

Power Spectrum Method for the Processing of the DNA in the Genome Sequencing

M. Valla, E. Gescheidtova, and P. Fiala

Department of Theoretical and Experimental Electrical Engineering, DTEEE, FEKT
Brno University of Technology, Brno, Czech Republic

Abstract— The fractal systems are broadly found throughout the nature, generally in any scale. For each dimension (1D, 2D ...) in any structures and in any mathematical algorithms or in any object (sequence, line, square) its dimension can be calculated. These dimensions are specific parameters for description of the DNA sequence characters ACGT strings. Generally a first step consists in the converting of a one-dimensional sequence into the image. In the second step, the method for calculating the dimensions for all scales is selected. For the calculations of the above mentioned dimensions the Power Spectrum (PS) method has been proposed and examined. The Power Spectrum method provides universal calculation of dimension and it allows to obtain the resulting multifractal coefficient. The multifractal coefficient represents the means rate of approximation to ideal power spectrum. It has to be emphasized that the multifractal coefficient is independent of any scale to be chosen. Moreover, the multifractal coefficient serves as an advanced parameter for mathematical description of analyzed specific sequence of the deoxyribonucleic acid (DNA) or the whole genome. The conversion of the sequence into the image as the first step of the pre-processing can be used also for the alternative imaging and the description of sequence and it is possible to choose another method for the processing and analyzing a fractal image (for example Box Counting method). In the paper, the method for the processing of the DNA in the genome sequencing, the Power Spectrum method, will be introduced. The results of the Power Spectrum methods will be presented also.

1. INTRODUCTION

Theory of fractal originated on basis of observation of natural structures. Application of this knowledge for DNA analysis leads to additional insight to genome. Term fractal is derived from Latin *fractus*, which means fraction. Fractals are geometrical shapes, which have non integer dimension and are similar for each other. Self-similarity is phenomenon, which occurs in case, when structure looks identical at whatever magnification. Mathematically, this phenomenon is called invariance against change of scale factor. For each object, its dimension can be calculated. There are many algorithms for calculation of this magnitude. For calculating itself, PS method (Power Spectrum method) was chosen. PS provides more possibilities of analysis at the cost of moderate increasing of calculating demands (for example optional possibilities of local segmentation). Our aim consisted in application of theory of fractals for analysis of images obtained from sequence of DNA code. Image output can be also presented as multidimensional transformation and record of genomic signal (from linear entry in line to planar in figure). All algorithms were developed in environment of matrix laboratory MATLAB.

Bovine virus diarrhea (BVD) is chosen, it is disease of animals, namely cattle. Are analyzed in three versions of the virus genome. Complete genome is downloaded from the NCBI public database.

Description of sequence “Bovine foam virus complete genome” is given by locus ID NC 001831 and length 12 002 bp. Metadata sequence — “A” base is represented by 302% (3 619 nucleotides), “C” base by 226% (2 717 nucleotides), “G” base by 22,8% (2 735 nucleotides) and “T” base by 244% (2 931 nucleotides).

Description of sequence “Bovine immunodeficiency virus complete genome” is given by locus ID NC 001413 and length 8 482 bp. Metadata sequence — “A” base is represented by 318% (6 699 nucleotides), “C” base by 212% (1 798 nucleotides), “G” base by 23,8% (2 017 nucleotides) and “T” base by 232% (1 968 nucleotides).

Description of sequence “Bovine leukaemia virus complete genome” is given by locus ID NC 001414 and length 8 419 bp. Metadata sequence — “A” base is represented by 22,0% (1 850 nucleotides), “C” base by 33,1% (2 790 nucleotides), “G” base by 21,0% (1 770 nucleotides) and “T” base by 239% (2 009 nucleotides). Bovine virus is closely related to human T — lymphotropic virus type HTLV-I.

2. METHODS

Processing of DNA chains is done in two steps, each step is one method used. The first method converts a linear chain DNA into an image (Chaos Game) and for converting the second image into a spectral region (Power Spectrum).

2.1. Chaos Game Representation of DNA

Linear record for DNA sequence (for example AGGCTGGAATGC) must be transformed into figure by following procedure. Square with four apexes — A, C, G, T (opposite pairs A-C and G-T) — and initial point, which is situated at the centre of square, are defined.

