Representación Gráfica del Juego del Caos de Genomas Completos

Gabriela Durán Meza[1], Jeanett López García [2], José Luis del Río Correa[1]
[1]Departamento de Física, UAM - Iztapalapa, Ciudad de México, México.
[2]División de Matemáticas e Ingeniería, UNAM-FES Acatlán, Estado de México.
E-mail: igabydu@xanum.uam.mx, jeanettlg@hotmail.com, jlrc@xanum.uam.mx

Resumen — **El algoritmo CGR (Chaos Game Representation) de Jeffrey revela la estructura fractal intrínseca del ADN y analiza fracciones del genoma de ciertos organismos, regularmente secuencias del orden de 10^6 pb.**
Proponemos una metodología, derivada de un proceso de codificación, que construye una representación numérica sin pérdida de información de una secuencia de ADN. Dicha secuencia numérica, también genera la representación gráfica del ADN, lo importante de nuestra metodología yace en que analizar genomas del orden de 10^9 pb.

Palabras Clave – **genoma, chaos game, multifractal**

Abstract — **Jeffrey's CGR (Chaos Game Representation) algorithm reveals the intrinsic fractal structure of DNA and analyzes fractions of the genome of certain organisms, regularly sequences of the order of 10^6 bp.**
We propose a methodology, derived from a coding process, that constructs a numerical representation without loss of information of a DNA sequence. This numerical sequence also generates the graphical representation of DNA, the important thing of our methodology is to analyze genomes of the order of 10^9 bp.

I. INTRODUCCIÓN

El ADN (Ácido Desoxirribonucleico) es una biomolécula, llamada habitualmente molécula de la vida porque posee toda la información necesaria en los organismos vivos para la creación de células, tejidos y órganos, y por ende, para perpetuar la vida misma. Desde mediados del siglo XX, el ADN es el eminente objeto de estudio de diversas ciencias. El trabajo de Jeffrey en 1990, inaugura el estudio de dicha macromolécula en el contexto de las matemáticas y física. Se convierte en un elemento peculiar para los sistemas dinámicos, para los procesos estocásticos e incluso para la geometría fractal.

En este trabajo se propone una Representación Genómica Binaria (RGB), que permite simplificar la codificación del genoma, debido a que la representación binaria es el lenguaje natural para el análisis de las secuencias genómicas, al expresar cada una de las bases como un par de números binarios, ó equivalente con un vector con componentes que solamente involucra ceros y unos, como se sigue de las coordenadas de Q dadas en (1). La RGB permite relacionar y extender dos propuestas de codificación aparentemente diferentes, la CGR de Jeffrey y la más reciente iCGR de Yin, además la RGB utiliza las ventajas de cada una de ellas como son la Representación Gráfica del Genoma (RGG) de la primera, y el uso de vectores enteros de la segunda. Ambas teorías asocian un punto en el plano a cada subsecuencia generada por la SG, la primera un punto dentro de Q, la segunda en un punto fuera de Q, en ambos casos el conocimiento de N el número de nucleótidos que conforman la subsecuencia y las coordenadas del punto asociado a ella, permiten el conocimiento de los nucleótidos que conforman la subsecuencia, que es la propiedad importante que debe tener cualquier método de codificación, la necesidad de esta propiedad se entiende en forma simple con la RGB.

La metodología de la representación del caos de Jeffrey denominada CGR por sus siglas en inglés (Chaos Game Representation), consiste en obtener un mapa en el cuadrado unitario que representa gráficamente una secuencia de ADN, la cual llamaremos secuencia genómica (SG). Basado en el juego del Caos de Barnsley, este es el elemento eje de esta nueva metodología. Dicho mapa ofrece un abanico de información de las secuencias de ADN, una de las más importantes es que las secuencias de ADN no son aleatorias y poseen una estructura estadísticamente autosimilar, en específico, tienen estructura multifractal [1].

La metodología CGR, emplea el cuadrado unitario Q como soporte de la estructura multifractal intrínseca del ADN. Es decir, en analogía con el juego del caos de Barnsley [2], donde el triángulo unitario funge como tablero o soporte del triángulo de Sierpinski, y el dado como generador de la secuencia aleatoria que construye el fractal [3]. En la metodología CGR, el cuadrado Q es el soporte, mientras se prescinde de un generador o dado, que es sustituido por la secuencia de ADN. Dichas secuencias, pueden descargarse de la bases de datos del National Center for Biotechnology Information.(www.ncbi.nlm.nih.govNCB).

