Hibridación de dos métodos de Codificación del Genoma

J.L. del Río-Correa[1], J. López-García[2], y G. Durán-Meza[3]

[1,3] Universidad Autónoma Metropolitana-Iztapalapa, Departamento de Física, México

[2] UNAM - FES Acatlán, División de Matemáticas e Ingeniería, México

jlrc@xanum.uam.mx, jeanettlg@comunidad.unam.mx, gdm.pfm@gmail.com

Tel. (52) 55 5804 4617

Resumen

Recientemente se han presentado dos métodos diferentes de codificación de Series Genómicas (**SG**), ambas basadas en el Juego del Caos de Jeffrey, en los cuales se hace el proceso de codificación-decodificación al encriptar una **SG** conformada por miles ó millones de nucleótidos. En este trabajo presentamos una hibridación de ambos métodos, manejada en base binaria, que permite manejar fácilmente el problema de codificación estableciendo la relación entre ambos esquemas y detectar cuando una **SG** presenta una o dos alteraciones por sustitución causada por efectos externos.

1. Introducción

Usando un Sistema de Funciones Iteradas, Jeffrey [1] [2] en 1990 introdujo una representación matemática del ADN, conocida como la CGR (Chaos Game Representation), que permite una representación gráfica genómica (RGG) única de Secuencias Genómicas (SG). La RGG está formada por puntos dentro de un cuadrado unitario Q, siendo el número de puntos igual al número de nucleótidos que contenga la SG, con la propiedad de que cada punto representa una cadena particular de nucleótidos, tal que conociendo las coordenadas del enésimo punto $\mathbf{P}_n$, se pueden encontrar las coordenadas del punto anterior $\mathbf{P}_{n-1}$, usando la relación: $\mathbf{P}_{n-1} = 2\mathbf{P}_n - Int(2\mathbf{P}_n) = Frac\,(2\mathbf{P}_n)$; y conociendo el último punto podemos conocer toda la SG; sin embargo, las coordenadas de $\mathbf{P}_n$ son números en [0,1] con una cantidad muy grande de dígitos, que se requieren conocer exactamente, lo que hace muy difícil su manejo analítico, razón por la cual la mayoría de los trabajos se enfocan a la RGG [3] [4]. Fue hasta 2019 que C. Yin [5] resolvió el problema usando una ecuación de recurrencia diferente a la de Jeffrey, donde son enteras las coordenadas asociadas a cada subsecuencia (SbS), introduciendo una representación alternativa llamada iCGR, donde las coordenadas del último nucleótido son números enteros muy grandes, que sin embargo presentan el problema de expresarlos computacionalmente de manera exacta, lo que solo permite manejar secuencias pequeñas, por lo aunque en principio el problema está resuelto, es compleja su implementación para SG con miles ó millones de nucleótidos. El trabajo propone un esquema de codificación más simple que los de Jeffrey y de Yin, que permite encontrar la forma como se relacionan estas descripciones, y un manejo sencillo tanto de los números en [0,1] como en el plano infinito; el esquema de codificación que se utiliza en este trabajo es una hibridación de los esquemas antes citados, que se implementa utilizando una descripción binaria, que permite una descripción numérica simple, ya que al usar la base binaria es sencillo el manejo computacional tanto de números en [0,1] así como de enteros muy grandes, en forma exacta.

2. Metodología

Métodos de Codificación de Jeffrey y Yin

Los métodos CGR e iCGR utilizan ambos una descripción con números binarios. El ADN es una cadena muy grande de 4 nucleótidos: Adenina (A), Citosina (C), Guanina (G) y Timina(T); en cada organismo el número y el orden en que aparecen los nucleótidos en la cadena son distintos. La codificación en el CGR [6], se hace construyendo un cuadrado unitario Q identificando sus vértices en base binario con los 4 nucleótidos, y su centro por los puntos:

