UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"COHERENCIA Y CARÁCTER LINGÜÍSTICO EN SECUENCIAS DE ADN `NO CODIFICANTE'"

T E S I S

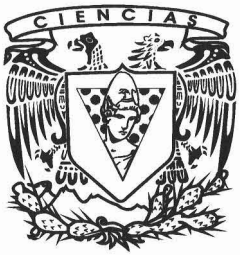
QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G A

P R E S E N T A:

BRENDA CANTÚ BOLÁN

DIRECTOR DE TESIS:
DR. ENRIQUE HERNÁNDEZ LEMUS



FACULTAD DE CIENCIAS
UNAM

2005

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:
"Coherencia y carácter lingüístico en secuencias de ADN no codificante"

realizado por *Cantú Bolín Brenda*

con número de cuenta *09133425 5* , quien cubrió los créditos de la carrera de: *Biología*

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	*Dr. Enrique Hernández Lemus*
Propietario	*Dr. Faustino Sánchez Garduño*
Propietario	*Biol. Blanca Rosa Hernández Bernal*
Suplente	*M.V.Z. Mario Javier Soriano Bautista*
Suplente	*Dr. Hever Raúl Aristóndo Marín*

Consejo Departamental de *Biología*

M. en C. Juan Manuel Rodríguez Chávez

FACULTAD DE CIENCIAS
UNIDAD DE ENSEÑANZA DE BIOLOGIA

DEDICATORIA

Para יהוה

Pampa ta quiichijqui yejcualtzin, ica tlalnamiquiliztli nochi catli onca para cuali: amotleno poloue huan amotleno mocahua. ¿Nihuelijyaya tlahpejpeni ce nescaiotl cachi cuali para nextilia moyejcualtzinliztli tlen nochi catli tichiuajki ipan tlaltipac huan motlahcuiloli? Ini eli para nehuan seyoc cualitlamantli para mitstlasotla ika notzonteco huan noyolo. Tlasocamati miyac pampa ta tiítztoc nechca nochipa huan pampa tieli para nehuan ce Teotlnochichicaualiztli, ce tata ika tlasotiliztli, ce tlasoiknimelahuac huan ce hueyitlapalehuini.

Para Nonatzin

Pampa tieli cuali ika nehuan, por nochi motlasotiliztli huan mochicaualiztli. Nineki uelis ixkopina monescaiotl. Nimitsltlasotla.

Para Notata

Pampa tipiajki chicaualiztli para techtemojtoc huan tinextixmati mocone, huan chiua ini mojmostla.

Para Nochanehua

Miyac tonalme amo nitlajtoa miyac, pero nianmechtlasotla ica nochi noyolo. Tlasocamati por nochi pampa anmopaleuia huan anmotlasotilistli.

Para Notlasoiknihua por nechtlaijyouiya huan nechtlasotla ika nochi notlamanme amocualme.

Agradezco a mis cinco sinodales: Dr.Enrique Hernández Lemus, Biol. Blanca Rosa Hernández, Dr. Faustino Sánchez, M.V.Z. Mario Soriano y Dr. Ramón Arzápalo por aceptar serlo, así como por revisar y corregir este trabajo y a todas las personas que con su apoyo, conocimiento y/o cariño me ayudaron a terminar este ciclo en mi vida y a comenzar uno nuevo.

COHERENCIA Y CARÁCTER LINGÜÍSTICO EN SECUENCIAS DE ADN '*NO CODIFICANTE*'

Tesis que para obtener el grado de Licenciado en Biología presenta

Brenda Cantú Bolán[1]

Director: Dr. Enrique Hernández Lemus

[1]Facultad de Ciencias, Universidad Nacional Autónoma de México.

Índice general

Introducción

Cuncta fecit bona in tempore suo et mundum tradidit disputationi eorum ut non inveniat homo opus quod operatus est Deus ab initio usque ad finem.
Ecl. 3:11

En las décadas que han transcurrido desde el descubrimiento de la estructura del código genético por Watson, Crick y otros, muchas hipótesis acerca del contenido informático de éste han sido propuestas. Algunos científicos, tras examinar el importante papel que la modificación tiene en la generación de proteínas, han concluido que éste es el papel fundamental de la información contenida en las secuencias genéticas. Tal hecho ha llevado a que muchos especialistas piensen que los segmentos de ADN que no participan en la síntesis proteica carecen por completo de importancia. Evidentemente, esta suposición conlleva implicaciones serias, en particular si consideramos que la mayor parte del genoma de los seres vivos (alrededor del 95 %) no codifica directamente para proteínas. Este mal llamado ADN *basura* o *ADN no codificante* posee, sin embargo, propiedades estadísticas trascendentes.