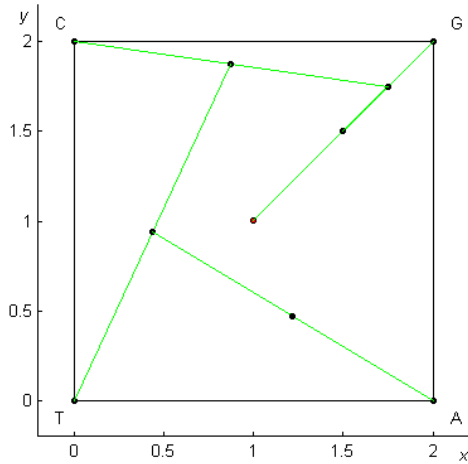


Figure 1: Initiation of attractor designed from first four nucleotides of sequence GCTA.

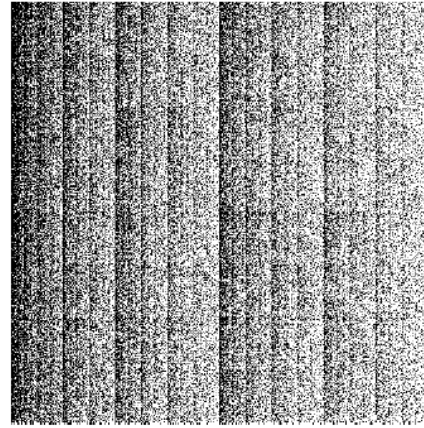


Figure 2: Chaos Game representation of DNA codes of sequence NG_013224 (ID NCBI database).

Square is filled by points by modified method *Chaos game* (description of method: initial point is connected with apex, which is given by character at given position in sequence and at one half of distinct between initial point and apex, point is projected; new point is initial for new interaction — new connection of point with next square apex is given by subsequent character of sequence; schematic demonstration first four steps methods in Figure 1).

2.2. Power Spectrum Method

Power Spectrum represents method of image evaluation by calculation of its power spectrum. Calculation is based on Fourier transformation (obtaining of real and imaginary part of image, marking $F(k_i)$), which is subsequently exponentiated; result is power spectrum P_i expressed by formula (1) [1],

$$P_i = |F(k_i)|^2 = \left(\sqrt{\text{Re}(F(k_i))^2 + \text{Im}(F(k_i))^2} \right)^2 = \text{Re}(F(k_i))^2 + \text{Im}(F(k_i))^2 \quad (1)$$

where $\text{Re}(F(k_i))^2$ is real part of frequency spectrum, $\text{Im}(F(k_i))^2$ is imaginary part of frequency spectrum and k_i is dimensional frequency at i th value. In addition, model of ideal power spectrum $\hat{P}_i$ must be considered.

$\hat{P}_i$ is obtained as model of one-dimensional fractal system (2)

$$\hat{P}_i = c \cdot \frac{1}{|k_i|^\beta} = c \cdot |k_i|^{-\beta}, \quad (2)$$

where c is constant of power spectrum and β is exponent related to fractal dimension D . Aim of this method is in determination of constant β , from which dimension is calculated. Calculation of constant β is realised in accordance with formula (3):

$$\beta = \frac{N \sum_{i=1}^N (\ln P_i) (\ln |k_i|) - \left(\sum_{i=1}^N \ln |k_i| \right) \left(\sum_{i=1}^N \ln P_i \right)}{\left(\sum_{i=1}^N \ln |k_i| \right)^2 - N \sum_{i=1}^N (\ln |k_i|)^2}, \quad (3)$$

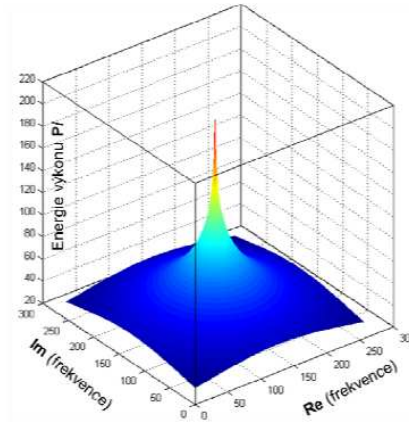


Figure 3: Ideal fractal $\hat{P}_t$ power spectrum of DNA image.

where N indicates number of samples of spectrum.

