.e. what value of N is required?

The right-hand side of Figure 3 attempts to answer this question. Here the taxonomic predictions for the highest ranked N scores were considered, where N was varied between 1 and 200. For example, if at least one of the N predictions had the correct species classification, then that counted as a correct classification of the species. All levels of the taxonomic hierarchy were checked for correctness and the prediction with the highest number of correct classifications was determined: only this prediction was used in the count of correct classifications. It is usually the case that if the strain's species is predicted correctly, then the higher taxonomic classifications are also correct. Thus, for the CGR box density property (with box size 1/16), the 10 highest ranked hits contain a prediction that is as accurate as that made by the top BLAST hit. It should also be noted that the top blast hit does not always predict the correct species. If the 50 highest ranked predictions made using the CGR box density property are considered, then that data set may be more accurate than just considering the top BLAST hit: thus the CGR box density property may be used to quickly reduce the 1550 genomes in the species-target dat set to around 50 genomes, which may then be more thoroughly searched using BLAST.

The fractal dimension and analogous specific heat capacity properties make less accurate predictions than the CGR box density property. The top 100 or 150 predictions need to be considered before finding a prediction as accurate as that made by BLAST. Thus these properties may still be used to filter large data sets, and it should be noted that they are most accurate with relatively large box sizes, which require the least effort to compute.

Example timings for the CGR box density and BLAST runs are given in Table 1 and Table 2. The timings in Table 1 demonstrate that the CGR box density property is a very fast way to identify the taxonomic identity of species when compared to BLAST, especially for larger sequence fragments. Table 2 gives the break down of the timings for the different fragment sizes: for the CGR box density size of 1/16 and for BLAST. Whereas the timings for BLAST increase rapidly for larger fragments, the timings for the CGR approach increase much more slowly, demonstrating its scalability.

These times do not include the timing to create the target databases: here BLAST is much faster than the CGR methods, taking almost 1 minute versus the 6 hours and 35 minutes hours required by the CGR methods when using a box size of 1/16. However, for all these experiments the CGR comparison software was implemented in Python and no significant effort was made to optimize the code for speed. Hence there is much scope for decreasing the runtime

of these comparisons. All timings were performed using a single core on a 2.93 Ghz Intel Xeon chip.

4. Discussion and future work

Microbial communities may be composed of many different species, and short read sequencing technologies are commonly used in metagenomic studies to generate the initial sequences. These reads are usually about 100 bp long, and once sequenced are assembled into contigs (contiguous sequences) that are orders of magnitude longer: it is often the case the longer contigs correspond to the most abundant species, whereas less common species may be represented by contigs of only 100s or maybe 1000s of base-pairs in length. Handling the relatively short contigs is a major challenge and interest for metagenomic studies because they may indicate the presence of rare and novel sequences – perhaps from microbial species unknown to science, with important commercial implications for industrial processes and biorefining. While this study has shown that the CGR box density property is an efficient approach for handling sequence fragments of 10 Kb to 100 Kb in length, such as the contigs output from metagenomic assemblies, it still has problems with processing either the unassembled reads or very small contigs from metagenomic assemblies.

A relatively simple scoring function is used in this study. One problem with is that for high resolution CGR box density plots (i.e. very small box sizes) there are many squares with no points inside, and these can skew the RMSD score. It also considers every box equally: an advantage of the CGR is that it creates a distinctive signature of the genome and hence the density of some boxes will be more useful for identifying that particular genome sequence than others. It should therefore be possible to extract key features from the CGR plot that are unique to each species, or even to other taxonomic levels (genus or family especially). Such features may represent combinations of nucleotides that are either over represented or under represented in a taxonomic classification.

A related active area of research is the statistical analysis of the frequency of occurrence of k-mers in genomic sequences. For example, Pride *et al* [16] investigated the evolutionary implications of biases in the usage patterns of 4-mers using 27 microbial genomes: they discovered a phylogenetic signal in the coding regions, and a significant variation between coding and non-coding regions. More recent researchers have developed methods based on the usage patterns of 4-mers to study microbial communities using metagenomics [17]. For instance, the HabiSign software is able to quickly identify subsets of sequences that are specific to a particular habitat [18], and MetaCluster 4.0 software [19] is used for the clustering or taxonomic binning of short metagenomic reads from 100 species. These methods based on k-mer usage patterns may be viewed as a subset of methods that can be applied to CGRs. CGRs are not limited to box sizes corresponding to k-mers, and a major advantage of CGRs is that they can be used for visualisation – a CGR is a type of fractal landscape that can be analysed using established methods of fractal analysis. Although it has not be explored here, it should be possible to use algorithms developed for computer vision to extract distinctive features from the CGR. These methods could be used with CGRs to extract features based on sets of boxes clustered together on the CGR that represent groups of k-mers that are either highly or lowly represented.