Por otro lado, los estudios lingüísticos indican que si se agrupan letras en determinadas combinaciones se obtienen palabras, si a su vez se unen palabras se genera una frase; dichas letras y dichas palabras, e incluso las frases no están dispuestas aleatoriamente, sino que tienen un orden tal que éstas entrañan un significado (es decir, poseen una cierta cantidad de información asociada). Si entonces se juntan frases, se produce un texto coherente que posee información ordenada. Igualmente es posible referirse a un genoma: si se agrupan nucleótidos en combinaciones específicas, se obtiene un gene, si se juntan genes se producirá una secuencia genómica y si se unen varias de éstas se produce un genoma. Los genes conllevan características que se expresan en el organismo a través de las proteínas y el genoma es comparable a un texto biológico donde se hace un recuento de cada rasgo del individuo [23, 6].

Siguiendo con misma idea, se sabe que las proteínas constan de combinaciones de 20 aminoácidos alineados siguiendo diversas ordenaciones en cadenas de longitud arbitraria. Los escritos, por otro lado, constan de combinaciones de 28 letras (en el alfabeto para hispanohablantes), más una serie de signos de puntuación, alineados en distintas ordenaciones en líneas que suelen plegarse para que estén dispuestos a lo largo de las páginas. Del mismo modo que las letras de nuestra lengua pueden combinarse en una enorme cantidad de modos, también cabe combinar los 20 aminoácidos de que se vale la vida en una ingente variedad de

proteínas distintas.

Se han realizado estudios aplicando este tipo de analogías a secuencias genómicas. Son sobresalientes los estudios del equipo de Eugene Stanley [75, 76, 77, 78], Vladimir Dokholyan [41, 42, 43, 99] y José Leonel Torres [82] respecto a las propiedades lingüísticas y estadísticas del ADN, particularmente de los fragmentos considerados no codificantes. Estos grupos de trabajo fueron pioneros en la aplicación de algoritmos de la mecánica estadística y de la lingüística matemática en secuencias genómicas, así como en lenguajes estructurados tanto naturales (Torres empleó el español y tres lenguas mesoamericanas) como artificiales (lenguajes de cómputo). Es apropiado también destacar el trabajo de Daniela Filippini [47] y su equipo de investigación, además del de Duga [46] sobre el papel que el ADN intrónico tiene en la formación y síntesis de estructuras nucleolares pequeñas en un tipo de ARN precursor necesarias para la síntesis de ARN ribosomal.

En este trabajo, se teoriza que, así como es posible darse cuenta en un texto coherentemente ordenado que las palabras no están dispuestas al azar, el orden en la disposición de nucleótidos en un genoma es prueba de la coherencia y, por lo tanto de la codificación (contenido informático) de éste en su totalidad. Así, se pretendió mostrar por medio de métodos estadísticos que el ADN no codificante tiene una distribución no aleatoria similar a la de un texto lingüístico coherentemente ordenado. Se procedió a estudiar las distribuciones de probabilidad asociadas con este ADNnc por medio de métodos estadísticos a fin de inferir algunas de sus propiedades lingüísticas. Se discutieron brevemente las posibles implicaciones de los resultados obtenidotes y sus aplicaciones.

Detallando un poco más: se realizó un análisis de algunas de estas propiedades estadísticas fundamentalmente relacionadas con el contenido informático, el grado de complejidad, y la coherencia (en sentido lingüístico) que poseen segmentos de secuencias de ADN vistos desde el punto de vista de un texto o código en el sentido lingüístico. Los análisis realizados tienen que ver con la distribución de frecuencias o repeticiones de ciertas unidades lingüísticas a través de enfoques como la ley de Zipf, que consiste en asignar un rango relacionado con la importancia relativa de una palabra dentro de un texto y estudiar la relación que esta función tiene con el significado y coherencia del texto en cuestión.

