UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS
INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS
CENTRO DE INVESTIGACIÓN EN CIENCIAS

"*ANÁLISIS ESTADÍSTICO DE SECUENCIAS GENÓMICAS*"

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA

ALBERTO CAMPOS AGUIRRE

DIRECTOR DE TESIS
DR. RAÚL SALGADO GARCIA

CUERNAVACA, MORELOS *NOVIEMBRE 2019*

Índice general

Capítulo 1

Introducción

Es sabido, respecto al ADN o el código genético que existen preguntas para las cuales todavía no se tiene una respuesta que nos permitan comprender con claridad los diversos fenómenos subyacentes de éste.Por ejemplo, existen genomas de diferentes seres vivos cuyo tamaño no guarda proporción con la complejidad del organismo, como por ejemplo una hormiga,la cual puede tener el doble de genoma que una abeja. Las especies de salamandras tienen entre cuatro y treinta y cinco veces más ADN que el ser humano mientras que cierta amabea (protista unicelular del género Amoeba), supera este doscientas veces etc, etc. Las razones de estas evidentes y aparentemente caprichosas diferencias, no se conocen aún completamente pero pueden tratarse de explicar acudiendo a fenómenos biológicos o realizando procesos estadísticos. Lo anterior motiva a proponer un modelo de estudio en base a la apicación de las matemáticas y sus respectivas ramas, de las propiedades que pudieran presentarse desde esta perspectiva y que sin embargo desde el marco teórico de la biología, no pudieran ser analizadas.

Motivados por las características de las secuencias formadas por las bases nitrogenadas (adenina,guanina, citosina y timina) encargadas de codificar para distintas proteínas y así regular las funciones del ser vivo en cuestión, podemos hacer inclución de la estadística y probabilidad , para analizar la forma en que se distribuyen las cadenas de ADN como ACTGCCATG...AT vistas como una secuencia de símbolos .Por ejemplo, atendiendo a este ultimo punto, en [1] se utiliza una medida estadística llamada entropía topológica para estimar la complejidad simbólica de distintas especies desde su genoma. Dicho trabajo encuentra de este modo una diferencia significativa entre distintas regiones del ADN (mas explicitamente llamadas regiones de exónes e introns) que consta en que las primeras están expuestas a una presión mucho mas selectiva que las segundas. Un ejemplo no menos importante en los sucesos que envuelven al código genético es presentado en [2], el cual a través de un mt́odo de correlaciones cuantificables son estudiados los nucleótidos del DNA y en donde es posible observar la ubiquidad de bajas frecuencias de ruido blanco, así como correlaciones a largo alcance y prominentes periodicidades de corto alcance, permitiendo realizar clasificaciones en el banco de datos de diversas especies o familias (primates, vertebrados etc.)

En efecto, en virtud de los ejemplos anteriores sobre las propiedades , el planteamiento del problema en cuestión y las justificaciones estadísticas que envuelven a ciertos procesos del ADN, en el presente trabajo se reitera el análisis a través de un modelo específico y de un concepto o herramienta matemática de vital importancia llamada Función complejidad(fc)

, la cual cuenta el número de palabras de una longitud dada sobre una secuencia de símbolos que descansan en un alfabeto en cuestión, para hacer de la forma mas exhaustiva posible, una comparación directa con las secuencias genómicas reales de diversas especies obtenidas de la base de datos de secuencias genéticas del NIH (*National Institutes of Health*), una colección de disponibilidad pública de secuencias de ADN.

Por supuesto, consecutivamente a la realización del modelo propuesto, se ve si es posible ajustarlo a la realidad, de donde sera se puede dar una explicación parcial de su comportamiento y en el caso contrario, descartar y reajustar sus paramétros para ver si así es ajustable. Como último punto de vital importancia, se podrá estudiar la estructura interna de las secuencias teóricas así como las secuencias reales,la similitud entre estas y la evolución que proveen, aunque evidentemente no obstante al hecho de buscar propiedades estadísticas que nos ayuden a encontrar una relación entre los modelos y los genomas reales, por su puesto es de interés justificarlas de la mano de ciertos aspectos biológicos complementarios.

Capítulo 2

Marco Teórico

2.1. El ADN

Aun que en el presente trabajo no se profundiza sobre aspectos muy específicos ligados al ADN en el sentido biológico, es prudente mencionar ciertos conceptos biológicos que se relacionan con secuencias de símbolos. Esto sera de utilidad más adelante cuando se establezcan los resultados del presente estudio.
Es sabido que las características que identifican a unos individuos de otros,se pasan de los organismos adultos a sus descendientes durante la reproducción. Dicho contenido de información se halla en una molécula llamada ácido desoxirribonucleico (ADN), la cual contiene las instrucciones biológicas que hacen de cada individuo algo único.
El ADN está formado por unos componentes químicos básicos denominados nucleótidos. Estos componentes incluyen un grupo fosfato, un grupo de azúcar y una de cuatro tipos de bases nitrogenadas alternativas. Para formar una hebra de ADN, los nucleótidos se unen formando cadenas, alternando con los grupos de fosfato y azúcar. Los cuatro tipos de bases nitrogenadas encontradas en los nucleótidos son: adenina (A), timina (T), guanina (G) y citosina (C). El orden, o secuencia, de estas bases determina que instrucciones biológicas están contenidas en una hebra de ADN. Por ejemplo, la secuencia ATCGTT pudiera dar instrucciones para ojos azules, mientras que ATCGCT pudiera indicar ojos de color café. Un gen es trozo de ADN que contiene la información necesaria para la síntesis de una molécula con una función especifica, habitualmente una proteína. Esta función puede estar vinculada con el desarrollo o funcionamiento de una función fisiológica. El gen es considerado la unidad de almacenamiento de información y herencia genética, pues transmite esta a la descendencia.

En el caso de los seres humanos la colección completa de ADN, o el genoma humano, consta de 3 mil millones de bases organizados en 23 pares de cromosomas, y conteniendo alrededor de 20,000 a 25,000 genes.

Una secuencia de ADN que contiene las instrucciones para elaborar una proteína se conoce

como gen. El tamaño de un gen puede variar desde aproximadamente 1,000 bases hasta 1 millón de bases en los seres humanos. Los genes solo forman aproximadamente el 1 por ciento de la secuencia de ADN. Otras secuencias reguladoras de ADN dictan cuándo, cómo y en que cantidad se elabora cada proteína. La mayoría de las secuencias del genoma humano no tienen una función conocida. Debido a la naturaleza altamente específica de este tipo de emparejamiento químico, la base A siempre forma pareja con la base T y, asimismo, la C con la G.

2.1.1. Intrones y Exónes

En 1977 Philip Sharp y Richard Roberts probaron por separado que los genes que codifican polipéptidos en las eucariotas se encuentran interrumpidos por secuencias no codificantes . En efecto, en las eucariotas (organismo un poco mas complejos con el ADN dentro de un núcleo) los genes tienen dos tipos de regiones: los exónes o regiones codificantes y los intrones o regiones no codificantes. De hecho las proteínas se codifican solo en los exónes, por lo tanto los intrones serán eliminados antes de que la información se convierta en proteína (expresión génica). Las procariotas, organismos mas simples sin núcleo, carecen de intrones.

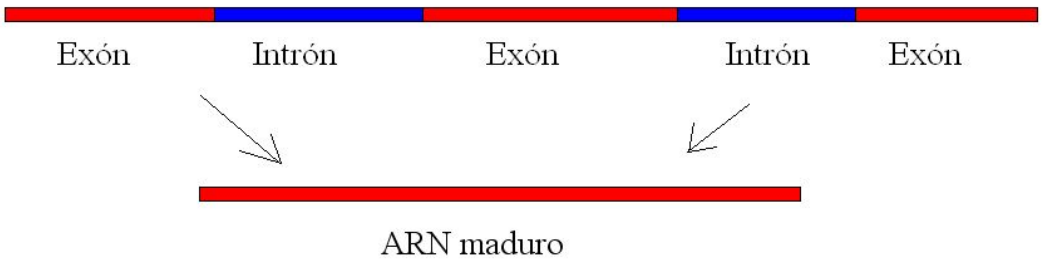

Figura 2.1: Esquema sencillo que representa un fragmento de la cadena de ADN la cual se subdivide entre dos regiones llamadas exónes e intrones.

Finalmente introducimos un aspecto de vital importancia exclusivo del ADN y que en el presente trabajo se hace alusión principalmente en los modelos evolutivos, se trata de las mutaciones

Mutaciones

Básicamente, las mutaciones pueden ser por sustituciones (que es cuando cambiamos uno de los nucleótidos por algún otro). De forma general, las sustituciones se denominan *transiciones* si suponen un cambio entre las bases nitrogenadas del mismo tipo químico, o *transversiones* si son un cambio de las purinas purina $\{A, G\} \rightarrow$ pirimidina $\{C, T\}$ o pirimidina $\rightarrow$ purina. Además, tenemos las deleciones o inserciones, las cuales son respectivamente la eliminación o adición de una determinada secuencia de nucleótidos, de longitud variable. Las grandes deleciones pueden afectar incluso a varios genes, hasta el punto de ser lo suficientemente grandes para apreciarse a nivel cromosomico con diversas técnicas de citogenética. Inserciones o deleciones de unas pocas pares de bases en una secuencia codificante pueden provocar desplazamiento del marco de lectura (o también llamado *frameshift*), de modo que la secuencia de nucleótidos del ARNm se lee de manera incorrecta. En general las mutaciones génicas afectan el genoma de la siguiente forma:

ADN codificante

Si el cambio en un nucleótido provoca en cambio de un aminoácido de la proteína la mutación se denomina no sinónima. En caso contrario se denominan sinónimas o silenciosas (posible porque el código genético es degenerado). Las mutaciones no sinónimas asimismo se clasifican en mutaciones con cambio de sentido si provocan el cambio de un aminoácido por otro, mutaciones sin sentido si cambian un codón codificante por un codón de parada (TAA, TAG, TGA) o con ganancia de sentido si sucede a la inversa.

ADN no codificante

Pueden afectar a secuencias reguladoras, promotoras o implicadas en el ajuste (*splicing*). Estas últimas pueden causar un erróneo procesamiento del ARNm, con consecuencias diversas en la expresión de la proteína codificada por ese gen.

2.2. El ADN y su estudio estadístico

Hemos hablado de ciertos estudios estadísticos los cuales envuelven al ADN visto como un proceso en términos de la dinámica simbólica. Por ejemplo, solo como un ejercicio sutil, si nosotros viéramos la correlación (a) entre los distintos monómeros o símbolos que descansan en el alfabeto típico de una muestra de de ADN ($\{A, C, T, G, \}$), tendríamos de forma lo que se muestra en la gráfica de la figura 2.2 .
Para este procedimiento hemos elaborado por una parte, un archivo generado aleatoriamente el cual descansa en el alfabeto $A = \{0, 1\}$ y con un tamaño finito de seis millones de símbolos. Por otra parte, elegimos una muestra arbitraria del mismo tamaño al primero pero obtenida

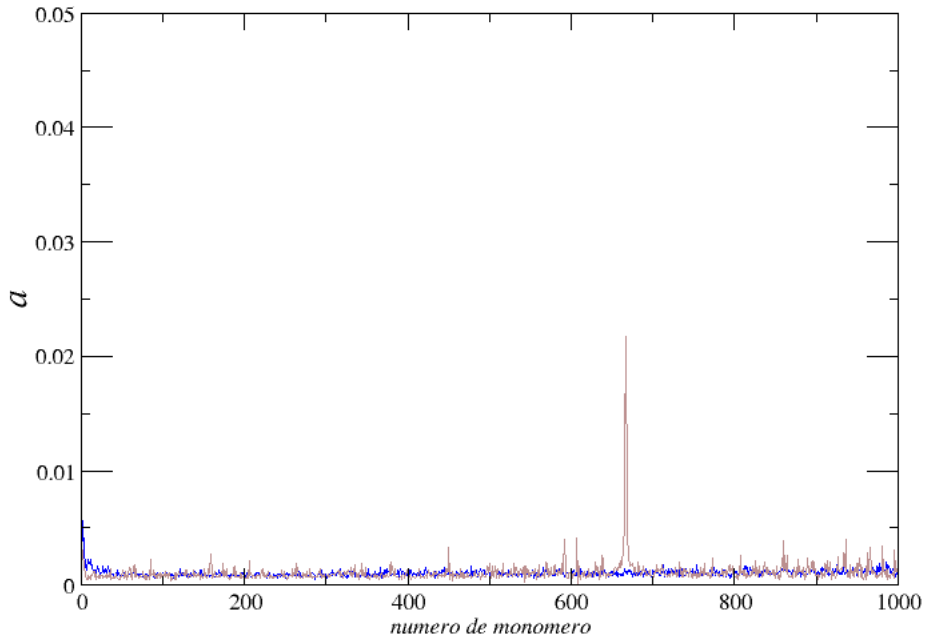

Figura 2.2: Gráfica que muestra la superposición de la transformada de Fourier de la correlación a contra la distancia l, entre los monómeros de una secuencia generada aleatoriamente y una muestra arbitraria del genoma de una especie.

del banco de datos genético.A continuación etiquetamos esta muestra de tal forma que un par de bases nitrogenadas (que llamaremos xx') sean reescritas como

$$xx' = \begin{cases} 0 & \text{si xx' = AG} \\ 1 & \text{si xx' = TC} \end{cases}.$$

En efecto, esto producirá una secuencia binaria por ejemplo como 10010100.... etc. Ahora procedemos a aplicar la formula de correlación dada por

$$a_n = \frac{\frac{1}{L-1}\sum a_n a_{n+1} - L^2 \sum a_n \sum a_{n+1}}{\frac{1}{L-1}\sum a_n^2 - \frac{1}{L^2}(\sum a_n)^2} \tag{2.1}$$

para los dos tipos de secuencias que hemos especificado y en donde L se toma como la longitud o distancia entre un monómero a otro y n, se trata de la etiqueta que designa a cada uno de esos símbolos o monómeros.
Es posible ver que la secuencia que fue elaborada aleatoriamente (linea de color azul) y la linea que corresponde a la secuencia tomada de la muestra real de un genoma (linea de color marrón), tienen una diferencia sutil y que consiste en un pico entre la distancia del primer monómero, y el monómero del intervalo entre seiscientos y ochocientos para la secuencia real, lo cual resulta en una evidente correlación .
Profundizando un poco mas en la descripción vista en la introducción de este trabajo, se menciono que en efecto es posible hacer una clasificación entre diversas especies aplicando

distintos conceptos estadísticos y reformulando una cadena genética en términos de secuencias simbólicas.En [3] por ejemplo, se discute el uso de la entropía topológica para analizar secuencias simbólicas que descansan en el alfabeto correspondiente a las bases nitrogenadas del ADN. Uno de los métodos propuestos se basa en lo siguiente. Dejando ser ω una secuencia formada con el alfabeto $\{A, C, T, G\}$ la entropía topológica resulta ser

$$H_T(\omega) = \lim_{n\to\infty} \frac{log_4(p_\omega(n))}{n} \tag{2.2}$$

donde $p_\omega(n)$ se trata del numero de diferentes n-longitud de sub-palabras que aparecen en ω. Por ejemplo, sea la secuencia simbólica $\omega = CGGGGGG...$, es fácil observar que las diferentes secuencias de longitud n es 2,así encontraremos que la correspondiente entropía topológica de esta secuencia de DNA sera :

$$H_T(\omega) = \lim_{n\to\infty} \frac{2}{n} = 0 \tag{2.3}$$

de esta forma la la entropía topológica es cero como n tiende a infinito. Utilizando este concepto ademas de otras formas de entropía como el método de Koslicki en [1] se logra estudiar esta propiedad de los cromosomas del homosapiens para las regiones que contienen exónes, así como para aquellas donde tenemos intrones.

En [4] tenemos un tipo de clasificación sobre estas regiones (exónes e intrones) en base a un método basado en caminatas aleatorias. Se define entonces, un desplazamiento neto $y(l) = \sum_{i=1}^{l} u(i)$ el cual es la suma de los pasos de un caminante regido bajo la siguiente regla: si ocurre una pirimidina entonces nos desplazaremos $u(i) = +1$, y en el caso contrario de encontrar una purina, el desplazamiento sera $u(i) = -1$ todo esto sobre una secuencia simbólica o cadena de ADN. Posteriormente, definimos a $F(l)$ como las fluctuaciones del promedio del desplazamiento. Finalmente, en [4] se estudia los posibles tipos de comportamiento de F, dentro de los cuales existe uno que resulta de interés y el cual sigue a una ley de potencia tal que

$$F(l) \sim l^\alpha \tag{2.4}$$

La virtud de este ultimo ejemplo en concreto, es el hecho de que al aplicar (2.4) a secuencias ricas en intrones o exones de una misma especie, obtenemos un α que distingue a cada región pudiendo clasificar estadísticamente estas.