$$A = (0,0); C = (0,1); G = (1,1); T = (1,0); V_0 = (0.1, 0.1). \quad (1)$$

La SG formada por N nucleótidos se denota por: $V_1V_2V_3 \ldots V_N$ con $V_i = \{A, C, G, T\}$. Para codificar una SG, Jeffrey introduce el siguiente SFI: $\mathbf{W} = \{w_{V_1}(\mathbf{X}), w_{V_2}(\mathbf{X}), w_{V_3}(\mathbf{X}), w_{V_4}(\mathbf{X})\}$; donde $w_{V_i}(\mathbf{X}) = \frac{1}{2}(\mathbf{X} + \mathbf{V}_i)$, obteniendo la secuencia de puntos asociados con la SG, por la ecuación de iteración: $\mathbf{P}_n = w_{V_n}\{\mathbf{P}_{n-1}\}$ con $\mathbf{P}_0 = \mathbf{V}_0$; donde $\mathbf{V}_n$ denota al enésimo símbolo de la secuencia. La SG se codifica asignando un punto diferente de Q, a cada SbS que se obtiene al ir leyendo nucleótido a nucleótido de la SG, con la siguiente regla de iteración:

$$\mathbf{P}_n = \frac{1}{2}(\mathbf{P}_{n-1} + \mathbf{V}_n) \quad \text{con} \quad \mathbf{P}_0 = \mathbf{V}_0; \quad (2)$$

de forma que a la SbS: $V_1V_2V_3 \ldots V_R$ formada por los R primeros nucleótidos, se le asigna el punto: $\mathbf{P}_R = 0.V_RV_{R-1}V_{R-2} \ldots V_1V_0$. De manera que es fácil ver que el conocimiento de $\mathbf{P}_R$ permite encontrar las coordenadas de cualquier SbS con $S < R$, la que está dada por: $\mathbf{P}_S = Frac\left(2^{R-S}\mathbf{P}_R\right); \mathbf{V}_S = Int\left(2^{R-S+1}\mathbf{P}_R\right) - 2Int\left(2^{R-S}\mathbf{P}_R\right)$; esta expresión muestra que conociendo exactamente las coordenadas de la secuencia completa se pueden obtener las coordenadas de cualquier SbS. El método de Yin para codificar la SG, parte de la ecuación de iteración: $\mathbf{P}_n = \mathbf{P}_{n-1} + 2^n\mathbf{V}_n$ con $\mathbf{P}_0 = \mathbf{V}_0$; esta ecuación permite encontrar el vector de posición asociado con cada SbS, en particular el vector de posición del último nucleótido de la secuencia, y por tanto encriptar toda la SG con tres números enteros: (N, X_N, Y_N), donde el primero corresponde al número de nucleótidos que conforman la SG, y los dos siguientes corresponden a las coordenadas enteras, que pueden ser positivas ó negativas, de manera que la nube de puntos correspondientes a la SG junto con todas sus SbS ocupan todo el plano, por lo que la figura no esta acotada, de manera que se pierde la representación gráfica de las SG, que ha mostrado ser una herramienta fundamental para el análisis de la SG del ADN. Por otra parte, con el método de Yin el proceso de decodificación es complicado debido a que las coordenadas tanto de la SG como de la SbS en general pueden ser enteros positivos y negativos (Figs. 1(a) y 1(b)).

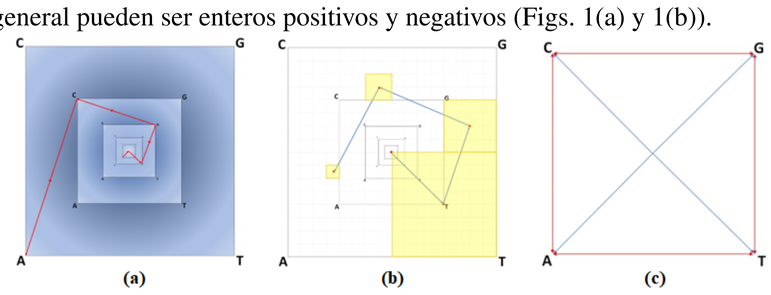

Fig. 1. (a) Codificación de la cadena TGCA con el método de Yin. **(b)** Codificación de la cadena TGCA con el método de Jeffrey. **(c)** Mutaciones cuando $T_X \neq F_X$ y $T_Y \neq F_Y$ en color azul, mutaciones con $T_X \neq F_X$ ó $T_Y \neq F_Y$ en color rojo

Lo anterior nos llevó a proponer una codificación de las SG [6] sin sacrificar la representación gráfica de Jeffrey, y adoptando el punto de vista de Yin de asociar tres enteros con toda la SG, así como con cada una de las SbS, pero ahora restringidos a enteros positivos, por lo que nuestra propuesta es una hibridación de ambos métodos de codificación, que permite un manejo simple de las SG.