Se encontró que en los lenguajes bien estructurados tal función toma la forma de una ley de potencias. Resultó significativo este hecho en las distribuciones de frecuencia estudiadas, pues como se sabe, el comportamiento del tipo ley de potencia en las funciones de distribución implica un fenómeno estadístico conocido como persistencia en la memoria relacionado con procesos estocásticos no Markovianos. Un análisis realizado sobre tales distribuciones de frecuencias nos indica a través de un estudio basado en geometría fractal el grado de complejidad de las secuencias genómicas estudiadas, dado fundamentalmente, por una cantidad conocida como la dimensión de Hausdorff. Se analizó bajo el enfoque de Shannon-Weaver el contenido de información que tales secuencias poseen, esto se hizo por las siguientes razones,

en primer lugar este modelo está libre de parámetros ajustables que otras medidas entropicas (i.e. todas las entropías de Renyí) si poseen, por otro lado se sabe que la entropía de Shannon constituye una cota mínima al contenido de información de una cadena de símbolos, por lo que, en todo caso el contenido informático medido por otros parámetros será simepre mayor. Adicionalmente, pudo comprobarse que la interacción bioquímica que generan en el ADN codificante las llamadas proporciones de Chargaff, también se halla presente, aunque de una manera distinta en el ADN no codificante [23].

Con la finalidad de estudiar un espectro amplio de seres vivos, este estudio se efectuó sobre el genoma completo de una bacteria (*Mycoplasma pneumoniae*), así como en fragmentos del ADN de la mosca de la fruta (*Drosophila melanogaster*), de un tipo de gato doméstico (*Felis catus*), del pino negro japonés (*Pinus thunbergii*), y sobre un generador aleatorio como control. Por otro lado, la comparación lingüística se llevó al cabo empleando como texto referencial una versión en español del libro bíblico de Isaías.

Todos estos estudios aportaron pruebas contundentes del sobresaliente grado de orden (en sentido estadístico), cantidad de información comunicada, coherencia, estructuración y complejidad presentes en el ADNnc. Es digno de destacar el hecho de que, debido a la naturaleza de los métodos de estudio empleados en este trabajo (cuyo carácter está basado en propiedades matemáticas y lingüísticas generales), las repercusiones a nivel específico, particularmente en el campo de la bioquímica y la biología molecular, serán necesariamente limitadas, por lo que será apropiado considerar las conclusiones de esta tesis como un auxiliar más que como un marco referencial preestablecido para estudia a en estas disciplinas.

El presente trabajo representa un análisis de Lingüística matemática y teoría de la Información aplicado a secuencias genómicas, por lo que las conclusiones derivadas de éste se circunscriben al ámbito global lingüístico, no aportando información concluyente respecto al comportamiento bioquímico de segmentos específicos del genoma. Este trabajo no pretende considerar el enfoque macroscópico dado por la Biología del desarrollo o la Fisiología. Tampoco está realizando un estudio a nivel minucioso a nivel microscópico como el de la Bioquímica. Más bien considera escalas de estudio intermedias que reflejan el comportamiento de fragmentos genómicos de distintos organismos. Tal como la estructura alfa hélice del ADN no se determinó por observación directa a partir de observaciones experimentales de las proteínas, sino a partir de consideraciones teóricas fundamentadas en el estudio de sustancias más sencillas, análogamente es posible inferir de manera teórica la no aleatoriedad del genoma.

Capítulo 1

Generalidades

1.1. Resumen

Usualmente, los especialistas tendían a pensar que un porcentaje importante (aproximadamente el 95 %) del ácido desoxirribonucleico (ADN) estaba constituido por lo que se conoce como *ADN no codificante* (ADNnc) [5], debido a que -hasta hace muy poco- no se conocían con exactitud los mecanismos de expresión genética de éste. El papel que se solía asignar a estos segmentos genómicos era el de meramente un resguardo sin función biológica propia contra el ataque aleatorio de elementos mutágenos [5]. Sin embargo, si se realiza un análisis estadístico global utilizando las técnicas de la lingüística estadística, es posible reconocer patrones coherentes y una estructura bien definida en el ADNnc [75, 76, 77, 78] que no podrían encontrarse si los nucleótidos que lo conforman estuviesen distribuidos al azar. Se utilizaron algunos métodos para cuantificar la correlación y la coherencia, tales como la *Entropía Informacional de Shannon*, la determinación de la *dimensión de Haussdorff*, la proporción de *bases complementarias en los genomas*, la construcción un *conjunto fractal característico* para cada secuencia dada y el análisis del *proceso de Markov* asociado con las distribuciones de frecuencia experimentales. Además, se realizó un estudio conocido como *Análisis de Zipf* a fin de comparar el carácter *lingüístico* de tales secuencias con el que posee un texto en un lenguaje natural *bien estructurado* (en este caso, el español). En este capítulo se describen la importancia de este trabajo desde el punto de vista teórico y práctico, así como algunas aplicaciones y se esboza su relación y pertinencia en el marco de la biología molecular contemporánea y ramas asociadas.