Más aun , C.-K. Peng y colaboradores en [5] como pregunta abierta a si las correlaciones a largo alcance están presentes en secuencias codificantes y no codificantes en el ADN, observaron que tras el análisis de 33301 secuencias de exónes y 29453 secuencias de intrones tomadas de las células eucariotas, las primeras presentan prácticamente no correlación alguna mientras que las segundas si que lo presentan, todo esto aplicando métodos matemáticos como la transformada estándar de Fourier. Ademas, en [5] para 874 secuencias codificantes y 1157 secuencias no codificantes con una longitud de más de 4096 de pares de bases nitrogenadas, se encontró que estas obedecen a un comportamiento regido por una ley de potencia en concordancia con (2.4).

2.3. La Función complejidad

Resulta conveniente profundizar matemáticamente desde el punto de vista de la dinámica simbólica. La función complejidad se puede ver como una medida clásica de desorden,para secuencias simbólicas de tamaño finito o infinito.Sin embargo, se introduce una descripción formal sobre el concepto de secuencia de la siguiente forma.

Espacios de secuencias

A continuación se propone una serie de definiciones básicas con el objetivo de ampliar la noción de secuencias simbólicas. Un alfabeto se trata de un conjunto finito, en donde cada elemento es designado como símbolo o letra.Un bloque o también una palabra sobre un alfabeto A no se tarta mas que de una sucesión finita de letras de A.Ahora, formalmente, una palabra u sobre A es una función u del conjunto $\{x \in \aleph : 1 \leq x \leq n\}$ en A (para algún entero positivo $n \geq 1$).Para cuando tenemos que $n = 0$, entonces se tendrá una función vacía o palabra vacía. La longitud o tamaño de la palabra en efecto sera n, es decir, la cantidad de términos de la sucesión y sera denotada por $\mid u \mid$. De ahora en adelante diremos que para una palabra no vacía u, el $k - esimo$ termino se denotara mediante u_k. Entonces, una palabra no vacía de largo n sobre A es $u_1u_2u_3....u_n$ con $u_i \in A$ para todo $i \in \{1, ..., n\}$. A su vez, un bloque de longitud n se llamara un $n - bloque$, dos palabras serán idénticas si tienen el mismo largo n y, para todo j entre 1 y n, entonces $u_j = v_j$.

El conjunto de todas las palabras que tienen de largo n sobre un alfabeto A se designa por A^n, y el conjunto de todas las palabras sobre A se designa por A^*. Es decir, $A^* = \cup_{n=0}^{\infty} A^n$. Notemos que para el caso cuando ≥ 0, A^n es un conjunto finito, en tanto que A^* es un conjunto que tiene una cantidad infinita de elementos.

Sea el alfabeto binario (y fundamental para la descripción de los modelos propuestos en el presente trabajo)$A = \{0, 1\}$, las palabras que pueden formarse a travéz de el son llamadas sucesiones binarias. Por ejemplo tenemos que A^0 es la sucesión cero. $A^1 = \{0, 1\}$, $A^2 = \{00, 01, 10, 11\}$, $A^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$, etc.Ademas, $A^* = \{\varepsilon, 0, 1, 00, 01, 10, 11\}$ en donde ε denota la palabra vacía.

Una vez definido los conceptos básicos que comprenden una secuencia simbólica, se enuncia la definición y características de los espacios de corrimiento completos o en su análogo los Full-shifts . Dado un alfabeto A, llamamos $full - shift$ sobre el alfabeto A, al conjunto de todas las funciones de Z en A. El full shift sobre A es denotado por A^Z.Una función de Z en A no es mas que una sucesión bi-infinita de símbolos de A, que puede escribirse de la siguiente forma:

$$x = (x_n)_{n \in Z} = ...x_{-1}x_0x_1x_2....$$

Como ejemplo de estas sucesiones bi-infinitas para $A = \{0, 1\}$ es la que por supuesto descansa el cero en sus términos impares y 1 en sus pares, es decir: ...,01010101010.....

es importante mencionar que A^Z tiene como elemento el nombre de puntos, los cuales constan de sucesiones bi-infinitas y ademas, que A^Z y A^* no tienen ningún elemento en común.
Dado $x = x_{-1}x_0x_1x_2...\varepsilon A^Z$, designamos por $x_{[i,j]}$ a la sucesión (finita) de los símbolos de x que van desde la coordenada i hasta j, ambas incluso.Es decir, $x_{[i,j]} = x_ix_{i+1}....x_{j-1}x_j$. Se adopta la convención de que para $i > j$, $x_{[i,j]} = \varepsilon$. Se denota por $x_{[i,\infty)}$ a la sucesión infinita a derecha mientras que $x_{(-\infty,i]}$ es la sucesión infinita a izquierda.
Una vez visto las principales características que envuelven a una secuencia simbólica, Procedemos entonces a formalizar los Espacios Shift a continuación.

DEFINICIÓN 2.1. Sea $\wp$ una colección de bloques sobre un alfabeto A, es decir, $\wp \subseteq A^*$. Designamos por $\chi_\wp$ al subconjunto de A^Z formado por todos aquellos puntos en los que no ocurre ningún bloque de $\wp$, Es decir

$$\chi_\wp = \left\{x \in A^Z : \forall f \in \wp, f \text{ no curre en x}\right\}$$

o también

$$\chi_\wp = \left\{x \in A^Z : \forall i, j \in Z, x_{[i,j]} \notin \wp\right\}$$

DEFINICIÓN 2.2. Un espacio shift sobre el alfabeto A es un conjunto $X \subseteq A^Z$ tal que existe un $\wp \subseteq A^*$ tal que $X = \chi_\wp$
Se piensa en general a $\wp$ como un conjunto de **bloques prohibidos** para X.

Lenguaje

Un lenguaje $\mathcal{L}$ sobre un alfabeto A es cualquier subconjunto de $A*$, es decir, el conjunto de todas las palabras de $A*$ que no pertenecen a $\mathcal{L}$. Si $\mathcal{L}$ se trata de un lenguaje entonces, se designara por $\mathcal{L}^c$ a su complemento en A^* , es decir, el conjunto de todas las palabras de A^* que no pertenecen a $\mathcal{L}$

Definamos pues, en base a los conceptos fundamentales de la dinámica simbólica previos, el concepto de Función Complejidad. Por conveniencia, dejemos ser caracterizada a una secuencia o cadena de símbolos de ahora en adelante como:

$$X := x_0x_1x_2x_3.....x_n \tag{2.5}$$

Esta cadena puede ser tanto finita (que particularmente la escribimos así en (2.1)) como infinita teniendo solamente una restricción sobre la misma y esto es , que le pediremos descansar sobre un alfabeto o un conjunto de caracteres de tamaõ finito.
Entonces, si designamos a la función $C_X(l)$ de un integrando positivo l como el numero de diferentes factores (o también como sub-cadenas consecutivas distintas) de longitud l de la cadena X, tendremos que la *Función Complejidad (fc)* de X con un alfabeto finito de tamaño k estará dada por

$$1 \leqslant C_X(l) \leqslant k^l \tag{2.6}$$

2.4. El estimador para la función complejidad

hemos mencionado las características principales de la función complejidad así como definido el concepto de secuencias simbólicas y los componentes que la conforman. puesto que vamos a proceder a realizar el calculo de las distintas palabras de longitud l sobre secuencias de aproximadamente un rango entre $2,000,000 \leq s \leq 250,000,000$ (en donde s es el tamaño de esta), se generaría un problema computacionalmente hablando en términos de tiempo,por lo cual es necesario implementar una aproximación que nos permita estimar la diversidad de palabras de esa longitud. Se introduce entonces el estimador de la función complejidad como sigue.

El genoma puede ser estudiado e interpretado como el resultado de un proceso estacionario y descrito por una medida que llamaremos $\mathbb{P}$. El respaldo de $\mathbb{P}$ es un sistema simbólico $X \subset A^{\mathbb{N}}$ caracterizado por su lenguaje $\mathcal{L}(X)$, esto es, el conjunto de todas las l-palabras ocurriendo como sub-bloques de secuencias admisibles sobre X. Si denotamos por $\mathcal{L}_l(X)$ el conjunto de todas las l-palabras ocurriendo como sub-bloques de secuencias admisibles sobre X, entonces la Función Complejidad de X es la función que asocia la cardinalidad de $\mathcal{L}_l(X)$ con cada $l \in \mathbb{N}$. Dentro de este marco de trabajo, el genoma de una especie o individuo puede ser considerado como la observación de una muestra finita aleatoria con finita precision, donde a su vez estará representada como una secuencia simple. En cualquier caso asumiremos que la muestra aleatoria sera típica con respecto de $\mathbb{P}$, la cual sera asumida ergodica con respecto al mapeo de corrimiento (shift map). Entonces, dada una muestra s con l-palabras no necesariamente deferentes, como fue sugerido al principio de este capitulo, estimaremos la función complejidad como $C(l)$.

Consideremos entonces una secuencia completa de ADN correspondiente a una especie cualquiera. Si nosotros quisiéramos saber el numero de palabras distintas dentro de esta secuencia, podemos reinterpretar el problema sutilmente de la siguiente manera. Proponemos una especie de urna la cual contendrá C bolas. Cada una de estas bolas están etiquetadas y ademas todas ellas serán explícitamente diferentes la una de la otra. El siguiente ejercicio el cual se trata de un experimento aleatorio, sera tomar M bolas de la urna estrictamente con remplazo,de tal forma que el espacio muestral sera:

$$\Omega = \{\omega = (\omega_1, \omega_1, \omega_1) : \omega_i \in \{1, 2, ...C\}\} \tag{2.7}$$

Por ejemplo, ω puede ser una colección dada por $\omega = (1, 1, 1, ..., 1)$ etc,etc. Proponemos

entonces dado el experimento anterior, la siguiente variable aleatoria.
sea

$$X = \sharp \text{ de bolas diferentes en la extraccion} \tag{2.8}$$

la variable aleatoria X la cual sera el numero de diferentes palabras de longitud l o la diversidad de pelotas obtenidas de la muestra tendrá como posibles valores por supuesto al conjunto $\{1, 2, 3, ..., min\{M, C\}\}$. Ahora, para nuestro problema en particular, la muestra de tamaño M extraída del DNA sera tal que: $M \ll C$

Asumiendo este principio, tendremos lo siguiente. $C(l)$ se trata del numero de l-diferentes palabras exactas de una secuencia (cantidad no conocida), o el numero de diferentes bolas contenidas en la urna.Puesto que no sabemos cuantos objetos hay dentro de esta urna y, ademas, cada uno de los objetos dentro de ella tiene la misma probabilidad de ocurrir, nuestro objetivo sera definir a la variable X que nos ayude a estimar el valor de $C(l)$.

Evidentemente la variable aleatoria X tiene una función de distribución la cual queremos o estamos interesados en calcular. Entonces sea $f(x) = \mathbb{P}(X = x)$, tras una serie de cálculos y, ademas $M < C$, desarrollando tendremos que la función de distribución obedecerá a:

$$f(x) = \frac{1}{H} \left(\begin{array}{c} M-1 \\ x-1 \end{array} \right) \left(\begin{array}{c} C \\ x \end{array} \right)$$

donde

$$H = \binom{C+M-1}{M}$$

Donde H se trata del numero total de diferentes subconjuntos de cardinalidad M que pueden ser compuestos de un conjunto de C diferentes objetos (donde ademas cada elemento puede ser repetido cuantas veces sea necesario).
Por lo tanto sustituyendo H en $f(x)$ es claro que tendremos

$$f(x) = \left(\begin{array}{c} M-1 \\ x-1 \end{array} \right) \left(\begin{array}{c} C \\ x \end{array} \right) \left(\begin{array}{c} C+M-1 \\ M \end{array} \right)^{-1}$$

.
Por supuesto podemos tomar el valor esperado de nuestra variable aleatoria que quedara en función de la función complejidad teórica C y el tamaño de la muestra M, entonces el valor esperado de X es fácilmente obtenido por

$$\mathbb{E}(X) = \frac{1}{H} \sum_{x=1}^{M} x \left(\begin{array}{c} M-1 \\ x-1 \end{array} \right) \left(\begin{array}{c} C \\ x \end{array} \right)$$

$$= \frac{C}{H} \sum_{x=1}^{M} x \left(\begin{array}{c} M-1 \\ x-1 \end{array} \right) \left(\begin{array}{c} C-1 \\ x-1 \end{array} \right)$$

$$= C \begin{pmatrix} C+M-2 \\ M-1 \end{pmatrix} \begin{pmatrix} C+M-1 \\ M \end{pmatrix}^{-1}$$

o finalmente

$$\mathbb{E}(X) = \frac{CM}{C+M-1} \tag{2.9}$$

Un desarrollo mas análitico de este estimador es posible verse en el apéndice A de [6]
La ecuación (2.6) muestra que $\mathbb{E}(X) \to C(l)$ cuando $M \to \infty$. Un conteo directo de de las diferentes l-palabras en la muestra M no es un buen estimador de la Función Complejidad y, Ademas, ya que $E[X] < C$ para todas las muestras de tamaño M, es posible observar que X no es un estimador imparcial para C. Es posible demostrar por medio de la varianza que X es pequeña cuando $C \gg s$. Esto significa que a mas realizaciones de X resulta un valor que no se desvía significativamente de $E[X]$ debido a las fluctuaciones aleatorias.Por lo tanto, en el caso que C sea mas grande que la muestra de tamaño M, entonces la variable aleatoria X nos daría información precisa acerca de C, la cual podemos extraer de $E[X]$ debido a las pequeñas fluctuaciones aleatorias. Entonces, despejando de (2.6) a C obtenemos que

$$C = \frac{(M-1)\mathbb{E}[X]}{M - \mathbb{E}[X]} \tag{2.10}$$

Se propone entonces de acuerdo al razonamiento anterior para encontrar a C, el siguiente estimador dado por:

$$C = \frac{M\mathbb{E}[X]}{M+1-\mathbb{E}[X]} \tag{2.11}$$

Finalmente escribiremos para nuestros fines el estimador de la función complejidad atendiendo la ecuación (2.8) de tal forma que

$$C(l) = \frac{Mk(l)}{M+1-k(l)} \tag{2.12}$$

De ahora en adelante nos concentraremos en (2.9) para efectuar una estimación de la función complejidad. Para evitar posibles confusiones denotaremos a $k(l)$ como el numero de palabras distintas o diversidad de palabras de longitud l, de la muestra seleccionada de tamaño M.
Para una mejor apresiación de como se ve el estimador (2.12), se muestra en la figura (2.3) una gráfica en donde superponemos el estimador de la función complejidad $C(l)$(hasta una $l = 50$), la función complejidad teórica $c(l)$ y el numero de palabras distintas de la muestra $M = 100,000$ que esta representado por $k(l)$.

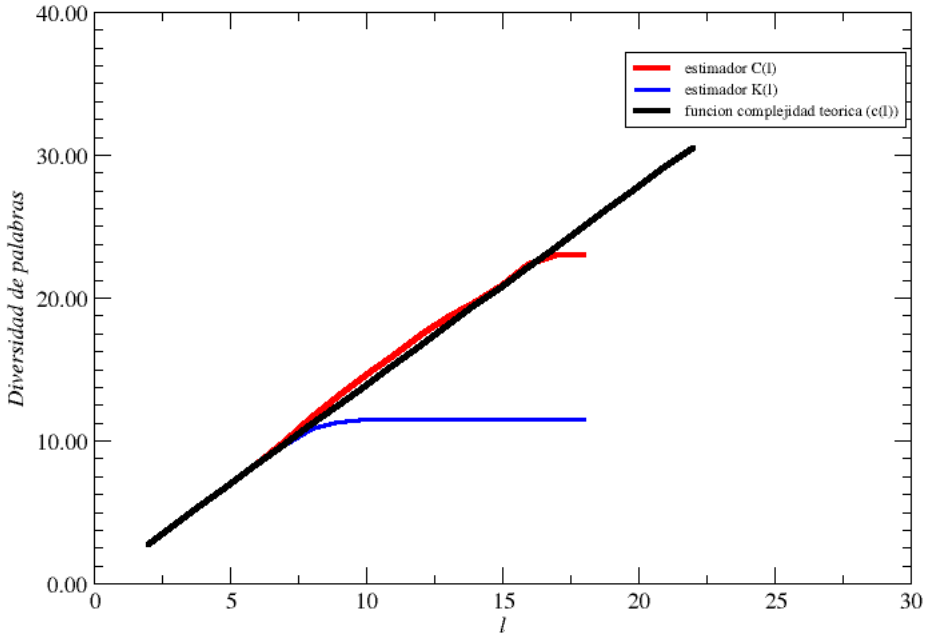

Figura 2.3: Diferencia gráfica entre el estimador $C(l)$, la fc teórica y el numero de palabras distintas k(l) para una secuencia simbólica s de tamaño 6,000,000 y calculado todo hasta una longitud l=50.

