complexCGR: Encoding DNA sequences as complex numbers $\mathbb{C}$, inspired in Chaos Game Representation.

JA

August 13, 2021

1 Related Work

The Chaos Game Representation of DNA was proposed in 1990 [2]. It is a recursive encoding that allows to represent a sequence $s \in \{A, C, G, T\}^N$, $N \in \mathbb{N}$ in three numbers (N, x, y), where N represents the length of the sequence, and (x, y) is some point in the square

$$S_{CGR} := \{(x, y) : -1 \leq x \leq 1, -1 \leq y \leq 1\}.$$

This triplet is a unique representation of a sequence, and can be decoded nucleotide by nucleotide in order to recover the entire genome.

Using CGR, in [1] they observe that subsequences of a genome exhibit the main characteristics of the whole genome, attesting to the validity of the genomic signature concept. That leads to an image representation based on k-mers called "Frequency Matrix CGR (FCGR)". Figures 1 and 2 provides examples of FCGR for random sequences in absence of nucleotide T.

In 2017 a similar approach to CGR was proposed, based on integers, called iCGR [3].

To define properly the CGR encoding, we need to assign each corner of the S_{CGR} square to a nucleotide. This can be done with $g : \{A, C, G, T\} \to \mathbb{R}^2$

$$g(n) := (g_x(n), g_y(n)) = \begin{cases} (1,1) & , n = A \\ (-1,1) & , n = C \\ (-1,-1) & , n = G \\ (1,-1) & , n = T \end{cases} \tag{1}$$

Definition 1 (CGR Encoding of a DNA sequence)**.** *The CGR encoding for the sequence* $s \in \{A, C, G, T\}^N$ *is given by*

$$\begin{aligned} x_i &= \frac{x_{i-1} + g_x(n_i)}{2} \\ y_i &= \frac{y_{i-1} + g_y(n_i)}{2} \end{aligned} \tag{2}$$

where the starting point $(x_0, y_0) = (0, 0)$*.*

Definition 2 (iCGR: integer CGR encoding of a DNA sequence)**.** *The iCGR encoding for the sequence* $s \in \{A, C, G, T\}^N$ *is given by*

$$\begin{aligned} x_i &= x_{i-1} + 2^{i-1} \cdot g_x(n_i) \\ y_i &= y_{i-1} + 2^{i-1} \cdot g_y(n_i) \end{aligned} \tag{3}$$

where $(x_0, y_0) = (0, 0)$

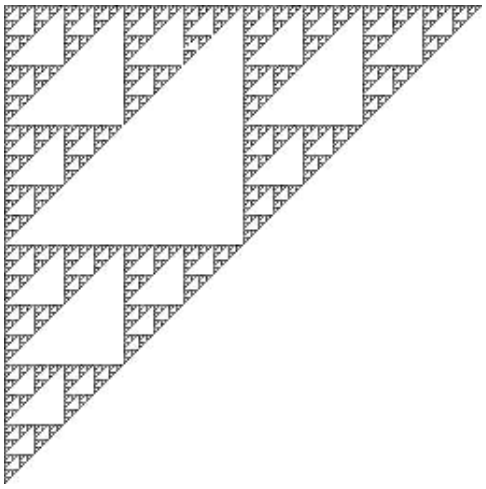

Figure 1: FCGA of a random sequence in absence of nucleotide T.

2 complexCGR

The complexCGR is a DNA sequence encoding that allows to represent an entire genome in two integers (k, N), where k correspond to the index of some of the 4^N roots of the complex number $z = 1$, and N is the length of the sequence.

Based on that every complex number z, $z \neq 0$, has exactly n different n-esim roots, given by

$$w_k = |z|^{1/n} cis\Big(\frac{\theta + 2k\pi}{n}\Big) \tag{4}$$

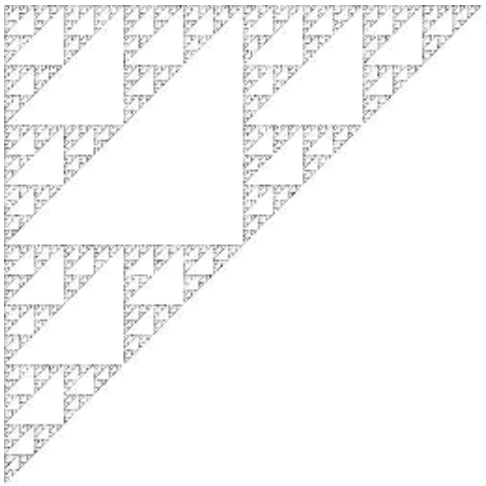

Figure 2: FCGA of a random sequence in absence of nucleotide T and lots of N.

where $\theta \in [0, 2\pi)$ is the argument of z. and in particular, for $z = 1$ we have $\theta = 0$ and $|z| = 1$, then

$$w_k = cis\Big(\frac{2k\pi}{n}\Big) \tag{5}$$

The sequences are assigned to the roots in alphabetic order based on their reverse sequences, starting from $\theta = 0$ (positive real axis in the complex plane), counterclockwise. This particular order is inspired by the CGR encoding, where sub-quadrants preserves the described order, as shown in Figure 3.

$$id(n) = \begin{cases} 0 & , n = A \\ 1 & , n = C \\ 2 & , n = G \\ 3 & , n = T \end{cases} \tag{6}$$

Definition 3 (complexCGR)**.** *Given a sequence $s \in \{A, C, G, T\}^N$, the complexCGR encoding of s is given by (k_N, N), where N is the length of the sequence and k_N is obtained as follows,*

$$k_i = id(n_i)4^{i-1} + k_{i-1} \tag{7}$$

where $k_0 = 0$.

