UACM

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO

NADA HUMANO ME ES AJENO

MAESTRÍA EN DINÁMICA NO LINEAL Y SISTEMAS COMPLEJOS

"SISTEMAS DE FUNCIONES ITERADAS Y VISUALIZACIÓN FRACTAL DEL DNA"

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN DINÁMICA NO LINEAL Y SISTEMAS COMPLEJOS

PRESENTA:

GUSTAVO CARREÓN VÁZQUEZ

DIRECTOR DE TESIS:

DR. PEDRO EDUARDO MIRAMONTES VIDAL

MÉXICO D. F. NOVIEMBRE DE 2007

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

Fractal geometry will make you see everything differently. There is danger in reading further. You risk the loss of your childhood vision of clouds, forests, flowers, galaxies, leaves, feathers, rocks, mountains, torrents of water, carpets, bricks, and much else besides. Never again will your interpretation of these things be quite the same.

Michael F. Barnsley

A mi Maryfer, mis padres y mis hermanos

Agradecimientos

Va de nuevo, a mis padres por su apoyo, a mi hermano Iván e Iridián por que de ellos también aprendo. Les agradezco a mis profesores de la maestría y de la facultad de ciencias por su dedicación al transmitirnos y sorprendernos con sus conocimientos, en especial a mi amigo Pedro Miramontes por creer y apoyarme en este proyecto.

Tambien le agradezco a los *fantásticos*, Jesús, Toño, Emmanuel y Vicente por que también de ellos aprendí mucho, y compartimos muchos logros académicos, por esos interminables instrumentos de certificación.

Al SES - Team, mis amigos Dora, Alejandra, Luis y en especial a Imanol Ordorika por apoyarme y darme espacio para términar este proyecto y otros más. De verdad se los agradezco.

Y el agradecimiento más grande para mi estrellita Maryfer, por comprender mi carrera, compartir mis logros, apoyarme en los momentos más difíciles y ser paciente. Gracias.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

El análisis de datos biológicos a partir de herramientas computacionales y matemáticas se ha convertido en un nuevo paradigma para el estudio y entendimiento de la dinámica intrínseca que genera esta información. En el campo de la bioinformática, especialmente al estudiar las cadenas de DNA o genomas, se ha observado que los análisis hechos con sistemas dinámicos no lineales son capaces de revelarnos visualmente particularidades, patrones y periodicidades, que con otros métodos son casi imposibles de discernir. Estos patrones a menudo son visualizaciones fractales que son bien conocidas por su estructura tan compleja pero que a la vez guarda una organización que los hace visualmente atractivos.

En 1993 Michael Barnsley propuso algunos métodos para el estudio de la geometría fractal, estos son los sistemas de funciones iteradas con probabilidades; es un conjunto de transformaciones afines con ciertas propiedades. La órbita del IFS genera un fractal determinado dependiendo de los valores de las transformaciones, este fractal es el atractor del sistema. Hay un IFS particular el cual tiene el nombre del *Juego del Caos*, en su representación básica consta de 3 transformaciones. En su generalización, el *Juego Circular del Caos*, existe una mayor diversidad de estructuras espaciales, plasmando la dinámica del sistema en un fractal dependiendo de las cualidades de los datos de entrada o de las probabilidades de elección de cada transformación; se deja una *huella digital* en el sistema.

En diversas publicaciones se ha utilizado el juego del caos con cuatro vértices para analizar secuencias de DNA, y tambien el juego circular del caos para

buscar similitudes y diferencias a partir de su medida invariante. Estos análisis han ayudado a entender correlaciones existentes en el DNA, así como, un método para la representación visual de genomas y cromosomas completos.

En el presente trabajo se propone un método usando IFS's para generar representaciones fractales de las secuencias de DNA, esto nos permitirá separar visualmente información de la secuencia a partir de su estructura fisicoquímica básica, es decir de sus 4 distintos nucleóticos, adenina, guanina, citosina y timina.

Los análisis con este método serán más confiables y precisos si se considera la totalidad del genoma hablando de bacterias y arqueobacterias y de cromosomas completos para eucariontes. La comparación entre distintas regiones dentro de una misma secuencia, nos permitirá obtener información de la naturaleza de la molécula de DNA, y permitir resaltar diferencias y similitudes; de la misma forma, se podrá comparar entre organismos pertenecientes al mismo grupo, y resaltar diferencias de grupos distintos.

Las herramientas y modelos matemáticos se describirán a lo largo del trabajo, en el siguiente capitulado:

En el capítulo 2 se encontrará una descripción detallada de la moléula de DNA, desde el punto de vista de la composición fisicoquímica y se abordará un poco la complejidad de las secuencias de DNA desde el punto de vista de la Teoría de la Información.

En el Capítulo 3 se darán las bases matemáticas para comprender el funcionamiento de lo Sistemas de Funciones Iteradas. Se dará una breve introducción a los espacios métricos, y se mostrarán dos algoritmos para calcular los atractores de los IFS, el *Algoritmo Determinístico*, y el *Algoritmo de Iteraciones Aleatorias*, fundado en la teoría de la ergodicidad y la teoría de la medida.

En el Capítulo 4 se aboradará putualmente un IFS particular el llamado *juego del caos* y su extensión el *juego circular del caos.*

En el Capítulo 5 se explicará el proceso que se realizó para el análisis bioinformático del DNA, se presentan los resultados que constan de visualizaciones fractales del DNA y algunas comparaciones entre las mismas imagenes fractales.

En el Capítulo 6 se hace una pequeña reflexión y una discusión del alcance y limitaciones del análisis hecho.

Capítulo 2

Ácido Desoxirribonucléico, DNA

2.1. La estructura química del DNA

Todos los seres vivos codifican la totalidad de su información genética en el ácido desoxirribonucléico. La molécula de DNA es un polímero que se forma por enlace covalente de miles de desoxinucleótidos. El desoxinucleótido esta formado por una parte constante que es una molécula de fosfato y desoxirribosa y una parte variable que corresponde a una base nitrogenada. El DNA contiene 4 bases nitrogenadas: adenina (A) y guanina (G), que derivan de la purina, y citosina (C) y timina (T) que derivan de la pirimidina.

Los desoxinucleótidos se unen para formar un polímero lineal llamado polidesoxinucleótido, el cual es una hebra donde se alternan moléculas de desoxirribosa y ácido fosfórico, mientras que las bases nitrogenadas se unen perpendicularmente a ésta (ver Figura 2.1).

En la estructura secundaria del DNA se forma una hélice de giro a la derecha conformada por dos hebras de polidesoxinucleótidos antiparalelos, en el exterior de la hélice se alternan moléculas de desoxirribosa y ácido fosfórico mientras que las bases nitrogenada se proyectan hacia el interior, las dos hebras se unen por puentes de hidrógeno que se establecen de manera específica o complementaria entre las bases de las dos hebras. Una molécula de adenina se une, por dos puentes de hidrógeno, a una timina, y una molécula de guanina se une por tres puentes de hidrógeno a una citosina. Debido a este principio de complementariedad la información contenida en el DNA se puede deducir de una sola rama de la doble hélice (ver Figura 2.2).

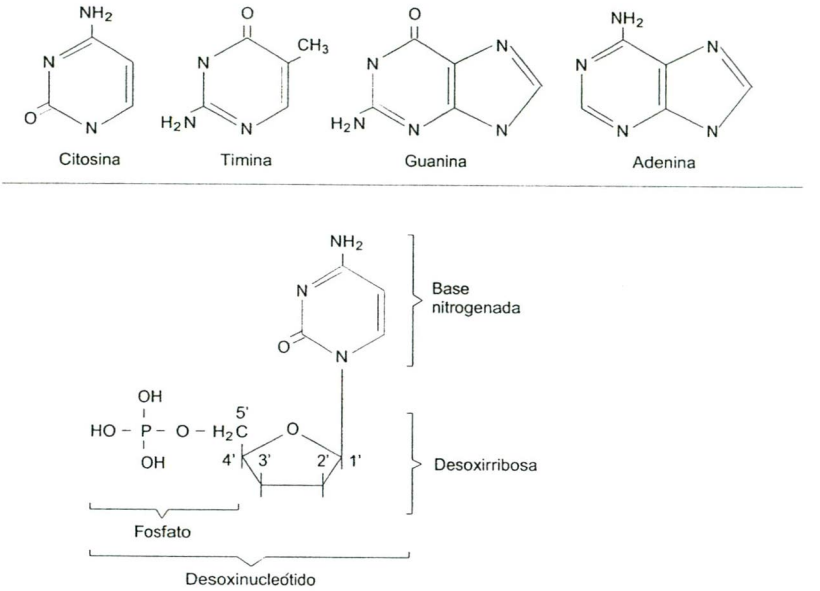


Figura 2.1: En la estructura química del DNA la molécula fundamental es el desoxinucleótido, el cual esta compuesto por tres elementos, un azúcar de 5 átomos (numerados en la imagen), un ácido fosfórico y una base nitrogenada.

La hélice tiene alrededor de 10 pares de bases por vuelta. En la estructura helicoidal, los desoxinucleótidos se organizan formando un surco mayor y un surco menor. Dependiendo del ángulo en que se observa, las 10 bases quedan hacia el surco mayor o hacia el surco menor. Existen proteínas específicas que interactúan con la secuencia de DNA a través del surco mayor ya que se encuentran más expuestas. Estas interacciones son muy importantes para la expresión genética (ver Figura 2.3).

En segmentos de la hebra donde la cantidad de bases nitrogenadas G y C dominen sobre las otras bases, la hebra será en promedio más *rígida* que donde dominan A y T, en el sentido de la energía necesaria para separar las dos ramas de la molécula, en el caso contrario sera *suave*. Por otra parte, la diferencia de tamaño entre las purinas y pirimidinas hace que la hélice no sea geométricamente uniforme, si de un solo lado se alternan purinas y

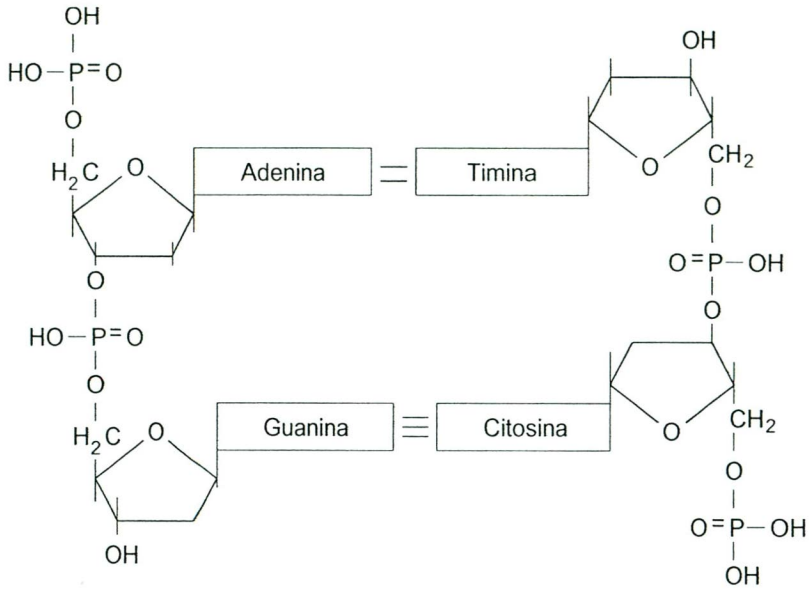


Figura 2.2: La molécula del DNA está formada por dos hebras antiparalelas unidas por puentes de hidrógeno entre bases complementarias, A-T y C-G.

pirimidinas, la doble hélice se encontrará dislocada, de manera que podemos afirmar que en este caso la molécula es en promedio *rugosa*. Si localmente se encuentran sólo purinas la cadena sera *lisa* [18].

2.2. Información genética en el DNA

En el DNA se encuentra codificada la información genética del organismo, su función principal es codificar las instrucciones esenciales para que, con los procesos biológicos que se desarrollan dentro de la célula, se pueda crear una nuevo individuo ya sea mezclándose con otra cadena de DNA o formando una copia idéntica del organismo. La descodificación de esta información se realiza en segmentos específicos del DNA, llamados genes, por medio de los proceso de replicación, transcripción y traducción (para abordar más sobre el tema vea [13]).

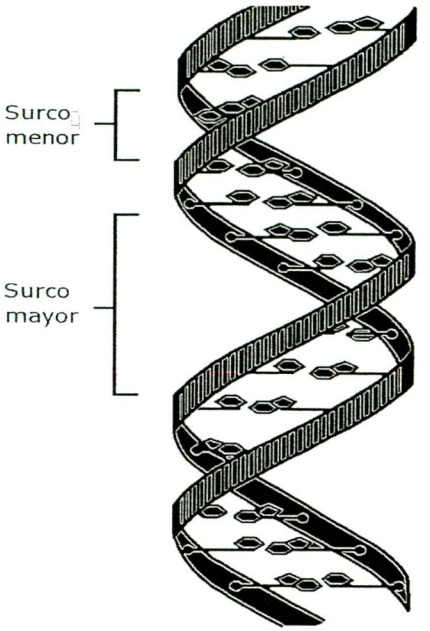


Figura 2.3: Se muestra la estructura helicoidal y los surcos mayor y menor que se forman.

En los procesos de replicación y transcripción puede suceder que se cambie una base nitrogenada por otra, a este fenómeno se le llama mutación, existen varios tipos de mutaciones, el caso anterior es el mas común. Esta mutaciones pueden ser corregidas por enzimas correctoras. La mayoría de las veces las mutaciones pueden ser desastrozas, pero son consideradas una fuente de variabilidad en los organismos.

Los genes se encuentran constituidos por regiones codificantes llamadas *exones* interrumpidas por regiones no codificantes llamadas *intrones* que son eliminadas durante el procesamiento del RNA.

Un gen contiene la información para la síntesis de una molécula de RNA que es complementaria a una de las dos hebras del DNA, el RNA sufre el proceso de traducción en los ribosomas con el cual se genera un proteína, dependiendo

de la secuencia del RNA se generará una cadena de polipéptidos que dara lugar a una proteína por medio del código genético. Al conjunto de todo el material genético de una celula se le llama *genoma*

2.3. Tamaños de los genomas

Las células se clasifican como procariontes (bacterias y arqueas) y eucariontes, las primeras no tienen un núcleo cuya pared separe el material genético del resto del protoplasma. Las celulas eucariontes, por otra parte, poseen un núcleo bien definido en el cual reside el DNA. La cantidad total de DNA presente en las células, o valor C, varía entre los diferentes organismos y se mide en pares de bases (pb). En los procariontes (bacterias y arqueas), el tamaño de los genomas varía de 6×10^5 en algunos parasitos intracelulares obligados (micoplasma), a más de 1×10^7 pb en varias cianobacterias. En general los genomas para procariontes tienen un tamano menor de 5×10^6 pb (datos tomados de [13]).

En los eucariontes los valores C son mayores que en los procariontes, aunque hay excepciones. La variación de los valores C en eucariontes es mucho mayor que en los procariontes, se encuentra en un rango de 8×10^6 a 7×10^{11} pb. [13]. La cantidad de pb en los organismos no tiene una relación directa con su complejidad y funcionamiento biológico o con el número de genes codificados en el genoma. Por otro lado organismos que son similarmente complejos presentan una gran variación en el valor C. La relación inexistente entre la complejidad del organismo y el valor C se le conoce como la *paradoja del valor C.* Aún midiendo la cantidad de genes con la complejidad del organismo no se nota una relación clara.

Por un lado es interesante conocer las longitudes de los genomas, pero por otro lado, la información que aporta acerca de la complejidad del organismo es baja, de igual forma cuando se secuencia las bases de un genoma tenemos materia en bruto que a través de algunas técnicas podemos encontrar información relevante del organismo.

2.4. Complejidad en genomas

La complejidad de los genomas se debe entender desde el punto de vista de cualquier interacción que exista en la molécula, considerando las redes genéticas y los procesos biológicos que se desarrollan, y no sólo sobre la cantidad de material genético que contenga el genoma. En la actualidad se pueden conseguir las secuencias completas de muchos organismos en los bancos de datos de genes, esto ha propiciado que las secuencias de DNA en su formato básico de 4 letras sean la materia prima para los análisis bioinformáticos. Se utilizan herramientas de las matemáticas, de las ciencias de la computación, de la estadística y de la teoría de la información para el estudio de la molécula, ayudando a encontrar similitudes y diferencias para la comprensión de la distribución de la información en la molécula y para comparar resultados entre un conjunto de secuencias distintas.

La proporción relativa de cada una de las bases A, G, C, y T, así como el orden o secuencia en que se encuentran, hace la variabilidad de los organismos. Usando la teoría de la información para discernir características y propiedades de la molécula, se han hecho diversos análisis donde se establece que la información que transporta el DNA pertenece a una zona característica, el nivel de información o entropía que contienen las secuencias, no estan en la zona de periodicidades o homogeneidades, ni en las zonas donde no hay patrones discernible como en el azar, se encuentran en un termino medio donde la cantidad de información se maximiza (ver Figura 2.4).

2.5. Análisis de secuencias de DNA

Con el nuevo enfoque para el estudio de las secuencias de DNA, se han encontrado comportamientos que con otras herramientas hubiera sido complicado o hasta imposible. Se pueden citar numerosos ejemplos de los resultados que se han obtenido sólo usando la sucesión de letras A, G, C y T que componen al genoma; la obtención de patrones estadísticos usando la función de información mutua en zonas que codifican y no codifican es ejemplo de ello [26].

Por otra lado también se han usado sistemas dinámicos discretos para el estudio de las secuencias de DNA y que han resultado de gran ayuda ya que

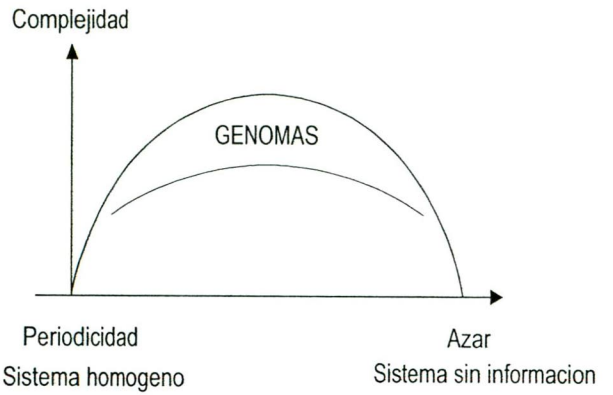


Figura 2.4: En la gráfica se representa esquematicamente la zona donde por sus propiedaes intrínsecas cae la complejidad de los genomas.

resaltan diferencias y similitudes entre conjuntos de secuencias y la secuencia particular de algún organismo. Como ejemplo tenemos el llamado juego del caos que es una categoría de los sistema de funciones iteradas (mapeos de contracción) en el cual se alimenta el sistema con cadenas de DNA, generando representaciones visuales compuestas con fractales y cuya medida invariante del fractal permite la diferenciación de características [5].

En el presente trabajo se utilizarán los sistemas de funciones iteradas para representar las diferencias fisicoquímicas que hay en la estructura básica del DNA, es decir solo con sus bases nitrogenadas, y crear representaciones visuales que nos permitan distiguir diferencias entre los dominios principales, eucarionte, bacterias y arqueobacterias. Para entender las bases matemáticas que utiliza este análisis se empezará por estudiar la teoría que hay atrás de los sistemas de funciones iteradas.

