International Journal of Biomedical Data Mining

Kaca et al., Int J Biomed Data Min 2018, 7:1
DOI: 10.4172/2090-4924.1000132

Research Article Open Access

Comparison of Genomes of *Pseudomonas Aeruginosa* Strains by using Chaos Game Representation

Wiesław Kaca[2*], Magdalena Nowak[1] and Grzegorz Czerwonka[2]

[1]*Department of Algebra, Geometry and Topology, Institute of Mathematics, Jan Kochanowski University, Kielce, Poland*

[2]*Department of Microbiology, Jan Kochanowski University, Kielce, Poland*

[*]**Corresponding author:** Kaca W, Department of Microbiology, Jan Kochanowski University, Kielce, Poland, Tel: +48 41 349 63 08; E-mail: wkaca@ujk.edu.pl

Received Date: February 28, 2018; **Accepted Date:** March 5, 2018; **Published Date:** March 12, 2018

Abstract

Chaos Game Representation method was used to compare full genomes of the several laboratory and clinical strains of *Pseudomonas aeruginosa*. The sensitivity of method was tested and results indicate that it can be helpful for clustering DNA base pair sequences and clonal differentiation of pathogenic Gram-negative bacterial strains.

Keywords: Chaos game representation; *Pseudomonas aeruginosa*; Genome

Introduction

Pseudomonas aeruginosa strains are one of human opportunistic pathogen. Strains are isolated on varied clinical settings. One of the most clinically important are lung infections of patients suffering on cystic fibrosis (CF) [1,2]. A differentiation clone of *P. aeruginosa* stains from CF patient's lung is crucial for medical treatment [3]. In our earlier studies of international reference panel *P. aeruginosa* strains we were presented that, except O-LPS structures, growth rate and Gallionella mallonela virulence CF and non-CF strains are not significantly different [3]. In presented studies we proposed new model for bacterial whole nucleoid DNA analysis based on Chaos Game Representation. For test one *P. aeruginosa* laboratory strain PAO 1 and two CF strain from international *P. aeruginosa* panel were chosen. In order to show sensitivity of the using method, we also put through an examination two modifications of PAO 1 genome: first one with colicin gene and the second one with phage lambda genome.

The algorithm of Chaos Game Representation (CGR) grown out of the branch of Mathematics called Fractal Theory. Fractals are objects of infinitely complex structure in a certain mathematical sense. Moreover, they are usually made of repeated pieces which have the same shape as the overall figure. It turns out that many natural objects like clouds, rivers, trees, blood vessels and bacteria colony, have fractal structure [4].

One of the simplest methods of describing fractals are Iterated Function Systems (IFS). An Iterated Function System is a finite set of functions (usually affine and contractive maps) which determines a specific fractal called the attractor of given IFS. Such attractors can be readily illustrated by using stochastic algorithm, known as the Chaos Game. In 1990 H.J. Jeffrey showed that its deterministic version may be used in data analysis to reveal patterns in long data strings such as DNA base pair sequences [5]. The method proposed by (Jeffrey 1990), called Chaos Game Representation, allows to represent whole genome in the unit square called CGR plot. Jeffrey's idea was later generalized by Fiser, Tusnady and Simon for sequences of arbitrary symbols [6]. This allows analyzing nucleotides in genomes, amino acids in proteins or words in languages and representing long symbolic sequences on a two-dimensional plot conserving their statistical properties.

For a given natural number , the CGR technique allows us to obtain the frequencies of all n-element sequences of nucleotides and also to compare and classified entire genomes of organisms. This method has been applied by Pandit A. and Sinha S. [7] to differentiate HIV-1 subtypes.

Materials and Methods

In our research we used CGR to examine genomes of five *Pseudomonas aeruginosa* strains, taken from NCBI Database and from [8]: laboratory strain – PAO 1 and its two insertions (with colicin gene and phage lambda genome), and two strains isolted from CF patients – DK2 and LES B58. The choosen clincal strains differs on growth density: low for LES B58 and high for DK2, pyocyanin production LES B58 and DK2 high and low, respectively; swimming and swarming abilities [3,8]. The length of considered genomes is the following: PAO 1 - 6 264 404 nucleotides, PAO 1 with colicin insertion – 6 265 958, PAO 1 with lambda insertion – 6 312 906, LES B58 - 6 601 757 and DK2 - 6 402 658 nucleotides. The algorithm described in this article was coded using program Mathematica.

