J Biochem Tech (2010) 2(5):S91-S93
ISSN: 0974-2328

Box counting method in recording, processing and evaluation of genomic signals

Martin Valla*, Jiri Nedved, Dusan Pavlik, Vojtech Adam, Jaromir Hubalek, Libuse Trnkova, Rene Kizek, Ivo Provaznik

Received: 25 October 2010 / Received in revised form: 13 August 2011, Accepted: 25 August 2011, Published: 25 October 2011

Abstract

Fractals are geometrical shapes with noninteger dimension and invariance against change of scale factor. For each geometrical shape, a single parameter - dimension - can be calculated. For calculation, Box Counting Method (BCM) was chosen. Determination of dimension on one level of resolution is not sufficient, it is necessary to subsequently process it and determine multifractal coefficient. Aim of our interest consists in analysis of images obtained from a linear sequence of DNA code. For analysis itself, sequences of *Homo sapiens* and *Citrobacter youngae* were chosen. Results of calculation are fractal coefficients derived from dimensions of generated structures. This result enables to introduce criterion for sequences determination. Graphical outputs may be also represented as multidimensional alternative transformation of linearly recorded genomic signal. Algorithms were developed in computational environment MATLAB. Data were downloaded from a public database EMBL (ESI) and GenBank (NCBI).

Keywords: Fractals, dimension, *Citrobacter youngae, Homo sapiens*

Martin Valla*, Jiri Nedved, Dusan Pavlik, Ivo Provaznik

Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Kolejni 4, CZ-61200 Brno, Czech Republic

*Tel: +420 541 149 411, Fax: +420 541 211 697
E-mail: valla@phd.feec.vutbr.cz

Vojtech Adam, Rene Kizek

Department of Chemistry and Biochemistry, Faculty of Agronomy, Mendel University in Brno, Zemedelska 1, CZ-613 00 Brno, Czech Republic

Jaromir Hubalek

Department of Microelectronics, Faculty of Electrical Engineering and Communication, Brno University of Technology, Udolni 53, CZ-602 00 Brno, Czech Republic

Libuse Trnkova

Department of Chemistry, Faculty of Science, Masaryk University, Kotlarska 2, CZ-611 37 Brno

Introduction

Theory of fractals originated on the 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 word fractus, which means fraction. Fractals are geometrical shapes, which have non-integer dimension and are similar for each other. Self-similarity is a 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 approaches for calculation of this magnitude. For calculation, Box Counting Method (BCM) method was chosen. BCM demonstrates simplicity and lucidity, but on the other hand, disadvantage in processing of colour, or black and white textures are evident. Determination of dimension on one level of resolution is insufficient; calculation must be necessarily carried out in its entirety. Resulting dimensions must be subsequently processed in accordance with chosen procedure; multifractal coefficient must be subsequently determined.

Aim of our object consists in application of theory of fractals for analysis of images obtained from sequence of DNA code. Sequences encoding two proteins (*Homo sapiens* and *Citrobacter youngae*) were chosen for our analysis. Results are fractals coefficients derived from dimensions of structures. They enable introducing of criterion for sequences differentiation. Output can be also presented as multidimensional transformation – imaging – and record of genomic signal (from linear entry in line to planar in Fig 1). All algorithms were developed in computational environment MATLAB.

Materials and methods

Image presentation of DNA

Linear record for DNA sequence (for example AGGCTGGAATGC) must be transformed into Figure 1 by following procedure. Square with four apexes - A, C, G, T (opposite pairs A-C and G-T) - and initial point, which is situated in the centre of square, are defined. The square is determinically divided by points by modified method of *Chaos Game*. The method defines an initial point 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 – a new connection of point with next square apex is given by subsequent character of sequence; schematic demonstration in Berthelsen 1993. First four steps of both methods are demonstrated in Figure 1.

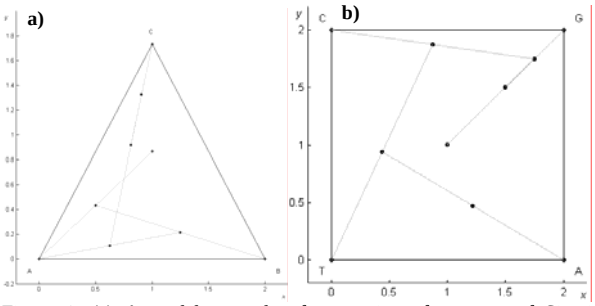

Figure 1: (a) A model example of generation of attractor of Sierpinski triangle by Chaos Game method (b) initiation of attractor designed from first four nucleotides of sequence GCTA

Fig 2a represents information carrying English definition "*Homo sapiens*, Cu++ transporting, alpha polypeptide (ATP7A) on chromosome X'" marked with locus NG_013224 and with length of 146 699 bp. Fig 2b is created from sequence "*Citrobacter youngae* ATCC 29220 C_sp-1.0.1_Cont0.7, whole genome shotgun sequence" marked with locus NZ_ABWL02000007 and with length of 274 831 bp. Lengths of sequences (bp) determine numbers of points marked in square.

ATP7A

Description of sequence „Homo sapiens ATPase, Cu++ transporting, alpha polypeptide (ATP7A) on chromosome X" is given by locus NG_013224 and length 139 699 bp. This sequence encodes transmembrane protein, which enables transport of copper(II) ions through cell membrane. A base is represented by 28,9 %, C base by 19,3 %, G base by 19,4 % and T base by 32,5 %.