El proceso es similar al juego del caos tradicional, cada esquina de Q está relacionada con una base nitrogenada {A,C,G,T}. Dada una secuencia de ADN, en un cuadrado unitario Q, se asocia a cada uno de sus vértices con una de las cuatro bases que constituyen el alfabeto de la secuencia

genómica. Denotando a las bases por las letras A,C,G y T, los vértices de Q son:

$$A(0,0),C(0,1),G(1,1),T(1,0) \quad (1)$$

Para obtener la representación gráfica de una secuencia genética, se sigue el siguiente proceso iterativo:

a) Se selecciona un punto semilla X_0 dentro del cuadrado, que en general puede ser arbitrario pero que por simplicidad geométrica Jeffrey selecciona X_0 como el punto medio de Q,

b) Al primer término de la SG se le asocia el punto X_1, que es el punto medio de la línea que une X_0 con el vértice del cuadrado cuya esquina coincide con el término de la secuencia genómica.

c) Al segundo término de la SG se le asocia el punto X_2, que es el punto medio de la línea que une el punto X_1 con el vértice de Q cuya esquina coincide con el término de la secuencia.

d) Repitiendo este proceso para toda la secuencia de DNA se generan tantos puntos como tenga la secuencia dada de DNA.

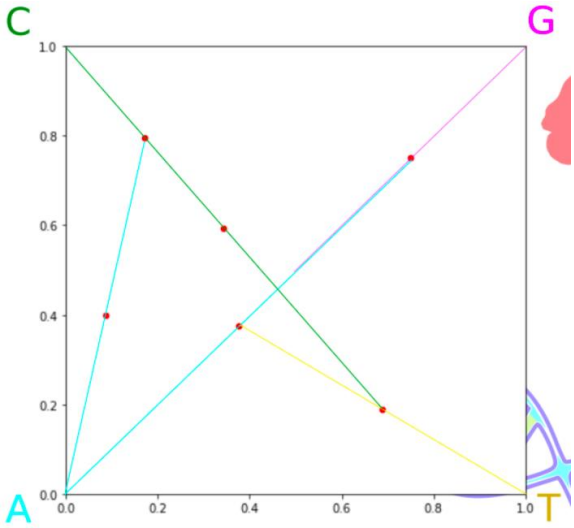


Figura 1. Representación del Juego del Caos de Jeffrey para una secuencia ejemplo **GATCCA**.

En la figura 1 mostramos un ejemplo pequeño, de una secuencia de 6 bases, donde los puntos rojos corresponden a los puntos medios que provienen del método CGR de Jeffrey, al que denominamos CGR canónico.

Para codificar una SG Jeffrey utiliza la regla de recurrencia

$$P_n = \frac{1}{2}\left(P_{n-1} + V_n\right) \; con \; P_0 = \left(\frac{1}{2},\frac{1}{2}\right); \quad (2)$$

y usando la SG genera N puntos dentro de el cuadrado unitario Q:

$$P_1 = \frac{1}{2}\left(V_1 + P_0\right); P_2 = \frac{1}{2}\left(P_1 + V_2\right); P_3 = \frac{1}{2}\left(P_2 + V_3\right); ..., P_N = \frac{1}{2}\left(P_{N-1} + V_N\right); \quad (3)$$

que en términos de los vectores V_i toman la forma:

$$P_1 = \frac{1}{2}\left(V_1 + P_0\right); P_2 = \frac{1}{2}V_2 + \left(\frac{1}{2}\right)^2\left(V_1 + P_0\right); P_3 = \frac{1}{2}V_3 + \left(\frac{1}{2}\right)^2 V_2 + \left(\frac{1}{2}\right)^3\left(V_1 + P_0\right)_1;$$

$$P_R = \frac{1}{2}V_R + \left(\frac{1}{2}\right)^2 V_{R-1} + \left(\frac{1}{2}\right)^3 V_{R-2} + ... + \left(\frac{1}{2}\right)^{R-2} V_3 + \left(\frac{1}{2}\right)^{R-1} V_2 + \left(\frac{1}{2}\right)^R\left(V_1 + P_0\right); \quad (4)$$

de donde se pueden demostrar las relaciones siguientes:

$$P_{R-S} = 2^S P_R - Int\left(2^S P_R\right); \qquad V_{R-S} = Int\left(2^{S+1} P_R\right) - 2Int\left(2^S P_R\right); \quad (5)$$

de forma que conociendo el punto asociado a la secuencia entera, se encuentran los puntos de cualquier subsecuencia, y el nucleótido que ocupa el primer lugar de ella.

La representación RGB[2], se originó a partir de la iCGR recientemente propuesta por Yin[4], donde la codificación del genoma se hace en forma vectorial, utilizando a los vectores:

$$C(-1,-1),T(-1,1),A(1,1),T(1,-1); \quad (6)$$

expresando cualquier subsecuencia como suma de múltiplos enteros de estos vectores, para lo cual Yin considera un cuadrado con centro en el origen de coordenadas, y sus vértices en los puntos C(-1,-1), T(-1,1), A(1,1),G(1,-1) y como se muestra en la figura:

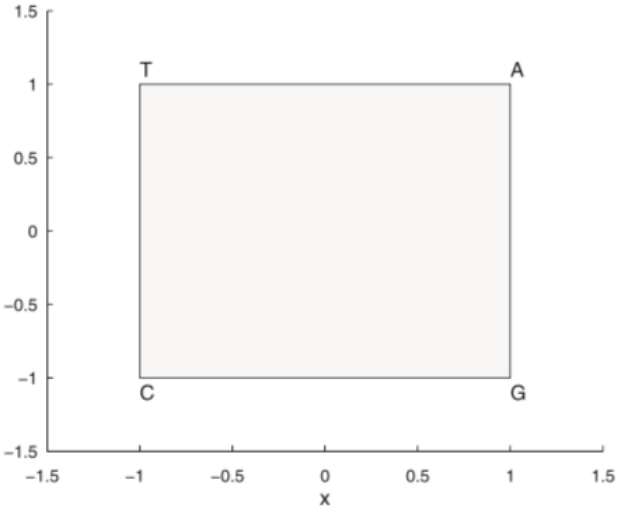


Figura 2. Propuesta de cuadrado con centro en el origen (metodología de Yin).