1.1.1. Objetivos:

1. Mostrar que el ADN *no codificante* tiene una distribución no aleatoria por medio de métodos estadísticos como 1) la Entropía Informacional de Shannon, 2) distribución de probabilidades [conteo de cúmulos y Repeticiones Diméricas en Fila (RDFs)], y 3) el cálculo de la dimensión fractal de Haussdorff, así como una posterior comparación

contra un control y un texto lingüístico coherentemente ordenado.

2. Estudiar las distribuciones de probabilidad asociadas con este ADNnc por medio de métodos estadísticos (Procesos de Markov, determinación de longitud de correlación y estudios de presencia de orden de largo alcance), con el fin de inferir algunas de sus propiedades *lingüísticas*.

3. Analizar brevemente las posibles implicaciones de los resultados obtenidos y sus aplicaciones.

1.1.2. Hipótesis

1. Si el ADNnc no tiene una distribución aleatoria (como generalmente se plantea), entonces contiene información aún no identificada que comunica algo y, por lo tanto, podría resultar útil para el organismo.

2. Por otro lado si es posible mostrar que la concentración de RDFs (localizadas en el ADNnc) tiene un comportamiento del tipo Ley de Potencia, entonces podemos concluir que posee orden de largo alcance a lo largo de sus nucleótidos. Este hecho apoyaría la hipótesis de no aleatoriedad del ADNnc

- **Importancia técnica**

  1. Actualmente algunos cálculos genómicos realizados por computadora (como el proyecto Genoma Humano [56]) consumen muchos recursos computacionales y tiempo. Si encontráramos peculiaridades *lingüísticas* y/o comunicativas (en sentido estadístico) en los genomas a estudiar, reduciríamos significativamente el tiempo de cómputo al optimizar los algoritmos de búsqueda [un ejemplo real: dado que los cúmulos de tamaño 5 en adelante de AT/TA no están presentes en el genoma del gato, desde el principio omitiremos buscarlos y no los tomaremos en cuenta para cálculos posteriores].

  2. En caso de que el ADNnc poseyera una distribución ordenada -tal como se teoriza en este trabajo- su estructura geométrica sería ordenada también y existiría en éste una regla general bien definida. Se podría entonces investigar sobre una sola maquinaria molecular que serviría para todos los procesos orgánicos. Veríamos entonces cómo la información dada por ADN codificante (ADNc) y ADNnc interactúa en ambas direcciones.

  3. Las características generales del ADNnc lo hacen especialmente útil para su aplicación a la identificación forense [6]. Como se puede deducir de su trascendente función, el ADN esencial está formado por secuencias altamente conservadas con muy pocas variaciones interindividuales e intergeneracionales, ya que de lo contrario se podrían ver afectadas funciones básicas para la vida de las personas. Los

mínimos cambios que tienen lugar, cuando son viables, aumentan el polimorfismo de proteínas y enzimas, aunque también pueden tener efectos negativos.

Por el contrario, el ADNnc presenta una gran variabilidad de unos individuos a otros, ya que estas secuencia no son conservadoras al no afectar sus cambios a la fisiología normal del individuo. Las variaciones debidas a cambios de bases sencillos, procesos de inserción-delección o de intercambio de ADN (recombinación) durante la formación de las células germinales (meiosis), hacen que se modifiquen el número de repeticiones o el orden de las bases de un determinado fragmento repetitivo, pudiendo producirse en un locus sencillo o en múltiples loci, siendo este el origen de la variación que hace que no haya dos personas, a excepción de los gemelos univitelinos, que tengan la misma secuencia del ADN.

