FRACTALS IN IMAGES OF DNA SEQUENCE DATA

Jiří Nedvěd
Bacheor Degree Programme (3), FEEC BUT
E-mail: xnedve00@stud.feec.vutbr.cz

Supervised by: Martin Valla
E-mail: xvalla00@stud.feec.vutbr.cz

ABSTRACT

This article describes methods for calculating the fractal dimension of images obtain from DNA sequence – the Box counting method (BCM). The fractal dimension is a non-integer parameter of textured objects that can be used for image analysis or segmentation. The fractal dimension estimation is based on the resolution invariance of images.

1. ÚVOD

Cílem projektu je uvedení do problematiky teorie fraktálů a aplikace těchto poznatků pro analýzu obrazů získaných ze sekvence DNA kódu. Termín *fraktál* je odvozen od latinského slova *fractus*, čili zlomek. Fraktály se nazývají útvary, které mají neceločíselnou dimenzi a jsou soběpodobné. *Soběpodobnost* je vlastnost, která se vyskytuje v případě, když struktura vypadá stejně v jakémkoliv zvětšení. Matematicky se tato vlastnost nazývá *invariance vůči změně měřítka.* U každého objektu lze také spočítat jeho dimenzi. Existuje mnoho algoritmů pro výpočet této veličiny. Byla vybrána metoda BCM (z angl. Box Counting Method) kvůli své názornosti a jednoduchosti. Jde vlastně o počítání černých pixelů (*N* ve vzorci 1) a porovnání s velikostí masky (*r* ve vzorci 1 – počet pixelů pod maskou).

2. APLIKACE

Využití je při porovnávání DNA sekvencí, kdy parametrem dané sekvence je vypočtená dimenze její obrazové reprezentace. Dimenze, i jejich proložení (fraktální koeficient), reprezentuje danou sekvenci. Samotná obrazová reprezentace už je parametrem a lze je tedy mezi sebou subjektivně (pohledem) porovnávat na různém stupni přiblížení. [3]

Základní vzorec pro výpočet dimenze je ve tvaru: [1]

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

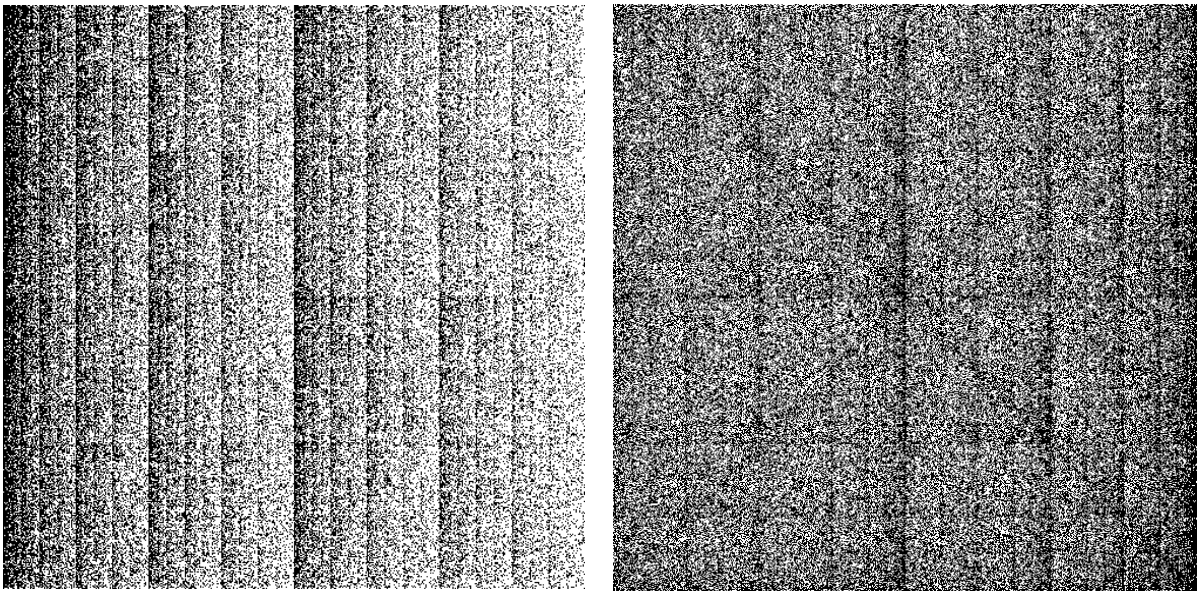
kde N – počet soběpodobných oblastí,

r – velikost oblasti,

D – dimenze objektu.

3. OBRAZOVÁ REPREZENTACE

DNA sekvenci nutno převést do obrazu následujícím způsobem. Je nadefinován čtverec s vrcholy A, C, G a T (protilehlé A, T a C, G) a počáteční bod, který leží uprostřed čtverce. Čtverec je zaplňován body metodou *Chaos game* (popis metody: počáteční bod se spojí s vrcholem, který je dán písmenem na dané pozici v sekvenci, a v polovině vzdálenosti mezi původním bodem a vrcholem se vykreslí bod; nový bod je vstupem pro další iteraci – opět spojení bodu s dalším vrcholem čtverce dán následujícím písmenem sekvence; názorná ukázka ve zdroji [2]). Pro vytvoření následujícího obrázku 1a byla použita sekvence nesoucí anglickou definici ‚Homo sapiens Atlase, Cu++ transporting, alpha polypeptide (ATP7A) on chromosome X', označena lokusem NG_013224 a o délce 146 699 bp (více informací ve zdroji [4a]). Druhý obrázek (1b) je vytvořen ze sekvence ‚Citrobacter youngae ATCC 29220 C_sp-1.0.1_Cont0.7, whole genome shotgun sequence', označena lokusem NZ_ABWL02000007 a má délku 274 831 bp (více informací ve zdroji [4b]). Délky sekvencí (bp) udávají počty bodů vykreslených ve čtverci.