Capítulo 3

Sistemas de funciones iteradas

En algunos sistemas dinámicos discretos el comportamiento del sistema depende de los datos con los que se alimenta, los Sistemas de Funciones Iteradas pertenecen a esta clase. Si los datos con los que alimentamos el IFS son secuencias de DNA obtenemos una *firma* o *huella digital* de cada conjunto de datos. Lo interesante en este caso es que al evolucionar el IFS converge hacia el atractor del sistema, un conjunto de puntos que tienen características fractales y que dependiendo de los datos de entrada se determiná su medida invariante. La visualización fractal de las cadenas de DNA resaltará propiedades químicas de la molécula y caracerísticas locales sobre la secuencia de nucleótidos.

Para entender el comportamiento de los IFS se empezará por definir algunos conceptos tales como el de un espacio métrico y el concepto geométrico de una tranformación afín. Para profundizar en el estudio de los IFS y en la geometría fractal se puede consultar [2, 9, 16, 19].

3.1. Espacio métrico y transformaciones

Los sistemas de funciones iteradas se desarrollan en un entorno que para los propósitos de este trabajo, será el *plano euclidiano* conocido también como $\mathbf{R}^2$; se definirá una *métrica o una medida de distancia* en este plano; puede ser, por ejemplo, la distancia euclidiana o cualquier otra que cumpla con las propiedades que se definirán. Así, con una medida de distancia y un entorno, se tiene un *espacio métrico.*

Denotemos al espacio métrico sobre el plano euclidiano junto con una función de distancia como $(\mathbf{R^2}, d)$ donde la función $d : \mathbf{R}^2 \times \mathbf{R}^2 \rightarrow \mathbf{R}$ mide la distancia entre pares de puntos x, y pertenecientes al plano euclidiano. La función d debe obedecer los siguientes axiomas.

1. $d(x, y) = d(y, x) \quad \forall\, x, y \in \mathbf{R}^2$
2. $0 < d(x, y) < \infty \quad \forall\, x, y \in \mathbf{R}^2, x \neq y$
3. $d(x, x) = 0 \quad \forall\, x \in \mathbf{R}^2$
4. $d(x, y) \leq d(x, z) + d(z, y) \quad \forall\, x, y, z \in \mathbf{R}^2$

Entonces decimos que la función d es una métrica sobre el espacio $\mathbf{R}^2$.

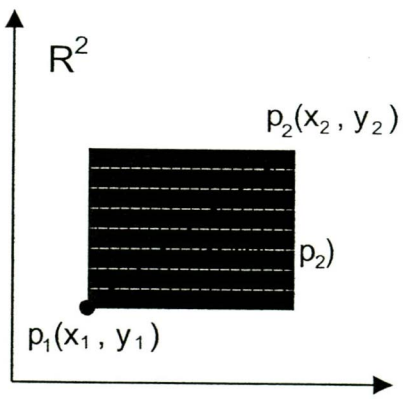


Figura 3.1: La distancia euclidiana es una métrica para el plano. Se ilustra la distancia entre los puntos p_1 y p_2.

En la Figura 3.1 se muestra el espacio métrico

$$(\mathbf{R}^2, d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}\ )$$

en donde d es la distancia euclidiana.

Definición 1 a) *Una* transformación *en* $\mathbf{R}^2$ *es una función* $f : \mathbf{R}^2 \to \mathbf{R}^2$. b) *Si* S *es un subconjunto de* $\mathbf{R}^2$, *entonces,* $f(S) = \{f(x) : x \in S\}$ *es la* imagen *de* S *bajo* f. c) *La función* f *es* uno-a-uno $f(x) = f(y)$ *implica* $x = y$. d) *La función* f *es* sobre *si* $f(\mathbf{R}^2) = \mathbf{R}^2$. e) *La función* f *es llamada* invertible *si es uno-a-uno y sobre; en este caso es posible definir una transformación* $f^{-1} : \mathbf{R}^2 \to \mathbf{R}^2$, *llamada* inversa *de* f, *tal que* $f^{-1}(y) = x$ *donde* x *es el único punto tal que* $f(x) = y$.

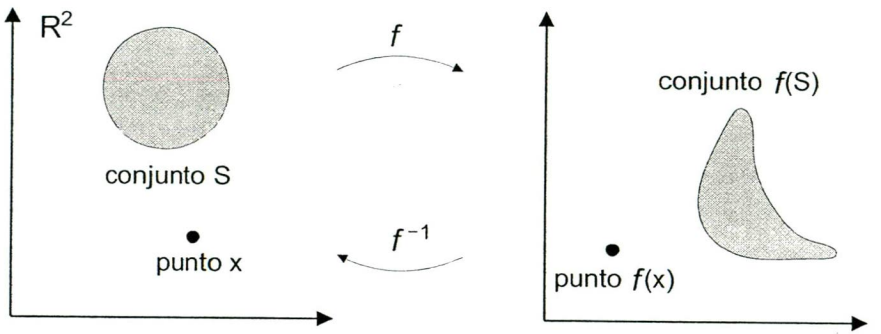


Figura 3.2: Transformaciones f y f^{-1} en el plano euclidiano. Se ilustra cómo se modifica el conjunto S al aplicarle una transformación f; así mismo, si aplicamos la transformación inversa al conjunto $f(S)$ obtenemos el conjunto original S; también se ejemplifica con un solo punto.

Ahora se definirán un par de conceptos más que son las iteraciones *delantera* y *trasera* de las transformaciones; esto nos servirá más adelante para explicar los sistemas de funciones iteradas.

Definición 2 *Sea* $f : \mathbf{R}^2 \to \mathbf{R}^2$ *una transformación sobre* $\mathbf{R}^2$. a) *La* iteración delantera *de* f *son las transformaciones* $f^{\circ n} : \mathbf{R}^2 \to \mathbf{R}^2$, *para* $n = 0, 1, 2, \ldots$. b) *Si* f *es invertible, entonces la* iteración trasera *de* f *son las transformaciones* $f^{\circ(-m)} : \mathbf{R}^2 \to \mathbf{R}^2$ *para* $m = 1, 2, 3...$ *(note que no se incluye el cero).*

Cuando se hace la *composición* de estas transformaciones quiere decir que se aplican a un conjunto de puntos en el plano (o solo un punto) tantas veces

como se indique, si se tiene n igual a cero, $f^{\circ 0}$ es la identidad y $f^{\circ 0}(S) = S$, se mantiene el mismo conjunto, es decir, la transformación no se aplica, si se tiene n igual a 1, la transformación se aplica una vez al conjunto de puntos, y así sucesivamente. Al aplicar la transformación inversa se regresa iteración tras iteración al conjunto original.

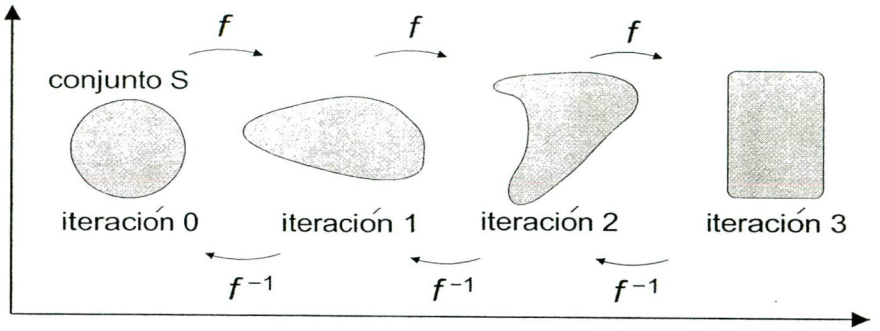


Figura 3.3: Composición de transformaciones.

En la Figura 3.3 se muestran tres iteraciones para un conjunto de puntos arbitrario, cuando se aplican todas las transformaciones delanteras se llega al conjunto marcado como tercera iteración $f^{\circ 3}(S)$; bajo cada iteración las imágenes $f(S)$ son distintas del conjunto original S sobre el cual se aplican. Ahora, si se aplican las iteraciones traseras, se regresa al conjunto inicial S.

En la siguiente subsección se estudiará un tipo de transformaciones con una estructura particular que denominaremos *transformaciones afines.*

3.2. Transformaciones afines

Las *transformaciones afines o lineales*, son la materia prima de los sistemas de funciones iteradas, construidas adecuadamente y bajo iteraciones delanteras se pueden formar conjuntos de gran belleza, con geometrías tan complicadas como las que existen en la naturaleza como flores, nubes, montañas o bosques. Para los propósitos de este trabajo, las propiedades de las secuencias de DNA formaran la estructuras de las trasformaciones afines, y por lo tanto generarán sus propios patrones.

Los cuatro tipos de transformaciones afines básicas son: traslación, escalamiento, reflección y rotación; y son transformaciones lineales. Son importantes por su simplicidad matemática, normalmente se expresan mediante la multiplicación de una matriz por un vector, más otro vector. Otro rasgo importante es, como vimos anteriormente, que podemos componer varias transformaciones; por ejemplo, una rotación con una traslación, para formar una sola transformación que hace exactamente lo mismo que las transformaciones originales pero en un solo paso.

Formalmente, una transformación afín es una función $w : \mathbf{R}^2 \rightarrow \mathbf{R}^2$ de la forma

$$w(x, y) = (ax + by + e, cx + dy + f), \tag{3.1}$$

donde a, b, c, d, e, y f son números reales.

Con la notación equivalente de matrices y vectores, tenemos lo siguiente:

$$w(\mathbf{x}) = w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} = \mathbf{A}\mathbf{x} + \mathbf{T}.$$

Aquí $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ es una matriz de 2×2 y $\mathbf{T}$ es un vector columna $\begin{pmatrix} e \\ f \end{pmatrix}$.

El vector columna con las componentes x y y representa un punto en el plano euclidiano. La estructura principal es la matriz, dependiendo de los valores de a, b, c y d los puntos cambiarán su distribución en el plano, veamos algunos de los comportamientos respecto a estos valores.

La *traslación* se aplica cuando se desea mover cada punto a una nueva posición que difiere de la posición original en un vector constante, es decir, sumándole a cada vector original el vector de los incrementos. En la Figura 3.4 se representa un cuadrado en el plano que se traslada a una nueva posición. Matemáticamente se describe de la siguiente manera

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} = \begin{pmatrix} x + e \\ y + f \end{pmatrix}$$

en donde e y f representan los incrementos sobre las componentes x y y.

El *escalamiento* de un conjunto de puntos se refiere a la contracción o expansión del conjunto. Se requiere de un factor de escalamiento que denotaremos como s y actúa sobre las componentes x y y de la siguiente manera

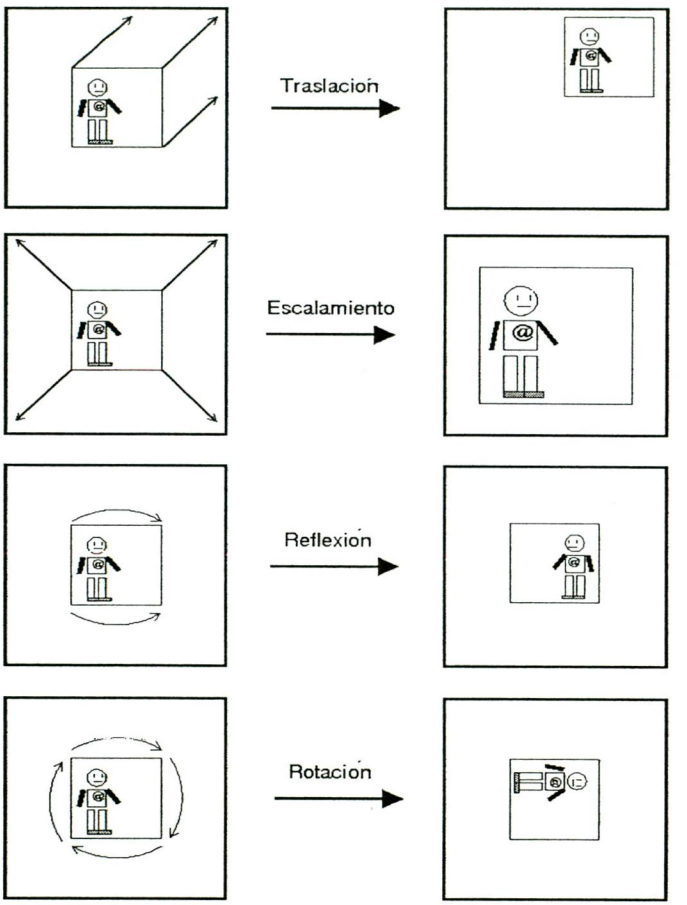


Figura 3.4: Transformaciones básicas.

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} s & 0 \\ 0 & s \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} sx \\ sy \end{pmatrix}.$$

Si el factor s es mayor que 1 entonces el conjunto se expande. Si s es menor que 1 entonces el conjunto se contrae. Ahora, si deseamos escalar un conjunto de tal forma que la escala horizontal y la vertical difieran, se debe escoger valores distintos para las componentes x y y

$$w\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}s_h & 0\\0 & s_v\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}s_h x\\s_v y\end{pmatrix}.$$

El factor s_h modifica la escala horizontal del conjunto mientras que s_v modifica la escala vertical. En la Figura 3.4 se representa el escalamiento de un cuadrado en donde $s_h = s_v$ por lo que la escala horizontal y vertical permanecen en la misma proporción. Cabe señalar que en los sistemas de funciones iteradas que nos interesan, el factor de escalamiento s deberá ser menor que 1.

Para la operación de *reflexión* se deben cambiar los signos de los valores de la diagonal de la matriz; si se quiere hacer una reflexión tomando como referencia el eje y, la primera entrada de la matriz debe ser negativa

$$w\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}-1 & 0\\0 & 1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}-x\\y\end{pmatrix}.$$

En la Figura 3.4 las líneas indican el movimiento de los puntos y la flecha representa la nueva posición del punto original. Si se requiere que la reflexión sea sobre el eje x, se debe hacer negativa la componente y. Por otro lado, si se requiere reflejar simultáneamente respecto al eje x y el eje y, se tiene lo siguiente

$$w\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}-1 & 0\\0 & -1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}-x\\-y\end{pmatrix}.$$

La *rotación* de un conjunto de puntos consiste en girar el conjunto alrededor de un punto. Por ejemplo, la expresión utilizada para una rotación de 90 grados, es la siguiente

$$w\begin{pmatrix}x\\y\end{pmatrix} = \begin{pmatrix}0 & 1\\-1 & 0\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} + \begin{pmatrix}0\\0\end{pmatrix} = \begin{pmatrix}y\\-x\end{pmatrix}.$$

En la Figura 3.4 se representa la rotación de la imagen en pasos de 90 grados.

La traslación, reflexión, escalamiento y rotación son transformaciones básicas que se pueden combinar para manipular el conjunto de puntos de una manera específica. Existen otro tipo de transformaciones afines que nos permite rotar con mayor libertad y que incluyen las transformaciones básicas. A estas transformaciones se les llama *transformaciones de semejanza* y se expresan de la siguiente manera

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} rcos\theta & -rsen\theta \\ rsen\theta & rcos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix},$$

$$w\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} rcos\theta & rsen\theta \\ rsen\theta & -rcos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix},$$

en donde θ es llamado el ángulo de rotación y toma valores en $0 \leq \theta < 2\pi$ y r es llamado el factor de escalamiento. Estas son una forma generalizada de las transformaciones básicas.

3.3. Propiedades de espacios métricos y subconjuntos

Describiremos ciertas propiedades topológicas que nos serán de utilidad para entender el comportamiento de los conjuntos generados por los sistemas de funciones iteradas que actúan sobre el espacio métrico $\mathbf{R}^2$, estas propiedades se pueden generalizar a n dimensiones pero nos enfocaremos en el plano euclidiano.

Los IFS's trabajan sobre un espacio que cumple con ciertas propiedades los cuales se conocen como *espacios métricos completos*, anteriormente se definió a $\mathbf{R}^2$ como un espacio métrico, ahora se explicará lo que es una sucesión de Cauchy para posteriormente definir a $\mathbf{R}^2$ como un espacio métrico completo.

Para entender las propiedades topológicas debemos enunciar como primer lugar lo que es una sucesión de puntos. Una sucesión es un conjunto de elementos que denotaremos de la forma $\{x_n\}_{n=1}^{\infty}$ los cuales son puntos en $\mathbf{R}^2$. Su ordenamiento esta dado por el índice que se denota como n y tiene la propiedad de preservar el orden.

Definición 3 *Una sucesión* $\{x_n\}_{n=1}^{\infty}$ *de puntos en el espacio métrico* $(\mathbf{R}^2, d)$ *es llamada una* sucesión de Cauchy *si para cualquier número dado* $\epsilon > 0$, *hay un entero* $N > 0$ *tal que*

$$d(x_n, x_m) < \epsilon \qquad \textit{para toda } n, m > N.$$

Esto quiere decir que, dada cualquier distancia positiva ϵ, es posible encontrar un índice N (que depende de ϵ) tal que para cualquier par de índices n y m, que superan a N, la distancia entre los puntos correspondientes de la sucesión es menor que ϵ.

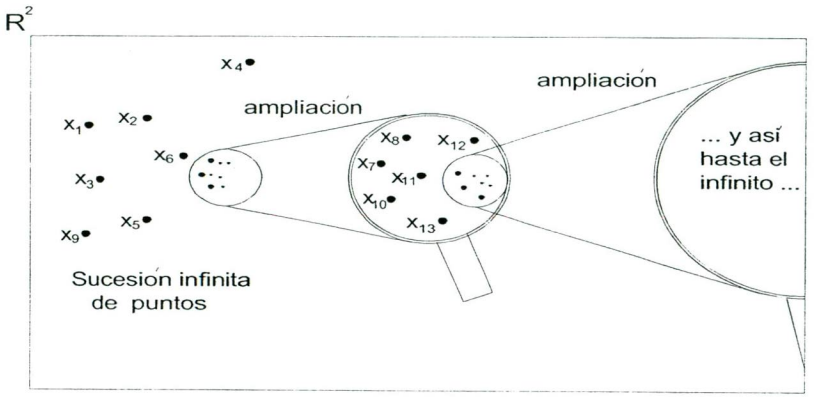


Figura 3.5: Imagen representando sucesivas ampliaciones sobre una sucesión de Cauchy.

En la Figura 3.5 se observa la representación de una sucesión de puntos en $\mathbf{R}^2$, conforme la serie va avanzando los puntos se juntan cada vez más (se representa haciendo ampliaciones sobre subconjuntos de puntos), pero esto no quiere decir que converjan a un punto específico. Para saber si converge a un punto o no introduciremos la siguiente definición.