- **Importancia práctica**

1. Una célula no utiliza la expresión de características propias de otro tipo de tejido para dividirse y dispersarse en el medio sin control [63, 86]. No cabe duda de que todo ello depende del proceso de transcripción a ácido ribonucleico (ARN) y luego a proteína de ciertas porciones del mensaje genético codificado en los cromosomas. La cuestión es descifrar cómo se conectan y desconectan las diversas porciones del mensaje genético. Esa línea de investigación suma la posibilidad obvia de comprender, primero y quizá más delante de controlar esta enfermedad en la cual las células dejan de obedecer las órdenes genéticas que dictan su comportamiento.
2. Se ha determinado que la presencia de cierta enfermedad (un tipo específico de cáncer) obedece a la existencia de RDFs de tamaño 11 de TAs. Sería posible ver en qué cromosoma pasa eso (cuál tiene más de estas repeticiones y usar terapia génica para curar esa enfermedad), atacar el gene enfermo,o cambiarlo [63]
3. Se sabe que las Repeticiones Diméricas en Fila (RDFs) en el genoma constituyen una importante fracción del ADNnc y son relativamente escasas en las secuencias que codifican para proteínas. Las RDFs son de considerable interés teórico y práctico debido a su alto polimorfismo. Por ejemplo: se sabe que las RDFs del tipo (CA)n se expanden debido al plegamiento en el proceso de replicación; estos errores frecuentemente son eliminados por la enzima reparadora MSH2. Sin embargo, una mutación en el gene MSH2 lleva a una expansión descontrolada de repeticiones que es una causa común del cáncer de ovario [86] y mecanismos similares se han atribuido para otros tipos de cáncer. Estas investigaciones sobre las RDFs en ADNnc podrían conducir a avances en el diagnóstico oportuno y posible prevención de enfermedades de esta clase. Las RDFs se denominan también ADN-satélite, porque con frecuencia pueden segregarse de la masa del ADN por centrifugación en cloruro de cesio. El número de RDFs de un tipo determinado puede variar entre distintos individuos, dando lugar a una huella única de ADN [75, 76, 77, 78].

4. Este trabajo pretende mostrar, entre otras cosas, por medio del análisis de la función de correlación entre *palabras genéticas*, que son los segmentos de ADNnc y no de ADNc aquellos más resistentes a cambios a lo largo del tiempo [77, 99]. Esto pudiera ser aplicado a investigaciones sobre tasas de mutación, ya que el investigador sabría con qué tipo de ADN trabajar desde el principio. Por otra parte, sería también posible determinar cuáles son los genes que codifican para ciertas adaptaciones. Las tasas de acumulación pueden cuantificar la tendencia de las secuencias del ADN a conglomerarse y pueden ser utilizadas en estudios posteriores de agregados de nucleótidos en las secuencias del código genético [76] . Por ejemplo, diferentes tasas de dímeros variados pueden sugerir diferentes tasas de mutación, específicas para cada dímero y organismo.

1.1.3. Aplicaciones y utilidad

Biología

- **Diferenciar entre especies similares**

  En Biología es sumamente común el encontrar dos organismos tan semejantes, que es difícil para el investigador concluir si se trata de una misma especie o ha descubierto una nueva. La comparación de datos obtenidos de los mismos algoritmos estadísticos que se emplean en esta tesis pudieran ser de utilidad para definir lo anterior.

- **Encontrar tasas de mutación**

  Al observar la distribución y la longitud de oligómeros en el genoma y aplicando posteriormente el método de conteo de cúmulos es posible determinar dichas tasas. Por otro lado, sería una ventaja hacer estos cálculos solo en el ADNc (que, como se sostiene en esta tesis, es el más susceptible a sufrir mutaciones) que es sólo un 5 % del genoma completo.

- **Iniciación de la replicación y la recombinación del Virus de Inmunodeficiencia Humana (VIH) en el sistema inmunológico**

  Las modificaciones de largo alcance de la estructura de la cromatina por facilitadores pueden también estar envueltos en la iniciación de la replicación y la recombinación de VIH en el sistema inmunológico [59]. Por medio del estudio más profundo del comportamiento del ADNnc (ya que, este aspecto se desarrollará con más detenimiento en el capítulo 5, es aquel que muestra una distribución del tipo ley de potencia) podrán comprenderse mejor tan importantes procesos biológicos.

- **Análisis de secuencias de ADN**

  Si podemos encontrar una regla general que nos indique cuáles son los exponentes característicos en la distribución de frecuencias de aparición de cadenas determinadas de nucleótidos en un genoma o fracciones de éste (Zipf, Shannon, Haussdorff, etc.), entonces podremos diferenciar entre ADNc y ADNnc para realizar las investigaciones

que nos convengan y no tendremos que trabajar con el genoma completo (Ilustremos la aplicación de la siguiente manera: la fórmula que determina la sucesión de Fibonacci no es trivial, por lo tanto, a menos que uno descifre la regla, los números resultantes de ella parecen aleatorios [15] sin embargo, una vez que se conoce el código es posible discriminar si un número pertenece o no a la serie. Lo mismo ocurre cuando sabemos qué secuencia corresponde a cuál función)este podría reducir significativamente el tiempo de cómputo utilizado y posibilitar el avance de la farmacología genómica, es decir el estudio con fines terapéuticos de las propiedades del códico genético.