Método Híbrido de Codificación

En esta sección se utiliza el método propuesto en la referencia [6], sin entrar en un análisis detallado, los cuales se encuentran en ella. Se muestra como obtener coordenadas enteras para la SG usando el método de Jeffrey, y cómo encontrar coordenadas enteras de SbS usando el método de Yin, además de cómo determinar el último nucleótido tanto de la secuencia entera, como de cualquier SbS de ella. El método de Jeffrey proporciona las coordenadas del punto dentro de Q asociado con cualquier SG dadas por (2), que pasamos a enteros dentro del primer cuadrante con el siguiente simple producto:

$$\mathbf{I}_R = \begin{pmatrix} I_{R_x} \\ I_{R_y} \end{pmatrix} = 2^{R+1}\begin{pmatrix} P_{R_x} \\ P_{R_y} \end{pmatrix} = \begin{pmatrix} V_{R_x}V_{(R-1)_x}V_{(R-2)_x}\ldots V_{2_x}V_{1_x}1 \\ V_{R_y}V_{(R-1)_y}V_{(R-2)_y}\ldots V_{2_y}V_{1_y}1 \end{pmatrix} \quad (3)$$

de forma que ahora podemos asociar los enteros $\left(R, I_{R_x}, I_{R_y}\right)$ con la SG. Es conveniente expresar (3) como una suma vectorial:

$$\mathbf{I}_R = \sum_{j=0}^{R} \mathbf{V}_j 2^j; \quad (4)$$

donde los vectores $\mathbf{V}_j$ están dados por (1). $\mathbf{I}_R$ es el vector de posición de un punto en el primer cuadrante del plano, asociado con la SG, y puede obtenerse directamente de la secuencia utilizando la ecuación de iteración, similar a la propuesta por Yin: $\mathbf{I}_n = 2^n\mathbf{V}_n + \mathbf{I}_{n-1}$; donde $\mathbf{I}_0 = 2\mathbf{V}_0 = 1$; esta expresión se obtiene directamente de (4), de donde es inmediato demostrar las siguientes relaciones:

$$\mathbf{V}_n = \begin{pmatrix} V_{nx} \\ V_{ny} \end{pmatrix} = \begin{pmatrix} Int\left(\frac{\mathbf{I}_{nx}}{2^n}\right) \\ Int\left(\frac{\mathbf{I}_{ny}}{2^n}\right) \end{pmatrix} \text{ ó en forma compacta } \mathbf{V}_n = Int\left(\frac{\mathbf{I}_n}{2^n}\right); \quad (5)$$

$$\mathbf{I}_{n-1} = \begin{pmatrix} I_{(n-1)x} \\ I_{(n-1)y} \end{pmatrix} = \begin{pmatrix} \left(\frac{\mathbf{I}_{nx}}{2^n}\right) \bmod 1 \\ \left(\frac{\mathbf{I}_{ny}}{2^n}\right) \bmod 1 \end{pmatrix} \text{ ó } \mathbf{I}_{n-1} = \left(\frac{\mathbf{I}_n}{2^n}\right) \bmod 1; \quad (6)$$