Definición 4 *Una sucesión* $\{x_n\}_{n=1}^{\infty}$ *de puntos en el espacio métrico* $(\mathbf{R}^2, d)$ *se dice que converge al punto* $\mathbf{x} \in \mathbf{R}^2$, *si para cualquier número dado* $\epsilon > 0$, *hay un entero* $N > 0$ *tal que*

$$d(x_n, \mathbf{x}) < \epsilon \qquad \text{para toda } n > N.$$

El punto $\mathbf{x} \in \mathbf{R}^2$ *es donde converge la sucesión y es llamado el* límite de la sucesión, lo denotaremos como

$$\lim_{n \to \infty} x_n = \mathbf{x}$$

Gráficamente lo podemos ver como un círculo de radio ϵ que satisface que dentro de él estarán todos los puntos de la sucesión a partir del índice N, donde N típicamente crece conforme ϵ es cada vez más pequeña.

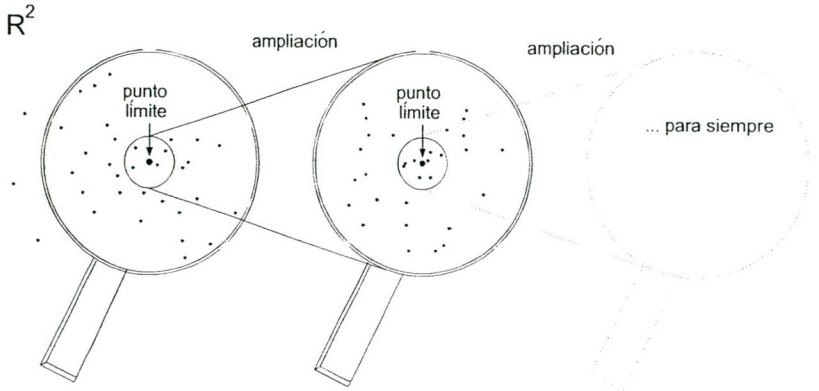


Figura 3.6: Ampliaciones alrededor del punto fijo.

Como se observa en la Figura 3.6 cada vez que hacemos una ampliación alrededor del punto fijo, son visibles más y más puntos, estas lupas son una buena analogía para los círculos mencionados anteriormente, ya que las lupas pueden desempeñar esta función, cuando hacemos un zoom es como si redujéramos el radio ϵ del círculo, y los puntos que están dentro de la lente amplificadora son los mayores a un índice N. En el infinito la serie convergerá al punto señalado como *punto límite.*

La relación entre una sucesión de Cauchy y una relación convergente es ésta: en cualquier espacio métrico, toda sucesión convergente es una sucesión de Cauchy pero sólo en ciertos espacios métricos es posible asegurar la afirmación inversa, que toda sucesión de Cauchy sea convergente. De hecho, cuando en el espacio métrico de que se trate se cumple que las sucesiones de Cauchy convergen, se dice que el espacio es *completo.*

En los cursos de análisis matemático se prueba que los espacios $(\mathbf{R}^n, d)$ son completos y es posible establecer, en ellos, las siguientes definiciones.

Definición 5 *Sea S un subconjunto de puntos del espacio métrico $(\mathbf{R}^2, d)$. Un punto $\mathbf{x} \in \mathbf{R}^2$ es llamado un* punto límite de S si hay una sucesión de puntos $\{x_n\}_{n=1}^{\infty}$ de $S \setminus \{\mathbf{x}\}$ tal que $\lim_{n\to\infty} x_n = \mathbf{x}$

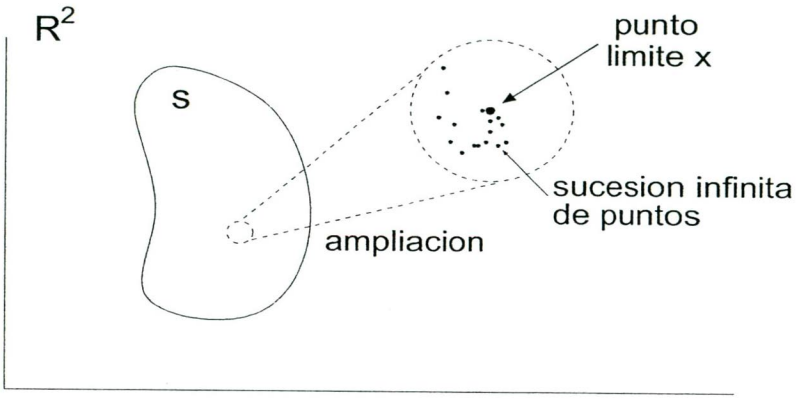


Figura 3.7: Punto límite del conjunto S.

En la Figura 3.7 se muestra el punto límite $\mathbf{x}$ que no pertenece a la sucesión de puntos. La sucesión converge al punto pero no lo incluye. Un punto límite $\mathbf{x}$ puede estar contenido en el conjunto S o no, y la siguiente definición explica esto.

Definición 6 *Sea S contenido en $\mathbf{R}^2$ un subconjunto de puntos del espacio métrico $(\mathbf{R}^2, d)$. La* cerradura (closure) de S, denotada como $\bar{S}$, se define como $\bar{S} = S \cup \{Puntos \ \ Límite \ de \ \ S\}$.

Definición 7 *S es* cerrado *si contiene todos sus puntos límite, esto es si $S = \bar{S}$*

Cuando S es un conjunto cerrado, sus puntos límite $\mathbf{x}$ están siempre dentro del conjunto, nunca en el exterior de S. Hay regiones en las que no hay puntos límite, como se muestra en la Figura 3.8, las ampliaciones nos revelan donde hay puntos límite, en la ampliación de la izquierda no hay puntos límite, mientras que en la otra tenemos dos. Si para cada punto $\mathbf{x}$ de S pudiéramos encontrar una sucesión de puntos de S que convergiera a dicho punto, tendríamos un conjunto de puntos límite con la misma cardinalidad que el conjunto S. La siguiente definición enuncia esta propiedad.

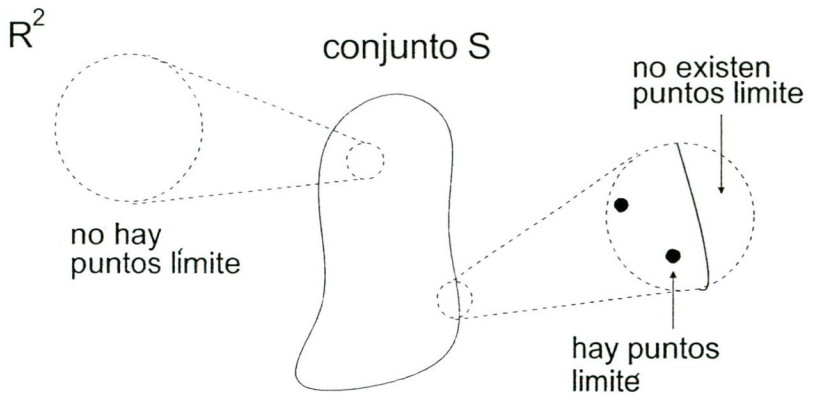


Figura 3.8: S es cerrado si sus puntos límite estan dentro del conjunto.

Definición 8 *El conjunto S es* perfecto *si tiene el mismo número de elementos que el conjunto de todos sus puntos límite.*

Se concluirá esta sección con la definición de conjunto compacto, este concepto es de gran importancia ya que los sistemas de funciones iteradas generan conjuntos compactos que al iterarlos generan geometrías fractales.

Definición 9 *Sea S contenido en $\mathbf{R}^2$ un subconjunto del espacio métrico $(\mathbf{R}^2, d)$. S es* compacto *si para toda sucesión infinita de puntos $\{x_n\}_{n=1}^{\infty}$ en S existe una subsucesión que tiende al límite en S.*

Podemos imaginar; por ejemplo, una sucesión infinita de puntos que esté conmutando entre dos puntos p_1 y p_2 en S, si tomamos una subsucesión con las posiciónes pares tendremos un conjunto que consta sólo de puntos p_2, y dicha subsucesión converge al punto p_2, análogamente si tomamos las posiciones impares. De esta forma se construyen las subsucesiónes para que tengan su límite en S. Otra forma de ver al conjunto compacto es que sea cerrado y acotado. Para una discusión profunda a cerca de estas definiciones véase [7], [3] y [20].

3.4. Transformaciones de contracción

Esta propiedad de las transformaciones es la clave para generar elementos con geometrías fractales, se empezará por definir el concepto de punto fijo.

Definición 10 *Sea* $f : \mathbf{R}^2 \to \mathbf{R}^2$ *una transformación en el plano euclidiano. Un punto* x_f *que pertenece al plano tal que al aplicar la transformación permanece en el mismo lugar,* $f(x_f) = x_f$*, es llamado* punto fijo *de la transformación.*

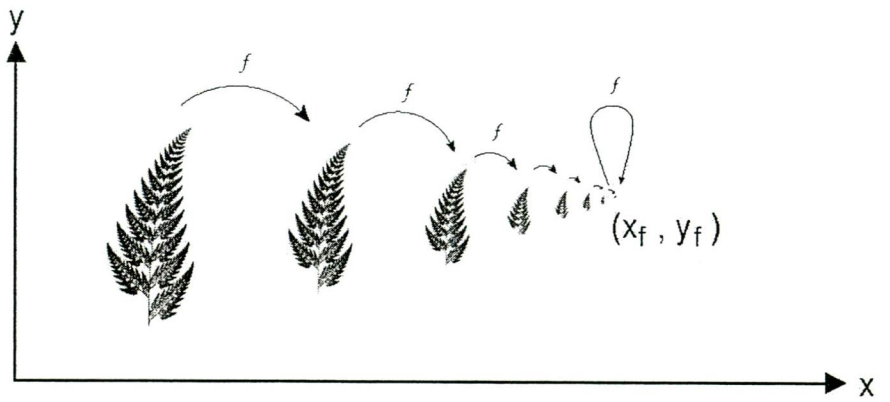


Figura 3.9: Ejemplo de punto fijo. Básicamente se le aplica una transformación de escalamiento.

En la figura 3.9 se ilustra cómo al conjunto inicial, que representa un helecho, se le aplica la transformación f y se va reduciendo iteración tras iteración hasta que el helecho se ha transformado en un pequeño conjunto, al tender estas iteraciones al inifinto el conjunto es ahora un punto que no cambia y es entonces cuando surge el punto fijo de la transformación f.

Se definirá lo que es una *transformación de contracción.*

Definición 11 *Una función* $f : \mathbf{R}^2 \to \mathbf{R}^2$ *sobre el espacio métrico* $(\mathbf{R}^2, d)$ *es llamada* transformación de contracción *si hay una constante entre* $0 \leq s < 1$ *tal que se cumpla la siguiente desigualdad para todo* x *y* y *en el plano:*

$$d(f(x), f(y)) \leq s \cdot d(x, y)$$

A la constante s se le llama factor de contracción para la transformación f.

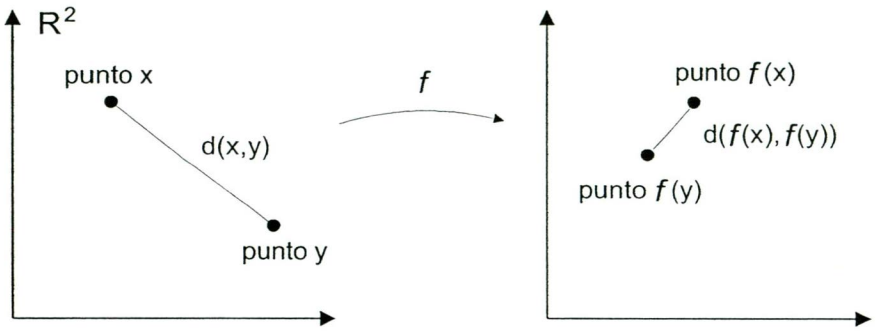


Figura 3.10: La transformación tiene la propiedad de contracción.

En la Figura 3.10 se aplica la transformación f a dos puntos cualesquiera en el plano, los nuevos puntos transformados tienen una distancia menor que los originales; es decir, están más cerca: se puede saber que tanto se acercaron con el factor de contracción; por ejemplo, si s fuera 0.5 quiere decir que los nuevos puntos transformados distan la mitad de su distancia original. Si s toma el valor de cero los puntos transformados quedan en una sola posición, es decir quedan encimados, por lo tanto su distancia es cero, el caso contrario es cuando s esta muy cerca del valor uno, la distancia entre los puntos transformados es casi igual que la distancia de los puntos originales.

Ahora se enunciará el *teorema del mapeo de contracción* el cual nos garantiza que al aplicar una transformación de contracción repetidamente a algún conjunto S que pertenece a $\mathbf{R}^2$, este converge a algún punto fijo.

Teorema 1 *El teorema del mapeo de contracción.*
Sea $f : \mathbf{R}^2 \to \mathbf{R}^2$ un transformación de contracción sobre el espacio métrico completo $(\mathbf{R}^2, d)$. Entonces f posee exactamente un punto fijo $x_f \in \mathbf{R}^2$, y más aún para cualquier punto $\boldsymbol{x}$ que pertenece a $\mathbf{R}^2$, la sucesión de puntos generados por las iteraciones delanteras $\{f^{\circ n}(\boldsymbol{x}) : n = 0, 1, 2, ...\}$ converge al punto fijo x_f; esto es,

$$\lim_{n\to\infty} f^{\circ n}(x) = x_f, \quad para\ cada\ x \in \mathbf{R}^2$$

Para la demostración del teorema véase [2].

Se han dado las herramientas necesarias para llegar a comprender el comportamiento de un sistema de funciones iteradas. En la siguiente sección se describirá un espacio en donde se pueden estudiar los conjuntos generados por los sistemas de funciones iteradas, de tal forma que podamos definir sencillas operaciones sobre estos conjuntos. Seremos capaces de entender estas dinámicas que dan como resultado los tan sencillos y complicados fractales.

3.5. Espacio de Hausdorff

Para estudiar los fractales que se generan con los sistemas de funciones iteradas es conveniente trabajar en el espacio de Hausdorff $\mathcal{H}$, ya que en este espacio los conjuntos pueden ser tratados como dibujos e imágenes binarias pues están consituidos por puntos negros sobre un fondo blanco, esto nos permitirá manipular los datos en la computadora y elaborar histogramas para nuestros análisis.

En el espacio de Hausdorff $\mathcal{H}$ podemos estudiar importantes subconjuntos de espacios métricos. Nos enfocaremos en el espacio métrico completo $\mathbf{R}^2$ con la distancia euclidiana, ya que es el espacio donde hemos definido todos nuestros conceptos. Empezaremos por definir formalmente este espacio.

Definición 12 *Sea* $(\mathbf{R}^2, d)$ *un espacio métrico completo. Entonces* $\mathcal{H}(\mathbf{R}^2)$ *denota el espacio cuyos puntos son subconjuntos compactos no vacíos de* $\mathbf{R}^2$.

En la sección anterior se definió lo que es un conjunto compacto, en términos sencillos es un conjunto que "está delimitado"; por ejemplo, una imagen. Estos conjuntos en el espacio de Hausdorff son puntos como se menciona en la definición, pero sobre el plano euclidiano son conjuntos.

A continuación se definirá una métrica que opere sobre el espacio de Hausdorff para posteriormente definirlo como un espacio métrico completo.

Definición 13 *Sea* $(\mathbf{R}^2, d)$, *sea* x *un punto en* $\mathbf{R}^2$ *y* B *un elemento en el espacio* $\mathcal{H}(\mathbf{R}^2)$, *entonces definiremos*

$$d(x, B) = min\{d(x, y) : y \in B\}$$

donde $d(x, B)$ *es llamada la distancia entre el punto* x *y el conjunto* B.

Para obtener esta distancia debemos calcular; por ejemplo, la distancia euclidiana entre todos los puntos del conjunto B y el punto x, esto genera un conjunto de números y el menor de ellos es la distancia que representa la definición anterior (ver Figura 3.11).

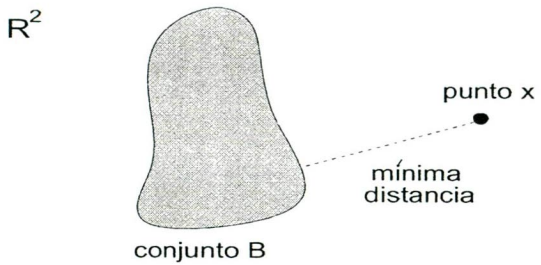


Figura 3.11: Se muestra la menor distancia que hay entre un punto x y el conjunto B.

Definición 14 *Sea* $(\mathbf{R}^2, d)$. *Sean* A *y* B *pertenecientes a* $\mathcal{H}(\mathbf{R}^2)$. *Definimos*

$$d(A, B) = max\{d(x, B) : x \in A\}$$

donde $d(A, B)$ *es llamada la distancia entre el conjunto* $A \in \mathcal{H}(\mathbf{R}^2)$ *y el conjunto* $B \in \mathcal{H}(\mathbf{R}^2)$.

Para obtener esta distancia se aplica la definición 13 para cada punto en A y el conjunto B, de esta forma se obtiene un conjunto de distancias mínimas; posteriormente se encuentra el elemento máximo para tener la distancia entre el conjunto A y el B; notemos que no es lo mismo calcular la distancia $d(A, B)$ y $d(B, A)$, esto servirá para la definición siguiente (ver Figura 3.12).

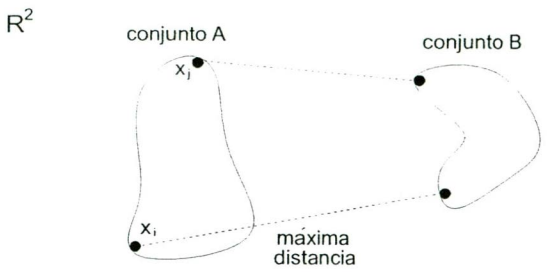


Figura 3.12: La distancia entre los conjuntos A y B se calcula tomando la distancia máxima de entre las distancias mínimas de los puntos de A al conjunto B.

Definición 15 *Sea* $(\mathbf{R}^2, d)$. *Entonces la* distancia de Hausdorff *entre los puntos* A *y* B *en* $\mathcal{H}(\mathbf{R}^2)$ *esta definida por*

$$h(A, B) = d(A, B) \vee d(B, A)$$

donde $x \vee y$ *significa el máximo de* x *y* y. *Llamaremos a* h *la métrica de Hausdorff sobre* $\mathcal{H}$.

Ya que se ha definido una métrica, se puede definir el espacio de Hausdorff como un espacio métrico completo, recordando que toda sucesión de Cauchy sobre este tipo de espacios converge.

Teorema 2 *Sea* $(\mathbf{R}^2, d)$. *Entonces* $\mathcal{H}(\mathbf{R}^2, h)$ *es un espacio métrico completo. Más aún, si* $\{A_n \in \mathcal{H}(\mathbf{R}^2) : n = 1, 2, ...\}$ *es una sucesión de Cauchy, entonces*

$$A = \lim_{n \to \infty} A_n \in \mathcal{H}(\mathbf{R}^2)$$

puede ser caracterizado como

$$A = \{x \in \mathbf{R}^2 \text{ } \textit{tal que} \textbf{ existe } \textit{una sucesión de Cauchy } \{x_n \in A_n\} \textit{ convergiendo a } x\}.$$

Para la demostración del teorema véase [2].