- **Solución a problemas de salud asociados con la exposición a factores que dañan el ADN**

  Estudios recientes de comparación de genes entre los genomas secuenciados de los diferentes organismos eucariontes animales demuestran que, en el caso de la mosca, existen en este organismo más del 65 % de los genes que en los humanos son responsables de las enfermedades congénitas hasta ahora identificadas en la especie humana [96]. En otras palabras, en la mosca se han encontrado al menos 177 genes que tienen equivalencia con genes humanos involucrados en enfermedades genéticas. Ejemplo de estos casos es el de una mutante de la mosca que presenta una patología similar a la que se observa en pacientes con la enfermedad de Parkinson. Indudablemente, el análisis del funcionamiento de estos genes y de los productos proteicos en este insecto permitirá en tiempos mucho más cortos, conocer con gran detalle las bases moleculares involucradas en muchas enfermedades genéticas humanas, a través del conocimiento de lo que ocurre en otros organismos modelo como la mosca, abriendo así posibilidades extraordinarias y novedosas para el tratamiento de este tipo de enfermedades en el humano.

  Los métodos lingüísticos-comunicativos y estadísticos empleados como se hace en este trabajo podrían ser una herramienta útil para comparar dichos genes y obtener así información que dirigiera a resolver este problema.

- **Generación de recombinantes**

  Los análisis de algunos tipos de mutación ofrecen nuevos enfoques al estudio detallado del proceder de la recombinación. Abren las puertas también a la generación artificial de esos mutantes [1, ?](por cruzamiento o por ingeniería genética) que podría aplicarse, por ejemplo, al control de las plagas de insectos y también a alimentos transgénicos y medicamentos como en el caso del factor coagulante *Novosen* [40]. El proceso de recombinación depende fuertemente de la estructura sintáctica en el polímero de ADN y esta estructura está determinada por las distribuciones de frecuencia para los oligómeros.

Lingüística

- **Estudio de la riqueza lexicográfica de un lenguaje**

Tomando a un genoma como control (dado que, según este trabajo, presenta un exponente de Zipf tan ordenado como el de cualquier lengua [90]) puede hacerse una comparación con un texto lingüístico y comparar la cantidad de palabras, longitud de éstas, frecuencia de aparición y exponentes de Zipf.

item **Comparación de complejidad entre idiomas**

Si se logra mostrar que el genoma es un texto biológico coherentemente ordenado (tanto como una lengua natural), sería posible tomar sus gráficas y exponentes como patrón para comparar contra un mismo texto lingüístico en distintos idiomas con la finalidad de conocer la complejidad de éstos.

- **Complejidad del lenguaje en niños y adultos y personas con diversas circunstancias**

  Se han estudiado las gráficas de Zipf que describen la aparición de elementos diferentes en un grupo dado como función de su rango y se ha definido la distancia entre dos gráficos de Zipf como característica de la diferencia entre los dos grupos lexicográficos [90, 82].

  Esta distancia es una medida cuantitativa de las diferencias de rango de los mismos elementos en ambos grupos y mediante la cuantificación de ésta pueden por ejemplo compararse dos secuencias genómicas constituidas por los mismos nucleótidos y calcular su diferencia. La diferencia entre ambos genomas se define como la distancia media cuadrática entre los rangos de todas sus palabras comunes. Esta medida puede ser útil para probar qué tan cerca (en rango) están secuencias de genomas diferentes o aún secuencias codificadas en diferente lenguaje.

- **Mejor calidad en la transmisión del mensaje con menor gasto lingüístico**

  Por medio de los algoritmos referentes a la entropía de Shannon, aplicados a los genomas, es posible hacer un cálculo de la manera más óptima de transmitir un mensaje [25] . Esto también puede compararse contra un genoma, pues, una vez más, este texto biológico resulta ser inmejorable en cuanto a maximización de la información utilizando menos *palabras*.