Capítulo 3

Los Modelos de Evolución Genómica

Para la construcción de la comparativa entre las secuencias genómicas ficticias generadas y las secuencias reales, se utilizaron tres modelos de evolución Genónima, sin embargo existen más [6] de los cuales se proporcionan las referencias citadas. Se hace un enfoque a través de los modelos: Expansión-Modificación, Zacks y Massip-Arndt. Los dos primeros de ellos obedecen a secuencias simbólicas generadas a partir de un alfabeto binario (0 y 1), y la última se trata de una secuencia que descansa sobre el alfabeto $\{A, C, T, G\}$ que es generada aleatoriamente y de la cual es sujeta a distintos procesos estadísticos. Enunciemos detalladamente los mencionados a continuación.

3.1. Modelo de Expansión Modificación

Uno de los Modelos mas tempranos para simular secuencias Genómicas reales fue el originalmente propuesto por Li[6]. Este sistema fue propuesto para comprender el origen de las correlaciones de largo alcance que también pueden ser observadas en cadenas o secuencias de genomas reales. Es importante señalar que existen ciertas características encontradas a través de este que resulta de sutileza observar. Esto es que Expansión - Modificación tiene una única medida estacionaria que es alcanzada por una condición inicial,que es respaldada por el llamado full-shift (o espacio de corrimiento completo), los cual significa que tenemos todas las l-palabras típicamente presentes en la secuencia generada por este sistema.De la misma forma es posible observar que la función de correlación obtenida tiende a decaer para casi todos los parámetros a los que la secuencia simbólica es envuelta.

El modelo de Expansión Modificación básicamente responde a dos procesos característicos bien definidos. Uno de ellos es el de Expansión y el otro, el de Modificación. Primeramente, se proporciona una semilla obtenida del vocabulario o el alfabeto finito Binario (esto es, entre 0 y 1).De forma consecutiva, podemos expandir esta semilla con una probabilidad p, o podemos modificarla remplazando por el símbolo complementario del alfabeto con una probabilidad de $1 - p$. Se ilustra con un ejemplo para una realización del modelo.

Ejemplo de una realización:

Supongamos que la semilla dentro del alfabeto dado por los símbolos $A := \{0, 1\}$ esta dado y que ademas en este caso optamos por el símbolo 0 . Entonces, suponiendo que tenemos una probabilidad p de ser expandida (3.1) y una probabilidad $1 - p$ de ser modificada,y ademas que ocurre un evento caracterizado por una determinada variable aleatoria la cual tiene como espacio muestral el alfabeto mencionado anteriormente para seleccionar, después de dicho evento y tras una sola iteración podemos tener por ejemplo la secuencia dictada por:

$$t_0 = 0.$$
$$t_1 = 1.$$

Donde hemos colocado en lugar de la semilla (dada por el tiempo o indice t_0), el siguiente símbolo tras el proceso estocástico o iteración correspondiente que sufrió el carácter inicial (dado por t_1). Evidentemente en el ejercicio hemos optado por escojer la opción de modificar la semilla 0 a 1 sin ninguna particularidad para notar como es este , y como concatenarlo a medida de que vamos obteniendo mas de estos. Ahora, por cada iteración se procede a realizar el mismo experimento para cada uno de los elementos de nuestra cadena actualizada;es decir, que cada uno de estos tendrán que someterse a la regla dictada por la ecuación (3.1).

$$x = \begin{cases} 0 & \mathbf{p} \\ 1 & \mathbf{1\text{-}p} \end{cases} \tag{3.1}$$

Por ejemplo, para $t_1 = 1$ (que se caracteriza con el símbolo x)del paso anterior como nos indica (3.1) podemos expandirlo con una probabilidad $1 - p$, de tal forma que para una iteración t_2 tendremos:

$$t_0 = 0.$$
$$t_1 = 1.$$
$$t_2 = 11.$$

y sucesivamente para un t_3 cada uno de los símbolos de t_2 podrá expandirse o modificarse según las condiciones forjadas, de tal forma que podemos realizar hasta n-número de ciclos para formar estructuras lo suficientemente grandes para su análisis posterior,es decir obtener secuencias como:

$$
\begin{aligned}
t_0 &= 0 \\
t_1 &= 1 \\
t_2 &= 11 \\
t_3 &= 00 \\
t_4 &= 0000 \\
t_5 &= 0000100 \\
t_6 &= 0000000011001 \\
&\vdots \\
t_n &= 01001100110010011.....
\end{aligned}
$$

De forma casi directa se observa que los parámetros que entran en juego para este modelo son tanto como la elección aproximada de la longitud de la secuencia simbólica, así como la probabilidad circunstancial que obedece a una distribución tipo Bernoulli p. Esto último es de una radical importancia, ya que precisamente procedemos asignar distintos valores computacionalmente hablando para observar tanto los factores de repetición de una palabra en cuestión, y las mutaciones engendradas en el transcurso de las iteraciones.

3.2. Modelo de Zacks

Pasemos al modelo de evolución propuesto por Zacks [6].Este puede verse como un sistema sustituto el cual es perturbado aleatoriamente.La forma más simple de una secuencia simbólica que podemos obtener del modelo de Zacks (es decir aquella en donde el sistema no sufre perturbaciones) se trata de la secuencia de Trhue-Morse, la cual se caracteriza por tener una función de complejidad lineal.Evidentemente cuando nosotros sometemos esta última a procesos aleatorios se presenta una modificación sobre la original lo cual implica que la complejidad de esta de igual forma lo hará. Resulta de utilidad mencionar que en las secuencias de ADN podemos encontrar no solo correlaciones de de largo alcance si no también la periodicidad de base tres [6] las cuales pueden estudiarse con este modelo.

Este sistema básicamente consta de lo siguientes pasos: sea la regla dictada por la relación (3.2) por cada paso de tiempo o por cada iteracion que hagamos sobre el sistema.

$$
x = \begin{cases} xx & \textbf{p} \\ xx' & \textbf{1-p} \end{cases} \tag{3.2}
$$

En donde el símbolo x o puede duplicarse con una probabilidad p, o puede mutar con una probabilidad $1-p$ siendo x' el carácter complementario del alfabeto mencionado. Una realización como ejemplo para observar el funcionamiento del presente sistema se expone a continuación.

Realización del Modelo de Zacks

Sea el alfabeto binario caracterizado por $A := \{0, 1\}$, entonces una posible realización puede ser una probabilidad $p = 1$ de duplicación, y una $q = p - 1 = 0$ de mutación. De esta forma por ejemplo si elegimos de manera arbitraria el símbolo 0 como semilla al tiempo t_0, entonces a un primer tiempo t_1 tendremos que:

$$0 = \begin{cases} 00 & \mathbf{p} \\ 01 & \mathbf{1\text{-}p} \end{cases} \tag{3.3}$$

y por lo tanto:

$$t_0 = 0.$$
$$t_1 = 00.$$

como la probabilidad de tener doble ceros en este sistema es de uno, para un tiempo t_2 debemos aplicar la regla (3.3) sobre cada símbolo de t_1, de tal forma que la resultante sera $t_2 = 0000$.En efecto si realizamos hasta la n-esima iteracion tendremos la secuencia obviada $t_n = 000000....,00_n$,sin embargo al invertir las probabilidades de tal forma que $p = 0$ y $q = 1-p = 1$ sobre la misma semilla,entonces para el primer tiempo tendremos que $t_1 = 01$,y para t_2 tendremos que perturbar cada uno de los símbolos del tiempo anterior.Para cero es claro por t_1 como evoluciona, mientras que para 1 la regla simplemente se invierte como:

$$1 = \begin{cases} 11 & \mathbf{p} \\ 10 & \mathbf{1\text{-}p} \end{cases} \tag{3.4}$$

y así tendremos que $t_2 = 1001$. Evidentemente podemos llevar hasta n realizaciones con la misma probabilidad siempre fija, y los resultados que obtendremos serán una sucesión simbólica de complejidad lineal conocida como Trhue-Morse y de los cuales sus primeras iteraciones corresponden a:

$$
\begin{aligned}
t_0 &= 1 \\
t_1 &= 10 \\
t_2 &= 1001 \\
t_3 &= 10010110 \\
t_4 &= 1001011001101001 \\
t_5 &= 10010110011010010110100110010110 \\
&\vdots \\
t_n &= 1001011001101001.....
\end{aligned}
$$

Resulta de gran interés, observar dadas las características que se presentan en el siste de Morse con una función de complejidad lineal, el cambio precisamente al hacer ligeras perturbaciones, hasta una variación de los parámetros significativamente drstica. Al igual que en el Modelo de Expansión-Modificación, todo el tiempo se tiene un control sobre el numero de iteraciones, y por supuesto los valores probabilísticos que encaran el concepto de duplicación y Mutación, procesos que en definitiva ocurren en las secuencias genómicas de los seres vivos.

3.3. Modelo de Massip-Arndt

Se abre paso al modelo que mejor parece reproducir las características encontradas en genomas reales en el presente trabajo, este es el propuesto por Massip y Arndt . Este sistema tiene dos mecanismos bastante importantes que son de hecho su principales característica.Se realiza una serie de pasos en vista de la importancia y de la cantidad de parámetros manipulables:

Proceso de Massip-Arndt

Como se había mencionado anteriormente, Massip-Arndt consta de dos pasos o mecanismos fundamentales que emulan los procesos que ocurren en las secuencias genomicas, estos son las mutaciones y duplicaciones. Sea entonces el alfabeto A de tamaño cuatro que va a representar las cuatro bases nitrogenadas que se encuentran el el genoma $A := \{A, G, C, T\}$, y sea una secuencia X de de longitud finita N de símbolos generada aleatoriamente tal que $X :=$

$x_0x_1x_2x_3.....x_N$ a la cual someteremos a constantes modificaciones, entonces el sistema se tiene que ajustar a lo siguiente:

Remplazamiento de Base Simple

Se procede a designar una probabilidad P_{SBR} asociada de remplazamiento sobre cualquiera de los elementos o símbolos de la secuencia inicial generada. Esta obedece a una probabilidad de tipo bernoulli similarmente al modelo de Zacks y el modelo de Expansión Modificación . Esto podemos ilustrarlo de la siguiente forma:

Sea A aquella base seleccionada del alfabeto descrito que representa el j-ésimo termino de la secuencia inicial, entonces cambiamos dicha base con probabilidad P_{SBR} por alguno de los los símblos complementarios C, T, G, cada uno de los cuales tienen la misma probabilidad $1/3$ de ser seleccionados

Duplicación de segmento

El siguiente mecanismo del presente modelo, tiene como objetivo replicar segmentos enteros de nuestra secuencia inicial a lo largo de la cadena simbólica generada aleatoriamente. Para esto, tomamos una longitud L fija la cual representara el segmento a duplicar, seguida de los números k y j que son las etiquetas indexadas o asociadas a los símbolos de la secuencia inicial X. Ahora, representaremos la probabilidad de duplicación como P_{SD} y en virtud de que $X = X_1X_2X_3.....X_N$ y ademas tenemos a μ y ϕ como dos fragmentos escogidos que empiezan a partir de los monómeros con etiqueta k y j (los cuales se escogen aleatoriamente), entonces :

$$X_k, X_{k+1}....X_{k+N} = \mu \tag{3.5}$$

$$X_j, X_{j+1}....X_{j+N} = \phi \tag{3.6}$$

y así , la secuencia ϕ sera remplazada por μ sobre la secuencia de tama´ no de símbolos X. Como nota aclaratoria, podemos resaltar como la cadena que generamos al principio del proceso siempre se va a mantener constante, únicamente ocurriendo modificaciones o alteraciones dentro de ella misma.

Es importante observar que no solo tenemos el poder de mover los dos mecanismos principales de Massip-Arndt que lo distinguen en esencia, si no que también entra en juego un par factores de relevante importancia que son las iteraciones que caracterizamos como t, y la longitud de palabra L de la que se componen los segmentos a ser duplicados. Estos dos parámetros

no obedecen a una probabilidad, si no que mas bién se varían progresivamente para ver el resultado de su cambio.

Capítulo 4

Metodología

Hemos mencionado las principales propiedades de la fc así como la problemática que se genera tras el análisis o cálculo computacional de esta. También profundizamos en una variante de la fc que consta de una aproximación que llamamos el estimador $c_X(l)$. En efecto, en el presente trabajo se presenta, cadenas de una longitud N aproximada de seis millones de bits o simbólos a las que hemos recurrido, para realizar las simulaciones de los modelos de evolución genómica,así como para algunas de las muestras de los genomas enteros de las diversas especies que estudiamos.En este apartado se muestra en particular, algunos experimentos controlados con cuatro secuencias de las cuales tres, descansan en un alfabeto binario, y una con las bases nitrogenadas del ADN.De esta forma procederemos a comparar algunas características de nuestro programa, con el trabajo establecido en particular en [6] para calcular $C(l)$, así podremos ver si nuestra rutina es funcional o no.

4.0.1. Pruebas de Control

Primera prueba de control

En virtud de que conocemos el valor teórico de la función complejidad formulada en (2.6), procedemos a integrar el alfabeto $A := \{A, A, T, C\}$, así como la muestra $M = 100,000$ la cual dejaremos constante en todos los procesos llevados a cabo, siendo $k(l)$ la complejidad o diversidad de palabras de la selección M de longitud de palabra l.
Ahora, se procede a generar una secuencia aleatoriamente como se indica al principio del capítulo compuesta por las cuatro bases nitrogenadas y con un tamaño $N := 6,000,000$, esto por motivos de eficiencia computacional. Entonces, habiendo hecho uso de los parámetros mencionados, se calcula la fc para así obtener el ajuste lineal de la siguiente forma :
Entonces, sea M el tamanõ de la colección de palabras equidistantes, tendremos que para una secuencia del tamanõ deseado N,cada palabra tiene que estar separada una longitud de

$$\frac{N}{M} = \frac{6,000,000}{100,000} = 60.$$

Ahora podemos hacer una colección sabiendo que la muestra que eligirémos sera tomada a lo largo de toda la secuencia simbólica envuelta en el proceso. Teóricamente, atendiendo a la ecuación (2.6) deberíamos encontrar un numero distinto de palabras $C(l)$ acorde a

$$C(l) = 4^l.$$

Esto quiere decir, que si tenemos una función que depende de l que sigue a una ley de potencia , entonces al tomar la gráfica semi-logaritmica tendremos

$$logC(l) = log4^l$$

$$logC(l) = l \cdot log4 = l \cdot \alpha \approx l \cdot 1{,}386$$

Después de encargarnos de generar la secuencia con el tamanõ deseado. procedemos a evaluar la pendiente α de la longitud de palabra (que se calculo hasta una $l = 50$) contra el logaritmo del estimador de la función complejidad y cuyo valor fue de $\alpha \approx 1{,}389$. Esto nos indica una buena aproximación en comparación con el valor teórico de la función complejidad con una pendiente $\alpha \approx 1{,}386$.

Segunda prueba de control

En seguida, para la segunda prueba de control se utilizo un archivo en formato binario que descansa sobre el alfabeto $\{0, 1\}$, y que corresponde al espacio de corrimiento completo de tipo Bernoulli (o full shift Bernoulli) el cual tiene como restricción la palabra $\{0, 0\}$. De esta forma, pasamos a realizar el cálculo de fc para una longitud desde $l = 2$ hasta una $l = 50$, con lo que se estimo un valor de la pendiente $\alpha = 0{,}72$ en concordancia con el valor obtenido en [6] de $\alpha = 0{,}72$.