Sea $(\mathbf{R}^2, d)$ un espacio métrico y sea $(\mathcal{H}(\mathbf{R}^2), h(d))$ la notación para representar su espacio de Hausdorff asociado, con la métrica $h(d)$ descrita anteriormente y donde d es la distancia euclidiana como la hemos manejado siempre.

La notación $h(d)$ muestra que d es la métrica que está por debajo o subyace a la métrica de Hausdorff h.

3.6. Mapeos de contracción en el espacio $\mathcal{H}$

Anteriormente se definieron los mapeos de contracción para $\mathbf{R}^2$ y se explicaron detalladamente sus propiedades, ahora llevaremos estos conceptos al espacio de Hausdorff.

Lema 1 *Sea* $w : \mathbf{R}^2 \to \mathbf{R}^2$ *un transformación de contracción sobre el espacio métrico* $(\mathbf{R}^2, d)$ *con factor de contracción* s. *Entonces* $w : \mathcal{H}(\mathbf{R}^2) \to \mathcal{H}(\mathbf{R}^2)$ *definida por*

$$w(B) = \{w(x) : x \in B\}, para\ todo\ B \in \mathcal{H}(\mathbf{R}^2)$$

es una transformación de contracción sobre $(\mathcal{H}(\mathbf{R}^2), h(d))$ *con factor de contracción* s.

Los sistemas de funciónes iteradas como su nombre lo dice constan de un conjunto de funciones o transformaciones que serán aplicadas a un conjunto, la forma de aplicarlas se explicará en el siguiente lema, y por lo tanto llegaremos al objetivo primordial de este capítulo que es dar una difinición formal de los sistemas de funciones iteradas sustentada en todos los conceptos que se han desarrollado.

Lema 2 *Sea* $\{w_n : n = 1, 2, ..., N\}$ *un conjunto de transformaciones de contracción sobre* $(\mathcal{H}(\mathbf{R}^2), h)$. *Sea* s_n *el factor de contracción para la transformación* w_n. *Definiremos* $W : \mathcal{H}(\mathbf{R}^2) \to \mathcal{H}(\mathbf{R}^2)$ *como*

$$\begin{aligned} W(B) &= w_1(B) \cup w_2(B) \cup ... \cup w_N(B) \\ &= \bigcup_{i=1}^{N} w_n(B), para\ cada\ B \in \mathcal{H}(\mathbf{R}^2) \end{aligned}$$

Entonces W *es una transformación de contracción con factor de contracción* $s = max\{s_n : n = 1, 2, ..., N\}$.

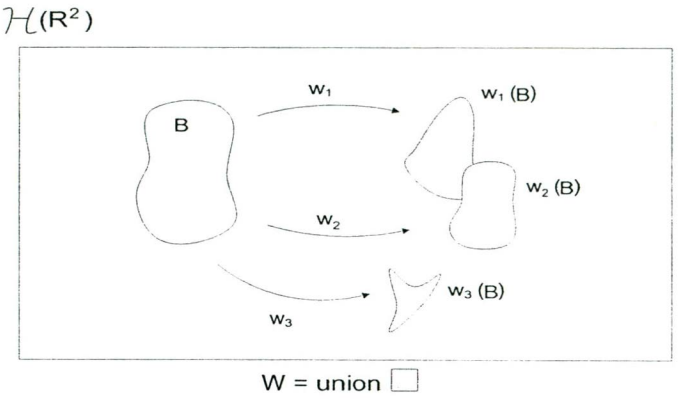


Figura 3.13: La aplicación de un conjunto de transformaciones de contracción al punto B en el espacio de Hausdorff $\mathcal{H}(\mathbf{R}^2)$, el resultado es la unión de los nuevos puntos.

3.7. Sistema de funciones iteradas

El nombre de *sistemas de funciones iteradas* (en inglés *iterated function system*) fue inventado o usado por primera vez por Barnsley [23]. Hay otras referencias relevantes como [16] y [28]. En esta sección se usará la definición dada por Barnsley [2] modificada ligeramente para el espacio $\mathbf{R}^2$.

Definición 16 *Un sistema de funciones iteradas consiste del espacio métrico completo* $(\mathbf{R}^2, d)$ *junto con un conjunto finito de transformaciones de contracción* $w_n : \mathbf{R}^2 \to \mathbf{R}^2$ *con su respectivo factor de contracción* s_n*, para* $n = 1, 2, ..., N$*. Usaremos la abreviación* "IFS" *para* sistemas de funciones iteradas. *La notación para los IFS es* $\{\mathbf{R}^2; w_n, n = 1, 2, ..., N\}$ *y su factor de contracción es* $s = max\{s_n : n = 1, 2, ..., N\}$.

Usaremos esta nomenclatura para denotar simplemente un conjunto de transformaciones finito actuando sobre el espacio métrico en cuestión, con la única condición de que sean transformaciones afines como las que se definieron anteriormente. El siguiente teorema es la parte central para describir la dinámica de un IFS.

Teorema 3 El teorema de los IFS
Sea $\{\mathbf{R}^2; w_n, n = 1, 2, ..., N\}$ *un Sistema de Funciones Iteradas con factor de contracción* s. *Entonces la transformación* $W : \mathbf{R}^2 \to \mathbf{R}^2$ *definida por*

$$W(B) = \bigcup_{n=1}^{N} w_n(B),$$

para todo $B \in \mathcal{H}(\mathbf{R}^2)$ *es un mapeo de contracción sobre el espacio métrico completo* $(\mathcal{H}(\mathbf{R}^2), h(d))$ *con factor de contracción* s; *esto es*

$$h(W(B), W(C)) \leq sh(B, C)$$

para todo $B, C \in \mathcal{H}(\mathbf{R}^2)$, *se tiene un único punto fijo,* $A \in \mathcal{H}(\mathbf{R}^2)$, *el cual obedece*

$$A = W(A) = \bigcup_{n=1}^{N} w_n(A),$$

y viene dado por $A = \lim_{n\to\infty} W^{\circ n}(B)$ *para cualquier* $B \in \mathcal{H}(\mathbf{R}^2)$.

Definición 17 *El punto fijo* $A \in \mathcal{H}(\mathbf{R}^2)$ *descrito en el teorema de los IFS se llama el* atractor del IFS.

Existen dos algoritmos para calcular los atractores de los sistemas de funciones iteradas, uno es totalmente determinístico y trabaja aplicando un conjunto de transformaciones a conjuntos compactos; el segundo, se aplican las transformaciones probabilísticamente y se calcula una orbita que corresponde a puntos en el atractor. Empezaremos por describir el algoritmo determinista.

3.8. Algoritmo determinista

El algoritmo determinista esta basado en la idea de calcular directamente las secuencias de conjuntos $A_n = W^{\circ n}(A)$ comenzando desde el conjunto inicial A_0. Formalmente:

Algoritmo 1 Algoritmo Determinista
Sea $\{\mathbf{R}^2; w_1, w_2, ..., w_N\}$ *un IFS. Se escoge un conjunto compacto* A_0 *que pertenece a* $\mathbb{R}^2$. *Se calculan sucesivamente conjuntos* $A_n = W^{\circ n}(A)$ *con la regla siguiente:*

$$A_{n+1} = \bigcup_{j=1}^{N} w_j(A_n) \text{ para } n = 1, 2, ...;$$

la cual construye un sucesión $\{A_n : n = 0, 1, 2, 3, ...\} \in \mathcal{H}(\mathbf{R}^2)$ *que converge al atractor del IFS en el espacio de Hausdorff.*

Aplicando el algoritmo anterior podemos generar el triángulo de Sierpinski. Primero definiremos el IFS adecuado. Sea $\{\mathbf{R}^2; w_1, w_2, w_3\}$ donde las w_i son transformaciones afines en $\mathbf{R}^2$ definidas como:

$$w_1 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$w_2 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 50 \\ 0 \end{pmatrix},$$

$$w_3 \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 50 \\ 50 \end{pmatrix}.$$

El triangulo de Sierpinski tendrá como vértices los puntos (0,0), (100,0), y (100,100). La figura 3.14 muestra el conjunto inicial que corresponde a un cuadrado de 100 por 100 pixels, tiene estas dimensiones para mostrar de una forma simple las siguientes 8 iteraciones.

Para manipular fácilmente el código podemos hacer una tabla que resuma las características del IFS anterior como la siguiente.

w	a	b	c	d	e	f	p
1	0.5	0	0	0.5	0	0	0.33
2	0.5	0	0	0.5	50	0	0.33
3	0.5	0	0	0.5	50	50	0.33

La útima columna representa una probabilidad asociada a cada transformación, en el ejemplo anterior no se utilizaron. En la siguiente sección se explica la forma en que trabajan los IFS con probabilidades.

También, se puede observar que todos los valores de las matrices son iguales, y solo cambia el vector de traslación (valor e y f), a este IFS particular se le conoce como el *juego del caos*. Si incrementamos el número de transformaciones y conservando la misma estructura de la matriz, se puede hablar

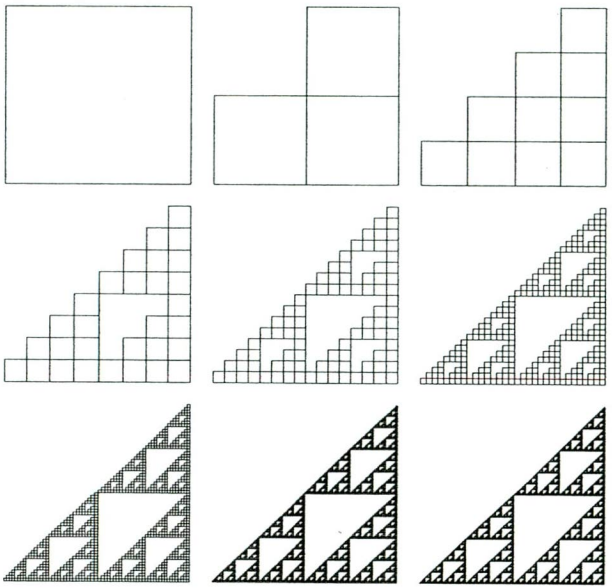


Figura 3.14: Iteraciones generadas por el IFS.

de un *juego del caos extendido*, en el próximo capítulo analizaremos este IFS particular, ya que por su construcción tan sencilla, y las complejidad de sus patrones o atractores, nos servirá para que datos de entrada plasmen su *huella digital.*

3.9. Algoritmo de iteraciones aleatorias

El algoritmo de iteraciones aleatorias esta fundado en la teoría ergódica, sus bases matemáticas se pueden consultar en [2].

Se difinirá un nuevo tipo de IFS que es parecido al anterior, ahora las transformaciones se aplicarán siguiendo una probabilidad, esto nos permitirá crear rápidamente una representación gráfica del atractor.

Definición 18 *Un* sistema de funciones iteradas con probabilidades *consiste*

de un IFS $\{\mathbf{R}^2; w_1, w_2, ..., w_N\}$ *junto con un conjunto ordenado de números* $\{p_1, p_2, ..., p_N\}$, *tal que*

$$p_1 + p_2 + ... + p_N = 1 \text{ y } p_i > 0 \text{ para } i = 1, 2, 3, ..., N.$$

La probabilidad p_i *esta asociada a la transformación* w_i.

Las probabilidades juegan un papel muy importante en el calculo de las imagenes que representan los atractores de los IFS, estas no tienen sentido en el Algoritmo Determinísitico.
El siguiente mecanismo nos permite asignar probabilidades a cada transformación y nos será de utilidad para el algoritmo que mostraremos posteriormente, en otros casos las probabilidades pueden asignarse empíricamente, dependiendo de lo que queramos como resultado.

$$p_i \approx \frac{|detA_i|}{\sum_{j=1}^{N} |detA_j|} = \frac{|a_i d_i - b_i c_i|}{\sum_{j=1}^{N} |a_j d_j - b_j c_j|} \text{ para } i, j = 1, 2..., N.$$

Se toma el símbolo $\approx$ ya que puede surgir el caso en que $A_i = 0$ entonces asignará una probabilidad cero y estaría violando la definición, por tal motivo debemos asignar un número positivo muy pequeño o ajustarlo según convenga. El mecanismo anterior será el que usaremos para signar probabilidades a las transformaciones dependiendo de los datos de la secuencia de DNA.

Algoritmo 2 El algoritmo de iteraciones aleatorias *Sea* $\{\mathbf{R}^2; w_1, w_2, ..., w_N\}$ *un IFS, donde las probabilidades* $p_i > 0$ *han sido asignadas para cada transformación* w_i *con* $i = 1, ..., N$ *donde* $\sum_{i=1}^{n} p_i = 1$. *Se escoge un punto cualquiera* $\mathbf{x}_0 \in \mathbf{R}^2$, *se escoge recursivamente y de forma independiente una transformación de acuerdo a su probabilidad. Se le aplica la transformación elegida y obtenemos un nuevo punto*

$$\mathbf{x}_n \in \{w_1(x_{n-1}), w_2(x_{n-1}), ..., w_N(x_{n-1})\} \text{ para } n = 1, 2, 3, ..., k,$$

donde la probabilidad del evento $\mathbf{x}_n = w_i(x_{n-1})$ *es* p_i. *Entonces se construye la sucesión* $\{\mathbf{x}_n : n = 1, 2, 3, ..., k\} \in \mathbf{R}^2$.

Esta sucesión $\{\mathbf{x}_n\}_{n=1}^{\infty}$ converge al atractor del IFS. Para poder visualizar el atractor debemos graficar los últimos n puntos ya que son los que están más próximos al atractor.

En el siguiente IFS se muestran los valores para generar el helecho creado por Barnsley, son de las atractores más representativos de los IFS's. Las probabilidades fueron tomadas empíricamente, o al menos no se comenta el método con las que fueron calculadas.

w	a	b	c	d	e	f	p
1	0	0	0	0.16	0	0	0.01
2	0.85	0.04	-0.04	0.85	0	1.6	0.85
3	0.2	-0.26	0.23	0.22	0	1.6	0.07
4	-0.15	0.28	0.26	0.24	0	0.44	0.07

En la Figura 3.15 se muestra el proceso de este algoritmo calculando el *helecho* (fern) de Barnsley.

Podemos ver que mientras se va iterando el algoritmo, se van agregando más y más puntos, en la primera imagen se han hecho 500 iteraciones, es decir hay 500 puntos en la imagen y aún no se observa el atractor completo, en la segunda hay 2000 puntos y ya se tiene más conocimiento de la forma del atractor, en la tercera hay 8000 puntos y en la última 100,000 puntos. La diferencia entre las últimas dos no es muy grande, se esta convergiendo cada vez más al atractor del IFS. Cuando se hace este proceso iterativo en la computadora basta con darle algunos miles de iteraciones para tener una visión cualitativa del atractor, que para una computadora no requiere más que unos segundos. Ahora, en la definición matemática la convergencia al verdadero atractor se obtiene en iteraciones infinitas, con la computadora estamos haciendo una abstracción de lo que realmente pasa en la teoría.

Algo interesante en el ejemplo anterior es que las probabilidades hacen que los puntos se acumulen en las orillas, es decir en la punta del helecho y en las terminaciones de las hojas principales, entonces si buscamos otra distribución de probabilidades los puntos se concentrarán en otras regiones, los puntos estarán "bailando" sobre la estructura del atractor, es inevitable, no pueden escapar a la dinámica de sus tranformaciones, aunque algunas regiones serán más probables de tocar que otras. Con las probabilidades podemos manipular la densidad de puntos en dichas regiones. A las distribuciones de puntos sobre los atractores se le conoce como medida invariante. Se estudiará en la siguiente sección.

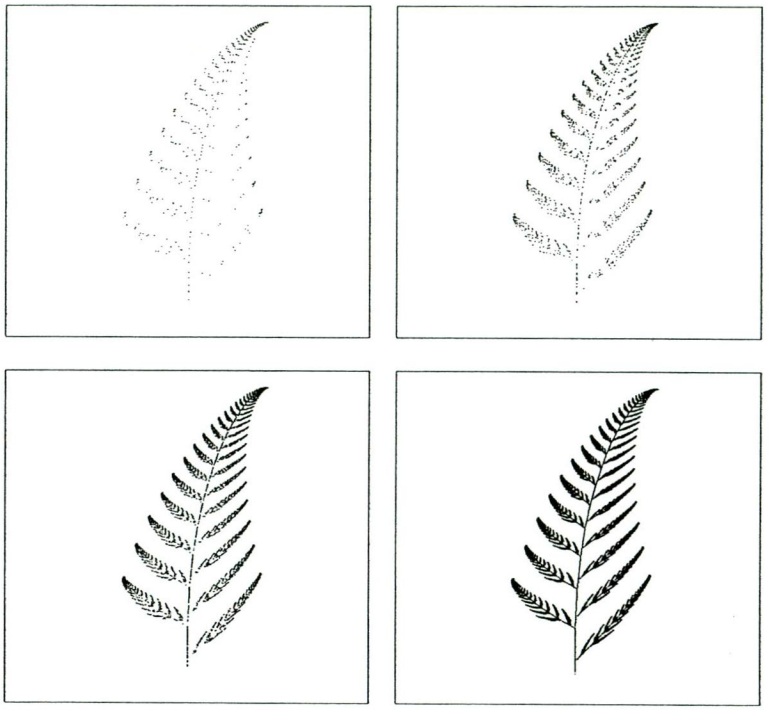


Figura 3.15: Helecho de Barnsley generado por un IFS con probabiliades. Se puede ver como iteración tras iteración el atractor toma forma.

3.10. Complejidad algorítmica para calcular el atractor

Los algoritmos presentados anteriormente son usados para calcular el atractor de un IFS, el algoritmo determinista estaba basado principalmente en las definiciones matemáticas; computacionalmente es muy costoso ya que el incremento en la evaluación de los puntos conforme se van iterando los conjuntos es de orden exponencial, por ejemplo, en la Figura 3.14, si empezáramos con una condición inicial computacionalmente nada costosa, un punto, (ya no un conjunto de puntos) despues de n iteraciones, se estarían

evaluando 3^n puntos, lo cual, después de un número grande de iteraciones es intratable. Por lo tanto el orden es de $k * m^n$ donde k es el número incial de puntos, m es el número de transformaciones y n es el número de iteraciones.

Con el algoritmo de iteraciones aleatorias, la sucesión de puntos se va construyendo tomando una transformación según su probabilidad, el punto inicial se va transformando y ese mismo punto alimenta el sistema en un paso de tiempo después, es un proceso de retroalimentación, la carga computacional esta en encontrar la transformación con la que se tiene que evaluar el punto, en el peor de los casos el algoritmo es cercano al orden cuadrado, es decir, el buscar entre las m transformaciones un número n de iteraciones y si m es tan grande como n se acerca al orden cuadrado, mientras que si m es pequeño, es de orden lineal multiplicado por una constante.

Por tal motivo el algoritmo de iteraciones aleatorias es el principal método para calcular el atractor del IFS, también es conocido como "el juego del caos". Para una explicación detallada de este algoritmo y de su comportamiento se puede consultar [2].