Denotando por $\vec{A},\vec{C},\vec{G},\vec{T}$ a los vectores de posición de los vértices anteriores, y dada la secuencia genómica $V_1V_2\ldots V_N$,

donde Vi ={A,C,G,T}, la secuencia se representa por N puntos generados al utilizar la ley de recurrencia:

$$\vec{P}_n = \vec{P}_{n-1} + 2^{n-1}\vec{V}_n \ \ para\ n \in \{1,2,...,N\},\ donde\ \vec{P}_0 = \vec{C} = (0,0); \tag{7}$$

así, los N puntos generados por la secuencia genómica $V_1V_2\ldots V_N$ son:

$$P_1 = \vec{V}_1, P_2 = P_1 + 2\vec{V}_2, P_3 = P_2 + 2^2\vec{V}_2, ..., P_N = P_{N-1} + 2^{N-1}\vec{V}_N \tag{8}$$

para codificar una SG, la fórmula de recurrencia se inicializa con el vector que corresponde a la esquina del cuadrado del primer nucleótido de la SG.

Es muy instructivo poner la forma de los vectores en términos de los vectores de las esquinas del cuadrado usado por Yin:

$$\begin{aligned} &\mathbf{P}_1 = \mathbf{V}_1, \mathbf{P}_2 = \mathbf{V}_1 + 2\mathbf{V}_2, \mathbf{P}_3 = \mathbf{V}_1 + 2\mathbf{V}_2 + 2^2\mathbf{V}_3, ..., \\ &\mathbf{P}_R = \mathbf{V}_1 + 2\mathbf{V}_2 + 2^2\mathbf{V}_3 + ... + 2^{R-1}\mathbf{V}_R; \end{aligned} \tag{9}$$

con componentes:

$$\begin{aligned} X_R &= V_{1x} + 2V_{2x} + 2^2V_{3x} + ... + 2^{R-2}V_{(n-1)x} + 2^{R-1}V_{Rx}; \quad V_{Rx} \in \{+1,-1\}; \\ Y_R &= V_{1y} + 2V_{2y} + 2^2V_{3y} + ... + 2^{R-2}V_{(n-1)y} + 2^{R-1}V_{Ry}; \quad V_{Ry} \in \{+1,-1\}; \end{aligned} \tag{10}$$

de donde se encuentra las propiedades:

$$signo(X_R) = signo(V_{Rx}); \quad signo(Y_R) = signo(V_{Ry}); \tag{11}$$

que permiten determinar cuál es el nucleótido del último término de la secuencia con R elementos.

El nucleótido anterior se encuentra utilizando (11) para determinar las coordenadas de la subsecuencia de R-1 nucleótidos, v.gr.

$$X_{R-1} = X_R - 2^{R-1}X_R; \quad Y_{R-1} = Y_R - 2^{R-1}Y_R; \tag{12}$$

y usando (6) se determina el (R-1) nucleótido:

$$signo(X_{R-1}) = signo(V_{xR-1}); \quad signo(Y_{R-1}) = signo(V_{yR-1}); \tag{13}$$

siguiendo este procedimiento sistemáticamente se encuentran todos los nucleótidos que conforman la secuencia.

El mapeo cerrado en Q que proponemos, se origina a partir del mapeo abierto a todo el plano propuesto por Yin [4], sin embargo nuestra propuesta difiere principalmente en la herramienta que denominamos *codificación en segmentos* de M nucleótidos, que consiste en separar la secuencia completa en subsecuencias. El cual provee una representación numérica sin pérdida de información de una secuencia considerablemente grande de ADN. Dicha secuencia numérica, proviene de un proceso de codificación binaria que proporciona una útil herramienta con la cual es posible obtener las representaciones CGR de organismos eucariotas con genomas del tamaño $123X10^6$ pb. Esta propuesta abre el panorama de aplicación de la metodología CGR, ya que podrían estudiarse organismos más diversos como es el extenso mundo de los organismo eucariotas.

II. METODOLOGÍA

La metodología CGR en secuencias de ADN es una técnica que se usa con más frecuencia en los últimos años [5,6], incluso esta metodología ha incursionado en otros procesos, tales como modelos de encriptación de imágenes dígitales [7]. Esta técnica es una herramienta eficiente para detectar estructuras, principalmente autosimilares en los distintos procesos de la naturaleza [1].