Tercera prueba de control

Para la tercera prueba de control se utilizo un archivo correspondiente al espacio corrimiento de longitud limitada (o run lenght limited shift),que se trata de una secuencia binaria con la restricción de un conjunto de palabras prohibidas finito. En este se obtuvo un ajuste lineal de la fc con una pendiente $\alpha = 0{,}37$ en concordancia con el valor extraído de [6] con una $\alpha = 0{,}38$.

Cuarta prueba de control

Para la cuarta prueba de control se utilizo un archivo correspondiente al espacio de Fibonacci (o *Fibonacci shift*), que se trata de una secuencia binaria con la restricción de un conjunto de palabras prohibidas finito. En este se obtuvo un ajuste lineal de la fc con una pendiente $\alpha = 0{,}37$ en concordancia con el valor extraído de [6] con una $\alpha = 0{,}38$. Se observa también en la gráfica de la Figura (4.1) la superposición de las tres ultimas pruebas de control con ánimos de ver como afecta la restricción de palabras de las secuencias generadas todas descansando sobre el alfabeto descrito en cada prueba de control.
Una vez finalizado satisfactoriamente las pruebas de control con respecto a la rutina de programación elaborada, en principio tenemos certeza acerca de que obtendremos un valor aproximado a la función complejidad teórica.

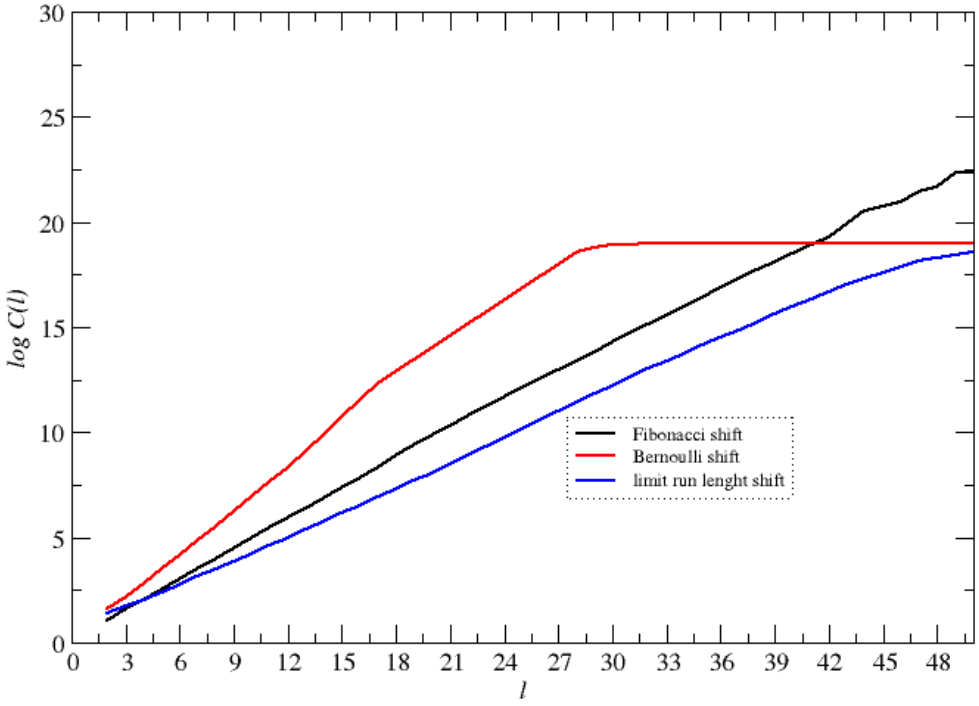

Figura 4.1: Gráfica de la longitud de palabra l contra el logaritmo de la función complejidad $C(l)$, de los tres archivos o secuencias que descansan en el alfabeto binario $\{0, 1\}$ y de los cuales realizamos un ajuste lineal α .

4.0.2. De Las Especies Utilizadas para la estimación de la función complejidad

hacemos una especificación mas clara en cuanto a las especies utilizadas para la estimación de la función complejidad. Primeramente, nos enfocamos en los 23 cromosomas del Homosapiens. Utilizamos como se ha reiterado en el presente trabajo, las secuencias completas y codificantes para cada cromosoma de esta especie. De la misma forma se toma los 23 cromosomas de la especie Gorilla gorilla gorilla y por último los tres cromosomas del Mosquito Aedes aegypti encargado de transmitir el dengue.

También se eligieron especies que caen en el dominio de la procariota (células que no contienen un núcleo definido y ademas no poseen regiones no codificantes o intrónes), y que específicamente se concentran en tres grupos de bacterias distintas las cuales son : Bacillaceae, Alteromonadaceae y Enterobacteriales. De cualquier forma, en el siguiente capitulo se muestra una descripción detallada con la información del cromosoma , así como el nombre de cada especie utilizada.

Muestro para la Estimación de la función complejidad

Para el calculo de la fc, en la elaboración de la rutina de programación, tenemos solo una restricción que se basa en la lectura de datos (que en nuestro caso de trata de cadenas (**char**) y que en comparación con el tamaño de algunos de los cromosomas de las especies estudiadas, resulta significativamente menor. Por ejemplo, para el cromosoma numero uno del Homosapiens, se estima que su tamaño es de aproximadamente docientos cincuenta millones de pares de bases o símbolos, mientras que el máximo valor que podemos registrar en nuestra rutina (en términos de una cadena de texto)es de seis millones (o también **char[6000000]**).Evidentemente si tomáramos solo una muestra en el primer cromosoma del homosapiens con la restricción mencionada , estaríamos discriminando una gran cantidad de información envuelta en este.De forma particular para observar una mejor comportamiento de la fc en esta especie, procedemos a realizar un segundo calculo con el mismo tamanño de secuencia de seis millones pero tomado aleatoriamente sobre el cromosoma completo de la especie , es de decir, tomaremos la muestra M en otro pedazo de esa secuencia .De esta forma podremos ver si los resultados previamente obtenidos son consistentes con la nueva muestra analizada.

4.0.3. Ajuste de los Modelos Evolutivos

Procedemos a realizar un ajuste que interviene directamente con la muestra seleccionada sobre cada simulación para el cálculo de la fc,

Modelo de Expansión Modificación

De los tres Modelos de Evolución Genómica, dos de ellos (como se ha indicado anteriormente) se centran sobre un alfabeto $\{0, 1\}$ de tamaño dos. Para beneficio de nuestro objetivo en general, procedemos a elaborar todas las secuencias correspondientes sobre un tamanño de $6,000,000$ de símbolos. De esta forma, eventualmente podemos tomar un colección de palabras de longitud l que es M, igual a cien mil palabras equiespaciadas o distribuidas con una distancia $\frac{N}{M} = \frac{6,000,000}{100,000} = 60$ de entre cada una de ellas, teniendo la seguridad de que la muestra embarque a toda la secuencia generada. Las simulaciones obtenidas para distintos valores de parámetros de los cuales se hablaron en el marco teórico y que son la probabilidad de expansión o modificación P, y la probabilidad complementaria $1-P$, fueron llevadas acabo entonces hasta alcanzar el tamanño deseado y especificado en este párrafo para después, tomar la colección de palabras y posteriormente disponernos a realizar la estimación de la función complejidad. Ya habiendo hecho el calculo computacional de la fc, lo que sigue es un análisis sobre las gráficas obtenidas de la longitud de palabra contra la fc. Así, convenientemente podremos ahora enfocarnos en encontrar características particulares de cada muestra bajo el parámetro modificado.

Modelo de Zacks

Para el modelo de Zacks el cual dispone de un alfabeto binario $\{0, 1\}$, utilizamos secuencias simbólicas obtenidas través de distintas simulaciones, cada una con una correspondiente modificación al parámetro que distingue a este modelo y los cuales son: probabilidad de mutación o duplicación P, y la probabilidad complementaria $1 - P$. Paramos todas las simulaciones hasta alcanzar una longitud de cadena de seis millones de símbolos y posteriormente realizamos una colección de $\frac{N}{M} = \frac{6{,}000{,}000}{100{,}000} = 60$ de palabras para el calculo computacional de la función complejidad. La similitud del modelo de Zacks último con el de espansión modificación resulta bastante en lo que a la generación de cadenas simbólicas se refiere,por lo cual ejecutamos el mismo procedimiento que consiste en tomar la colección de tamaõ M y posteriormente estimar la fc para después hacer un análisis estadístico sobre las gráficas de la longitud de palabra contra la función complejidad.

Modelo de Massip-Arndt

para el modelo de Massip-Arndt quien tiene su alfabeto sobre la base A, C, T, G, se realizaron simulaciones variando cuatro parámetros distintivos que mas explicitamente son: la variación de t que puede verse como el paso del tiempo discreto, la longitud utilizada para copiar y pegar segmentos como se ha explicado en el marco teórico, y por supuesto las probabilidades de duplicación de segmento P_{SD} y de remplazamiento de base simple P_{SBR} que corresponden a una probabilidad de tipo Bernoulli.Como ajustamos el modelo para realizar las simulaciones fue de la manera siguiente.

Primero generamos una secuencia de símbolos que descansan sobre el alfabeto que distinguen a las cuatro bases nitrogenadas del ADN y que coinciden con el alfabeto de Massip-Arndt. Paramos hasta alcanzar un tamaño de secuencia símbolos de seis millones y ademas se resalta que cada uno de estos, tiene la misma probabilidad de ocurrir a lo largo de la cadena generada (es decir cada uno tiene una probabilidad de 1/4 de ocurrir), con lo que garantizamos que sea completamente aleatoria su realización.

Lo siguiente es someter a los procesos descritos tras la variación de los cuatros parámetros, a la secuencia generada aleatoriamente, misma que consecutivamente usamos para hacer la estracción de una muestra de tamaño M, y la cual es seleccionada a través del mismo procedimiento $\frac{N}{M} = \frac{6{,}000{,}000}{100{,}000} = 60$ que los modelos anteriores, es decir tomamos M palabras equiespaciadas una longitud de 60 caracteres o símbolos.Lo que sigue, es evidentemente hacer el calculo de la diversidad de palabras y, un análisis de las gráficas longitud de palabra contra función complejidad para un posterior análisis sobre cada gráfica.

4.0.4. Del análisis de las gráficas de la función complejidad

Es conveniente adelantar un poco el análisis propuesto de las grafícas de la función complejidad tanto obtenidas para la especies reales, así como las obtenidas de las simulaciones de

nuestros modelos evolutivos. Posterior a la resultante del perfil de cada gráfica de la fc, para aquellas las cuales muestren un comportamiento definido de tipo lineal , estaremos interesados en obtener la pendiente α de dicha curva, así como la definida longitud de cruce l_c (que se trata de la longitud de palabra l a partir de la cual tenemos un cambio de comportamiento en la fc) a partir de la cual ocurre este comportamiento. El motivo de extraer dichas propiedades a cada curva es debido al hecho de poder encontrar una diferencia o similitud a cada especie, familia o clase que nos permita tener una un aproximado a alguna clasificación similar a la taxonónomica pero desde el punto de vista estadístico.

Capítulo 5

Resultados

5.1. Función complejidad para los cromosomas del Homosapiens

Haciendo uso del estimador (2.12) visto en el marco teórico, se procede a calcular con una muestra $M = 100,000$ de palabras de longitud l, sobre las secuencias simbólicas formadas por cada uno de los cromosomas del homosapiens quienes descansan en el alfabeto $\{A, C, T, G\}$. Se muestra en la figura (5.1) el despliegue de las gráficas de la función complejidad de la longitud l contra el logaritmo de $C(l)$ que corresponden a los 23 cromosomas analizados.

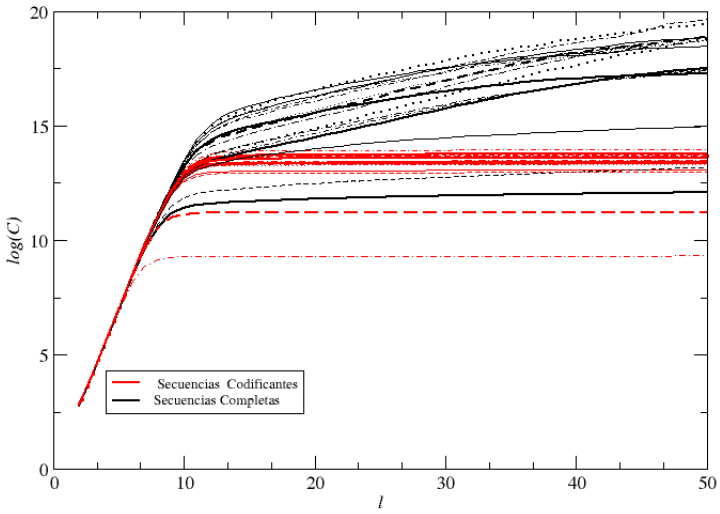

Figura 5.1: gráfica de l contra el $logC(l)$ de los 23 cromosomas del homosapiens.Mostramos la familia en color rojo para las secuencias codificantes, mientras que la familia en color negro pertenece a las secuencias completas.

Ademas, mostramos en la figura 5.2, el despliegue de las gráficas correspondiente a la familia de cromosomas propuesta en su version l contra $C(l)$.

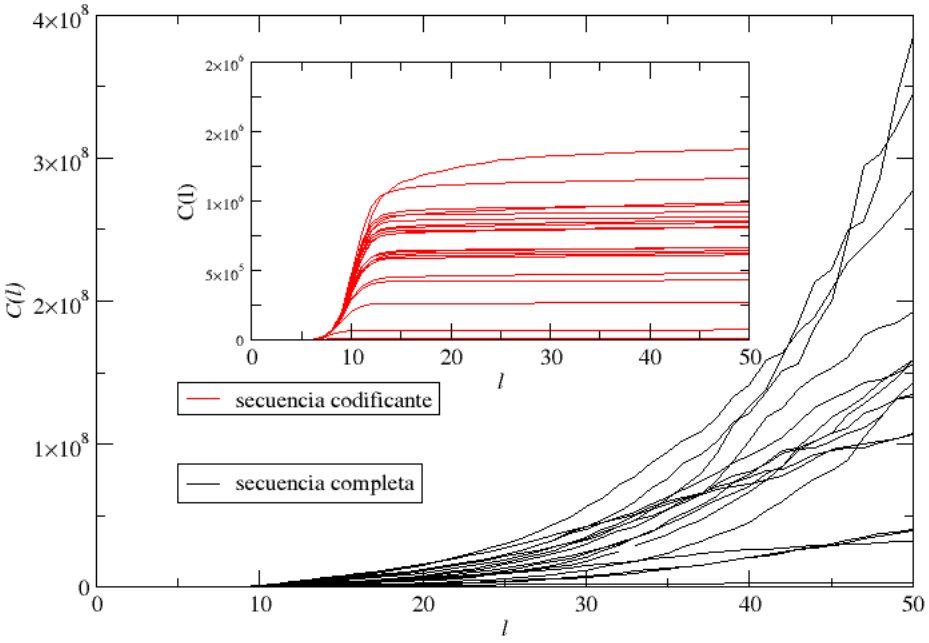

Figura 5.2: gráfica de l contra el $C(l)$ de los 23 cromosomas del homosapiens.Mostramos la familia en color rojo para las secuencias codificantes, mientras que la familia en color negro pertenece a las secuencias completas.

Una de las principales características encontradas que resulta evidente en la figura (5.1), es el hecho de poder distinguir dos familias que pertenecen a las secuencias codificantes, y a las no codificantes. En efecto, para las secuencias que tienen solamente exones (las gráficas de color rojo), es posible observar que presentan un comportamiento sutil que consiste en una tendencia exponencial aproximadamente para longitudes pequeñas ($2 \leq l \leq 15$) , y para longitudes grandes $15 \leq l \leq 50$ observamos una tendencia lineal.