CGR plot become on the unit square [0,1]^2 Where each vertex is assign as the four nucleotide bases A = (0,0), T = (0,1), G = (1,0) and C = (1,1). We initialize, by taken the first point x_0 in the middle of the square. Reading given DNA sequence we obtain the next point by the formula: $x_{k+1} = 0.5*(x_k + n)$

Where $n \in \{A, T, C, G\}$ Is $(k+1)$th nucleotide of the DNA sequence. In other words, the point x_{k+1} is plotted half way between the point x_k and the vertex n. The process is repeated for the complete sequence and the entire genome is plotted in a two-dimensional plot. Darkness of appropriate areas on the square correspond to the frequencies of different word of a given lengths. To obtain the frequences of all the k-letter words, CGR plot must be divided into a $(2^k \times 2^k)$ grid and the number of occurrences in each box of the grid must be counted. We can compare such grids of frequencies using appropriate metric and determine the subspecies of organisms basis on

the DNA sequence, like in [7]. Here we use a standard taxi cab metric given by the formula: $d(A,B) = \sum_{i,j=1}^{m} |a_{ij} - b_{ij}|$ for grids (matrices) $A = [a_{ij}]_{(i,j=1)}^{m}$, $B = [b_{ij}]_{i,j=1}^{m}$ and m is the size of the grids A and B.This metric gives us an information about the sum of differences between corresponding positions in grid A and B Comparing all given genomes between each other we obtain the matrix of distances. This allows us to find relationships between DNA sequences.

Results and Discussion

Here we present results of analyze of *Pseudomonas aeruginosa*:PAO 1 (and its two modifications), LES B58 and DK2. First, we generated CGR grids and CGR plots (Figure 1) for each genome with respect to oligonucleotides of length 10 (Figure 1).

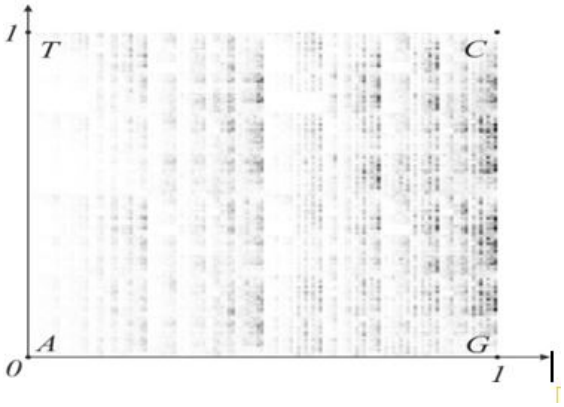

Figure 1: CGR plots for DNA sequences of each genomes *Pseudomonas aeruginosa* (DK2, LES B58 and PAO 1).

We found six the most often repeated words in each DNA sequence (Table 1).

Oligonucleotides of the length 10	Places on the top six lists for each genome				
	DK2	LES B58	PAO 1	PAO 1 + colicin	PAO 1 + lambda
GCCGGCGGCG	1st	1st	2nd	2nd	2nd
CGCCGCCGGC	2nd	2nd	1st	1st	1st
GCCGCCGGCG	3rd	3rd	6th	6th	6th
CGGCGCCGGC	4th	6th	4th	4th	4th
GCCGGCGCCG	5th	4th	5th	5th	5th
CGCCGGCGGC	6th	5th	3rd	3rd	3rd

Table 1: The most often oligonucleotides in genomes *P. aeruginosa* DK2, LES B58 and PAO 1 strain (and its two mutations) and their places on the top six lists for each genome.

Table 1 shows places which are taken by the six the most often 10-nucleotides words in the top lists for all five genomes of Pseudomonas spp.