Inspirados en el mapeo iCGR de Yin [4], en la que se representa una secuencia por tres números enteros (positivos y negativos), proponemos una representación alternativa, en donde todos los enteros son positivos, y los puntos asociados a las secuencias están restringidos al primer cuadrante sin perder la representación gráfica de Jeffrey, aunado a esto permite extender dicha representación. Utilizando la representación binaria de las coordenadas de los puntos (X,Y) del cuadrado unitario, se propone el arreglo de la figura 3.

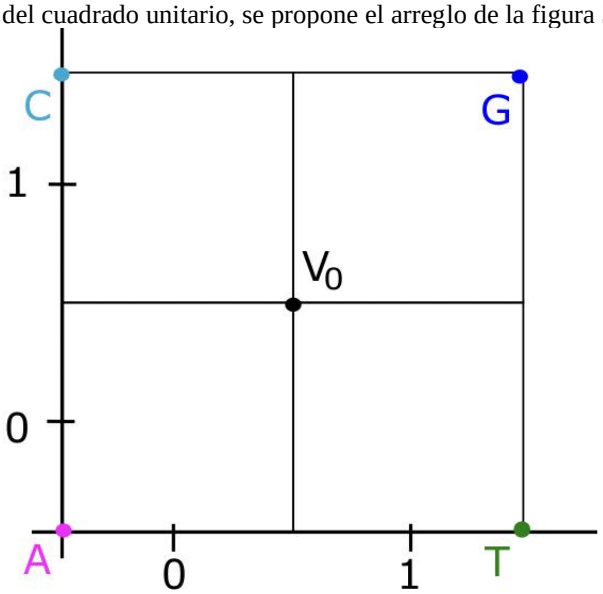


Figura 3. Representación binaria de las coordenadas de los puntos (X,Y). Representación gráfica de los puntos con su respectiva base.

En la figura 3 se muestra el cuadrado unitario y el punto medio de él V_0, el *pseudonucleótido*. Cuyas coordenadas son:

$$A=(0,0);\ C=(0,1);\ G=(1,1);\ T=(1,0);\ P_0=(0.1,0.1) \tag{14}$$

De manera que la información anterior se compacta en la expresión (15):

$$\begin{pmatrix} & P_0 & A & C & G & T \\ X & 0.1 & 0 & 0 & 1 & 1 \\ Y & 0.1 & 0 & 1 & 1 & 0 \end{pmatrix} \tag{15}$$

donde P_0 corresponde a un punto inicial o punto semilla de coordenadas $P_0 = (0.1,0.1)$.

El objetivo principal de la codificación en segmentos es comprimir una secuencia de ADN, es decir; dada una secuencia de ADN se genera una secuencia más pequeña pero que simboliza totalmente a la secuencia original. Sin embargo, se logra una particularidad más, a través de la codificación, se obtiene un objeto matemático, al que denominamos secuencia numérica, que es una representación efectiva y completa de la secuencia de ADN.

Para codificar una SG de N nucleótidos, utilizamos los resultados mostrados en [2], que son dados a continuación.
Se considera la Secuencia Genómica Extendida (SGE), que se obtiene agregando a la SG dada el seudo-nucleótido V_0. La SGE se divide en p partes cada una conteniendo M nucleótidos, y una parte adicional formada por q<M nucleótidos, dando esta información en el par de enteros (N+1,M), que permiten encontrar p y q:

$$\begin{aligned} p &= Int\frac{N+1}{M}; \\ q &= \frac{N+1}{M}\bmod 1; \end{aligned} \tag{16}$$

El siguiente paso consiste en asociar a cada uno de los segmentos en que se ha separado la SGE, un punto con coordenadas enteras en el primer cuadrante C1.
Lo cual se encuentra usando el hecho de que la SGE tiene asociada el vector [2]:

$$\mathbf{I}_\alpha = \sum_{j=1}^{M} 2^j \mathbf{V}_j^\alpha; \tag{17}$$

que se reescribe en la siguiente forma:

$$I = \sum_{\alpha=0}^{p-1} 2^{\alpha M}\tilde{N}_\alpha + 2^{pM}\sum_{K=0}^{q} 2^K \tilde{N}_p; \tag{18}$$

donde

$$\tilde{N}_\alpha = \sum_{\alpha=0}^{M-1} 2^K V_{\alpha M+K}; \quad \tilde{N}_p = \sum_{\alpha=0}^{q} 2^K V_{\alpha M+K}; \tag{19}$$

De manera que la SGE y sus p+1 subsecuencias, están codificadas por los siguientes vectores de componentes enteras:

$$\tilde{\mathbf{N}}_0, \tilde{\mathbf{N}}_1, ..., \tilde{\mathbf{N}}_{p-1}, \tilde{\mathbf{N}}_p \tag{20}$$

y por el vector con componentes (N,M), que se requiere para encontrar p y q.