Evidentemente existen excepciones para las dos familias propuestas que son importantes de analizar. De la familia de gráficas para las secuencias completas,hay algunas de ellas que tienen una función de complejidad menor que el promedio. Aunque en principio no sabemos si la naturaleza de estas radica en el sentido biológico, cabe resaltar que para el cromosoma 15, 12 y 13 disponen de un tamaño de cromosoma de aproximadamente mas de 100M,mientras que el cromosoma Y tiene una tamaño de 58M y del cual de tiene una función complejidad mayor que los anteriores descritos.Esto puede hacernos pensar tal vez que la función complejidad no dependa del tamaño del cromosoma al menos en base a la toma de muestra seleccionada. De la misma forma, mientras que para el cromosoma 15 tenemos un porcentaje de introns de 91 %, para el cromosoma 13 tenemos un porcentaje de 95 % de introns. Puesto que al menos uno de los cromosomas tienen un porcentaje menor de introns y ademas, obtenemos un perfil de la función complejidad mayor que para 15 y 13

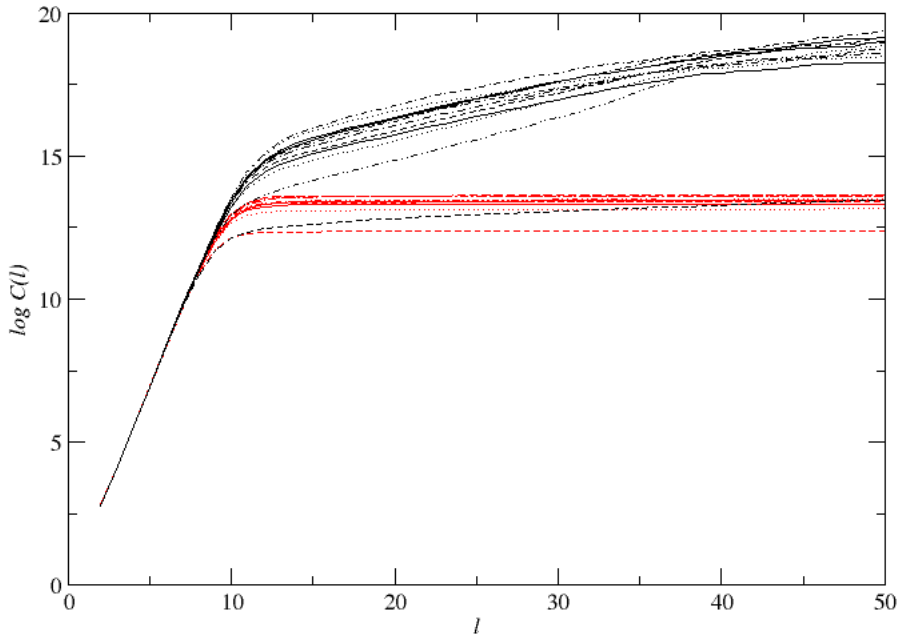

Figura 5.3: gráfica de l contra el $logC(l)$ para 13 cromosomas del homosapiens.Mostramos la familia en color rojo para las secuencias codificantes, mientras que la familia en color negro pertenece a las secuencias completas sobre un nuevo despliegue, utilizando un recorrido de ventana o desplazamiento de aproximadamente seis millones de pares de bases nitrogenadas .

(recordando que estos son los de un perfil de gráfica menor de la fc), tal vez el porcentaje de los intrones no intervenga directamente de acuerdo a los resultados obtenidos para esta muestra.

De la familia de las secuencias codificantes, podemos encontrar dos gráficas que tienen un perfil de la función complejidad menor al promedio de la familia descrita. Por ejemplo, el perfil con la fc mas baja (de la familia en color rojo) que se muestra en la figura (5.1), corresponde a la secuencia codificante del cromosoma Y.Una de las posibles causas de este comportamiento registrado, podría ser debido a que el tamaño de esta secuencia sea de 772K, situación de que el cromosoma tenga solo 1 % de exones en ella. De hecho, la información útil que podemos encontrar en este cromosoma ocurre hasta una longitud aproximadamente igual a $l \leq 13$, ya que a partir de esta se observa una saturación.

Es posible ver ademas en la Figura 5.3, un nuevo despliegue de gráficas de esta especie.La forma en que se recolectaron estos, fue desplazándonos o haciendo un recorrido a partir de la muestra previamente tomada , es decir, tomamos una nueva cantidad M usando una secuencia

muestral completamente nueva. En esta, se utilizo un recorrido de ventana en las secuencias codificantes y completas pertenecientes a los cromosomas descritos del 1 al 14. Lo que se observa en general para las dos familias es nuevamente la evidencia del comportamiento de la fc con na tendencia exponencial para las primeras longitudes de palabra, y un carácter lineal para longitudes grandes (en el caso de las secuencias codificantes). Es posible ver de igual forma que las secuencia completas conservan un comportamiento exponencial evidente y que en general resultan de una fc mucho mayor a las secuencias que solo contienen exónes. Finalmente hacemos una descripción tanto del tamaño de las secuencias usadas, así como de la longitud de cruce l_c y la pendiente α correspondiente al comportamiento lineal de todos los cromosomas estudiados sobre el cuadro 5.1 .

5.2. Función complejidad para Bacterias

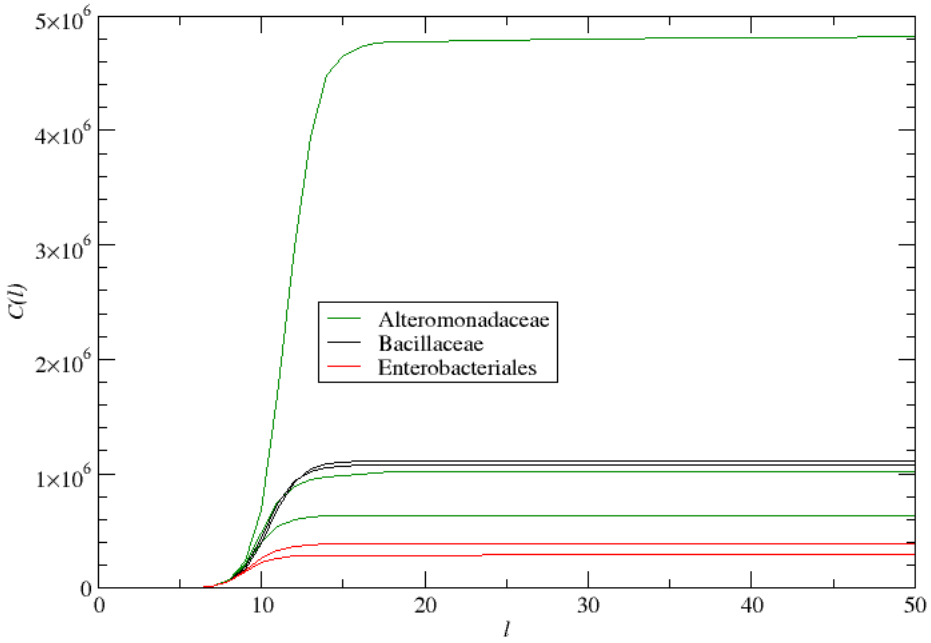

Figura 5.4: Gráfica que representa el perfil de la función complejidad en la versión l contra $C(l)$, para tres familias distintas del régimen bacteriano

En la siguientes secciones se dispone a calcular la fc para distintas especies desde el régimen bacteriano, mas específicamente enfocándonos en la familia bacillaceae (de la cual se tomaron dos especies), enterobacteriales (en total dos especies analizadas) y alteromonadaceae (de la cual se tomaron tres especies). De acorde a lo visto anteriormente en el ajuste de los

N.Crom.	t. Crom.	t. S.Cod.	α S.Cod.	% Intrones	% Exones	l_c
1	252M pb	28M pb	6.31	88 %	12 %	15
2	245M pb	25M pb	6.64	89 %	11 %	14
3	201M pb	18M pb	6.49	94 %	6 %	14
4	192M pb	11M pb	6.80	89 %	11 %	14
5	184M pb	11M pb	6.55	94 %	6 %	14
6	173M pb	14M pb	7.18	91 %	9 %	17
7	161M pb	12M pb	6.67	92 %	8 %	14
8	147M pb	10M pb	6.50	93 %	7 %	14
9	140M pb	12M pb	6.68	91 %	9 %	14
10	135M pb	13M pb	6.47	89 %	11 %	15
11	137M pb	14M pb	6.91	90 %	10 %	14
12	135M pb	14M pb	6.47	89 %	11 %	14
13	115M pb	5M pb	5.37	95 %	5 %	14
14	108M pb	8M pb	6.70	92 %	8 %	14
15	103M pb	9M pb	6.89	91 %	9 %	14
16	91M pb	10M pb	7.67	89 %	11 %	14
17	84M pb	14M pb	6.86	83 %	17 %	15
18	81M pb	51M pb	5.24	94 %	6 %	13
19	59M pb	14M pb	7.87	77 %	23 %	15
20	65M pb	6M pb	5.35	91 %	9 %	15
21	47M pb	2M pb	2.60	94 %	6 %	14
X	158M pb	8M pb	6.67	95 %	5 %	15
Y	58M pb	772K pb	1.18	99 %	1 %	11

Cuadro 5.1: Tabla del valor de la pendiente α del ajuste lineal y el tamaño de los cromosomas para la especie Homosapiens.Se muestra de igual forma la longitud de cruce l_c, así como la cantidad de exónes e intrones dentro del cromosoma.

Bacteria.	t. Crom.	α	l_c
Virgibacillus	4M pb	1.99	18
Bacillus cererus	5M pb	2.04	16
Salmonella	4Mpb	1.84	19
Yersinia	4M pb	2.99	16
Alteromona	5M pb	4.62	14
Marinobacter	4M pb	6.79	16
Paraglaciecola	5M pb	4.27	24

Cuadro 5.2: Tabla del valor de la pendiente α del ajuste lineal y el tamaño de los cromosomas pertenecientes a la familia bactereana compuesta por las Enterobacteriales,Alteromonadaceae y Bacillaceae ,donde también se muestra el tamaño de la secuencia codificante, así como la longitud de cruce l_c.

N.Crom.	t. Crom.	t. S.Cod.	α S.Cod.	% Introns	% Exones	l_c
1	310M pb	15M pb	6.83	95 %	5 %	17
2	474M pb	28M pb	5.95	94 %	6 %	17
3	409M pb	25M pb	7.41	94 %	6 %	18

Cuadro 5.3: Tabla del valor de la pendiente α del ajuste lineal y el tamaño de los tres cromosomas pertenecientes al mosquito Aedes aegypti,donde también se muestra el tamaño de la secuencia codificante, así como la longitud de cruce l_c.

N.Crom.	t. Crom.	t. S.Cod.	α S.Cod.	% Introns	% Exones	l_c
1	222M pb	10.5M pb	7.42	95 %	5 %	15
2A	107M pb	3.3M pb	2.00	97 %	3 %	15
2B	127M pb	4M pb	2.24	97 %	3 %	15
3	193M pb	6.5M pb	7.27	96 %	4 %	14
4	187M pb	3.7M pb	2.55	98 %	2 %	15
5	160M pb	7.2M pb	8.17	96 %	4 %	16
6	168M pb	4.9M pb	3.89	97 %	3 %	15
7	148M pb	4.1M pb	2.74	97 %	3 %	18
8	140M pb	3M pb	1.77	98 %	2 %	14
9	107M pb	4M pb	3.38	96 %	4 %	15
10	127M pb	4M pb	2.66	97 %	3 %	14
11	118M pb	6M pb	7.66	95 %	5 %	16
12	128M pb	5M pb	5.28	96 %	4 %	15
13	92M pb	2M pb	0.62	98 %	2 %	14
14	87M pb	3M pb	1.45	97 %	3 %	17
15	76M pb	3M pb	1.39	96 %	4 %	19
16	74M pb	4M pb	3.07	96 %	4 %	16
17	90M pb	2M pb	0.80	98 %	2 %	16
18	73M pb	1.6M pb	0.33	98 %	2 %	15
19	48M pb	5M pb	7.06	90 %	10 %	15
20	60M pb	2M pb	7.32	97 %	3 %	14
21	32M pb	1M pb	0.86	97 %	3 %	13
22	32M pb	2M pb	5.12	94 %	6 %	15
X	158M pb	3M pb	5.20	98 %	2 %	17

Cuadro 5.4: Tabla del valor de la pendiente α del ajuste lineal y el tamaño de los cromosomas pertenecientes al Gorilla gorilla gorilla,donde también se muestra el tamaño de la secuencia codificante, así como la longitud de cruce l_c.

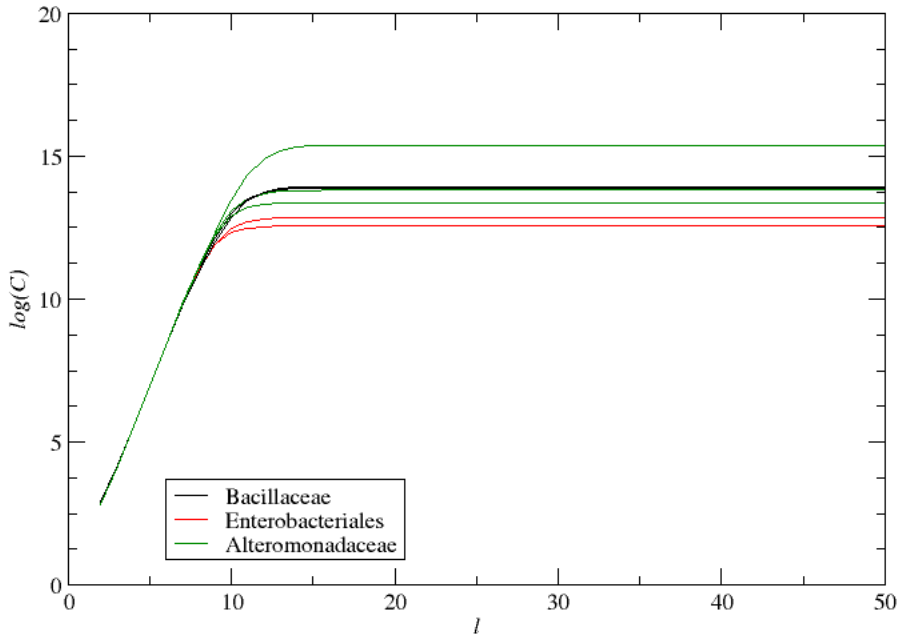

Figura 5.5: Gráfica que representa el perfil de la función complejidad en la versión l contra $logC(l)$, para tres familias distintas del régimen bacteriano

modelos evolutivos, todas las secuencias de ambas familias corresponden a cadenas formadas únicamente por exónes debido a la naturaleza de la célula procariota. El énfasis de presentar las secuencias codificantes para estas familias de bacterias, es debido al hecho de que nuevamente como en el caso de los 23 cromosomas del homosapiens, tenemos un comportamiento sobre el perfil de la función complejidad, de exponencial a lineal como puede apreciarse en la Figura (5.4) y Figura (5.5).

5.3. Función complejidad para Gorilla gorilla gorilla

En virtud de que de los cromosomas del homosapiens así como los de la bacterias estudiadas en este trabajo, presentan un comportamiento con un cambio bien definido en la función de complejidad, procedemos a realizar un despliegue sobre los 22 cromosomas del Gorilla gorilla gorilla, de la fc para las secuencias codificantes o aquellas con únicamente exónes y para las secuencias completas. En efecto es posible ver que en la Figura 5.6 y en la Figura 5.7 (cual muestra la versión l contra $logC(l)$ tenemos a excepción del cromosoma 19 que aparece como una linea punteada, un comportamiento con las características antes mencionadas de la fc en las secuencias de exónes. Es de importancia observar, que a diferencia de varios de los cromosomas del homosapiens, los cromosomas del gorilla gorilla gorilla tienen apenas una cantidad de información que codifica entre el 2 % y el 5 % a excepción nuevamente del cromosoma 19 el cual posee una cantidad de 10 % de exónes (y que también presenta el único perfil de la fc que no se comporta de forma exponencial-lineal).

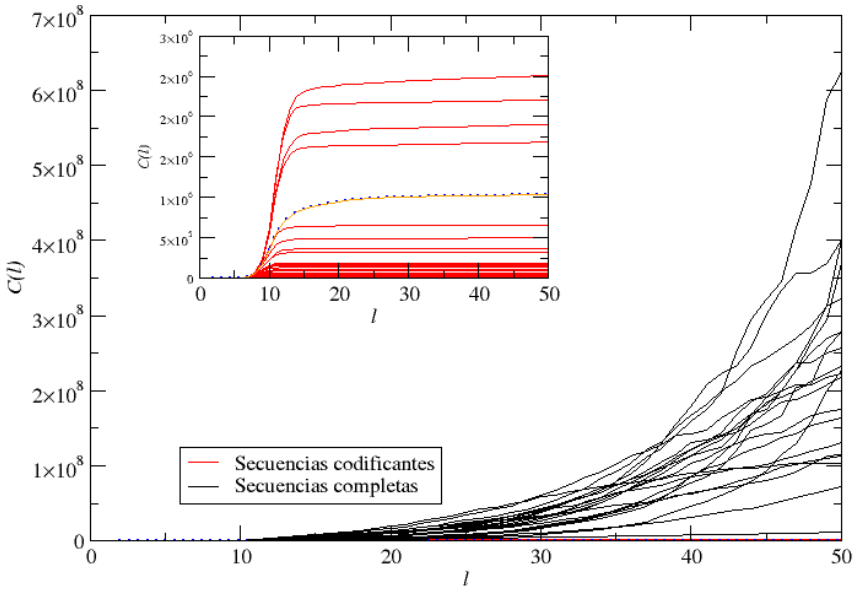

Figura 5.6: Gráfica de los 23 cromosomas del Gorilla gorilla gorilla para las secuencias codificantes y secuencias completas, en la versión l contra $C(l)$.