- **Identificación de autores**

  La importancia de los métodos de la lingüística estadística convierte a esta disciplina en una herramienta útil para el análisis de secuencias de ADN, por ejemplo: el exponente de Zipf α de los lenguajes ha sido utilizado para identificar autores [12, 76, 82], debido a que, un escritor, particularmente en una etapa de su vida como tal tiene un conjunto definido de palabras que conoce y utiliza en su obra y tiene, además preferencia por algunas de éstas para expresar sus ideas. Tal hecho induce una distribución de frecuencias de palabras características de este autor en ese período. De igual manera este índice puede ser utilizado para relacionar regiones del segmento genómico con sus funciones biológicas.

1.2. Situación del trabajo en el marco de la Biología contemporánea

1.2.1. Ramas de la Biología involucradas

- **Genética**: Esta ciencia es la base sobre la cual se sustenta este trabajo. A través del aislamiento y caracterización del material genético de los organismos vivos, podemos conocer mucho más a fondo el funcionamiento de la célula. Las contribuciones fundamentales en la genética han permitido desarrollar la capacidad para manipular el ADN de las células y recientemente determinar la secuencia de genomas.

- **Genómica**: Incursionamos en esta nueva ciencia dado que las secuencias genómicas de los organismos empleados tienen un papel esencial en el trabajo. En un futuro próximo conoceremos cómo están organizados y localizados los genes en los cromosomas, determinaremos los mapas genéticos de los seres vivos y saber cómo se regulan y en qué tipo de procesos celulares participan los genes y sus productos proteicos.

  Hoy en día se ha logrado ya determinar la secuencia nucleotídica de todo el genoma de varias bacterias y también de varios organismos eucariontes y con esto nace la ciencia genómica, que nos permite el análisis comparativo de las secuencias de todos los genomas.

  Eventualmente seremos capaces de entender en detalle la complejidad de toda interacción genética y proteica que subyace al más simple de los procesos de funcionamiento y desarrollo. Así pues, obtener el mapa genético así como la secuencia del genoma del organismo vivo es elemento primario para la mejor comprensión de la vida.

- **Biofísica**: La Bioenergética (considerada como un apéndice de la Biofísica) del organismo se ocupa del estudio de la adecuada conversión de nutrientes dentro de los procesos metabólicos. La bioenergética se ve optimizada por la estructura altamente ordenada del código genético. A su vez, el ADN se replicará de manera más eficiente si el manejo de energía biológica es óptimo; por otro lado, la eficiencia máxima en la información transmitida es resultado directo dichos procesos orgánicos.

  Por otro lado, sin la mecánica cuántica no existiría hoy la biología molecular, parte importante de este proyecto. Comprender bien la física importa en biología, pues constituye un prerrequisito esencial para la interpretación de la estructura de los átomos y de su enlace en la formación de moléculas. La biología depende de la química y ésta a su vez de la física cuántica: la vida depende del funcionamiento de un tipo concreto de molécula, la doble hélice de ADN y la moderna descripción de la acción de los procesos vitales de dicha estructura. No cabe duda de que la vida depende del comportamiento ordinario de átomos y moléculas de acuerdo con las mismas leyes que gobiernan la

materia inerte, leyes que se asientan sobre la física. La unión entre los dos filamentos de la hélice se mantiene gracias a los efectos cuánticos que generan la atracción que es el puente de hidrógeno. Todos los procesos vitales que se desarrollan en el interior celular pueden interpretarse como la interacción entre sustancias químicas complejas en obediencia de las leyes de la física cuántica.

La física cuántica ofrece una interpretación de la vida en el nivel molecular. Esta explica cómo se ensamblan y actúan las proteínas en la célula y cómo se enrollan y desenrollan las hélices de ADN para elaborar ARN mensajero o para replicarse a sí mismas para transmitir el mensaje hereditario, el código genético, de generación en generación.

- **Biomatemáticas**: Un principio de *orden* común en la biología contemporánea es que muchos procesos biológicos obedecen a relaciones de escalamiento, de tal modo que al presentarse cambios en las magnitudes implicadas, es frecuente que la forma de las relaciones entre estas magnitudes se preserve. A este comportamiento suele denominársele alometría [7, 8]. Podemos apreciar que la estructura lingüística del genoma es alométrica. Prácticamente todos los métodos que se utilizaron en el presente trabajo, tienen que ver con esta rama de la Biología (Caminante aleatorio de Lévy, análisis de Zipf, cálculo de la dimensión de Haussdorf por series de Mandelbrot, Entropía de Shannon), así como las herramientas empleadas para darles seguimiento.