Resumiendo, en este trabajo se utiliza la codificación en segmentos propuesta en la ref. [2]. Este proceso divide una secuencia de ADN de largo N en varios segmentos de M bases, de la siguiente forma:

$$\underbrace{V_0 V_1 \cdots V_{M-1}}_{\alpha=0}\ \underbrace{V_M V_{M+1} \cdots V_{2M-1}}_{\alpha=1}\ \cdots$$

$$\underbrace{V_{(p-1)M} V_{(p-1)M+1} \cdots V_{pM-1}}_{\alpha=p-1}\ \underbrace{V_{pM} V_{pM+1} \cdots V_{pM+q}}_{\alpha=p} \tag{21}$$

donde V_i denota alguna de las cuatro distintas bases o caracteres V_i = A,C,G,T y V_0 es el pseudonucleótido, que se agrega al inicio de la secuencia completa.

Dada una secuencia de ADN con N bases, se codifica por una secuencia de p + 2 vectores de la siguiente forma:

$$[(N+1,M),(N_{0X},N_{0Y}),(N_{1X},N_{1Y}),...,(N_{pX},N_{pY})] \tag{22}$$

El primer par ordenado (N+1,M), la primera entrada corresponde al largo total de la secuencia, más el pseudonucleótido V_0 y la segunda entrada es la longitud de cada segmento. El primer par es de naturaleza global, dando información sobre la SG y el tamaño de los segmentos, además de no considerarse en la representación gráfica.
Para mostrar el procedimiento anterior consideraremos un caso simple. Ejemplificando el proceso antes descrito, consideremos la codificación de la secuencia extendida de 30 nucleótidos mas un seudonucleótido:
V_0ACTGGTATTGTACTCCTAACCCTAGGTTGC
deseamos formar grupos de 8 nucleótidos, así el primer vector es (30+1,8), con el que determinamos p y q dados por () , encontrando:

$$p = Int\frac{31}{8} = 3, q = 31\bmod 8 = 7;$$

así tendremos 3 grupos de 8 y uno de 7 nucleótidos

una secuencia con 31 nucleótidos (incluyendo V_0) y grupos de M=8.Por lo que el primer vector es (31,8), como p y q
Tomamos la Secuencia Genómica extendida, colocando debajo de cada nucleótido su vector correspondiente dado por (15), obteniendo el siguiente arreglo:
:

SG	V_0ACT	$GGTA$	$TTGT$	$ACTC$
X	1001	1110	1111	0010
Y	1010	1100	0010	0101

SG	$CTAA$	$CCCT$	$AGGT$	TGC
X	0100	0001	0111	110
Y	1000	1110	0110	011

(23)

Para el grupo con α=0, obtenemos

$$N_{0X} = 2^0 + 2^3 + 2^4 + 2^5 + 2^6 = 121$$

$$N_{0Y} = 2^0 + 2^2 + 2^4 + 2^5 = 53$$

Para el grupo con α=1

$$N_{1X} = 2^0 + 2^1 + 2^2 + 2^3 + 2^6 = 79$$

$$N_{1Y} = 2^2 + 2^5 + 2^7 = 164$$

Para el grupo con α=2

$$N_{2X} = 2^1 + 2^7 = 130$$

$$N_{2Y} = 2^0 + 2^4 + 2^5 + 2^6 = 113$$

Para el grupo con α=3

$$N_{3X} = 2^1 + 2^2 + 2^3 + 2^4 + 2^5 = 62$$

$$N_{3Y} = 2^1 + 2^2 + 2^5 + 2^6 = 102$$

por lo que la secuencia genómica:
V_0ACTGGTATTGTACTCCTAACCCTAGGTTGC
esta caracterizada por los enteros (8,3,7) y los pares (121,53),(79,164), (130,113), (62,102).
Los vectores que caracterizan a las subsecuencia de los primeros 8,16.24 y 31 nucleótidos son:

$$I_8 = \begin{pmatrix} 121 \\ 53 \end{pmatrix}; I_{16} = \begin{pmatrix} 121 + 2^8 x79 \\ 53 + 2^8 x164 \end{pmatrix}; I_{24} = \begin{pmatrix} 121 + 2^8 x79 + 2^{16} x130 \\ 53 + 2^8 x164 + 2^{16} x113 \end{pmatrix};$$

$$I_{31} = \begin{pmatrix} 121 + 2^8 x79 + 2^{16} x130 + 2^{24} x62 \\ 53 + 2^8 x164 + 2^{16} x113 + 2^{24} x102 \end{pmatrix};$$

También se puede calcular fácilmente el vector para cualquier subsecuencia intermedia, por ejemplo si se requiere la de 21 nucleótidos, se requieren los primeros 5 nucleótidos de α=2, que son:

SG	$CTAA$	C		
X	0100	0	=	2
Y	1000	1	=	1

de donde se encuentra:

$$I_{21} = I_{16} + 2^{16} \begin{pmatrix} 2 \\ 1 \end{pmatrix};$$

este ejemplo muestra la potencia y versatilidad del método que proponemos.
Ya que por una parte permite manejar un número mucho menor de puntos a graficar, que en este ejemplo son 4 puntos, en lugar de los 30 requeridos en el método Jeffrey, que permite hacer la representación grafica de SG mucho mayores que las habituales, sin perder información.