Subsequently, calculation of PS multifractal coefficient dimension MCD is easy and is represented by formula (4):

$$MCD_{PS} = \frac{5 - \beta}{2} \quad (4)$$

Whole detailed deduction is described in [1], or also with example and figures of individual steps of calculation in [2].

3. RESULTS

Bovine virus is closely related to human T — lymphotropic virus type HTLV-I.

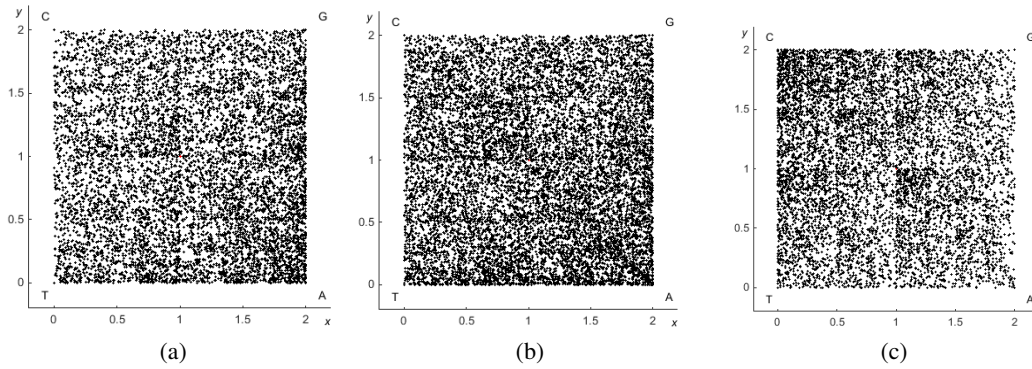


Figure 4: Chaos game representation of image of sequence Bovine foamy virus. (a) Bovine immunodeficiency virus. (b) Bovine leukaemia virus (c).

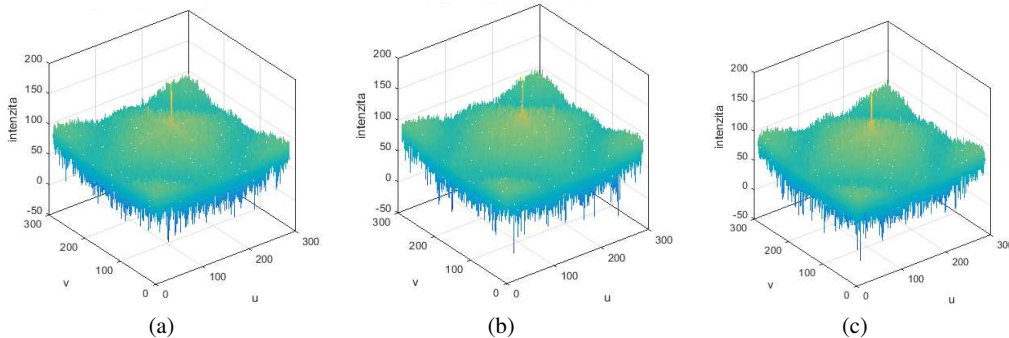


Figure 5: Power spectrum of image of sequence Bovine foamy virus. (a) Bovine immunodeficiency virus. (b) Bovine leukaemia virus (c).

Table 1: Resulting multifractal coefficients model sequences calculated by PS method.

<i>Name</i>	<i>ID sequence</i>	<i>MCD_{PS} [-]</i>
Bovine leukaemia virus	NC 001414	1,813
Bovine immunodeficiency virus	NC 001413	1,818
Bovine foamy virus	NC 001831	1,880

4. CONCLUSIONS

PS method provides universal calculation of dimension and obtains resulting multifractal coefficient MCD. Multifractal coefficient, it means rate of approximation to ideal power spectrum, serves as parameter for mathematical description of analysed sequence. Conversion of sequence to image can be used also for alternative imaging and description of sequence. Fractal coefficient can be used for parameterization and mutual comparing of DNA sequences.

ACKNOWLEDGMENT

Research described in this paper was financed by the by the project of Czech Science Foundation 15-08803S and National Sustainability Program under grant LO1401. For the research, infrastructure of the SIX Center was used.

REFERENCES

1. Turner, J. T., *Fractal Geometry in Digital Imaging*, ISBN 0-12-703970-8, Academic Press, Leicester, 1998.
2. Berthelsen, C. L., "Fractal analysis of DNA sequence data," 160, The University of Utah, 1993.