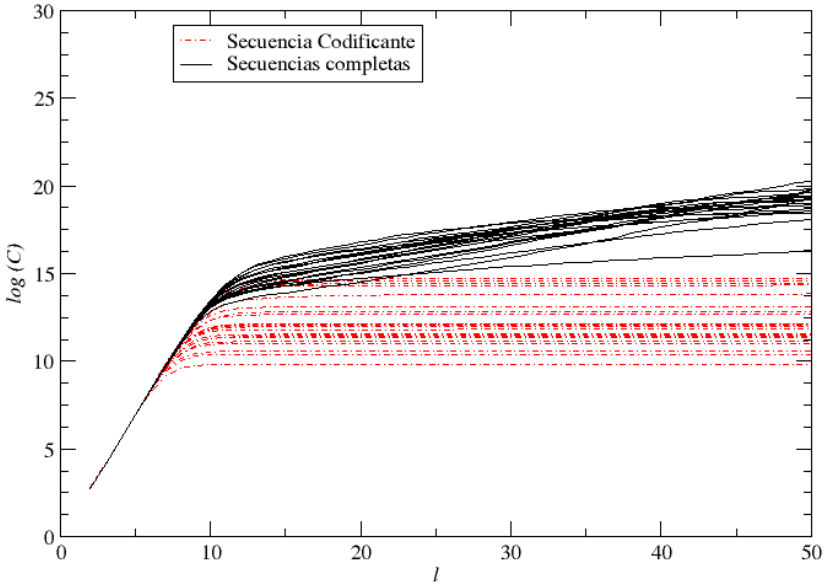

Figura 5.7: Gráfica de los 23 cromosomas del Gorilla gorilla gorilla para las secuencias codificantes y secuencias completas en la versión l contra $logC(l)$.

5.4. Función complejidad para el mosquito Aedes aegypti

Continuando con el criterio del comportamiento en la fc de las secuencias codificantes, procedemos a elaborar finalmente las gráficas de la función complejidad de los tres cromosomas del mosquito Aedes aegypti, el cual puede ser un portador del virus del dengue y de la fiebre amarilla.Este descansa en la familia culicidae y es de la clase insecta. En la Figura 5.8 se puede observar las distintas secuencias para ambas familias, con el comportamiento el cual estamos interesados (exponencial-lineal) presente en las secuencias codificantes, en su versión longitud de palabra l contra $C(l)$ .Es posible de igual forma ver en la Figura 5.9 la versión de l contra el $logC(l)$.

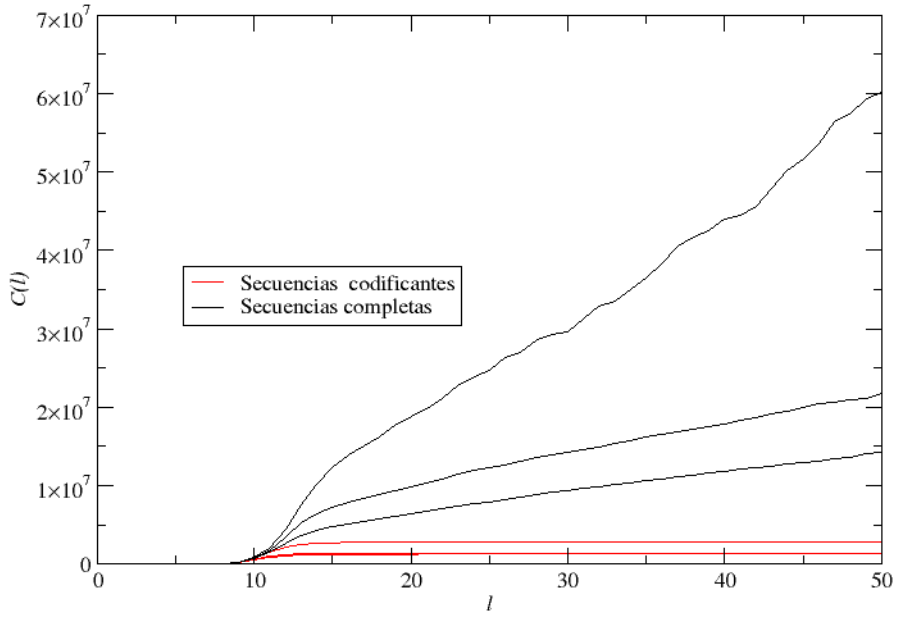

Figura 5.8: Gráfica de los tres cromosomas del mosquito aedes, para las secuencias codificantes y secuencias completas de la l contra $C(l)$.

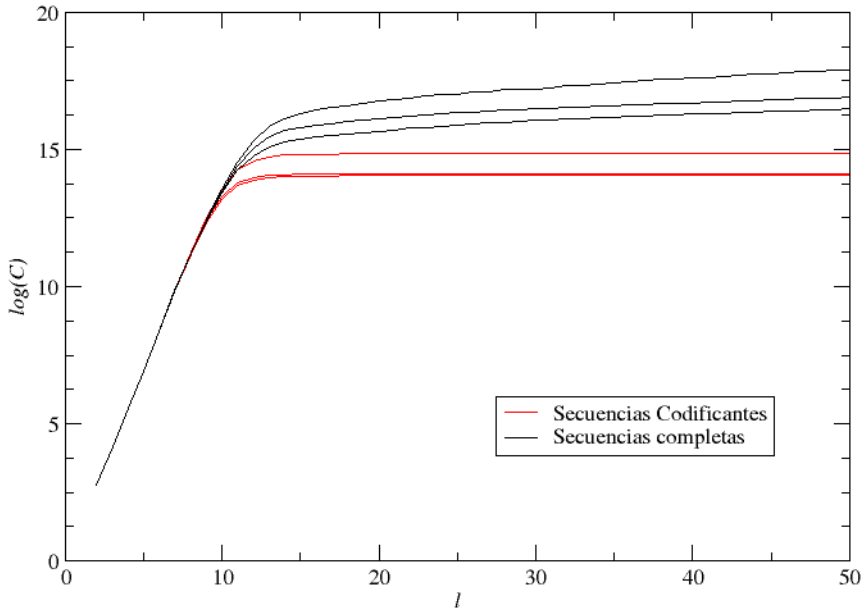

Figura 5.9: Función complejidad para los tres cromosomas del mosquito Aedes aegypti, correspondiente a las secuencias codificantes y secuencias completas en su versión longitud de palabra l contra $logC(l)$.

5.5. Modelo Expansión Modidicación

Como explicamos en el marco teórico, existen dos parámetros que identifican al modelo de expansión modificación, los cuales son la expansión o modificación P, y su probabilidad complementaria $1 - P$

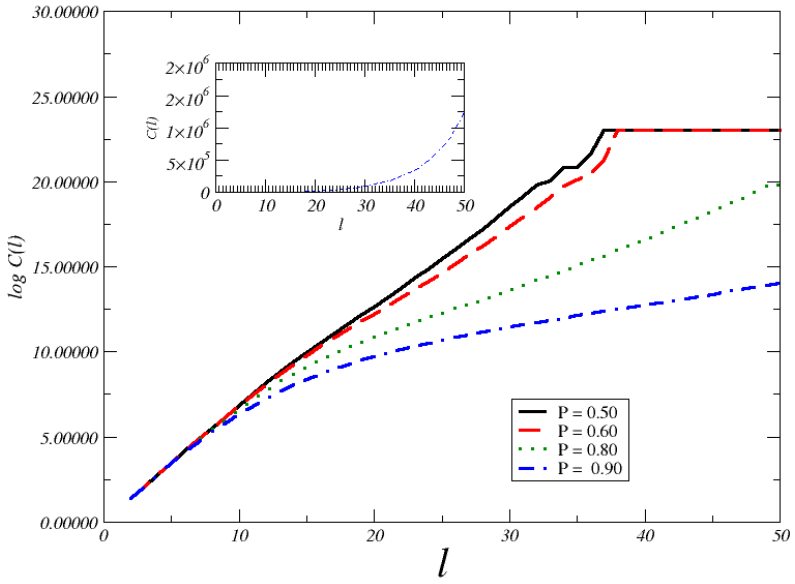

Figura 5.10: Se muestra el perfil de la función complejidad para distintos valores del parámetro P, en su version l contra el logaritmo de $C(l)$

El objetivo de ahora en adelante tras haber estimado la función complejidad de los cromosomas del homosapiens así como las especies del régimen bacteriano, sera aproximar un perfil similar a través de los tres modelos evolutivos propuestos. Se despliega entonces, las simulaciones obtenidas para la variación de los parámetros mencionados a continuación. En la figura (5.10) quien representa el logaritmo de la función complejidad contra la longitud de palabra, podemos observar cuatro simulaciones distintas para un valor de parámetro P de expansión tal que $P = \{0{,}5, 0{,}6, 0{,}8, 0{,}9\}$.

Entonces, uno de las suspicacias mas evidentes encontradas en este modelo,es el hecho de que en cuanto aumentamos significativamente la probabilidad de expansión, nos encontramos con una disminución del perfil de la gráfica a medida que vamos aumentando la longitud de palabra, para una l aproximadamente entre $10 \leq l \leq 50$.Una posible explicación de la disminución de la fc en cuanto aumentamos a P, podría ser el hecho de que estamos dando una probabilidad alta de que la cadena se expanda, entonces tendremos símbolos repetidos con mucho mayor frecuencia que símbolos distintos sobre la probabilidad complementaria $1 - P$ de modificación.

Otro aspecto que se puede notar sobre dos de las cuatro gráficas de la figura 5.10, es que para los valores de $P = 0{,}5$ y $P = 0{,}6$ para una longitud l tal que $l \geq 35$, la función complejidad bajo la muestra de tamaño $M = 100,000$ alcanza una saturación la cual evidentemente nos deja de proporcionar información alguna, lo cual sugiere que a valores de P pequeños, mas rápidamente vamos a saturar la función complejidad.

En cualquier caso, podemos observar que incluso para valores de P muy altos (recordemos

que se trata de una probabilidad de tipo Bernoulli con un valor extremo de P idéntico a 1), o en su equivalente para una probabilidad complementaria $1 - P$ de modificación muy baja, la curva de la fc la cual es mostrada en su version $C(l)$ contra l, obedece todo el tiempo a un comportamiento puramente exponencial.Esto ultimo implica que evidentemente no fue posible reproducir el perfil de la gráficas para la fc de los cromosomas del homosapiens, ni para cualquiera de las bacterias que fueron analizadas en este trabajo por lo menos para las secuencias codificantes de las cuales se tiene un comportamiento que para valores pequeños de l es exponencial, y para valores grandes obtenemos un comportamiento lineal.

5.6. Modelo de Zacks

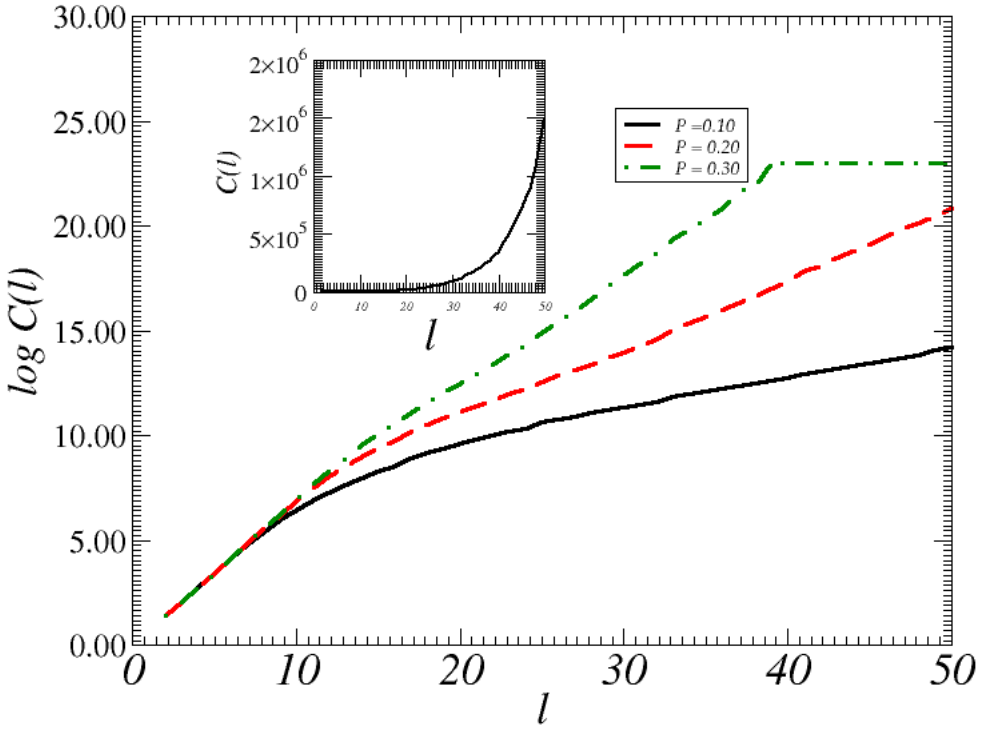

Figura 5.11: gráfica que muestra a distintos valores de parámetro P, el perfil de la función complejidad de l contra el logaritmo de $C(l)$, para el modelo de Zacks.

Pasamos a realizar las simulaciones del modelo de Zacks el cual tiene un alfabeto binario que descansa en $\{0, 1\}$ y que se puede interpretar como una cadena de símbolos donde pueden

existir duplicaciones, así como mutaciones como fue explicado en el marco teórico. Se despliega entonces las simulaciones obtenidas para este modelo las cuales pueden ser apreciadas en la figura 5.11. De todas las gráficas de la fc obtenidas en este modelo, utilizamos valores pequeños del valor de parámetro P correspondiente a duplicaciones.El motivo de usar estos valores de parámtero, consiste en observar el comportamiento de la función complejidad para ver si es posible obtener un perfil similar a las especies reales utilizadas. Se observa, que incluso para un valor de duplicación pequeño ($P = 0{,}1$ o en su defecto una probabilidad complementaria $1 - P = 0{,}7$ de mutación), la grafica que también puede verse en la figura 5.11 de la longitud de palabra l contra $C(l)$ resulta todo el tiempo ser de carácter exponencial. Esto por supuesto no logra reproducir las características encontradas en los genomas reales para las secuencias codificantes, que conservan un comportamiento exponencial,y después lineal.

5.7. Modelo Massip y Arndt

Pasamos a obtener las simulaciones obtenidas a través del modelo que mejor parece reproducir las característias en la gráfica de la fc de secuencias de ADN reales estudiadas en el presente trabajo, y que se trata del modelo de Massip-Arndt.

Como se indico en el marco teórico, para efectuar cada una de las simulaciones presentes en este modelo, variamos en total cuatro parámetros esencialmente.En el cuadro 5.5, se presenta nueve simulaciones en total para distintos valores de l parámetro temporal t.Ademas, se muestra en la figura 5.12 las gráficas obtenidas correspondientes del logaritmo dela fc contra la longitud de palabra.