3.11. Medida invariante

A través del algoritmo de iteraciones aleatorias podemos obtener distintas distribuciones de puntos, variando las probabilidades de asignación de cada transformación, los puntos no pueden escapar de la dinámica del sistema, sin embargo, se pueden aglomerar en regiones y ver denso el fractal, por otro lado pueden parecer zonas *flacas* donde hay muy pocos puntos. Para estudiar estos comportamientos se recurre a la medición de cantidad de puntos sobre los atractores. Los sistemas de funciones iteradas se basan en un proceso infinito iterativo, asi que, por cada iteración se agregan más puntos al sistema, si se itera un número mayor de veces el sistema se acercará a una medida difinitiva, esto se le conoce como *medida invariante.*

La medida invariante es una herramienta muy útil en este trabajo, ya que dependiendo de las probabilidades asignadas a cada transformación es como cambiará las densidades de puntos, de esta forma, por cada genoma se tendrá una distribución de puntos diferente, representativa del organimo. A continuación se presentarán los conceptos básicos.

A los subconjuntos de $\mathbf{R}^2$ que contengan puntos los llamaremos *subconjuntos de Borel* sobre $\mathbf{R}^2$ y los denotaremos como $\mathcal{B}(\mathbf{R}^2)$. Los subconjuntos de Borel de $\mathbf{R}^2$ incluyen los subconjuntos compactos no vacíos de $\mathbf{R}^2$ así que $\mathcal{H}(\mathbf{R}^2)$ está contenido en $\mathcal{B}(\mathbf{R}^2)$. Si $\mathcal{O}$ es un subconjunto abierto de $\mathbf{R}^2$ entonces $\mathcal{O} \in \mathcal{B}(\mathbf{R}^2)$

Para medir la densidad de puntos sobre el atractor definiremos regiones B que lo cubran totalmente. B donota una bola cerrada en $\mathbf{R}^2$. La densidad estará dada por un sistema que produce una sucesión de puntos $\{Z_n\}_{n=1}^{\infty}$. Definamos a

$$\mathcal{N}(B,n) = \text{número de puntos en } \{z_0, z_1, ..., z_n\} \cap B \text{ para } n = 0, 1, 2, ...$$

La medida invariante la denotaremos como μ, la cual es una función de $\mathbf{R}^2 \to [0,1] \subset \mathbf{R}$ y la definiremos como:

$$\mu(B) = \lim_{n\to\infty} \frac{\mathcal{N}(B,n)}{(n+1)}$$

Esto indica que la densidad en la bola B es la proporción de puntos producidos por el sistema que caen en la bola. Por ejemplo si A es un atractor entonces $\mu(A) = 1$ y $\mu(\varnothing) = 0$, también $\mu(\mathbf{R}^2) = 1$, lo que nos dice que todo el espacio $\mathbf{R}^2$ tiene la misma densidad que el atractor del sistema, por lo que el sistema no genera puntos que no estén sobre el atractor.

Para comprender estos conceptos defininamos una bola $B \in \mathcal{B}(\mathbf{R}^2)$ en el triángulo de Sierpinski generado por el juego del caos, como se muestra en la Figura 3.16 y aplicamos la medida μ a la bola con diferentes cantidades de iteraciones n, los resultados se muestran en el Cuadro 3.1

Si se tiene una medida sobre un conjunto de Borel entonces diremos que esta medida es de Borel, por lo que μ es una medida de Borel.

Definición 19 *Sea μ una medida de Borel sobre una región $\mathcal{D} \subset \mathbf{R}^2$. Si $\mu(\mathcal{D}) = 1$, entoces decimos que μ esta normalizada.*

Sea $\mathcal{B}$ que denota los subconjuntos de Borel de $\mathcal{D}$. Sea $w : \mathcal{D} \to \mathcal{D}$ continua y definamos $w^{-1} : \mathcal{B} \to \mathcal{B}$. Si v es una medida normalizada de Borel sobre $\mathcal{D}$ entonces $\nu \circ w^{-1}$ también lo es. Utilizaremos estas construcciones para la siguiente definición.

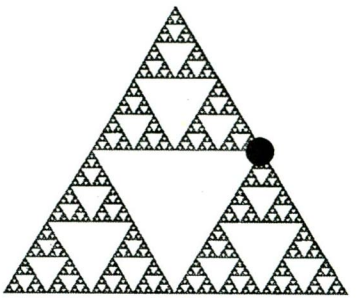


Figura 3.16: La bola negra sobre el triángulo de Sierpinski representa una bola.

Definición 20 *Sea* $\{\mathcal{D}; w_1, w_2, ..., w_N; p_1, p_2, ..., p_N\}$ *un IFS con probabilidades. El operador de Markov asociado con este IFS se define como*

$$M(\nu) = p_1 \nu \circ w_1^{-1} + p_2 \nu \circ w_2^{-1} + ... + p_N \nu \circ w_N^{-1}$$

La medida de Borel μ es invariante bajo el operador de Markov $M(\nu)$. En otras palabras $M(\nu)$ define una nueva medida de Borel normalizada sobre $\mathcal{D}$. Esta nueva medida de Borel la evaluaremos sobre un subconjunto dado $B \in \mathcal{D}$ que es el atractor del IFS.

n iteraciones	$\mathcal{N}(B_1, n)/n$	$\mathcal{N}(B_2, n)/n$
10000	0.010900	0.013900
50000	0.009500	0.014540
100000	0.009700	0.014010
500000	0.009008	0.013204
1000000	0.009145	0.013098
5000000	0.009106	0.013101

Cuadro 3.1: Medidas de dos bolas con distintas iteraciones, se puede ver que entre más iteraciones la proporción de puntos converge a un valor.

La iteración sobre el operador de Markov construirá las densidades sobre subconjuntos de Borel y lo podemos representar mediante histogramas. En particular existe una única medida η tal que al aplicarle el operador de Markov no varía, es decir $M(\eta) = \eta$ y cumple con

$$\lim_{n \to \infty} M^{\circ n}(\nu) = \eta$$

A η le llameremos el *punto fijo* del operador de Markov.

Definición 21 *Sea η el punto fijo del operador de Markov. Entonces η es llamada la* medida invariante del IFS con probabilidades.

Capítulo 4

El juego del caos

El juego del caos tradicional fue propuesto por Barnsley [2] como un algoritmo muy sencillo para calcular el atractor de un IFS. Este juego esta fundamentado en el algoritmo de iteraciones aleatorias, así que desde las primeras iteraciones se está calculando parte del atractor. Las probabilidades son muy importantes ya que a partir de estas se aplicará la transformación de contracción correspondiente. La idea fue extendida por Takashi Tsuchiya [34], al incrementar el número de lados del juego del caos a un número n, bajo este esquema las probabilidades de cada transformación están determinadas por los valores de las sucesiones de mapeos unidimensionales; formando un atractor único para cada conjunto de datos.

El juego del caos nos genera imágenes con geometrías fractales dependiendo de la composición de los datos, así que es de gran ayuda para conocer la complejidad, homogeneidad y aleatoriedad de los datos introducidos.

4.1. Un IFS particular

El juego del caos es un IFS con una estructura particular, en su versión estándar consta de 3 transformaciones cuyos valores podemos ver en la siguiente tabla.

w	a	b	c	d	e	f	p
1	0.5	0	0	0.5	e_1	f_1	0.33
2	0.5	0	0	0.5	e_2	f_2	0.33
3	0.5	0	0	0.5	e_3	f_3	0.33

donde las componentes en e y en f describen la traslación de lo puntos. Si hacemos $\mathbf{e} = \frac{e}{2}$ y $\mathbf{f} = \frac{f}{2}$ solo estamos trasladando los puntos a la mitad de distancia de lo que anteriormente estarían, esta traslación de los puntos no afecta cualitativamente el atractor del IFS, entonces las transformaciones quedan de la siguiente forma

$$w_i(x,y) = (\frac{x+\mathbf{e_i}}{2}, \frac{y+\mathbf{f_i}}{2}), \qquad \forall i \in [1,3] \tag{4.1}$$

que es la definición de distancia media entre dos puntos en el plano. Las componentes e_i y f_i delimitan el área donde actuará el IFS. Así que pueden ser considerados como vértices del polígono, en este caso del triángulo. Este sistema es más sencillo de manejar y manipular, permacen las propiedades de los IFS's, pero con reglas más sencillas. Veamos como se juega el *juego del caos.*

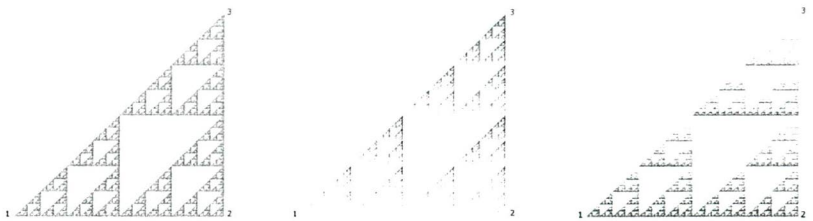
4.2. El juego del caos tradicional

Lo primero que hay que definir son las posiciones de las componentes e y f, en este sentido es mejor si forman un triángulo equilátero. Como vimos anteriormente los IFS's y en particular el juego del caos, necesita de una condición inicial, así que dentro del tríangulo se genera una condición inicial $p0$, que será el primer elemento de nuestra sucesión de puntos, ahora aplicamos el algoritmo de iteraciones aleatorias, es decir, se elige cualquiera de los 3 vértices con igualdad de probabilidad $(\frac{1}{3})$ y se aplica la la transformación, la distancia media entre la condición inicial y el vértice elegido, esto nos genera un punto $p1$ y se repite el proceso hasta tener una sucesión de n puntos, si gráficamos todos estos puntos obtendremos el atractor para este sistema en particular, el triángulo de Sierpinski.

Aunque las probabilidades son escogidas aleatoriamente, el sistema mismo tiene una dinámica que debe de respetarse, si cargaramos la probabilidad hacia un vértice específico, se obtendría una medida invariante distinta pero dentro de la dinámica del sistema.

Es un método útil representar en gama de colores las distintas densidades que aparecen sobre el atractor ya que facilita el reconocimiento de estructuras que puedan surgir con distintas distribuciones de probabilidades. En la

Figura 4.1 se calcula el atractor del juego del caos para 3 vértices con distintas probabilidades de aplicación de las transformaciones, se aprecia que la medida invariante en cada sistema es distinta, la gama de colores nos indica la densidad de puntos que ocurre en cada zona. Las zonas en azul corresponden a una densidad baja de puntos respecto a la totalidad de puntos en el sistema, las zonas en verde corresponden a una densidad media respecto a la contidad de puntos y las zonas en rojo son las densidades máximas de puntos. En la imagen (A) se observa que por tener la misma probabilidad para cada vértice la distribución de puntos es homogénea, no se carga hacia ningún lado, mientras que la imagen (B) los puntos se concentran en el vértice superior de tal forma que en casi todo el atractor la densidad de puntos es baja, en la imagen (C) se consigue otra distribución de puntos muy distinta a las otras.

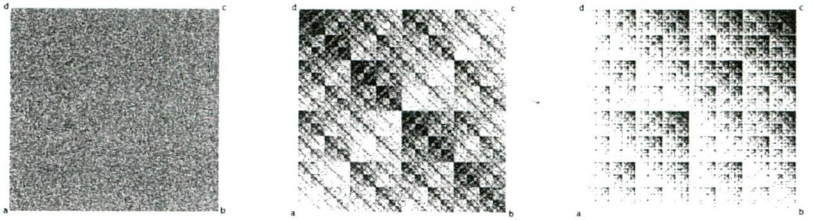


(A) (B) (C)

Figura 4.1: El juego del caos con distintas probabilidades de elección para los 3 vértices. (A) Todos los vértices tienen un tercio de probabilidad. (B) La probabilidad para el vértice 1 es 0.1, para el vértice 2 es 0.1 y para el vértice 3 es 0.8. (C) La probabilidad para el vértice 1 es 0.45 para el vértice 2 es 0.45 y para el vértice 3 es 0.1.

Por la construcción del sistema es fácil incrementar el número de vértices. Si se incrementa el juego del caos a 4 vértices el sistema cambia de dinámica, si se utilizan distintas distribuciones de probabiliades también cambiará la medida invariante del atractor, en la Figura 4.2 se muestran las distribuciones de puntos para 4 vértices.

Los atractores estan sujetos a las mismas reglas, solo que difieren sus probabiliades, y por lo tanto su medida invariante. Así que, se puede alimentar el



(A) (B) (C)

Figura 4.2: (A) El juego del caos con la misma probabilidad de elección para los 4 vértices. (B) Con una distribución de probabilidad distinta para la elección de los vértices; a=0.1, b=0.3, c=0.1 y d=0.5. (C) Con la siguiente distribución a=0.1, b=0.1, c=0.7 y d=0.1.

juego del caos con secuencias de datos y formar una representación visual, este mecanismo se puede utilizar para convertir una secuencia de símbolos que es unidimensional a una representación visual en dos dimensiones. A través de este mecanismo es posible categorizar cadenas de símbolos dependendiendo de su estructura cualitativa y cuantitativa basada en la distribución espacial de su información.

Este tipo de análisis se hizo con éxito para un alfabeto de 4 símbolos las cuales representan las bases nitrogenadas del DNA.

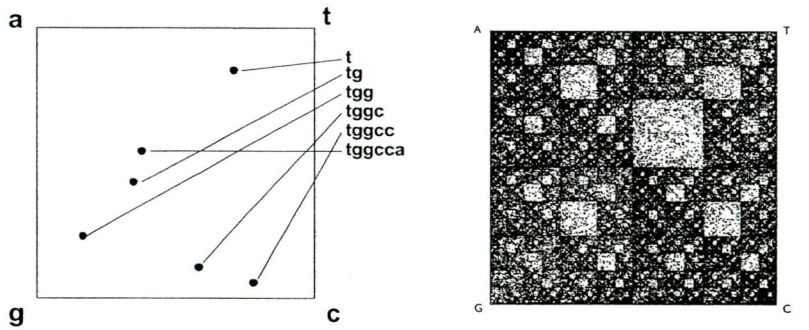
4.2.1. El juego del caos para el DNA

En 1990 H. Joel Jefrey [29] [30] uso por primera vez el juego del caos como mecanismo para generar representaciones visuales en dos dimensiones de secuencias de DNA. Posteriormente otros autores [22, 27, 31, 25, 32] utilizaron el juego del caos como parte de sus análisis para secuencias de DNA, también ha sido utilizado para representaciones visuales de secuencias de proteínas considerando un alfabeto con mayor número de símbolos [24].

Las representaciones del juego del caos conocidas con el nombre *CGR* por sus siglas en inglés, es una técnica para convertir una secuencia de símbolos unidimensional a una representación de dos dimenensiones de tal forma

que preserva la estructura de las subsecuencias construyendo una representación visual. El DNA puede ser manipulado como una cadena de símbolos compuesta por el alfabeto A, G, C y T. Cada vértice del juego del caos es etiquetado con una de estas letras. La aplicación de la tranformación dependerá de la posición de la letra sobre la secuencia, se tomará el punto medio entre el punto interno y el vértice etiquetado con la letra correspondiente. La composición de letras de las subsecuencias formarán un atractor particular para cada secuencia de DNA. Cuando predomienen subsecuencias con cierta estructura en el atractor aparecerán regiones densas, mientras que si hay subsecuencias que aparecen muy poco estas regiones se verán dispersas.

En la Figura 4.3 aparecen las primeras iteraciones del juego del caos para la secuencia del *H. Sapiens*, posteriormente se muestra el resultado de leer toda la secuencia completa.



(A) (B)

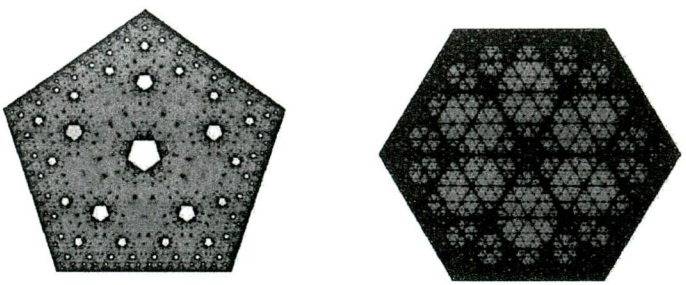
Figura 4.3: (A) Las primeras 6 iteraciones. (B) Todo la secuencia del Homo Sapiens.

A cada organismo se le puede asociar una representación visual, cada una de estas tendrá una medida invariante diferente por tal motivo el método puede servir para diferenciar a los organismos y observar similitudes en la estructura de sus subcadenas. Este método sirve para alfabetos de 4 símbolos y se puede extender a n símbolos; se tendría que considerar vértice por símbolo por lo tanto incrementar el número de lados del polígono. Hacer crecer el número

de lados nos genera otro sistema, a éste se le conoce como el *el juego circular del caos.*

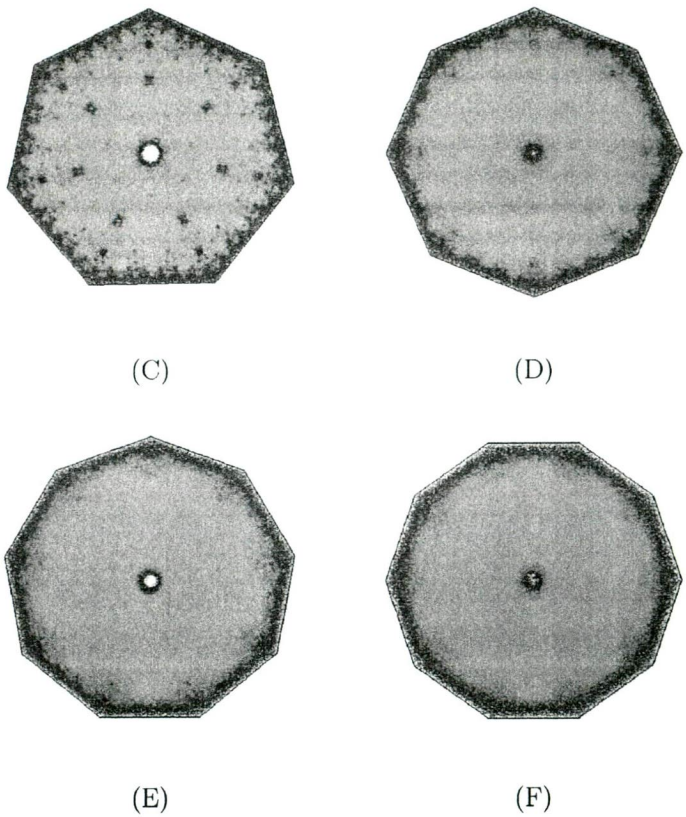
4.3. El juego circular del caos

Como se observo anteriormente conforme aumentamos el número de lados el atractor del juego del caos cambia, llegará un momento en que el atractor se estabilice en un estructura, es entonces, cuando tenemos un nuevo sistema. En la Figura 4.4 y 4.5 se muestra la evolución del sistema aumentando el número de lados. En todas los atractores, se usuaron probabilidades homogéneas, es decir no hay preferencia por ningún vértice, para el sistema de cinco lados hay $\frac{1}{5}$ de probabilidad para cada vértice, para seis lados hay $\frac{1}{6}$ y así sucesivamente. La gama de colores indica la densidad de puntos que hay sobre el atractor, las zonas azules corresponden a cantidades bajas en el conteo de puntos, las zonas verdes presentan mayor cantidad de puntos y las zonas rojas son donde esta el máximo conteo de puntos. Se puede notar que las distribuciones de puntos sobre los atractores se van homogenizando cada vez más. Si tendemos el número de lados hacia infinito obtendremos un nuevo sistema que se llama juego circular del caos.