Obrázek 1a, 1b (vpravo): Bodová reprezentace části DNA kódů. Vlevo (a) je sekvence NG_013224 a vpravo (b) je sekvence NZ_ABWL02000007.

4. ZPRACOVÁNÍ

Tato metoda je založena na tzv. „sčítání bodů" pod danou maskou. Velikost masky určuje měřítko zobrazení. Pro každé přiblížení se vypočte velikost dimenze obrázku podle vztahu (1). Pro větší počet přiblížení lze získat více dimenzí a při vynesení do společného grafu pak lze tyto dimenze proložit přímkou. Směrnice přímky je parametr obrázku a nazývá se multifraktální koeficient. Rovnici (1) lze upravit do tvaru

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

tento tvar take vyjadřuje obecnou rovnici přímky

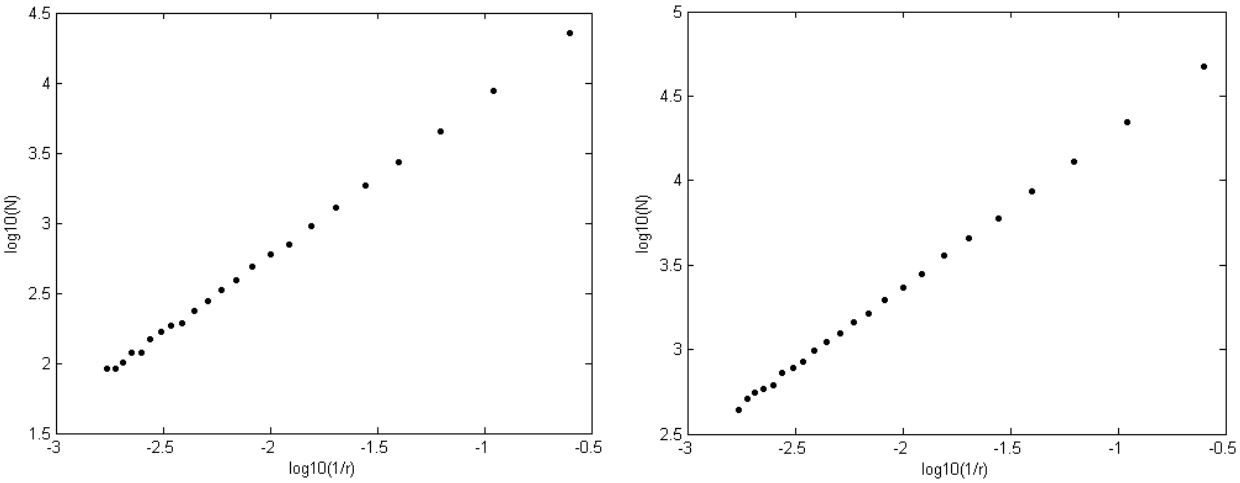
$$y = k \cdot x\,, \tag{3}$$

kde $y = \log N$, udává počet zjištěných bodů pod maskou,

$k = D$,spíše posloupnost dimenzí; nazývá se fraktální koeficient,

$x = \log\left(\frac{1}{r}\right)$, zlogaritmovaná převrácená hodnota velikosti masky v pixelech.

Výsledek vytvořené funkce pro obrázek 1a (popř. 1b) v prostředí MATLAB je na obrázku 2a (popř. 2b). Dimenze jsou vykresleny zprava doleva tak, že prvnímu přiblížení odpovídá bod vpravo nahoře (velikost masky 2 x 2 pixel), druhému přiblížení bod od něj nalevo (velikost masky 3 x 3 pixely), atd. Směrnice přímky, která proloží body v obrázku 2a a 2b, po 24 přiblíženích je hledaný multifraktální koeficient hodnotu $f_k = 1{,}1075$ pro obr. 2a a $f_k = 0{,}9397$ pro obr. 2b.



Obrázek 2a a 2b: Výsledek BCM. Vlevo (a) výsledek BCM s multifraktálním koeficientem f=1,1075 pro sekvenci NG_013224 a vpravo (b) s koeficientem f=0,9397 pro NZ_ABWL02000007.

5. ZÁVĚR

Metoda BCM poskytuje výpočet dimenzí na různých stupních přiblížení a získává výsledný multifraktální koeficient, který slouží jako parametr pro matematický popis dané sekvence (směrnice přímky, která proloží zjištěné dimenze). Fraktální koeficient lze tedy využít pro parametrizaci sekvencí DNA.

LITERATURA

[1] Turner, J., T.: Fractal Geometry in Digital Imaging, Leicester, Academic Press 1998, ISBN 0-12-703970-8

[2] Sierpinski Gasket Generator. Pomoc pro vytvoření funkce. Dostupné z WWW: < http://www.shodor.org/master/fractal/software/Sierpinski.html >

[3] BERTHELSEN, Ch. L.: Fractal analysis of DNA sequence data. The University of Utah, 1993. 160 s.

[4] Sekvence z veřejné databáze NCBI. Přímé odkazy na použité sekvence:
a < http://www.ncbi.nlm.nih.gov/nuccore/262231850 >
b < http://www.ncbi.nlm.nih.gov/nuccore/NZ_ABWL02000007.1?ordinalpos=3&itool=EntrezSystem2.PEntrez.Sequence.Sequence_ResultsPanel.Sequence_RVDocSum >