Con la notación anterior podemos encontrar al vector $\mathbf{I}_S$ de componentes enteras para cualquier subsecuencia S de la secuencia con R nucleótidos, expresando (4) en la forma: $\mathbf{I}_R = \sum_{j=0}^{R} 2^j\mathbf{V}_j = \sum_{j=0}^{S} 2^j\mathbf{V}_j + 2^{S+1}\mathbf{V}_{S+1} + \ldots + 2^R\mathbf{V}_R$; $\frac{\mathbf{I}_R}{2^{S+1}} = \frac{\mathbf{I}_S}{2^{S+1}} + \sum_{j=S+1}^{R} 2^j\mathbf{V}_j = \frac{\mathbf{I}_S}{2^{S+1}} + Int\left(\frac{\mathbf{I}_R}{2^{S+1}}\right)$; resultando: $\mathbf{I}_S = \mathbf{I}_R - 2^{S+1}Int\left(\frac{\mathbf{I}_R}{2^{S+1}}\right)$. Para encontrar el último nucleótido de la secuencia con vector entero $\mathbf{I}_S$, usamos (5), con lo que se obtiene: $\mathbf{V}_S = \frac{\mathbf{I}_S}{2^S} = 2\left(\frac{\mathbf{I}_S}{2^{S+1}}\right) = 2\left[\left(\frac{\mathbf{I}_R}{2^{S+1}}\right) \bmod 1\right]$. Este resultado junto con (1), permite conocer cuál es el nucleótido que ocupa el lugar S en la sub-SG caracterizada por $\mathbf{I}_R$. Así, se ha mostrado como haciendo una hibridación de los esquemas de Jeffrey y Yin, se obtiene una codificación con tres números enteros, sin perder la representación gráfica de Jeffrey, con la ventaja de que es más simple su implementación numérica, como se muestra en [7].

3. Resultados

Mutación por sustitución de un nucleótido en la SG, su localización e identificación

En esta sección se considera una SG en la que por alguna razón un nucleótido ha sufrido una mutación, se requiere conocer el lugar donde ha ocurrido la mutación e identificar el par de nucleótidos involucrados en ella. Cuando ocurra cualquier mutación en la SG el número de nucleótidos no cambia, pero sí los vectores asociados con cada secuencia. Denotando por **T** la SG original y por **F** la SG donde hay mutaciones. Para el caso donde haya una mutación que involucra solamente a un nucleótido, los vectores de coordenadas enteras que caracterizan ambas secuencias son: $\mathbf{T} = \sum_{j=0}^{R-1}\mathbf{V}_j2^j + \mathbf{V}_R2^R + \sum_{j=R+1}^{N}\mathbf{V}_j2^j$; $\mathbf{F} = \sum_{j=0}^{R-1}\mathbf{V}_j2^j + \mathbf{E}_R2^R + \sum_{j=R+1}^{N}\mathbf{V}_j2^j$; donde se denota por R al lugar donde ocurrió la mutación que cambió el nucleótido $\mathbf{V}_R$ por el $\mathbf{E}_R$. La diferencia entre estos vectores es:

$$\mathbf{T} - \mathbf{F} = (\mathbf{V}_R - \mathbf{E}_R)\,2^R; \quad (7)$$

como el factor numérico es positivo, de (7) se sigue: $sign\,(T_x - F_x) = sign\,(V_{Rx} - E_{Rx})\,; sign\left(T_y - F_y\right) = sign\left(V_{Ry} - E_{Ry}\right)$; esta propiedad tiene las siguientes implicaciones:

$$\begin{aligned} T_x > F_x \Rightarrow V_{Rx} > E_{Rx} \Rightarrow V_{Rx} = 1,\ E_{Rx} = 0; \\ T_x < F_x \Rightarrow V_{Rx} < E_{Rx} \Rightarrow V_{Rx} = 0,\ E_{Rx} = 1; \end{aligned} \quad (8)$$

que también se aplican análogamente a las componentes T_y, F_y. La ecuación (7) permite localizar el lugar donde ha ocurrido la mutación, y determinar el par de nucleótidos $\mathbf{V}_R$ y $\mathbf{E}_R$ involucrados en la mutación. La localización se encuentra que si hay un cambio la componente X y/o Y de $\mathbf{V}_R$ y $\mathbf{E}_R$ solamente difieren en ± 1, así se tiene:

$$|T_x - F_x| = |V_{Rx} - E_{Rx}|\,2^R = 2^R; \quad (9)$$

de manera que el lugar R donde ocurre la mutación es:

$$R = \log_2 |T_x - F_x|\,; \quad (10)$$

Utilizando las propiedades dadas en (8), se encuentran los nucleótidos involucrados en la mutación. A continuación, se enlistan los diferentes casos posibles cuando **T** y **F**, tienen diferentes dos de sus componentes:

$$\begin{aligned} &T_x - F_x > 0; T_y - F_y > 0; V_R = (1,1) = G, E_R = (0,0) = A; \\ &T_x - F_x < 0; T_y - F_y < 0; V_R = (0,0) = A, E_R = (1,1) = G; \\ &T_x - F_x > 0; T_y - F_y < 0; V_R = (1,0) = T, E_R = (0,1) = C; \\ &T_x - F_x < 0; T_y - F_y > 0; V_R = (0,1) = C, E_R = (1,0) = T; \end{aligned} \quad (11)$$

El procedimiento anterior para encontrar los nucleótidos involucrados en la mutación funciona cuando las dos componentes de los vectores **T** y **F** son diferentes, cuando esta condición no se cumple, se requiere un paso adicional, puesto que solamente una de sus componentes es diferente, v. gr. $T_X \neq F_X$, en este caso primero se determina R usando (10), a continuación, se determina V_{RY} usando (**??**), ya que conocemos T_Y, de manera que:

$$V_{RY} = E_{RY} = Int\left[2\left[\left(\frac{T_y}{2^{R+1}}\right) \bmod 1\right]\right]; \quad (12)$$

en tanto que V_{RX} y E_{RX} se determinan utilizando (11), así los nucleótidos involucrados son de la forma (V_{RX}, V_{RY}) y (E_{RX}, V_{RY}), que se determinan explícitamente utilizando (1). En este tipo de mutación se encuentran los siguientes pares de nucleótidos que solamente difieren en una de sus componentes por una unidad: **A** $\leftrightarrow$ **C**; **A** $\leftrightarrow$ **T**; **C** $\leftrightarrow$ **G**; **G** $\leftrightarrow$ **T**; que junto con las transiciones dadas en (11): **A** $\leftrightarrow$ **G**; **C** $\leftrightarrow$ **T**; proporcionan todas las posibles mutaciones, que son ilustradas en la Fig. 1(c).

4. Discusión

Mutación por sustitución de dos nucleótidos en la SG, su localización e identificación

A continuación, se analiza el caso en que han mutado dos nucleótidos, en este caso cuando se conoce que una SG caracterizada por **T** cambia en dos nucleótidos y pasa a ser **F**, se requiere determinar la localización de los lugares R y S, con $R > S$, en que han ocurrido los cambios, así como los nucleótidos involucrados. Para resolver este problema se definen las siguientes secuencias auxiliares: $A_x = Max\,(T_x, F_x)\,;\ B_x = Min\,(T_x, F_x)\,;\ D_x = A_x - B_x$; que expresadas en binario son:

	$R+1$	R			S	
A_x	X X		1	X X	A_{Sx}	X
B_x	X X		0	X X	B_{Sx}	X
$D_x^{\pm}$	0	0	1	0 0	± 1	0

(13)

donde se utilizó que $A_x > B_x$, implica que $A_{Rx} = 1$, $B_{Rx} = 0$; sin embargo, para el sitio S se pueden tener dos casos posibles: $I : A_{Sx} > B_{Sx} \Rightarrow A_{Sx} = 1, B_{Sx} = 0$ y $II : A_{Sx} < B_{Sx} \Rightarrow A_{Sx} = 0, B_{Sx} = 1$. Sin embargo, en ambos casos la localización tanto de R como de S es la misma, por lo que supondremos que se cumple el primer caso, de manera que:

$$D_x^1 = 2^R + 2^S \quad \Rightarrow \quad 2^{R+1} > D_x^1 > 2^R; \quad (14)$$

así se tiene que: $R = Int\left(\log_2 D_x^1\right) = Int\left(\log_2 |T_x - F_x|\right)$. Para encontrar S, nótese que por (14): $D_x^1 - 2^R = 2^S$; por lo que: $S = \log_2\left(|T_x - F_x| - 2^R\right)$. Para el segundo caso en lugar de (14) se tiene: $D_x^1 - 2^R = -2^S \Rightarrow S = \log_2\left(\left|D_x^1 - 2^R\right|\right)$; teniendo en cuenta que $D_x^1 = \mid T_x - F_x \mid$ obtenemos el mismo resultado que para el caso I.