5. Conclusions

The Chaos Games Representation of genome sequences was first introduced almost 2 decades ago. It is a convenient way to visualise and compare genome sequences and has been used for a number of alignment-free methods of sequence comparison. While recent papers have investigated its use for analysing small data sets using fractal based methods (i.e. less than 100 sequences), it has not been clear from the literature if it is able to resolve differences between sequences in the large data sets currently being produced by new sequencing technologies. Also, relatively little interest has been given to using the CGR plots themselves as a basis of comparison rather than derived fractal properties. This study has addressed these issues; it used the Genometa data set of microbial species, which was divided into one data set of 1550 target species genomes and queried with sequences derived from an additional 977 genomes – each of these queries is a different strain to the species targets. It has shown that the CGR box density property can be used as a fast method for filtering large data sets and that it is more superior to those based on the fractal properties: whereas BLAST takes 25 hours to search this data set, the method introduced here can be used to prefilter this data set, reducing the 1550 target genomes to just 50, in about 15 minutes. This is a promising result, and approach described here can be further developed – for example with more sophisticated scoring functions.

References

[1] E. R. Mardis, A decade's perspective on DNA sequencing technology, Nature 470 (7333) (2011) 198–203.
[2] E. C. Hayden, Nanopore genome sequencer makes its debut, Nature News and Comment (2012) 17 Feb.
[3] I. Holmes, R. Durbin, S. Centre, W. Trust, G. Campus, Dynamic programming alignment accuracy, J. Comput. Biol 5 (1998) 493–504.
[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool., Journal of molecular biology 215 (3) (1990) 403–410.
[5] S. Vinga, J. S. Almeida, Alignment-free sequence comparison-a review, Bioinformatics 19 (4) (2003) 513–523.
[6] H. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Research 18 (8) (1990) 2163–2170.
[7] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences, Mol. Biol. Evol. 16 (10) (1999) 1391–1399.
[8] Z.-G. Yu, V. Anh, K.-S. Lau, Measure representation and multifractal analysis of complete genomes, Phys. Rev. E 64 (2001) 031903.
[9] Z.-G. Yu, V. Anh, K. Lau, Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome, Physica A 301 (2001) 351–361.
[10] P. Moreno, P. Velez, E. Martinez, L. Garreta, N. Diaz, S. Amador, I. Tischer, J. Gutierrez, A. Naik, F. Tobar, F. Garcia, The human genome: a multifractal analysis, BMC Genomics 12 (1) (2011) 506.
[11] A. Pandit, A. Dasanna, S. Sinha, Multifractal analysis of hiv-1 genomes, Mol Phylogenet Evol 62 (2) (2012) 756–63.
[12] S. S. Mande, M. H. Mohammed, T. S. Ghosh, Classification of metagenomic sequences: methods and challenges, Briefings in Bioinformatics 13 (6) (2012) 669–681.
[13] J. S. Almeida, J. A. Carriço, A. Maretzek, P. A. Noble, M. Fletcher, Analysis of genomic sequences by chaos game representation, Bioinformatics 17 (5) (2001) 429–437.
[14] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, B. I. Shraiman, Fractal measures and their singularities: The characterization of strange sets, Phys. Rev. A 33 (1986) 1141–1151.
[15] C. Davenport, J. Neugebauer, N. Beckmann, B. Friedrich, B. Kameri, S. Kokott, M. Paetow, B. Siekmann, M. Wieding-Drewes, M. Wienhofer, S. Wolf, B. Tommler, V. Ahlers, F. Sprengel, Genometa - a fast and accurate classifier for short metagenomic shotgun reads., Plos One 7 (5) (2012) e41224.
[16] D. T. Pride, R. J. Meinersmann, T. M. Wassenaar, M. J. Blaser, Evolutionary Implications of Microbial Genome Tetranucleotide Frequency Biases, Genome Research 13 (2) (2003) 145–158.
[17] J. Droge, A. C. McHardy, Taxonomic binning of metagenome samples generated by next-generation sequencing technologies, Briefings in Bioinformatics 13 (6) (2012) 464–655.
[18] T. Ghosh, M. Mohammed, H. Rajasingh, S. Chadaram, S. Mande, Habisign: a novel approach for comparison of metagenomes and rapid identification of habitat-specific sequences, BMC Bioinformatics 12 (Suppl 13) (2011) S9.
[19] Y. Wang, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, MetaCluster 4.0: A novel binning algorithm for NGS reads and huge number of species, Journal of Computational Biology 19 (2) (2012) 241–249.