(A) (B)

Figura 4.4: (A) El juego del caos para 5 lados, se observa una superposición de estructuras pentagonales. (B) Para 6 lados se observa otro atractor y se superponen estructuras exagonales.



(C) (D)

(E) (F)

Figura 4.5: Conforme se aumenta el número de lados se pierde estructura en el atractor del IFS. (C) Juego del caos para 7 lados. (D) Para 8 lados. (E) Para 9 lados. (F) Para 10 lados.

Al incrementar el número de lados llegamos al juego circular del caos, matemáticamente tiene esta forma

$$w_i(x,y) = (\frac{x+\mathbf{e_i}}{2}, \frac{y+\mathbf{f_i}}{2}), \qquad \forall i \in [1,N] \tag{4.2}$$

Donde w_i corresponde a la transformación i-ésima, $\mathbf{e_i}$ y $\mathbf{f_i}$ describen la posición de los i vértices que corresponden al polígono.

Este sistema presenta un comportamiento especial, es la superposición de estructuras circulares, la zona en rojo tenue es donde se intersectan los círculos, y como se observa, la misma dinámica hace que las orillas sea de díficil acceso. Las probabilidaes de aplicación de la transformación es la misma para todos los vértices, así que esta distribución es la más homogénea que se pueda conseguir. Ver Figura 4.6.

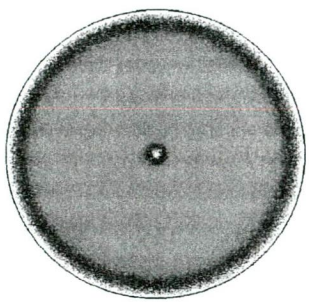


Figura 4.6: El juego circular del caos con 1000 lados y probabilidades homogéneas.

El alimentar el juego del caos con sucesiones de datos, nos permite crear nuevas atractores sobre el mismo sistema, explícitamente se crea una nueva dinámica que consta de dos cosas, la estructura del juego del caos (transformaciones y probabilidades), y, los datos con los que alimentamos al juego del caos. Los datos influyen de una manera fuerte sobre la huella digital en el sistema.

4.4. Series de datos sobre el juego circular del caos

Los mapeos unidimensionales pueden ser usados para representar estas dinámicas, el proceso es el siguiente; tenemos V vértices para comenzar el juego, entonces se divide el intervalo $[0, 1]$ en V subintervalos cada uno etiquetado con un número desde 1 hasta V respectivamente de la siguiente manera

$$1 : [0, \tfrac{1}{V}),\ 2 : [\tfrac{1}{V}, \tfrac{2}{V}) \ , \ldots ,\ V : [\tfrac{V-1}{V}, 1]$$

El paso siguiente es escoger una condición inicial x_0 y obtener una sucesión de números proveniente de algún mapeo que nos ayudará a seleccionar el vértice, se calcula la distancia entre el i-ésimo vértice seleccionado y la condición inicial x_0, se obtiene un nuevo punto x_1; en la siguiente iteración se vuelve a seleccionar otro vértice y se calcula la distancia entre el nuevo vértice y el punto x_1; este proceso se repite un número suficientemente grande de veces. Al final tendremos el atractor del juego del caos para V vértices; si V es suficientemente grande, se tendrá el atractor del juego circular del caos, el cual es único para cada mapeo.

En las Figuras 4.7 se muestran los atractores de los siguientes mapeos:

$$x_{k+1} = \lambda x_k(1 - x_k) \quad \lambda = 4 \tag{4.3}$$

$$x_{n+1} = \begin{cases} 2x_n & \text{si } 0 \leq x \leq \frac{1}{2} \\ 2 - 2x_n & \text{si } \frac{1}{2} < x \leq 1 \end{cases} \tag{4.4}$$

$$x_{n+1} = \begin{cases} 1 - 2x_n & \text{si } 0 \leq x \leq \frac{1}{2} \\ 2 - 2x_n & \text{si } \frac{1}{2} < x \leq 1 \end{cases} \tag{4.5}$$

$$x_{n+1} = \begin{cases} 2x_n & \text{si } 0 \leq x \leq \frac{1}{2} \\ 2x_n - 1 & \text{si } \frac{1}{2} < x \leq 1 \end{cases} \tag{4.6}$$

La ecuación (4.3) corresponde al mapeo logístico en su dinámica caótica, (4.4) es el mapeo tienda de apache, (4.5) es el mapeo r-adic inversa, y por último (4.6) que corresponde al mapeo diente de sierra (equivalente al shift de Bernoulli).

El juego del caos permite una gran variedad de patrones dependiendo de su estructura interna, es decir, de su matriz de transformación, en este capítulo se analizó una estructura particular en la cual la diagonal de la matriz consta de valores 0.5, pero si cambiamos esta condición se obtienen estructuras diversas. La superposición de estructuras de la que se habló se separan y permite observar mejor el atractor. Ver Figura 4.8

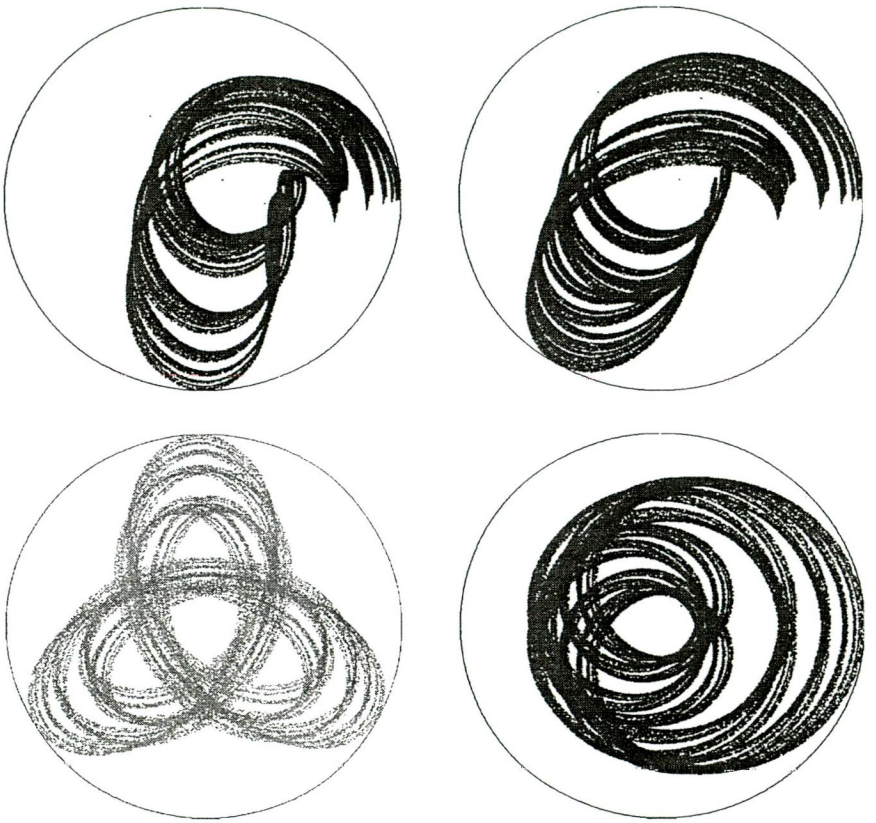
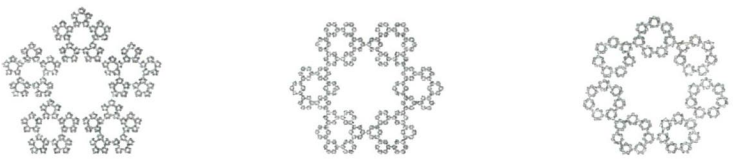


Figura 4.7: Dependiendo de los datos de entrada se genera un atractor específico.

Se han hecho análisis tomando datos de sistema biológicos como en [6], se toman cadenas de DNA codificadas en forma binaria y se alimenta el juego circular del caos, por cada organismo se obtiene una medida invariante.

El juego circular del caos ha demostrado ser una herramienta eficiente en trasladar la dinámica de los datos en imágenes con geometrías fractales, la visión cualitativa de los datos nos permite comprender la calidad de los datos



(A) (B) (C)

Figura 4.8: (A)El juego del caos con 5 lados y con $r = \frac{3}{8}$ de valor en la diagonal de la matriz. (B) 6 lados y $r = \frac{1}{3}$. (C) 7 lados y $r = \frac{1}{4}$.

y la cantidad de información que contiene la serie. Datos periódicos no podrán generar estructuras tan ricas de igual forma datos aleatorios no generarán ningún patrón discernible.

En la próxima sección se explicará el procedimiento que se uso para la construcción de los IFS's a partir de las secuencias de DNA, y como obtener una visualización fractal de la cadena, mostrando las diferencias químicas que pueda tener la molécula de DNA.

Capítulo 5

Genomas sobre sistemas de funciones iteradas y su visualización fractal

En los capítulos anteriores se ha mostrado que las herramientas utilizadas para este análisis bioinformático está fundamentado en marcos teóricos robustos. Con ayuda de las matemáticas y de las ciencias de la computación se pueden encontrar información en las secuencias de DNA, que por medio de otros métodos parecería difícil. Trasladar la información contenida en el DNA hacia algún método visual compacta la información y permite una inspección la cual puede revelar sutilezas en la información.

El objetivo del análisis es mostrar que las cadenas de DNA presentan una configuración especial en su información, por medio de los IFS's podemos identificar estas propiedades. De la misma forma será una herramienta útil para diferenciar secuencias de DNA entre organismos representativos de los dominios principales, eucariontes, bacterias y arqueobacterias.

El análisis se basa en una pequeña diferenciación de la estructura química del DNA, se deja a un lado configuraciones espaciales, procesos biológicos que involucran al DNA, y regiones codificantes y no codificantes, para sólo quedarnos con la naturaleza química y energética de los nucleótidos. Precisando, formaremos 2 grupos el de las bases nitrogenadas adenina y timina, y el de citosina y guanina; estos grupos se forman en base de los enlaces energéticos y del principio de complementariedad que rige en el DNA. Se observará

que con sólo esta diferenciación podemos encontrar una gran cantidad de información en el DNA.

5.1. Genomas: materia prima

De los bancos de datos de genes se consiguieron secuencias de DNA en formatos *fasta* [1] que son las secuencias completas compuestas con el alfabeto A, G, C y T. Para eucariontes cada archivo representa la secuenciación de un cromosoma completo y para procariontes es la totalidad de su genoma.

A partir de la lectura de archivos fasta se construye un IFS's para cada organismos. El proceso es el siguiente, se divide la secuencia de DNA en n segmentos iguales proporcionales al tamaño del genoma, con cada segmento se construirá una transformación afin.

... A G C T T | G C A A T A G C T T C A G A T | T C G C G ...

Para cada segmento se contarán las cantidades de A, G, C y T, obteniendo su frecuencia relativa, se normalizarán los datos al tamaño de la ventana cumpliendo la siguiente ecuación

$$f_{A_i} + f_{G_i} + f_{C_i} + f_{T_i} = 1 \qquad \forall i \in [i, N] \tag{5.1}$$

En la matriz de las transformaciones afines el componente a representará la frecuencia relativa de bases A que contiene el segmento i, el componente b representará la frecuencia relativa de bases G, el componente c representaran la frecuencia relativa de bases C y por último el componente d representarán las bases T. Las transformaciones de nuestro IFS, toman la siguiente forma:

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f_{A_i} & f_{G_i} \\ f_{C_i} & f_{T_i} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} v_{x_i} \\ v_{y_i} \end{pmatrix} \tag{5.2}$$

En donde v_{x_i} y v_{y_i} representan los vértices del polígono en el juego circular del caos.

[1] Existen varios formatos que describen diferentes características de la secuencia de DNA.

Esta matriz puede tener muchas configuraciones según la asignación de la frecuencia relativa de la base con respecto al componente de la matriz. Es muy importante como se acomoden estos componentes, ya que la asignación de las probabilidades para cada transformación dependerá de la matriz y estará relacionada directamente con las propiedades químicas de la molécula de DNA. La probabilidad para cada transformación se asignará dependiendo de la estructura de la matriz, esto es conveniente por que el mismo sistema ajustará su dinámica interna. Se utilizará el mecanismo propuesto por Barnsley para asignar probabilidades a un IFS probabilístico.

$$p_i \approx \frac{|detA_i|}{\sum_{j=1}^{N}|detA_j|} = \frac{|a_i d_i - b_i c_i|}{\sum_{j=1}^{N}|a_j d_j - b_j c_j|} \text{ para } i,j = 1,2...,N.$$

Por la construcción del sistema se resaltará la similitud entre las bases de 2 enlaces (A y T) y de 3 enlaces (G y C) dado que el determinante de estas matrices es:

$$A * T - G * C$$

lo que significa, que las frecuecias relativas de las bases A y T serán comparadas con las de G y C para obtener una medida respecto a las proporciones de las diferentes bases en el segmento de DNA. Cabe destacar que en la molécula de DNA, A es la base complementaria de T y G es la base complementaria de C, así que nuestra configuración también coincide con el principio de complementariedad.

Ahora entonces, se puede observar que segmentos de genoma que contenga gran cantidad de bases A y T respecto a las otras bases G y C, se le asignará una probabilidad mayor a la transformación afin, y viceversa.

Ya construido el IFS, se usará el algoritmo de iteraciones aleatorias para constuir las visualizaciones fractales. Para tener un punto de comparación antes se analizará, lo que pasaría con una secuencia compuesta de letras A, G, C y T construida al azar.

5.2. Del JCC al JCC *modificado*

Supongamos que tenemos una cadena infinita generada al azar con el alfabeto A, G, C, T y la dividimos en segmentos de tamaño n grande, las frecuencias

relativas de las letras tenderán a ser iguales mientras mas grande sea la ventana. La construcción de un IFS con probabilidades bajo este esquema estará representado por las matrices afines:

$$w_i \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0{,}25 & 0{,}25 \\ 0{,}25 & 0{,}25 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} v_{x_i} \\ v_{y_i} \end{pmatrix} \tag{5.3}$$

Donde v_{x_i} y v_{y_i} representan los vértices del polígono de n lados.

El valor que el mecanimos de las probabilidades le asigna a cada transformacion afin es cero, por tal motivo asignamos empíricamente un numero pequeño positivo a cada transformación, en este caso asignamos $0{,}001$ a cada w_i y luego normalizamos las probabilidades para que la suma sea uno. Se aplica el algoritmo de iteraciones aleatorias y el resultado se puede ver en la Figura 5.1.

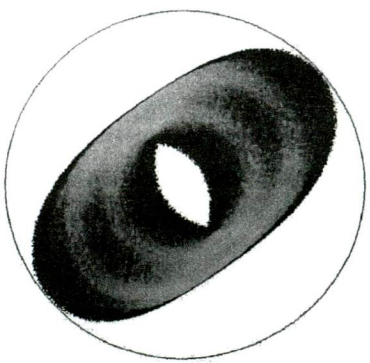


Figura 5.1: Atractor del juego circular del caos modificado, matriz con valores $a = 0{,}25$, $b = 0{,}25$, $c = 0{,}25$, $d = 0{,}25$.

La modificación de la matriz influye directamente en el comportamiento del sistema, la estructura ovalada es producto de los valores de la matriz, los puntos ahora son escalados a proporción de un $\frac{1}{4}$ por las componenetes a y c, son rotados y reflejado por b y d. Para comprender mejor la dinámica se ralizó un histograma, la gama de colores nos dice mucho a cerca de su medida invariante, los colores nos indican donde se aglomeraron más puntos,

la escala empieza por tonos de azul donde hay pocos puntos, luego sigue a tonos color verde y al llegar a rojo son los zonas donde hay una gran cantidad de puntos respecto a otras zonas. A este sistema le llamaremos *juego circular del caos modificado.*

En las Figuras 5.2 se muestra la evolución del juego circular del caos tradicional con $n = 1000$ lados, hacia el juego circular del caos $modificado$ al ir variando los valores de las matrices.

En las figuras anteriores se observa que mientras vamos avanzando hacia los valores del juego circular del caos $modificado$ empiezan a distinguirse unos arillos para comprender esta dinámica se muestra en la Figura 5.3 como converge el sistema respecto al numero de lados.

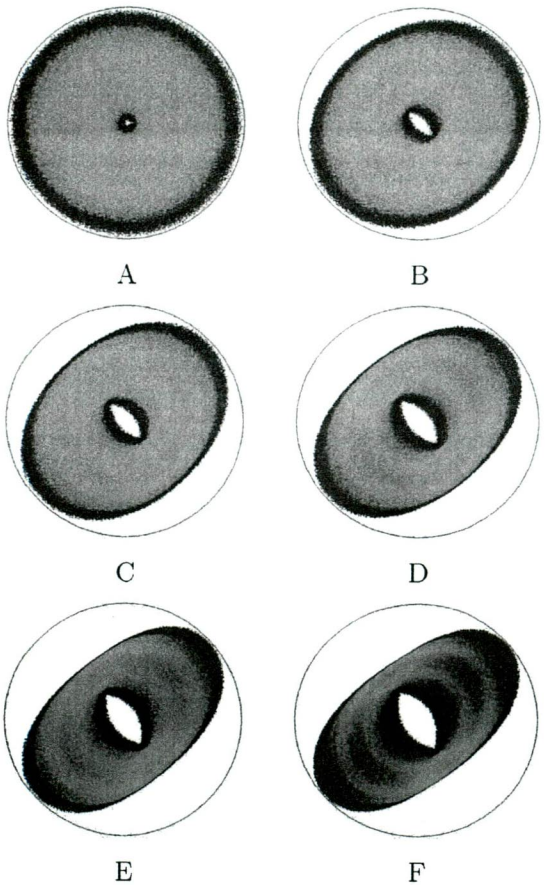


Figura 5.2: (A) Matriz con valores $a = 0{,}5$, $b = 0{,}0$, $c = 0{,}5$, $d = 0{,}0$ (B) Matriz con valores $a = 0{,}45$, $b = 0{,}05$, $c = 0{,}05$, $d = 0{,}45$ (C) Matriz con valores $a = 0{,}40$, $b = 0{,}10$, $c = 0{,}10$, $d = 0{,}40$ (D) Matriz con valores $a = 0{,}35$, $b = 0{,}15$, $c = 0{,}15$, $d = 0{,}35$ (E) Matriz con valores $a = 0{,}30$, $b = 0{,}20$, $c = 0{,}20$, $d = 0{,}30$ (F) Matriz con valores $a = 0{,}25$, $b = 0{,}25$, $c = 0{,}25$, $d = 0{,}25$

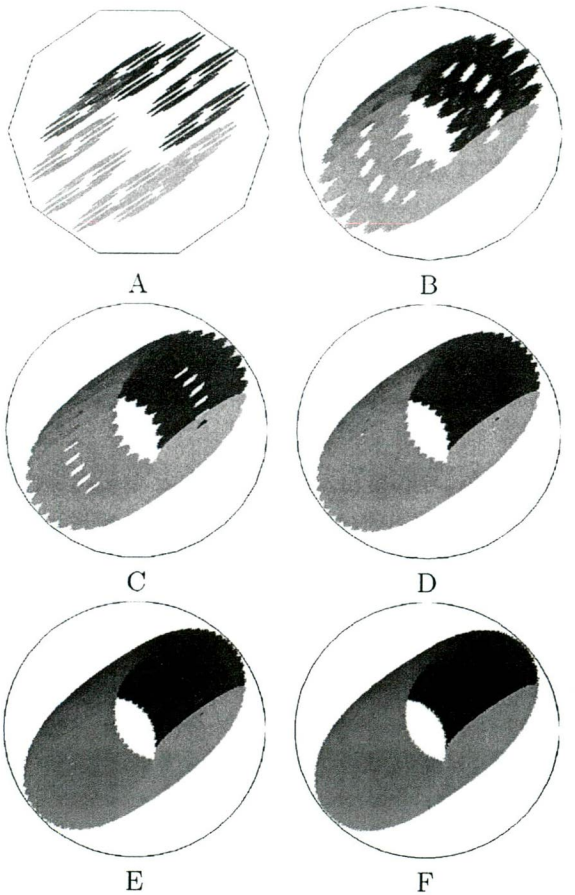


Figura 5.3: (A) El juego del caos modificado con la matriz $a = 0{,}25, b = 0{,}25, c = 0{,}25, d = 0{,}25$ y 10 lados. (B) Con 20 lados. (C) Con 30 lados. (D) Con 40 lados. (E) Con 50 lados. (F) Con 100 lados.