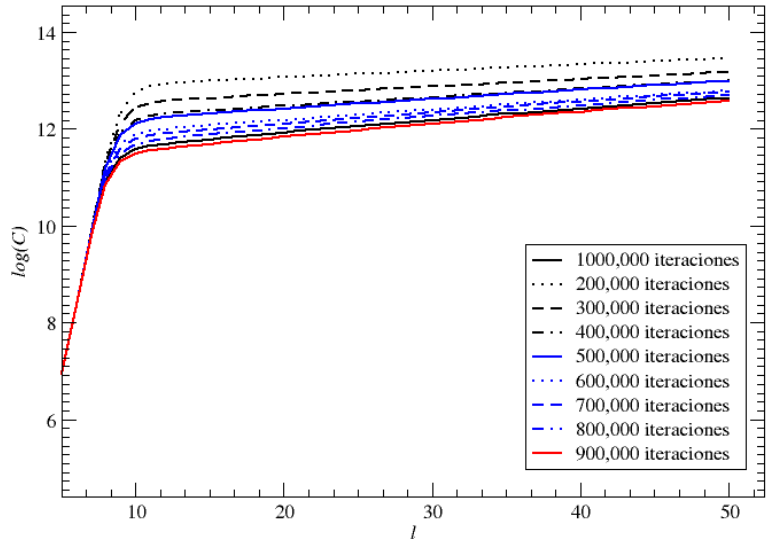

Figura 5.12: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro t, en su version de la longitud de palabra contra $C(l)$

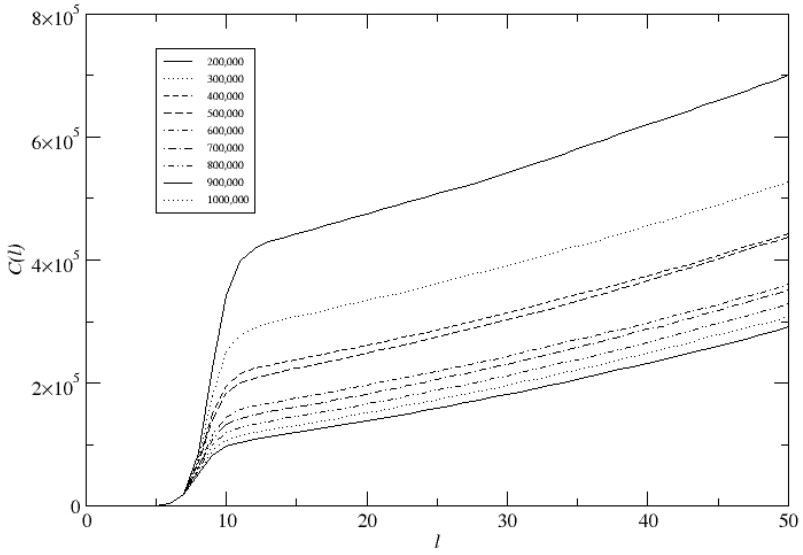

Figura 5.13: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro t, en su version de la longitud de palabra contra el logaritmo de $C(l)$.

t	α	l_c
200,000	9.03	10
300,000	8.87	10
400,000	8.83	9
500,000	8.87	9
600,000	8.72	9
700,000	8.75	9
800,000	8.72	9
900,000	8.66	8
1,000,000	8.68	8

Cuadro 5.5: Modelo Massip-Arndt.Variaciones en iteraciones (t) ,con numero de PSD = 0.5 y numero de PSBR = 0.06 (ambos constantes) y longitud $l = 1000$,en donde ademas α es la pendiente tras esa variación.

Es posible observar dos tipos de comportamiento sobre el perfil de todas las simulaciones obtenidas para el parámetro t, que consisten en un comportamiento exponencial para longitudes pequeñas, y un perfil de tendencia lineal para longitudes grandes. Una mejor vista de las características presentes de estas simulaciones es posible verse en la figura 5.13, de la longitud de palabra contra $C(l)$.

Sobre el análisis a través de la variación de t, tomando a l_c como la longitud de palabra de cruce del comportamiento exponencial a lineal (previamente definido), registramos un valor l_c de 8 a 10. Un aspecto importante encontrado en estas simulaciones, es que a medida de que incrementamos el numero de iteraciones t, es posible observar una disminución significativa de la función complejidad , lo que podría en principio darnos intuitivamente una idea de como poder ajustar tanto el perfíl de fc así como la longitud de cruce a medida que vamos variando t.

Por ultimo, para calcular la pendiente de la región de la fc donde su comportamiento es lineal, disponemos a tomar la α para un intervalo $13 \leq l \leq 50$. En el cuadro 5.5 es posible ver la α correspondiente a cada variación de t , de donde encontramos un incremento de esta a medida de que incrementamos el valor del parámetro t.

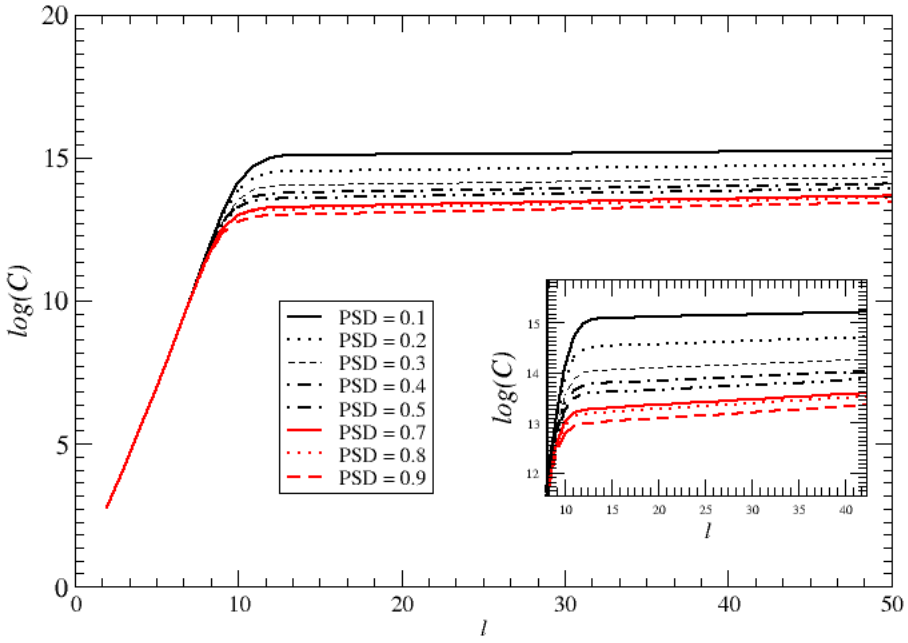

Figura 5.14: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro PSD, en su version de la longitud de palabra contra el logaritmo de $C(l)$.Es posible observar también un acercamiento para apreciar el momento en el cual tenemos un cambio de comportamiento de exponencial a lineal en todos los perfiles de las simulaciones obtenidas

Presentamos ahora el calculo del ajuste lineal, así como la longitud de cruce y el perfil la función complejidad, del resultado de variar el parámetro de la duplicación de segmento. Es posible ver en el cuadro 5.6 los valores obtenidos en α correspondientes a cada cambio en PSD. Ademas, se ilustra en la gráfica 5.14 una vista general del perfil de cada simulación en su version l contra el logaritmo de $C(l)$, y en la gráfica 5.15 la version l contra $C(l)$. Obtuvimos en la longitud de cruce en promedio de $l_c = 10 \pm 3$. Podemos observar que para este parámetro seleccionado, la variación de la pendiente a medida que incrementamos la probabilidad de duplicación de segmento se va incrementando ligeramente, por lo cual este suceso podría en principio, darnos una herramienta que nos permita ajustar α acorde a nuestras necesidades.

PSD	α	l_c
0.1	9.81	13
0.2	9.61	12
0.3	9.41	12
0.4	9.28	11
0.5	9.20	11
0.7	9.06	11
0.8	9.04	11
0.9	8.94	10

Cuadro 5.6: Modelo Massip-Arndt.Variaciones en PSD con numero de PSBR = 0.0 constante y longitud $l = 1000$ sobre 100,000 iteraciones.Ademas se muestra también α que es la pendiente tras esa variación.

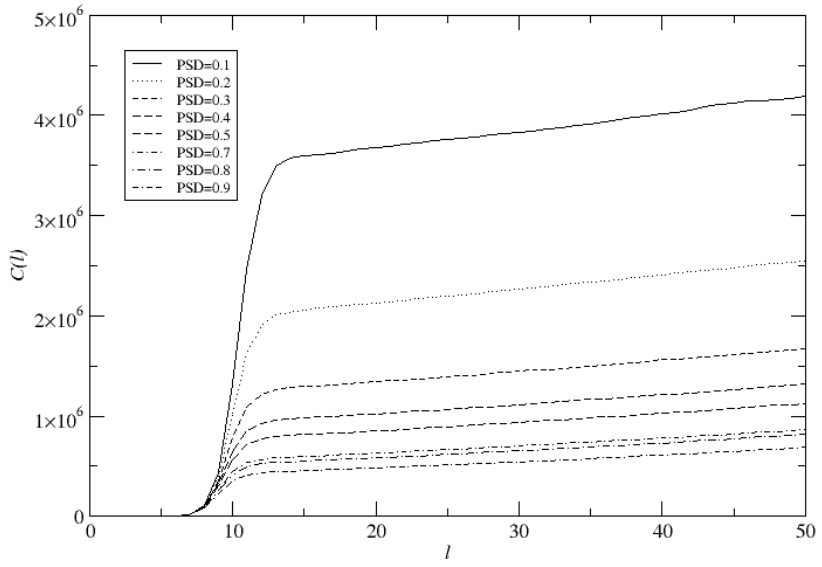

Figura 5.15: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro PSD, en su version de la longitud de palabra contra $C(l)$ para el parámetro descrito

Consecutivamente, en el cuadro 5.7 desplegamos los resultados obtenidos para la variación del parámetro de remplazamiento de base simple y el calculo de la pendiente del ajuste lineal . La gráfica 5.16 nos muestra el perfil de la fc en su version l contra el logaritmo de $C(l)$, y en la gráfica 5.17 la version l contra $C(l)$.Obtuvimos en la longitud de constante e igual a 11. Podemos observar que para este parámetro , la variación de la pendiente a medida que incrementamos la probabilidad $PSBR$, se va incrementando también ligeramente hasta llegar a pequeñas variaciones.Otro factor importante y sobresaliente suscitado, consiste en que la l_c parece no tener cambio tras la variación de este parámetro .En efecto , este suceso podría también en principio, darnos una herramienta que nos permita ajustar a α, sin tener mucha variación en la longitud de cruce.

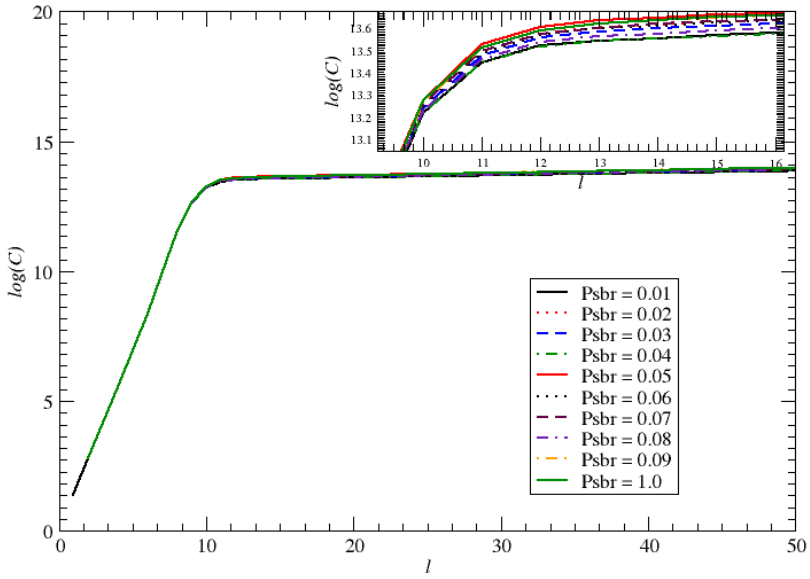

Figura 5.16: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro $PSBR$, en su version de la longitud de palabra contra el logaritmo de $C(l)$. Hacemos un acercamiento para observar la brecha o abanico de valores de longitud en el cual tenemos un comportamiento exponencial a lineal

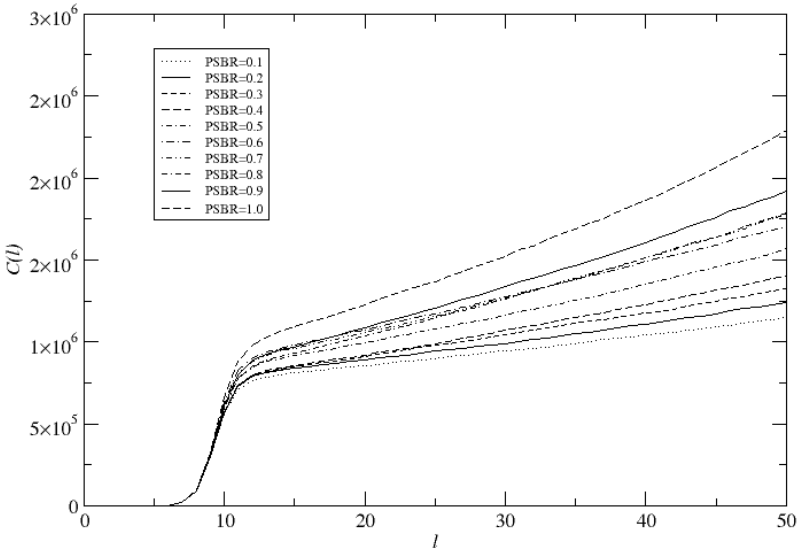

Figura 5.17: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro $PSBR$, en su version de la longitud de palabra contra $C(l)$.

PSBR	α	l_c
0.01	9.06	11
0.02	9.14	11
0.03	9.17	11
0.04	9.14	11
0.05	9.24	11
0.06	9.23	11
0.07	9.25	11
0.08	9.23	11
0.09	9.32	11
0.10	9.34	11

Cuadro 5.7: Modelo Massip-Arndt.Variaciones en PSBR con numero de PSD = 0.5 fijo y l = 1000 sobre 100,000 iteraciones, en donde ademas α es la pendiente tras esa variación.

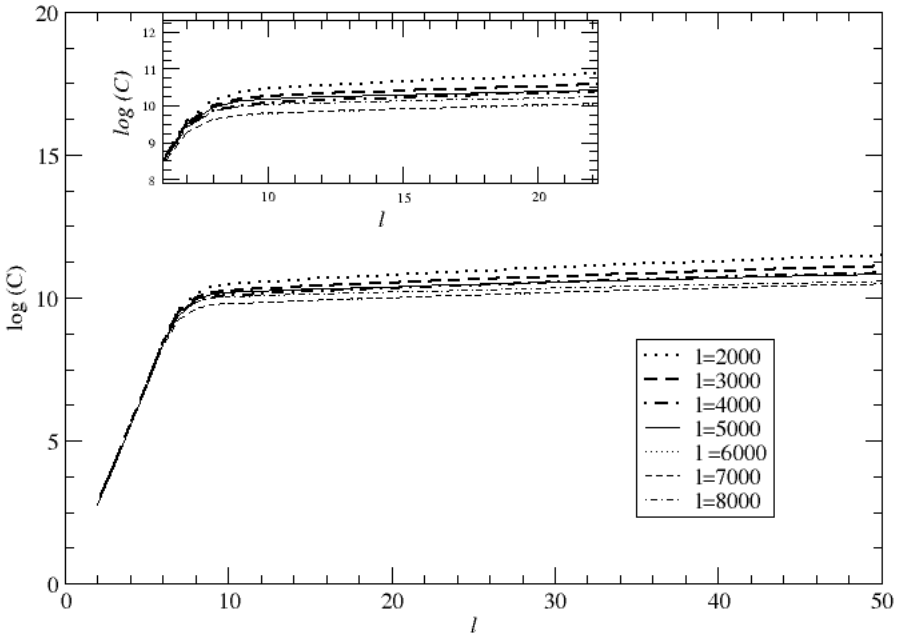

Figura 5.18: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro l, en su version de la longitud de palabra contra el logaritmo de $C(l)$.también se hace un acercamiento para poder observar la longitud a partir de la cual tenemos un cambio de comportamiento de exponencial a lineal.