Lemma 1. *Given $s \in \{A, C, G, T\}^N$ and, $\bar{s}$ his complementary sequence. If (k, N) is the complexCGR encoding for s, then $(4^N - k - 1, N)$ is the complexCGR encoding for $\bar{s}$*

Proof. (By induction) For $N = 1$, we have $w_0 \to A$,$w_3 \to T$ and $w_1 \to C$, $w_2 \to G$, then the result is true for this case.

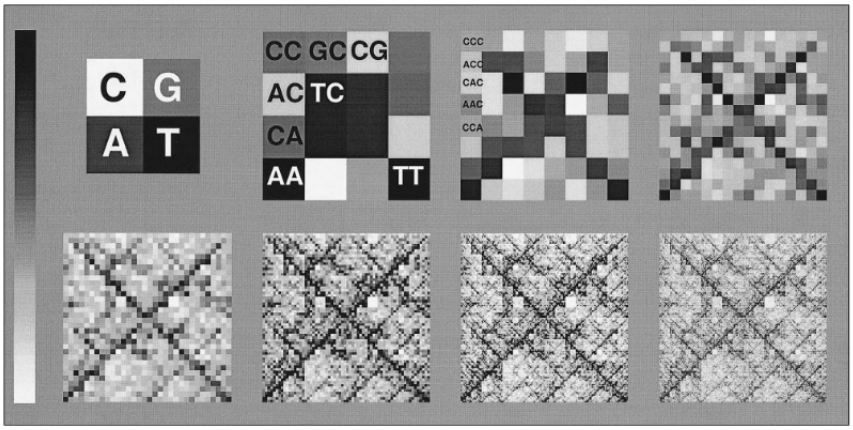

FIG. 1.—The fractal nature of chaos game representation (CGR) images. The frequencies of words up to eight letters long used by the archaebacteria *Archeoglobus fulgidus* are represented from left to right and from top to bottom. For single-letter words, frequencies of letters from only one strand are represented. The gray scale is fitted to the frequency values in order to use its full range of variation for each CGR image.

Figure 3: From [1], page 2.

Now, given $s \in \{A, C, G, T\}^N$ with encoding (k_N, N), we need to prove that the encoding for $\bar{s}$ is $(4^N - k_N - 1, N)$. Let's define $s[N]$ as the last nucleotide of s.

$$\begin{aligned}
\bar{k}_N &= id(\bar{s}[N])4^{N-1} + \bar{k}_{N-1}, \text{ by definition of } k_N \\
&= id(\bar{s}[N])4^{N-1} + 4^{N-1} - k_{N-1} - 1, \text{ by induction hypothesis} \\
&= id(\bar{s}[N])4^{N-1} + 4^{N-1} - k_N + id(s[N])4^{N-1} - 1, \text{ by definition of } k_N \\
&= 4^{N-1}\Big(id(s[N]) + id(\bar{s}[N]) + 1\Big) - k_N - 1 \\
&= 4^N - k_N - 1, \text{by (6) complementary nucleotides sums 3}
\end{aligned}$$

□

Corollary 1. *Let's define $\Delta := 2\pi/n$, the angle between consecutive n-esim roots. If we rotate all the roots by $\Delta/2$, we can affirm that if z is the root for s, then the conjugate of z, $\bar{z}$ is the root for the complementary sequence $\bar{s}$.*

Proof. If z is the root for the N-long sequence s, then there exists $k \in \{0, ..., 4^{N-1}\}$ such that

$$z = cis\Big(2k\pi/4^N\Big)$$

By Lemma 1 the complementary sequence $\bar{s}$ is represented by

$$w = cis\Big(2(4^N - k - 1)\pi/4^N\Big)$$

Now, rotating w by $\Delta/2$ and then computing the conjugate should led to z (also rotated by $\Delta/2$),

$$\overline{z \cdot cis(\Delta/2)} = w \cdot cis(\Delta/2)$$

For w,

$$\begin{aligned} w \cdot cis(\Delta/2) &= cis\Big(\frac{2(4^N - k - 1)}{4^N}\pi\Big) \cdot cis(\Delta/2) \\ &= cis\Big(2\pi - \frac{2k\pi}{4^N} - \frac{2\pi}{4^N}\Big) \cdot cis(\pi/4^N) \\ &= cis\Big(2\pi - \frac{2k\pi}{4^N} - \frac{2\pi}{4^N} + \frac{\pi}{4^N}\Big) \\ &= cis\Big(-\frac{2k\pi}{4^N} - \frac{\pi}{4^N}\Big) \end{aligned} \tag{8}$$

On the other hand, for z,

$$\begin{aligned} z \cdot cis(\Delta/2) &= cis\Big(\frac{2k\pi}{4^N}\Big) \cdot cis(\Delta/2) \\ &= cis\Big(\frac{2k\pi}{4^N} + \frac{\pi}{4^N}\Big) \end{aligned} \tag{9}$$

Then is clear that

$$\overline{z \cdot cis(\Delta/2)} = w \cdot cis(\Delta/2)$$

□

References

[1] P J Deschavanne, A Giron, J Vilain, G Fagot, and B Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Molecular Biology and Evolution*, 16(10):1391–1399, 10 1999.

[2] H.Joel Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Research*, 18(8):2163–2170, 04 1990.

[3] Changchuan Yin. Encoding dna sequences by integer chaos game representation. *Journal of Computational Biology*, 26, 12 2017.