Se hace presente la fractalidad del juego del caos en las primeras imágenes ya que la estructura en forma de elipse o pseudo-elipse, contiene a la vez otras pseudo-elipses y así sucesivamente, al tener un número grande de lados estos se superponen y se pierden en la complejidad de la estructura. Los puntos son coloreados según la asignación del vértice; se puede distiguir el área que puede abarcar el punto. Como todos los vértices tienen la misma probabilidad, las estructuras elipsoidales son regulares y espaciadas homogéneamente.

En las secuencias de DNA no tendremos probabilidades homogéneas, la medida invariante nos servirá para detallar el compartamiento y con una visualización adecuada podremos ver las diferencias químicas de la molécula.

5.3. Asociación de un genoma a un IFS único

Los organismos que se eligieron para el análisis son los mostrados en el Cuadro 5.1

(A) *Arabidopsis thaliana.* Pequeña planta perteneciente al grupo eucarionte.

(B) *Bacillus subtilis.* Organismo perteneciente al grupo de bacterias.

(C) *Caenorhabditis elegans.* Nemátodo de suelo perteneciente al grupo de los eucariontes.

(D) *Chlamydia trachomatis.* Parásito intracelular, clasificado como bacteria.

(E) *Escherichia coli.* Bacteria que se encuentra en los intestinos de humanos y animales.

(F) *Homo sapiens.* Especie terrestre que habita en casi todo el planeta, pertenece al grupo de los eucariontes.

(G) *Methanococcus jannaschii.* Arqueobacteria acuática que se encuentra en chimeneas marinas de 2600 m de profundidad.

(H) *Methanococcus maripaludis.* Arqueobacteria.

(I) *Mycoplasma pulmonis.* Bacteria.

(J) *Pan Troglodytes.* Eucarionte, chimpancé.

(K) *Sulfolobus solfataricus.* Arqueobacteria termofílica que se encuentra en los volcanes terrestres a temperaturas entre los 70 y 90 grados C.

(L) *Thermatoga maritima.* Bacteria que se encuentra en volcanes marinos con una temperatura de 80 grados C.

(M) *Thermoplasma volcanium* Arqueobacteria que crece óptimamante en temperaturas mayores a los 50 grados C.

Cuadro 5.1: Organismos representativos de los dominos Eukarya, Bacteriea y Archaea.

Aplicamos los procedimientos anteriores para asociar un IFS a un organismo.

Partiremos en 1000 segmentos las secuencias de DNA, por lo tanto el IFS constará de 1000 transformaciones, cada una describiendo el comportamiento en relación a sus enlaces de hidrógeno de la molécula en ese segmento. Las probabilidades de las transformaciones estarán sujetas a los valores de la matriz.

Una vez construido el IFS usamos el algoritmo de iteraciones aleatorias con 2 millones de iteraciones dentro del juego circular del caos modificado, obtendremos una aproximación muy buena respecto su medida invariante.

5.4. Resultados

5.4.1. Análisis para eucariontes

En primer instacia se analizará la secuencia de DNA del cromosoma 21 del *Homo sapiens*.

Figura 5.4: Atractor del IFS del *H. sapiens*.

En la Figura 5.4 se muestra la representación visual del *H. sapiens*. Las distribuciones de puntos no estan repartidos homogéneamente sobre el atractor, las zonas díficiles de alcanzar por la dinámica son los bordes internos y bordes externos representados por color azul. En el interior de la elipse se distinguen regiones también difíciles de alcanzar esto es debido a las probabilidades de

asignación. La zona roja pertenece a una aglomeración de puntos, es en esta región donde las probabilidades de los vértices es máxima. Para obtener aún más información se graficará el $det(A * T - G * C)$ contra el vértice correspondiente, ya que se relaciona directamente con la probabilidad.

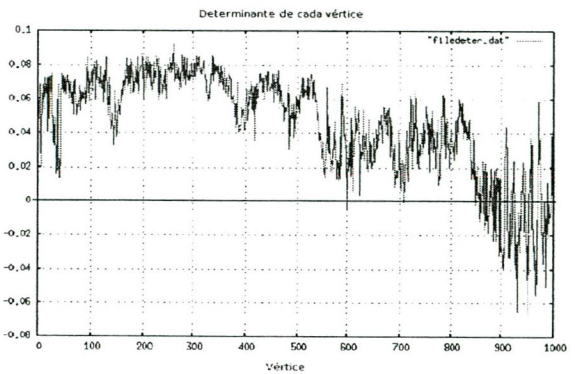


Figura 5.5: Diferencia de proporciones entre moléculas de triple enlace y de doble enlace.

En la Figura 5.5 se describe cualitativamente lo que pasa en la dinámica del juego del caos modificado, si las proporciones de moléculas de doble enlace A y T es mayor que las de triple enlace G y C, el determinante es positivo, si ocurre lo contrario el determinante es negativo. Específicamente con el *H. sapiens* hay una caída en las proporciones de A y T, por lo tanto se puede decir que es más *rígida* la molécula en su terminación.

Se puede conjuntar la información de la Figura 5.5 con el histograma de la Figura 5.4 creando una imagen tridimensional (ver Figura 5.6), anteriormente se explicó que estas estructuras eran la superposición de pseudo elipses, si agregamos una coordenada más al juego del caos, Z, se podrán diferenciar, la cual representará la probabilidad de asignación de cada vértice, así que, las pseudo-elipses que se encuentren en el extremo superior corresponden a transformaciones que tengan mayor cantidad de bases A y T, y, pseudo- elipses que se encuentren en la parte inferior corresponderán a transformaciones con mayor cantidad de bases G y C, en ambos casos las elipses estarán bien

definidas. Las pseudo-elipses borrosas corresponde a probabilidades cercanas a cero por tal motivo los puntos "caen" muy poco en esas regiones.

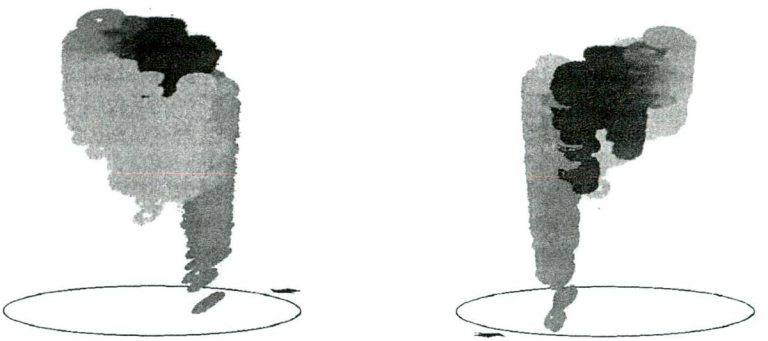


Figura 5.6: La flecha indica la posición del vértice 1, el recorrido es en sentido contrario a las manecillas del reloj. Los puntos coloreados con gama de azul a verde corresponden a los vértices [1,500], los puntos coloreados de verde a rojo correspone a los vértices [501,1000]. Se muestran dos imagenes, una frontal y otra trasera para observar la distribución de los anillos.

En la Figura 5.7 se muestra el atractor del juego del caos para el cromosoma IV del organismo ecuarionte A. *thaliana* generado a partir de la distribución de probabilidad de la Figura 5.8. El atractor tiene una estructura ancha en comparación del atractor que cabría esperar por azar, también está bien definido, eso nos indica que las variaciones de porcentajes de moléculas de doble y triple enlace que se presentan a lo largo de la cadena no son muy grandes. En la gráfica de la Figura 5.8 se observa en general que las variaciones están entre el rango $[0{,}4, 0{,}9]$, podemos decir que la molécula en promedio es *suave*.

En la Figura 5.9 se muestra el atractor del organismo C. *elegans*, se puede apreciar la similitud que tiene con la A. *thaliana*. Su distribución de probabilidad, Figura 5.10, comparte un rasgo muy distintivo con el organismo anterior, el porcentaje de A y T siempre es mayor que el de G y C. Este

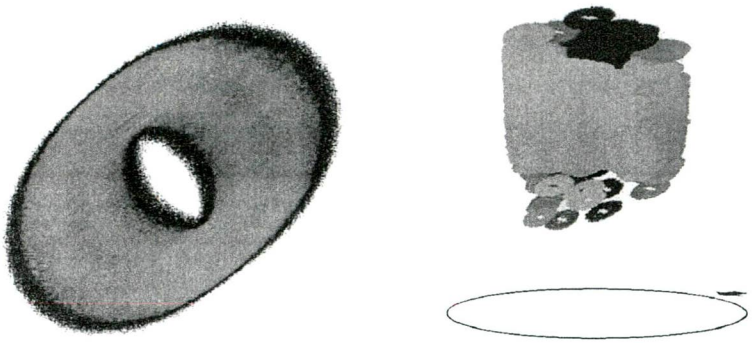


Figura 5.7: *A. thaliana.* Histograma y superficie en 3D.

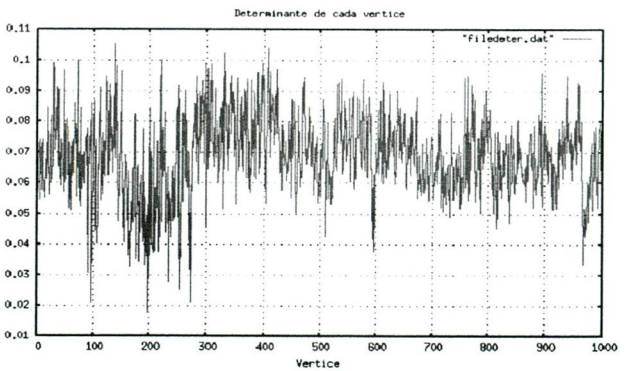


Figura 5.8: *A. thaliana.* Diferencias entre las bases AT y GC. Se puede apreciar que en ninguna ventana las frecuencias relativas de G y C son mayores que las de A y T.

organismo también tiene poca variabilidad en sus porcentajes que son principalmente en el intervalo [0,4, 1,0]. Se aprecian dos fluctuaciones importantes en proporción a las demás variaciones, un mínimo cerca de la ventana número 300 y un máximo cerca de la ventana 700, en la representación visual se pue-

den notar los anillos aislados correspondientes a estas pequeñas fluctuaciones.

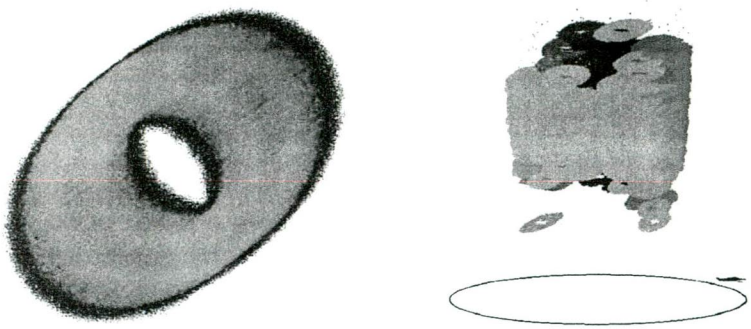


Figura 5.9: *C. elegans.* Histograma y superficie en 3D.

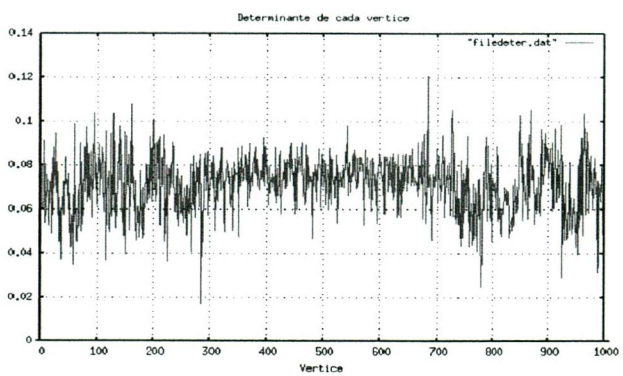


Figura 5.10: *C. elegans.* Predominan las bases con enlaces dobles.

En este análisis se presentaron 3 organismos representativos del grupo de los eucariontes, a simple vista se observa una gran similitud entre los atractores, a diferencia del H. *sapiens* que tiene una particularidad al final de su

secuencia. Las distribuciones de los anillos son muy parecidas, comparten características en las frecuencias relativas de sus bases de doble y triple enlace. Las cantidades de A y de T predominan a lo largo de toda la secuencia, en ninguna región las proporciones de G y C son mayores, esto hace que los anillos se aglutinen en las partes superiores en la representación visual. Entonces se puede decir que, para estos organismos en particular, su molécula del DNA es *suave* y con poca variabilidad respecto a las bases.

5.4.2. Análisis para bacterias

Se realizó el mismo análisis para bacterias y se muestran los resultados más representativos. En la Figura 5.11 se observa un atractor un poco irregular en su parte interna, esto nos demuestra que las frecuencias relativas varían considerablemente en algunas regiones de la secuencia. También es un poco más angosto que los eucariontes, esto se debe a que en la matriz de transformación los valores deben de ser muy parecidos.En su distribución de probabilidad, ver Figura 5.12, se observa que predominan los porcentajes de bases de enlaces triples, G y C, a lo largo de la secuencia, su rango de variabilidad esta entre $[-0{,}03, 0{,}03]$, como estas diferencias estan muy cercanas a cero los porcentajes son proporcionalmente parecidos, se notan 3 máximos muy acentuados que pertenecen a porciones de la secuencia donde predominan bases A y T, mientra que hay un mínimo alrededor de la ventana 50 donde la cantidad de G y C es máxima en proporción a las demás ventanas. En la representación visual en 3D se observa que entre más alejados esten los anillos del cero, sobre el eje z, son más definidos, esto por que es más probable que los puntos se aglomeren en estas zonas.

En la Figura5.13 se aprecia un atractor irregular en su contorno y en su parte interna, esto nos indica que hay variabilidades muy grandes a lo largo de la secuencia; son tan acentuadas que destruyen la forma elipsoidal del contorno. Su distribución de probabilidades mostrada en la Figura 5.14 nos permite observar un comportamiento bastante irreguar, a lo largo de la secuencia predominan las bases A y T, sin embargo hay una gran variabilidad de estas, tiene unos máximos bastantes pronunciados. Un comportamiento atípico aparece alrededor de la ventana 500 donde hay una gran concentración de basea A y T, permaneciendo por 50 ventanas aproximadamente.

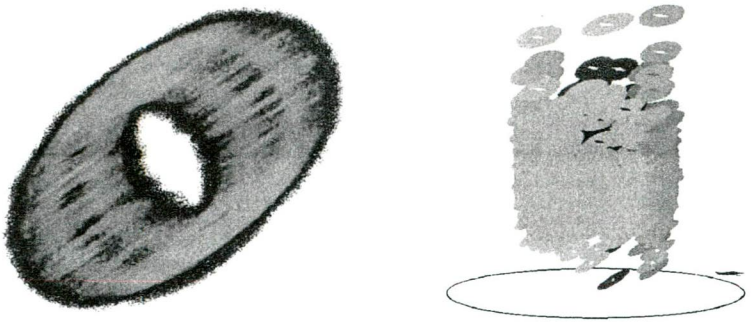


Figura 5.11: *E. coli.* Histograma y superficie en 3D.

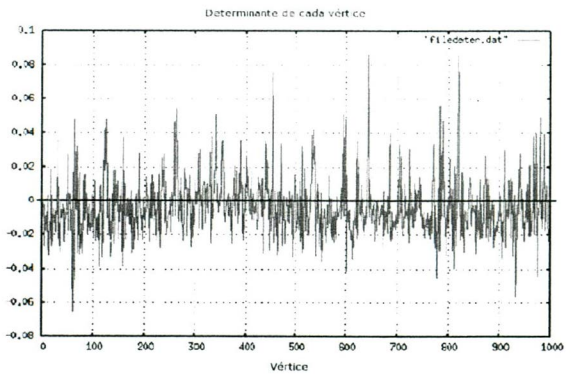


Figura 5.12: *E. coli.* La gráfica permanece alrededor del cero, significa que las bases se presentan casi en proporciones iguales.

En general para este grupo se observó una distribución menos uniforme y con mayor variabilidad, las frecuencias relativas variaron mucho, alrededor del cero y por arriba de él, en la representación visual en 3D las bacterias presentaron los anillos más definidos ya que los saltos en sus distribuciones se reflejan directamente con la formación bien definida de anillos.

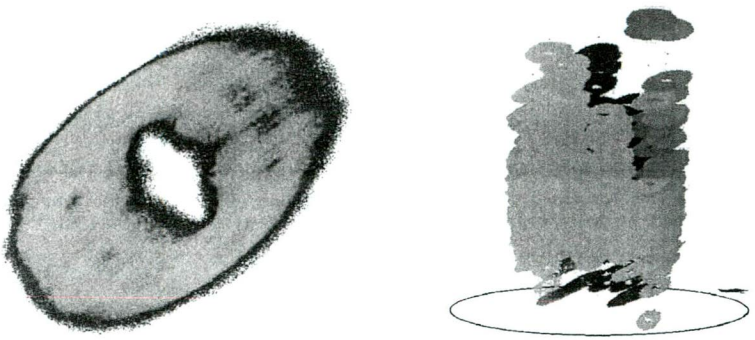


Figura 5.13: *B. subtilis.* Histograma y superficie en 3D.