El tratamiento anterior corresponde al proceso de codificación de secuencias genómicas presentado en la referencia [2], en este mismo trabajo, se puede estudiar el proceso de decodificación. Para efectos de este trabajo, nosotros sólo consideramos el proceso de codificación con la intención de obtener (4).

IV. DISCUSIÓN

El vector (22) deviene a consecuencia del proceso de codificación. Dicho proceso prevalece como una forma compacta que representa a la secuencia de ADN original. Sin embargo, el vector (22) es una secuencia numérica que representa los patrones y caracteres intrínsecos y significativos de la secuencia original de ADN. Este hecho incita a realizar un diverso y amplio análisis, gráfico, numérico y geométrico/fractal, de las secuencias numéricas que representan a las secuencias de ADN.

El objetivo de este trabajo consiste en utilizar al vector (22) como una herramienta útil y poderosa que calcula CGR de organismos eucariotas, los cuales presentan genomas mayores de 10^6 pb. En las referencias [1,9], se calculan CGR para organismos principalmente procariotas, por ejemplo de virus y bacterias. En las referencias mencionadas, no sólo las secuencias procariotas son el objeto de estudio, las secuencias eucariotas son analizadas, pero sólo fracciones del genoma, es decir, presentan los CGR de cromosomas o genes del genoma Humano, sin embargo las secuencias estudiadas no superan el orden de 10^6 pb.

El algoritmo tradicional de Jeffrey, es una técnica poderosa que considera la representación gráfica de las secuencias de ADN vía el Juego del Caos. Esta metodología revela la autosimilaridad propia de las secuencias, es decir, las secuencias de ADN no son aleatorias empero tienen estructura multifractal y su estructura es claramente visible en los CGR derivados del mapeo canónico de Jeffrey como se muestra en la figura 4.

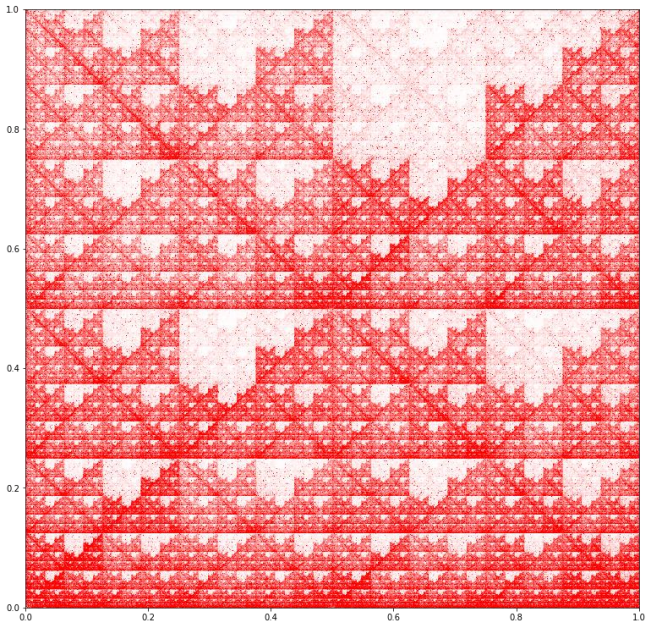


Figura 4. Fracción del genoma de la Ballena Jorobada (*Megaptera novaeangliae*) , $40X10^6$ pb.

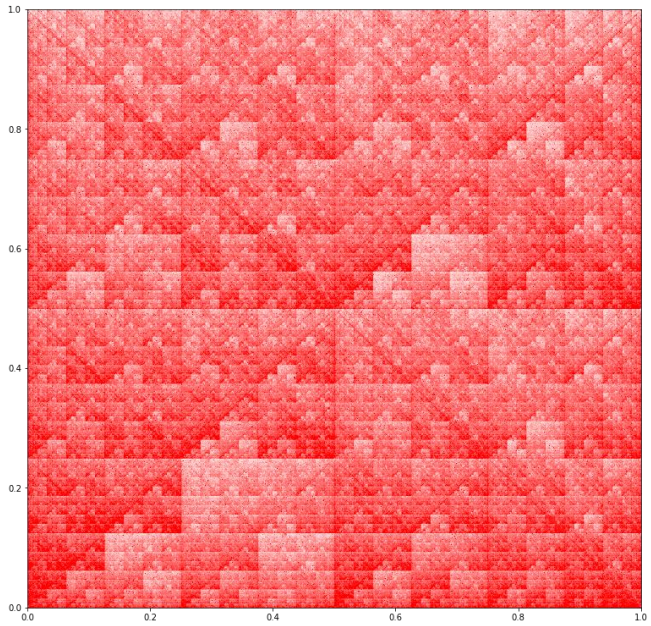


Figura 5. RGB del genoma completo de *Psilocybe cubensis*.