Teniendo los valores R y S donde han ocurrido los cambios, se determinan los nucleótidos para la SG correspondiente al vector **T**, utilizando (12) para las componentes (V_{Rx}, V_{Ry}), para encontrar (E_{Rx}, E_{Ry}) se requiere utilizar las coordenadas enteras del vector **F**, se calculan de la siguiente manera: Si $T_x - F_x > 0 \Rightarrow V_{Rx} > E_{Rx} \Rightarrow V_{Rx} = 1$, $E_{Rx} = 0$. Si $T_x - F_x < 0 \Rightarrow V_{Rx} < E_{Rx} \Rightarrow V_{Rx} = 0$, $E_{Rx} = 1$. Si $T_x - F_x = 0 \Rightarrow V_{Rx} = E_{Rx}$. Y finalmente se hace un análisis similar para las componentes Y del vector **D**. Combinando ambos resultados se obtienen los nucleótidos involucrados en el cambio sufrido en el lugar R. Para encontrar los nucleótidos en el lugar S se hace un análisis similar, obteniéndose: Si $|T_x - F_x| - 2^R > 0 \Rightarrow V_{Sx} > E_{Sx} \Rightarrow V_{Sx} = 1$, $E_{Sx} = 0$. Si $|T_x - F_x| - 2^R < 0 \Rightarrow V_{Sx} < E_{Sx} \Rightarrow V_{Sx} = 0$, $E_{Sx} = 1$. Si $|T_x - F_x| - 2^R = 0 \Rightarrow V_{Sx} = E_{Sx}$.

5. Conclusiones

Un análisis cuidadoso de los casos de mutación estudiados en el presente trabajo nos permite hacer la generalización para casos mas complejos en donde existen q mutaciones y se requiere conocer los lugares donde ocurrieron y los nucleótidos involucrados. Lo anterior se logra por medio del proceso iterativo en el cual se encuentran todos los lugares $R_1, R_2, \ldots, R_q$ donde han ocurrido las q mutaciones, así como los nucleótidos involucrados en todas ellas. La solución del problema planteado muestra claramente la potencia del método híbrido no sólo para codificar y decodificar secuencias genómicas, sino que también permite recuperar la representación gráfica, como se muestra en otro trabajo que se presentará en esta reunión [7].

Referencias

[1] H. J. Jeffrey, *Chaos game representation of gene structure*. Nucleic Acids Res, Vol. 18, No. 8, pp. 2163–2170, 1990.

[2] H. J. Jeffrey, *Chaos Game visualization of sequences*, Comput. & Graphics. 16, pp. 25-33, 1992.

[3] 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*. Mol. Biol. Evol. 16, pp. 1391-1399, 1999.

[4] G. Durán-Meza, J. López-García, and J. L. Del Río-Correa, *The self-similarity properties and multifractal analysis of DNA sequences*. Applied Mathematics and Nonlinear Sciences, Vol. 4, No. 1, pp. 267–278, 2019.

[5] C. Yin *Encoding and Decoding DNA Sequences by Integer Chaos Game Representation*. Journal of Computational Biology Vol. 26, No. 0, pp. 1-9, 2019.

[6] J. L. Del Río-Correa, Y. A. Álvarez-Ballesteros, y G. Durán-Meza, *Codificación y Decodificación de Secuencias Genómicas*, en Memorias de la XXV Reunión Nacional Académica de Física y Matemáticas, pp. 225-231, 2020.

[7] G. Durán-Meza, J. López-García, y J. L. Del Río-Correa, *Representación gráfica del juego del caos de genomas completos*. Por presentarse en la XXVII Reunión Nacional Académica de Física y Matemáticas, 2022.