Results for *P. aeruginosa* clinical DK2 and LES B58 are closer to each other than to PAO1 laboratory strain. Results for every mutation of PAO 1 are the same as for the basic genome. Nevertheless, CGR plots indicted for all five DNA sequences of Pseudomonas are very similar (Figure 1) and we practically do not see any difference between them. Differences occur in grids of DNA frequencies of genomes of three strains. We compared such grids using taxi cab metric and obtained the matrix of distances between genomes (Table 2). Once again we noticed that PAO 1 diverges from the others – the distances between PAO 1 and two other CF strains (1 746 390 and 1 789 650) are more than two times greater than the distance (808 8981) between DK2 and LES B58. The same results can be obtained during CGR analyze genomes with respect to the words of length less than 10 but greater than 5. For comparison, the distance between PAO 1 and the same genome with colicin gene (approx. 1500 nucleotides more) is 1 572. The distance between any genome and the same genome with only one different nucleotide is 20 (when we analyze them with respect to the words of length 10). In our distance matrix (Table 2).

	DK2	LES B58	PAO 1	PAO 1 + colicin	PAO 1 + lambda
DK2	0	808 981	1 735350	1735640	1746390
LES B58	808 981	0	1 781010	1 781200	1 789650
PAO 1	1 735350	1781010	0	1 572	48 516
PAO 1 + colicin	1 735640	1781200	1 572	0	49 860
PAO 1 + lambda	1 746390	1789650	48 516	49 860	0

Table 2: Distance matrix for genomes *pseudomonas* (DK2, LES B58 PAO 1 and its two mutations) with respect to the taxi cab metric.

Conclusion

We can see that every mutation of PAO 1 strain is much more close to the basic genome than to the other strains. Moreover CGR analysis allowed us to presented differences among two clinical isolates. The distance 808 8981 between LES B58 and DK2 shows a sum of differencies between corresponding positions in grids for these two strains and might reflect differences on virulence factors among them. That supposition needs to be confirmed by higher number of DNA sequences studies by CGR methods.

Acknowledgements

We would like to thank Wiesław Kubiś for his fruitful comments and discussions.

Contributors

All authors have read and approved the final article. WK, MN and GC participated in the conception and design of the study and analysis and interpretation of data. WK and MN drafted the manuscript. All authors have final approval of the version to be submitted.

Funding Information

The MG was supported by National Science Centre grant DEC-2012/07/N/ST1/03551 and WK by BS grant 2018 from UJK.

Compliance with Ethical Standards

Isolates were obtained as participants of UE project COST BM 1003 as part of routine activity and were analysed anonymously in a retrospective manner. Ethical approval and informed consent were thus not required.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Campodónico VL, Gadjeva M, Paradis-Bleau C, Uluer A, Pier GB (2008) Airway epithelial control of Pseudomonas aeruginosa infection in cystic fibrosis. Trends Mol Med 14: 120–133.
2. Valderrey AD, Pozuelo MJ, Jiménez PA, Maciá MD, Oliver A, et al. (2010) Chronic colonization by Pseudomonas aeruginosa of patients with obstructive lung diseases: cystic fibrosis, bronchiectasis, and chronic obstructive pulmonary disease. Diagn Microbiol Infect Dis 68: 20–27.
3. Cullen L, Weiser R, Olszak T, Maldonado RF, Moreira AS, et al. (2015) Phenotypic characterization of an international Pseudomonas aeruginosa reference panel: strains of cystic fibrosis (CF) origin show less in vivo virulence than non-CF strains. Microbiology 161: 1961–1977.
4. Barnsley MF (1993) Fractals everywhere (2nd edn.). Academic Press Professional, Boston, MA.
5. Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Re 18: 2163-2170.
6. Fiser A, Tusnady G, Simon I (1994) Chaos game representation of protein structures. J Mol Graph 12: 302-304.
7. Pandit A, Sinha S (2010) Using genomic signatures for HIV-1 sub-typing. BMC Bioinformatics 11: S26.
8. De Soyza A, Hall AJ, Mahenthiralingam E, Drevinek P, Kaca W, et al. (2013) Developing an international Pseudomonas aeruginosa reference panel. Microbiologyopen 2: 1010–1023.