En la figura 4 se muestra la representación gráfica CGR de sólo $40X10^6$ pb de un total de $2.74X10^9$ pb del genoma completo de la Ballena Jorobada, con nombre científico *Megaptera novaeangliae*. Dicho CGR se distingue por su incuestionable estructura fractal, los patrones repetidos a distintas escalas, las zonas poco visitadas y las diagonales que se llenan, son una evidencia de que la secuencia de dicho organismo posee características intrínsecas en su genoma.

La mosca de la fruta, con nombre científico *Drosophila melanogaster*, es un animal que sirve como modelo genético. Por su rápida reproducción, se pueden estudiar varias generaciones en un corto tiempo. Desde principios del siglo XX se utiliza frecuentemente en experimentación genética, recientemente se sabe que algunos genes humanos vinculados con enfermedades tienen su homólogo en el genoma de la mosca de la fruta [10, 11]. Su genoma tiene $165X10^6$ pb, con cerca de 13600 genes. Por lo anteriormente mencionado, dicho organismo surge como un incuestionable objetivo del algoritmo canónico CGR. Sin embargo, es imposible obtener la representación gráfica CGR del genoma completo de la Drosophila melanogaster. El algoritmo canónico CGR sólo puede representar organismos con genomas casi 50% más pequeños. Si bien, no es posible medir el CGR canónico del genoma completo de organismo tan grandes como el de la ballena jorobada o de la mosca de la fruta, es posible calcular el CGR de organismos más pequeños, como se muestra a continuación.

El genoma completo del hongo basidiomycota *Psilocybe cubensis*, muy importante por sus potenciales aplicaciones médicas, especialmente en trastornos psiquiatricos y neurológicos [12], tiene el tamaño de $46.6X10^6$ pb. En su mapeo CGR que se observa en la figura 5, manifiesta la estructura estadísticamente autosimilar, en el estilo propio del organismo.

Otro ejemplo, con un genoma más grande de $80X10^6$ pb es la planta carnívora acuática *Utricularia gibba*, en este caso el CGR es apenas visible, se observa la estructura cercanamente autosimilar pero los puntos comienzan a llenar Q. Esto es una consecuencia debido a la enorme cantidad de puntos y no por una aleatoriedad propia de la secuencia, ver figura 6.
El vector (22) resuelve las limitaciones del mapeo canónico CGR, ya que funge como la representación numérica de la secuencia total de ADN en su versión comprimida, es decir, gracias al proceso de codificación, la secuencia numérica (4) contiene toda la información de la secuencia de ADN original en cuestión. En particular, el vector (22) puede representar el genoma completo de organismos eucariotas, por ejemplo el genoma de la *Drosophila melanogaster*. Graficando los pares ordenados de (22), omitiendo el primer par ordenado, se reproduce la estructura multifractal intrínseca del organismo, ver figura 7.

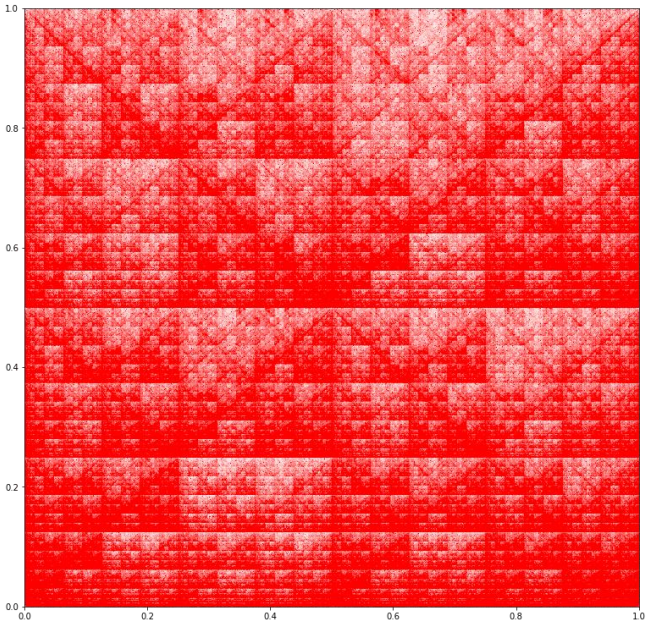
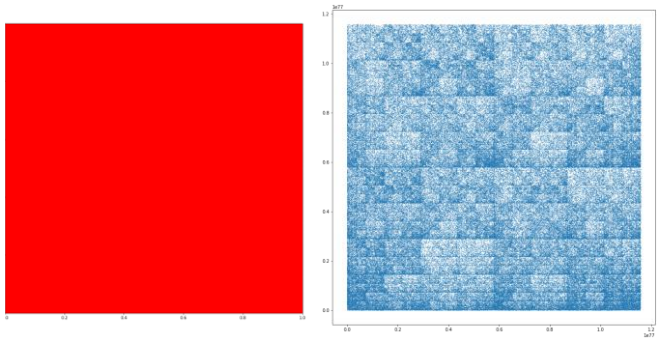


Figura 6. CGR del genoma completo de *Urticaria Gibba*.



(a) (b)

Figura 7. (a) CGR canónico y (b) RGB del genoma completo de la *Drosophila melanogaster*.