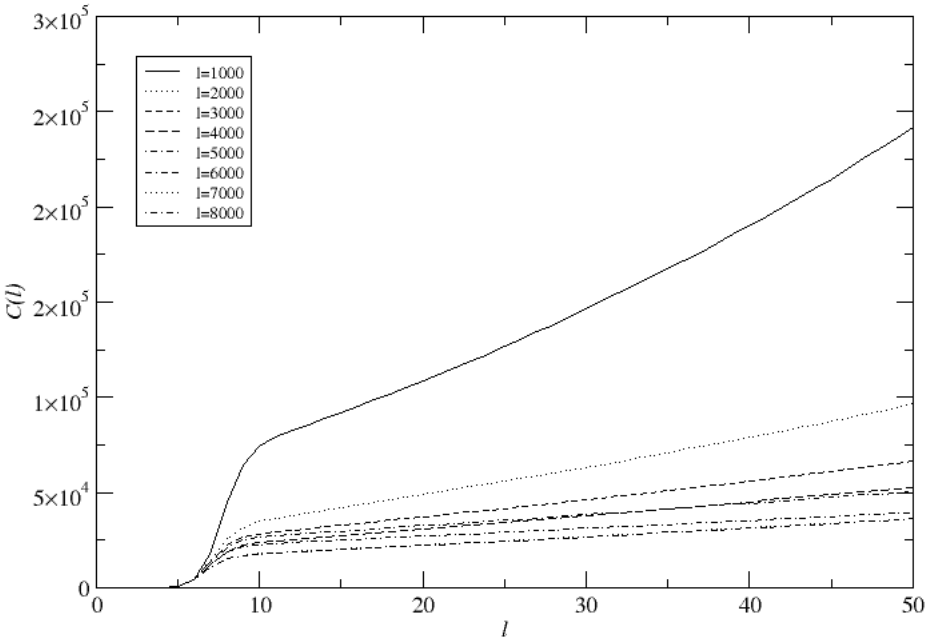

Figura 5.19: gráfica de los distintos perfiles de la fc obtenidos para distintos valores de parámetro l, en su version de la longitud de palabra contra $C(l)$

Por ultimo, en el cuadro 5.8 desplegamos los resultados obtenidos para la variación de la longitud de palabra l y el calculo de la pendiente del ajuste lineal que corresponde a cada valor de esta. La gráfica 5.18 refleja el perfil de la fc en su version l contra el logaritmo de $C(l)$, y en la gráfica 5.19 la version l contra $C(l)$.Obtuvimos en la longitud de cruce un valor de $l_c =$ igual a 8 y 7 . Es posible observar que para este parámetro , la variación de la pendiente a medida que incrementamos l , provoca una disminucion en esta ligeramente .Para la l_c, al igual que en el caso $PSBR$, parece haber un cambio mínimo a diferencia de PSD y de t.En efecto ,nuevamente esto podría darnos una herramienta mas que nos ayude intuitivamente a ajustar α y sin modificar en exceso a la longitud de cruce.

5.7.1. Comparación entre las especies y los Modelos evolutivos

Comparación entre las especies obtenidas

Procedemos a realizar una comparación entre las especies analizadas en este trabajo bajo las propiedades estadísticas obtenidas computacionalmente. Básicamente podemos encontrar el orden comparativo de la siguiente forma jerarquizada:

l	α	l_c
1000	8.52	8
2000	7.47	8
3000	6.97	8
4000	6.80	7
5000	6.62	7
6000	6.39	7
7000	6.38	7
8000	6.29	7

Cuadro 5.8: Modelo Massip-Arndt.Variaciones en longitud con un PSBR = 0.0 y un PSD = 0.6 (ambos constantes) sobre 1000,000 iteraciones. l se trata del cambio de longitud mientras que α es la pendiente tras esa variación.

1. ***Comparación entre Homosapiens, Gorilla gorilla gorilla y Mosquito Aedes*** .

De la familia Hominidae la cual consta de 4 géneros y 8 especies vivientes y, de donde hemos seleccionado 2 de ellas tenemos lo siguiente: mientras que el homosapiens presenta un tamaño de secuencia codificante entre 772K y 28M de pares de bases nitrogenadas, el Gorilla gorilla gorilla presenta un tamaño de entre 1M a 10.5M de pares de bases. Para la longitud de cruce se observa que mientras el gorilla posee una l_c tal que $13 \leq l_c \leq 18$ con una frecuencia mayor para la longitud de tamaño 15. El homosapiens posee un l_c ta que $11 \leq l_c \leq 17$ con una frecuencia mayor para la longitud de tamaño 14.Finalmente para la pendiente de la región lineal observamos en promedio una $\overline{\alpha} = 3{,}76$ para el gorilla y una $\overline{\alpha} = 6{,}18$ para el homosapiens. Para el mosquito aedes el cual solo consta de tres cromosomas, se obtuvo una pendiente $\overline{\alpha} = 6.73$ mientras que la longitud de cruce fue mayor que la de los otros miembros del dominio del reino Animalia.

2. ***Comparación entre familias Bacterianas .***

De las bacterias estudiadas, Es posible observar que dentro de las tres familias propuestas , las Enterobacteriales presentan una menor funcíon de complejidad,sin embargo, para la longitud de cruce (Cuadro 5.2) mientras que la Yersenia es igual a la Bacillus cererus y Marinobacter (es decir tenemos la misma l_c para tres especies de familias distintas), la Salmonella presenta una l_c mayor que la de Virgibacillus y Marinobacter .Para la pendiente α de la parte lineal de cada una de las especies analizadas obtuvimos que las Enterobacteriales poseen en promedio un $\overline{\alpha} = 2.42$, mientras que las Baccillaceae tienen un $\overline{\alpha} = 2.02$ y finalmente la familia de Alteromonadaceae un $\overline{\alpha}$= 5.22.

3.***Comparación entre algunas especies de reinos distintos***. Finalmente de la comparación entre Reinos mientras que las Enterobacteriales y las Baccillaceae tiene una pendiente promedio menor a las eucariotas descritas anteriormente, las Alteromonas poseen en promedio una pendiente mayor al homosapiens. Esto último podría darnos evidencia de que la parte lineal de la función complejidad para las secuencias codificantes, pueda estar sujeta a diversas características genómicas que nos puedan ayudar a obtener similitudes independientemente del reino entre especies .

De la comparación entre los modelos evolutivos y las secuencias genómicas reales

A continuación mostramos en la gráfica de la figura 5.20 la superposición entre los cromosomas del homosapiens,gorilla gorilla gorilla aedes y el regimén bacteriano con las simulaciones que mejor se ajustaron al perfil de estas secuencias reales. El barrido que mejor pareció ajustarse a las secuencias reales fue el de duplicación de segmento con las características descritas en el cuadro 5.6. El conjunto de cromosomas dentro de este despliegue de simulaciones corresponden a Virgibacillus y Bacillus cererus, los cromosomas del Homosapiens $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 19, 20, x\}$, los cromosomas del gorilla gorilla gorilla $\{1, 3, 5, 11, 12, 19\}$, y finalmente los tres cromosomas del Aedes aegypti. El cromosoma que se encuentra en extremo inferior de estas simulaciones corresponde al cromosoma 10 del homosapiens, mientras que el cromosoma que se encuentra en el extremo superior de la gráfica 5.20 se trata del mosquito Aedes numero 3. Resulta de interés la información que podemos obtener de este último hecho. El extremo inferior del despliegue de las simulaciones se trata de una duplicación de segmento de 0.9 a una longitud constante de $l = 1000$ sobre 100,000 iteraciones, mientras que el extremo superior se trata de una probabilidad duplicación de 0.1 de palabras .Esto podría darnos certidumbre sobre los principales mecanismos de evolución como la replicación de secuencias que codifican para proteínas, evolución del sistema y ademas una comprensión sobre los parámetros que necesitamos fijar para ajustar tanto la pendiente α, así como la longitud de cruce l_c.Por ejemplo en la figura 5.20, sabemos que si el cromosoma 10 antes mencionado se encuentra a probabilidades próximas a un PSD=0.9, entonces a un a longitud de palabra determinada y constante sobre el mismo numero de iteraciones,tendríamos que disminuir a 0.6 el PSD para obtener un perfil de complejidad similar al cromosoma 6 del homosapiens o al cromosoma 2 del mosquito Aedes,poniendo de manifiesto algunas diferencias sutiles en base a la información estadística que poseemos. En cualquier caso, parece que el modelo de Massip Arndt nos da una buena aproximación de los perfiles descritos en este trabajo.

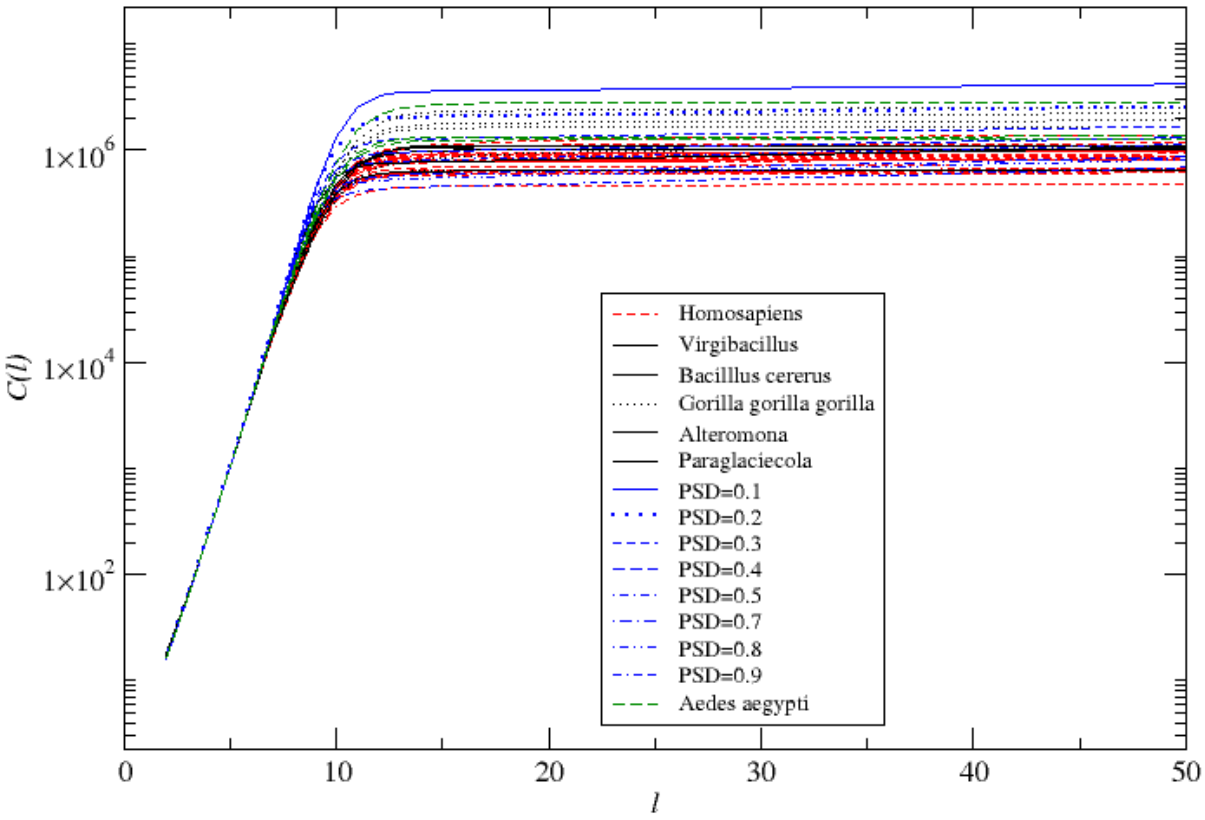

Figura 5.20: Gráfica que nos muestra la comparación de todas las especies utilizadas,superpuestas con el abanico de valores de Massip Arndt para la duplicación de segmento (PSD).

Capítulo 6

Conclusiones

De las gráficas de la función complejidad obtenidas para los 23 cromosomas del homosapiens, fue posible observar una diferencia evidente entre las dos familias (secuencias codificantes y secuencias completas), las cuales constan de un comportamiento exponencial-lineal para aquellas que tienen exónes, y un comportamiento puramente exponencial para aquellas que tienen intrones.

A su vez, de los tres modelos de evolución descritos en el presente trabajo, Al parecer para Expansión-Modificación no fue posible encontrar un perfil de la función complejidad similar al de las secuencias reales utilizadas.De la misma forma para el modelo de Zacks, no fue posible encontrar para los valores de parámetro manipulados, resultados consistentes con genomas reales. Sin embargo para el modelo de Massip-Arndt, fue posible reproducir un perfil de la fc similar al menos para las secuencias codificantes encontradas en los 23 cromosomas del homosapiens,el gorilla gorilla gorilla, el mosquito Aedes y algunas especies de las tres familias de bacterias estudiadas.Desde luego, esto resulta importante ya que variando ciertos parámetros de Massip-Arndt,ahora podríamos tener una idea más especifica de la diferencia entre cromosomas de la misma especie, así como especies de distintos reinos o familias en concreto como se especificara en el siguiente punto.

A través de la variación de los parámetros utilizados para el modelo de Massip-Arndt, obtuvimos distintos perfiles que ademas de seguir un comportamiento exponencial-lineal, pudimos modificar la altura o el valor del logaritmo de $C(l)$, la longitud de cruce l_c, y la pendiente en donde tenemos la región lineal en el perfil de la fc, todo esto en la version de la longitud de palabra l contra el logaritmo de $C(l)$ (para las secuencias codificantes sin embargo). De igual forma, a través de la manipulación en general de estos parámetros, fue posible observar que a medida que incrementábamos por ejemplo la probabilidad de remplazamiento de base simple, la pendiente aumentaba ligeramente, mientras que si aumentábamos la probabilidad de duplicación de segmento, lo que obteníamos era una disminución de la pendiente.Esto podría indicarnos, que en general α este fuertemente vinculada con la complejidad resultante tras incrementar las mutaciones o multiplicar las palabras,pudiendo darnos información clasificadora sobre que especies en concreto tienen un porcentaje o concentración mayor de complejidad.

Esto ultimo resulta de vital importancia, ya que posiblemente al conseguir ajustar a través de las simulaciones obtenidas con los valores de parámetro adecuados el perfil de la función complejidad real, tal vez podríamos extraer de igual forma , información estadística o bilógica que nos ayude a comprender como es la estructura de las secuencias genómicas de ciertas especies en la región donde se presentan exónes mas específicamente. Por ejemplo,resulta interesante observar que por lo menos uno de los cromosomas del mosquito Aedes tiene un perfil de la fc mayor a algunos de los cromosomas de las demás especies utilizadas, lo cual podría sugestionar que a medida de que las especies tengan menos cromosomas, mayor podría ser las mutaciones en ellas (recordando que uno de los cromosomas del mosquito se aproxima al perfil de la simulación con probabilidad de duplicación de segmento de tan solo 0.1 . De la misma forma, fue posible observar una diferencia significativa de la pendiente obtenida, para cada perfil de nuestras especies del comportamiento lineal resultante. Mientras que para el homosapiens obtuvimos una $\overline{\alpha} = 6.18$, para el gorilla obtuvimos una $\overline{\alpha} = 3.76$ (especies de la misma familia), que resulta al menos mayor que dos de las familias Bacterianas estudiadas pero no de las Alteromonadaceae con una $\overline{\alpha} = 5.22$. En general, aunque en principio seria conveniente un análisi mas exhaustivo sobre mas especies de cada reino, en principio podríamos establecer una clasificación al menos entre familias correspondientes a cada especie en concreto. Esto debido a que los valores reportados al menos en la pendiente, nos dio una dinánimca en general que nos permitiría en principio aproximarnos a una clasificación precisa desde el ámbito estadístico .De cualquier forma, parece que nuestra heramienta (fc) nos puede ayudar en encontrar diferencias sutiles, que tienen que ver con concentración de mutaciones o duplicaciones a cada especie en concreto.

Capítulo 7

Referencias

[1] David Koslicki,(2011)Topological entropy of DNA sequences.

[2] Richard F.Voss, (1992)Evolution of long-Range Fractal Correlations and $1/f$ Noise in DNA Base Sequences.

[3] Shuilin Jin, Yon Wang (2013) A Generalized Topological Entropy for A nalyzing the Complexity of DNA Sequences.

[4] C.-K.Peng (1992) Long-Range Correlations in nucleotide Sequences.

[5]S.V. Buldyrev.(1995) Long-Range Correlation properties of coding and noncoding DNA sequences:GenBank analysis.

[6] R Salgado García and E Ugalde (2016) Symbolic Complexity for nucleotide sequences: a sign of the genome structure.

DR. VICTOR BARBA LÓPEZ
COORDINADOR DEL POSGRADO EN CIENCIAS
PRESENTE

Atendiendo a la solicitud para emitir DICTAMEN sobre la revisión de la TESIS titulada **"Análisis Estadístico de Secuencias Genómicas"** que presenta el alumno **Alberto Campos Aguirre (10009534)** para obtener el título de **Maestro en Ciencias.**

Nos permitimos informarle que nuestro voto es:

NOMBRE	DICTAMEN	FIRMA
Dr. Federico Vázquez Hurtado CINC-UAEM	Aprobado	
Dr. Joaquín Escalona Segura CINC-UAEM	Aprobado	
Dr. Hernán Larralde Ridaura ICF-UNAM	Aprobada	
Dr. Gustavo Martínez Mekler ICF-UNAM	Aprobado	14/11/19
Dr. Raúl Salgado García CINC-UAEM	Aprobado	