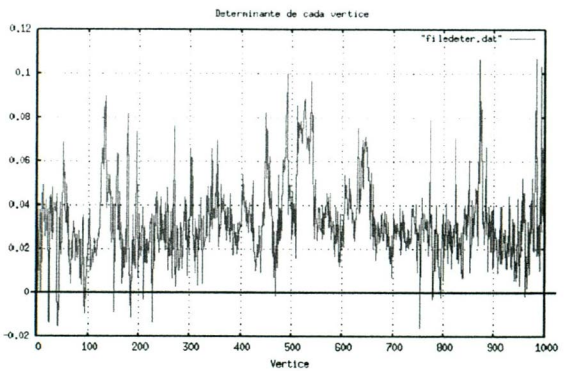


Figura 5.14: *B. subtilis.*

5.4.3. Análisis para Arqueobacterias

Se presenta el análisis para arqueobacterias, se tomarón dos organismos de este grupo y se observo lo siguiente. En la Figura 5.15 perteneciente al atractor del M. *jannaschii* se aprecia una forma de atractor como la de los eucariontes, un poco ensanchada, e internamente regular, no es tan intrincada como las bacterias. Ahora su distribución de probabilidad mostrada en la Figura

5.16 nos revela propiedades importantes, la primera y que salta a la vista son los dos mínimos que se encuentran a lo largo de la secuencia, ahí las bases G y C predominan fuertemente sobre A y T haciendo la molécula mucho más rígida de lo que es en promedio. La otra propiedad es el rango de variación es bastante pequeño de [0,05, 0,12], lo que nos dice que las proporciones de AT y GC permanecen con poca variabilidad a lo largo de la secuencia. En la visualización en 3D se notan los mínimos y la aglomeración de anillos en la parte superior.

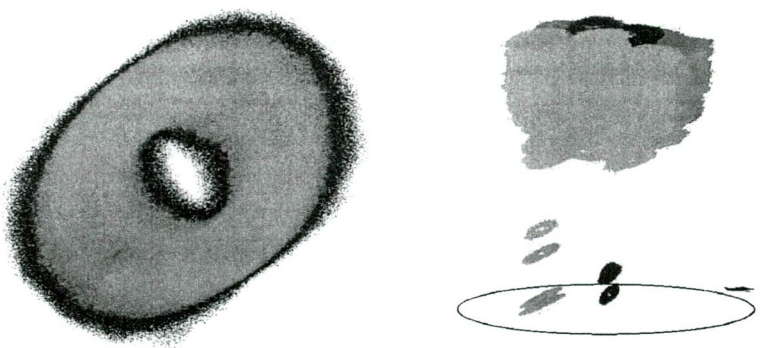


Figura 5.15: *M. jannaschii.* Histograma y superficie en 3D.

El atractor de la arqueobacteria S. *solfataricus* presenta rasgos muy parecidos a la de la arqueobacteria M. *jannaschii*, es ancha y no tiene cambios muy bruscos en su estructura interna. En su distribución de probabilidades mostrada en la Figura 5.18 se presenta un comportamiento cualitativo parecido a la de la arqueobacteria anterior, se distingue un mínimo muy acentuado donde predominan concentraciones grandes de moleculas de triple enlace, mientras que en el resto de la secuencia predominan moléculas de doble enlace, su rango de variabilidad es mayor que el del M. *jannaschii*, se encuentra entre [0,4, 1,0], parecido al rango de los eucariontes. En la representación en 3D se observan los mínimos y una abultamiento de anillos en la parte superior, en general la estructura es parecida a la del organismo anterior.

Con estos dos casos representativos de las arqueobacterias descubrimos particularidades que posiblemente compartan el grupo de las arqueobacterias, los

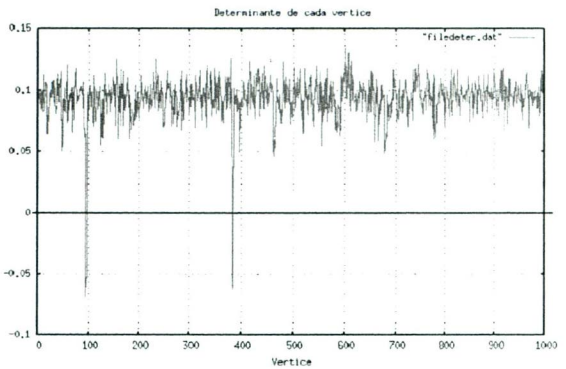


Figura 5.16: *M. jannaschii.*

rangos de variación en las proporciones de grupos de bases no es muy grande, pero presentan un comportamiento particular, en algunos segmentos de la secuencia se disparan las proporciones de bases, haciéndola anómalamente rica en bases G y C.

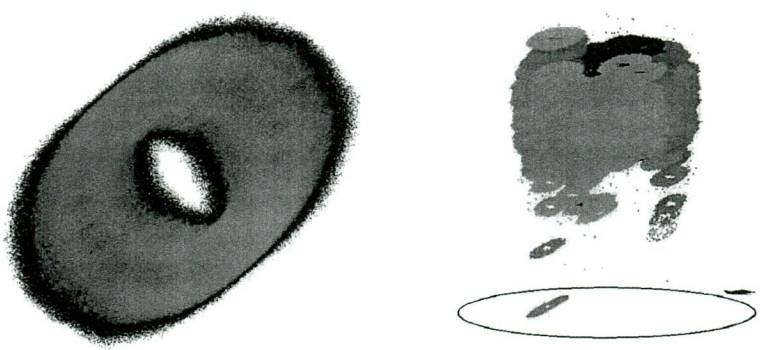


Figura 5.17: *S. solfataricus.* Histograma y superficie en 3D.

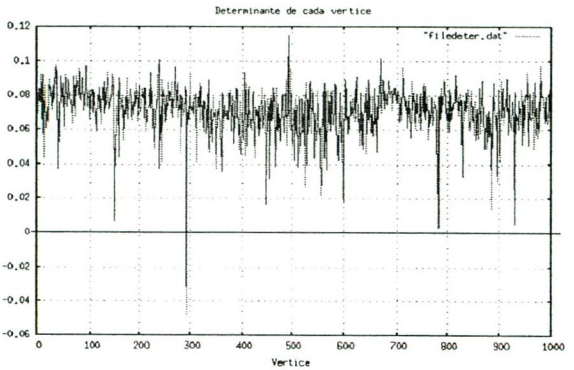


Figura 5.18: *S. solfataricus.*

5.5. Comparación de resultados

Una ventaja del juego circular del caos es que normaliza la información del DNA en una imagen, en el análisis anterior nuestra información tiene magnitudes de orden que va de 1×10^6 pb en algunas bacterias hasta 3×10^7 pb en el cromosma 21 del homo sapiens. Las imágenes de este trabajo tiene un tamaño de 600 x 600 pixeles, pero se puede reducir o aumentar según la precisión que se necesite. Esto nos permite comparar las imágenes y crear dendrogramas que nos haga notar las diferencias y similitudes entre las secuencias de los distintos organismos.

Para encontrar las diferencias y similitudes entre estas imágenes, definimos la distancia entre imágenes, como la distancia euclidiana entre vectores del tamaño de la imagen. Estas imagenes constan de 600x600 pixeles, así que calculamos las distancias entre vectores de tamaño 360,000. Esto nos genera una tabla de distancias simétrica respecto a su diagonal.

Se utiliza el análisis de cúmulos para encontrar las grupos caracterísicos según sus propiedades, el método de clasificación jerárquica que se utilizó para generar el dendrograma es el *método de pesos ponderados.*

En la Figura 5.19 se representa esquemáticamente los grupos generados a partir de la similitud entre sus imágenes, es notable ver que los eucariontes

forman 2 subgrupos ajenos muy distantes entre sí, existe un grupo donde predominan las bacterias estas son *T. maritima*, *E. coli*, *C. trachomatis* y *B. subtilis*, y existe dentro de este grupo la arqueobacteria *T. volcanium*, por otro lado existe un grupo de arqueobacterias con una bacteria presente en el mismo.

La clasificación tiene resultados aceptables, ya que solo se esta considerando una particularidad de la molécula de DNA; como se vio en el Capítulo 2, la molécula de DNA consta de una maquinaria con muchos componentes y de muchos procesos biológicos complejos. Por otro lado la clasificación no es objetivo primordial del trabajo, si no, permitir resaltar las propiedades de las secuencias de DNA para entender el comportamiento interno usando la información menos posible para maximizarla con otros proceso dinámicos, como el juego circular del caos.

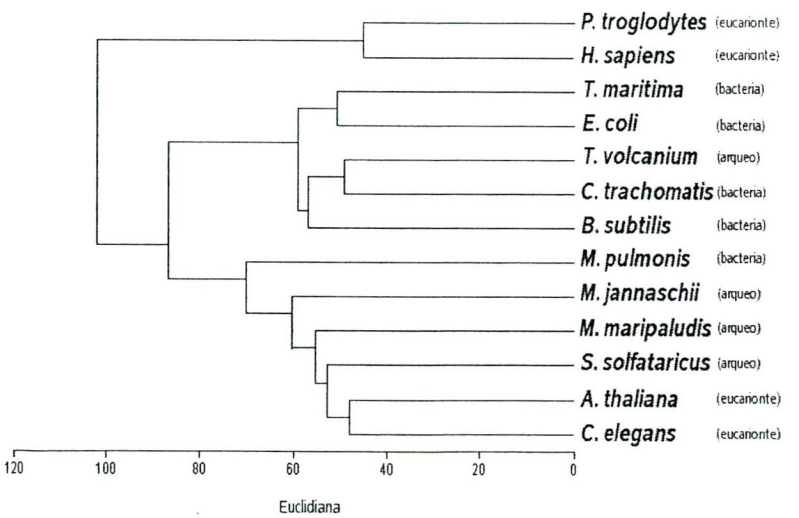


Figura 5.19: Dendrograma generado a partir de la matriz de distancias de las imagens utilizando el análisis de *pesos ponderados*.

Capítulo 6

Discusión

Dentro de la estructura fisicoqímica del DNA existe una gran variedad de características por resaltar. Biológicamente la molécula tiene sus mecanismo para decodificar la información, tiene instrucciones precisas, que a partir de esta se puede construir un organismo. La complejidad de las instrucciones para descodificar esta información involucra una gran cantidad de procesos. En el análisis presentado en este trabajo, se utilizaron estas instrucciones para alimentar un sistema dinámico discreto y esperar un comportamiento que diferenciara las secuencias de DNA.

Como primer punto, se demostró que las secuencias de DNA caen en una región particular en el universo de la complejidad de gramáticas de 4 símbolos, no son azarosas, ni periódicas, ni estadísticamente homogéneas, los niveles de información esta distribuidos sobre la molécula, maximizando su información en ciertas regiones y minimizandola en otras, es la naturaleza de la molecula, para los procesos biológicos estas diferencias deben ser importantes.

Por otro lado, el análisis bioinformático sirvió para mostrar que con un mínimo de información sobre la molécula se puede realizar una diferenciación cualitativamente robusta entre organismos, esto se logró solamente resaltando la estructura fisicoquímica considerando los enlaces dobles y triples que tienen las bases nitrogenadas.

El que en promedio los organimos tengan mayor porcentaje de bases A y T, esto considerando los 3 grupos estudiados, eucariontes, bacterias y arqueobacterias, nos dice que la molécula de DNA debe ser flexible ante cambios y presiones del medio, aunque a veces deben de existir regiones en las que

se necesita que la molécula sea más rígida, como paso en el grupo de las arqueobacterias, posiblemente en esas regiones la molécula debe de ser rígida necesariamente para el buen funcionamiento del organismo, ya que estos organismos viven en medios hostiles y con fuertes presiones del medio. El comportamiento del H. *sapiens* también resulto particularmente interesante, la abundancia de bases G y C al termino del cromosoma 21 debe de obedecer a algún funcionamiento particular de este cromosoma, que posiblemente debe ser más robusto en su parte final debido a algunos procesos que suceden ahí. Con el análisis propuesto solo se pueden hacer conjeturas a cerca de las implicaciones biológicas que pueda tener, se requiere de análisis más profundos y metículosos para explorar las regiones que el anáisis nos muestra como anómalas o diferentes al comportamiento cualitativo que predomina. El análisis sólo nos dice que deben de existir razones del por qué del comportamiento tan particular de las moléculas.

Por otro lado se trabajo únicamente con las frecuencias relativas de los segmentos de DNA, por lo que se perdió la correlación espacial, es decir la posición de cada base A, G, C, T; pero se ganó una descripción estadística muy básica de los segmentos. Bajo este proceso se normalizó la cantidad de datos y se alimentó al IFS. Como se explicó en el Capítulo 4, este sistema ha servido para plasmar huellas digitales de los datos introducidos, esto mismo proceso se realizó para detectar diferencias y similitudes a través de una vista cualitativa y cuantitativa del atractor. Las imagenes generadas nos permitieron conocer algunas particularidades de las cadenas a simple vista.

Se puede conjeturar que éste trabajo se encuentra muy cerca del posible mínimo teórico de información que se puede extraer contando únicamente las frecuencias relativas de las cuatro bases nitrogenadas.

El método propuesto sirve para cualquier cadena compuesta por un alfabeto de 4 símbolos.

Por último, y como trabajo futuro, surge la idea de buscar nuevas formas en la construcción del IFS a partir de las secuencias de DNA; por ejemplo, variando los valores de asignación para las matrices de las transformaciones, otra posibilidad podría ser formar un juego del caos con probabilidades recurrentes, es decir calculando las frecuencias relativas para dos símbolos o más, esto se ha hecho antes, pero no se tiene conocimiento de aplicarlo pa-

ra secuencias de DNA. Otra posibilidad podría ser introducir la correlación espacial como maquina aleatoria para la elección del vértice. .

La visión interdisciplanaria del trabajo permitió conjuntar una serie de herramienas y procesos utilizados en los sistemas dinámicos discretos no lineales, de las ciencias de la computación, en la teoría de la información y en la estadística, el resultado fue la abstracción de un problema en el terreno de la biología abordado y comprendido con el enfoque global de los sistemas complejos.

Bibliografía

[1] Angel Edward, *OpenGL A Primer*, Addison Wesley, United State of America, 2002.

[2] Barnsley Michael F., *Fractals Everywhere, second edition*, Academic Press, Inc.,United States of America, 1993-1998.

[3] Bartle Robert G., Sherbert Donald R., *Introducción al análisis matemático de una variable*, Limusa, México, 1989.

[4] Brown T. A. *Genomes*, Wiley-Liss, United States of America, 1999.

[5] Carreón V. Gustavo, Hernández Z. Jesus E., *Tesis Conjunta de Licenciatura: El juego circular del Caos en el DNA y compresión fractal de imágenes*, Universidad Nacional Autónoma de México, Facultad de Ciencias, México, D.F., 2004.

[6] Carreón V. Gustavo, Hernández Z. Jesus E., Miramontes V. Pedro, *DNA Circular Game of Chaos* en Statistical Physics and Beyond, AIP Conference Proceedings Vol. 757, Mellvile, Nueva York, 2005.

[7] Chinn, W. G., Steenrod N. E., *Primeros conceptos de topología*, Alhambra, Madrid,1975.

[8] Devaney, R. L., *An Introducction to Chaotic Dynamical Systems, second edition*, Addison-Wesley, Redwood city, 1989.

[9] Flake Gary William, *The computational beauty of nature: computer exploration of fractals, chaos, complex systems, and adaptation*, The MIT Press, United States of America, 1998.

[10] Fogel Gary B., Corne David W., *Evolutionary computation in Bioinformatics*, Morgan Kaufmann Publishers, Elsevier, 2003.

[11] Página web *GeneQuiz*, http://jura.ebi.ac.uk:8765/ext-genequiz, 2003.

[12] Gray, Frank, *Pulse code communication*, 1953, U.S. Patente No. 2,632,058

[13] Jiménez Luis Felipe, Merchant Horacio, *Biología celular y molecular*, Prentice Hall, México, 2003.

[14] Joyanes Aguilar Luis, *PROGRAMACION EN C++. Algoritmos, estructuras de datos y objetos*, Mc Graw Hill, España, 2000.

[15] Mandelbrot Benoit B., *Fractals: From, Chance and Dimension*, W. H. Freeman and Co., San Francisco, 1977.

[16] Mandelbrot Benoit B., *Fractal Geometry of Nature*, W.H.Freeman and Co., New York, 1982.

[17] Marsden Jerrold E., Tromba Anthony J., *Calculo Vectorial Cuarta Edición*, Addison Wesley Longman, Estado de México, 1998.

[18] Miramontes P., Medrano L., Cerpa C., et al.,*Structural and thermodynamic properties of DNA uncover different evolutionary histories*, Journal of Molecular Evolution, 40(6):698-704, jun 1995.

[19] Peitgen Heinz-Otto, Jürgens Hartmut, Saupe Dietmar, *Chaos and Fractals New Frontiers of Science*, Springer-Verlag, New York, 1992.

[20] Rudin Walter, *Principles of Mathematical Analysis Third Edition*, Mc Graw Hill, New York, 1976.

[21] Woo Mason, Neider Jackie, et al, *OpenGL Programming Guide Third Edition*, Addison Wesley, United States of America, 1999.

Artículos

[22] Almeida, Jonas S., *et. al.*, *Analysis of genomic sequence by Chaos Game Representation*, Bioinformatics vol. 17 no. 5, 429-437, 2001.

[23] Barnsley M., Demko S., *Iterated Functions System and the global construcction of fractals*, Proceedings of the Royal Society of London, A399, 1985, 243-275.

[24] Basu Soumalee, *et. al*, *Chaos game representation of proteins*, Journal of Molecular Graphics and Modelling 15, 279-289, 1997.

[25] Dutta C. Das J., *Mathematical characterization of Chaos Game Representation. New algorithms for nucleotide sequences analysis*, J. Mol. Biol., 228, 715-719.

[26] Grosse Ivo, Herzel, Hanspeter, Buldyrev Sergey, Stanley H. Eugene, *Species independence of mutual information in coding and noncoding DNA*, Physical Review E, Volumen 61, Número 5, mayo 2000.

[27] Gutiérrez J. M, Rodríguez M.A., Abramson G., *Multifractal analysis of DNA sequences using a novel chaos-game representation*, Physica A 300, pp. 271-284, 2001.

[28] Hutchinson J., *Fractals and Self-Similarity*, Indiana University of Journal of Mathematics, 30, 1981, 713-747.

[29] Jefrey, H. Joel, *Chaos game representation of genetic sequences*, Nucleic Acids Research Vol. 18, No. 8, 2163-2170, 1990.

[30] Jefrey, H. Joel, *Chaos game visualization of sequences*, Computer & Graphics Vol. 16, No. 1, 25-33, 1992.

[31] Joseph Jijoy, Sasikumar Roschen, *Chaos game representation for comparison of whole genomes*, BMC Bioinformatics 2006, 7:243.

[32] Oliver, J.L., Bernaola-Galvan, P., *et. al*, *Entropic profiles of DNA sequences through chaos-game-derived images*, J. Theor. Biol., 160, 457-470

[33] Shannon C. E., *A mathematical theory of communication*, Bell Systems Technical Journal, Vol. 27, 1948, 379-423, 623-656.

[34] Tsuchiya Takashi, *Circular Chaos Game Representation of 1-D Chaos and its relation to the Complex Weierstrass Function*, Internation Journal of Bifurcation and Chaos, Vol 9, No. 10, 1999, 2069-2080.