Con el mapeo binario RGB, se puede analizar el genoma completo del mamífero acuático *Megaptera novaeangliae* con 2.74X10^9 pb como se muestra en la figura 8, empleando segmentos con 1023 pb, esto implica que M = 1023.

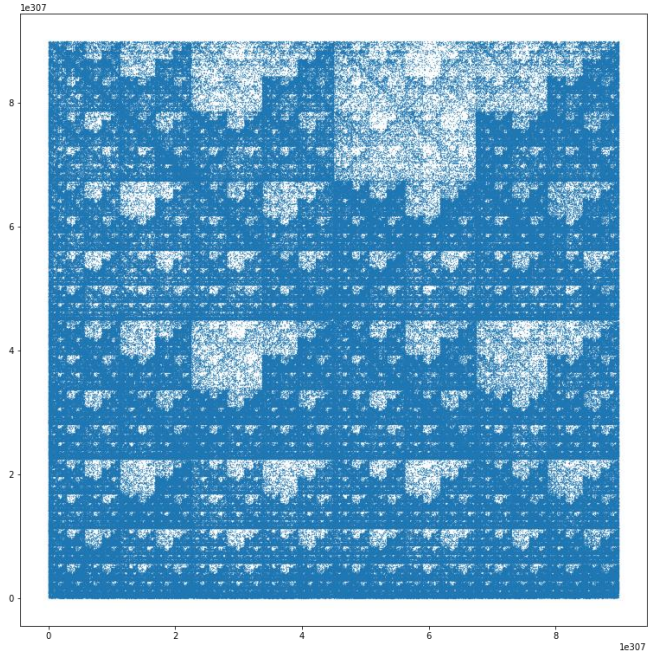


Figura 8. Mapeo RGB del genoma completo de *Megaptera novaeangliae*.

V. CONCLUSIONES

Se concluye que la nueva representación binaria propuesta RGB, simplifica el proceso de codificación – decodificación, introduce un nuevo proceso de codificación, de grano grueso dado en (22), que es una eficiente herramienta para generar la representaciones gráficas de genomas completos de organismo mucho mayores que 10^6 pb. Como se muestra en la figura 7 (a), el mapeo canónico CGR es limitado, por lo tanto no puede analizar genomas más grandes y de gran interés como lo es el genoma de la *Drosophila melanogaster*. Sin embargo, el mapeo RGB, figura 7 (b) muestra que el genoma tiene estructura, este resultado es el opuesto al obtenido en la figura 7 (a), donde da la apariencia de que la secuencia es aleatoria por llenar todo el cuadrado unitario. Sin embargo esta apariencia resulta de las limitaciones del algoritmo y no de la estructura propia de la secuencia.

También se muestra que el algoritmo RGB es apto para medir genomas mayores que 10^9 pb, como lo observamos en la figura 8. Donde presentamos la representación gráfica del juego del caos del inmenso genoma de la ballena jorobada.

Lo anterior es posible gracias al mapeo binario de los nucleótidos y al proceso de codificación en segmentos de M nucleótidos, por lo que esta metodología surge como una herramienta eficiente y poderosa que permitirá analizar varios

grupos de organismos sin limitaciones en cuestión de tamaños de las secuencias.

REFERENCIAS

[1] Durán-Meza, G., López-García, J., & del Río-Correa, J. L. (2019). The self-similarity properties and multifractal analysis of DNA sequences. *Applied Mathematics and Nonlinear Sciences*, *4*(1), 267-278.

[2] Del Río-Correa, J. L., Álvarez-Ballesteros, Y. A., & Durán-Meza, G. Codificación y Decodificación de Secuencias Genómicas.

[3] Barnsley, M. F. (2014). *Fractals everywhere*. Academic press.

[4] Yin, C. (2019). Encoding and decoding DNA sequences by integer chaos game representation. *Journal of Computational Biology*, *26*(2), 143-151.

[5] Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A., & Fletcher, M. (2001). Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, *17*(5), 429-437.

[6] Löchel, H. F., Eger, D., Sperlea, T., & Heider, D. (2020). Deep learning on chaos game representation for proteins. *Bioinformatics*, *36*(1), 272-279.

[7] Ayubi, P., Setayeshi, S., & Rahmani, A. M. (2020). Deterministic chaos game: a new fractal based pseudo-random number generator and its cryptographic application. *Journal of Information Security and Applications*, *52*, 102472.

[8] Yin, C. (2019). Encoding and decoding DNA sequences by integer chaos game representation. *Journal of Computational Biology*, *26*(2), 143-151.

[9] Jeffrey, H. J. (1990). Chaos game representation of gene structure. *Nucleic acids research*, *18*(8), 2163-2170.

[10] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., ... & Saunders, R. D. (2000). The genome sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185-2195.

[11] Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., & Bier, E. (2001). A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster. *Genome research*, *11*(6), 1114-1125.

[12] Chafee, H. (2022). *The Therapeutic Efficacy of Psilocybin in a Preclinical Model of Depressive-and Anxiety-Like Symptomology* (Doctoral dissertation, Open Access Te Herenga Waka-Victoria University of Wellington).