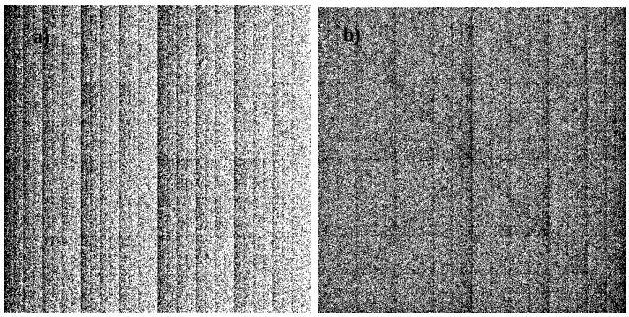

Figure 2: point presentation of part of DNA codes. On the left (a) is sequence NG_013224 and on the right (b) is sequence Z_ABWL02000007 *ATCC 29220*

Sequence is in database described as "*Citrobacter youngae* ATCC 29220 C_sp-1.0.1_Cont0.7, whole genome shotgun sequence" and is characterized by locus NZ_ABWL02000007 and length 274 831 bp. Sequence originates in genome of bacterium *Citrobacter youngae* strain ATCC 29220. Sequence contains 23,7 % of A base, 27,7 % of C base, 25,2 % of G base and 23,4 % of T base.

For more well-arranged illustration, NG_013224 can be projected in colour by association with one colour to one nucleotide - A (yellow) C (green) G (blue) T (red). See Figure 3a in black and white presentation, Figure 3b of the same sequence in colour presentation.

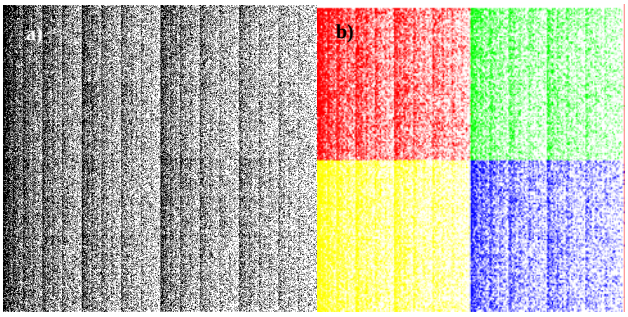

Figure 3: Point presentation of DNA codes of sequence NG_013224. On the left (a) in black and white presentation, on the right (b) in colour presentation of points depiction - G-blue, C-green, T-red, A-yellow.

After transformation of one-dimensional sequence into image, it is necessary to process this image in next step. Image presentation itself represents parameter, so image presentations can be subjectively compared with each other (by insight) on various level of approximation (Berthelsen, 1993). Next possibility consists in mathematical description and evaluation of image.

Box Counting Method processing

Box Counting Method was used for image analysis. Method is based on multiple calculation of classical Hausdorff dimension. Basic formula for calculation is in the form (Turner 1998)

$$D = \frac{\log N}{\log\left(\frac{1}{r}\right)} \qquad (1)$$

Where N is number of self-similar areas, r is size of area and D has significance as dimension of object.

BCM is based on principle of points summation below given mask. Size of mask determines scale of imaging. For each approximation, size of dimension is calculated according formula (1). For higher number of approximations, more dimensions can be obtained; these dimensions can be interlaid by line at imaging in graph. Direction of line is parameter of image and is called multifractal coefficient. Equation (1) can be modified as

$$\log N = D \cdot \log\left(\frac{1}{r}\right) \qquad (2)$$

This form also represents universal equation of line

$$y = k \cdot x \qquad (3)$$

Where $y = \log N$ determines number of points below mask, $k = D$ sequence of dimensions (multifractal coefficient), $x = \log (1/r)$, reverse value of mask size (in pixels).

Results and discussion

Results of functions from Fig 2a and Fig 2b are demonstrated in Figure 4. Dimensions are projected dextrosinistrally so that point on high right corresponds to first approximation (size of mask 2 x 2 pixels), point on the left to this point corresponds to second approximation (size of mask 3 x 3 pixels), etc. Direction of line, which interlays points after 24 approximations, is the searched multifractal coefficient in Figure 4.

After summarizing of fractal coefficients in table, result is following:

Table 1: Multifractal coefficients of two model sequences calculated by BCM method

ID sequence	D_{BCM} *[-]*
NG_013224	1.1445
NZ_ABWL02000007	0.9562

Conclusion

BCM provides calculating of dimensions on different levels of approximation and obtains resulting multifractal coefficient, which serves as parameter for mathematical description of given sequence (direction of line interlaying obtained dimensions). Fractal coefficient can be used for parameterization and mutual comparing of DNA sequences.

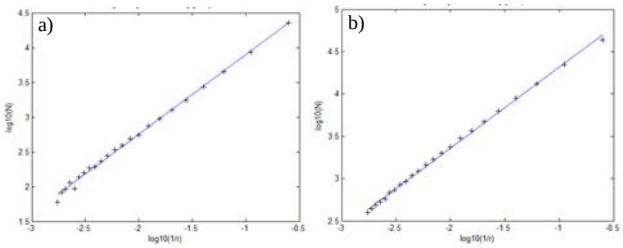

Figure 4: Result of BCM. (a) Result of BCM with multifractal coefficient 1.1445 [-] for sequence NG_013224, and (b) result of BCM with multifractal coefficient 0.9562 [-] for NZ_ABWL02000007.

Acknowledgement

This work was supported by grants FRVS TO A/2894, GACR 102/09/H083, GA AV KAN208130801, and LC06035.

References

Berthelsen ChL (1993) Fractal analysis of DNA sequence data. The University of Utah

Turner, JT (1998) Fractal Geometry in Digital Imaging, Leicester, Academic Press, ISBN 0-12-703970-8