- **Bioquímica y Biología Molecular**: Los objetos vivos están formados por moléculas sin vida; si pretendemos comprender los procesos químicos de la vida hemos de entender los de las moléculas que no la tienen, así como las leyes de la física que subyacen a toda interpretación de la química.

La mayoría de los compuestos químicos de los organismos vivos contienen carbono, de ahí que se denomine química orgánica al estudio de los compuestos de carbono. De forma característica, las moléculas dotadas de alguna importancia para la vida contienen igualmente átomos de hidrógeno, oxígeno, fósforo y nitrógeno. La bioquímica constituye, en gran medida, el estudio de las enzimas, y éstas son proteínas globulares que son, precisamente, el producto final de los genes.

Las proteínas durante los complejos procesos bioquímicos en los que participan no mantienen necesariamente la estructura estacionaria que se estudia con métodos como los aquí expuestos pero que están estrechamente relacionados, de hecho una de las hipótesis de trabajo relaciona la estabilidad estructural de las proteínas y su capacidad de adaptación a ambientes químicos diversos con *reglas gramaticales* [82].

Existe también una fuerte dependencia bioquímica en el tamaño de las repeticiones de oligómeros en los genomas de diferentes especies, de ahí la diferencia en el funcionamiento metabólico de diferentes seres.

Otra teoría en este trabajo es el estudio del proceso de reciclaje (llevado a cabo por el ADNnc) de productos de desechos celulares para producir materia prima para el organismo.

- **Taxonomía**: Comparando las distintas gráficas de Zipf y las de entropía de Shannon es posible distinguir entre especies cercanas.

Por otro lado, comparando la acumulación de cúmulos y RDFs de oligómeros en el ADN de los genomas y las gráficas que pueden obtenerse con ayuda de los métodos estadísticos aquí utilizados sería posible determinar a qué grupo de seres vivos pertenece un organismo.

1.2.2. Ramas de influencia

- **Lingüística**: Este trabajo tiene impacto en esta ciencia en cuanto a la comparación de cantidad y longitud de palabras, así como coherencia, complejidad, optimización en un texto lingüístico o bien en un lenguaje y su posterior comparación con otros. El código genético de cuatro letras es más restrictivo que el polipeptídico de 20 caracteres y ciertamente mucho menos que el Morse, el cual Schrödinger cita como el arquetipo de código simple [20] donde, por primera vez se presenta al público esta noción.

También podría resultar útil comparar gráficas de la distribución de frecuencias de aparición de palabras y tener un indicio más de cuáles idiomas son derivados de otros, cómo están conformados los dialectos y cuál es el origen etimológico de ciertas palabras.

- **Medicina**: El presente trabajo pretende defender que el ADNnc, conocido muchas veces como ADN *basura*, sirve a algún propósito (que sólo podrá ser descifrado por completo en el laboratorio con ayuda de procedimientos bioquímicos), para lo cual se demostrará que tiene una distribución ordenada dentro del cromosoma. Esto confirmaría lo que se ha demostrado a través de siglos de estudio en la rama de la medicina, que el organismo del ser humano está diseñado con un orden específico sin importar el grupo étnico correspondiente al individuo. Todos responden a una estructura anatomoclínica semejante por lo que los medicamentos y modelos experimentales aplicables a un grupo social se pueden reproducir en otro. El descubrimiento de la función del ADNnc, revolucionará el desarrollo de la medicina genómica. Por ejemplo en los casos de Síndrome de Down, o trisomía del par 21, Síndrome de Turner, Síndrome de Klinefelter, padecimientos que se hallan relacionados con la presencia excesiva de RDFs. Actualmente sin solución médica definitiva [96].

Igualmente, las técnicas de ingeniería genética han permitido desde 1973, el aislamiento de muchos genes humanos y su utilización para la construcción de organismos transgénicos para la producción de proteínas humanas recombinantes que hoy en día se utilizan en diferentes problemáticas clínicas y para el tratamiento y la prevención de enfermedades. Asimismo, con el avance de la ciencia genómica -y particularmente con el desciframiento de la secuencia del genoma humano- tenemos una visión más avanzada de la forma en que están organizados los seres humanos, y también de las diferencias, los polimorfismos genéticos, que existen en todos y cada uno de los genes humanos y que son responsables de nuestra individualidad genética y por ello también de nuestra predisposición genética a enfermedades.

Esta nueva información, aunada a las técnicas de ingeniería genética, permiten hacer un uso más sofisticado de los genes humanos, no sólo para producir proteínas específicas en organismos transgénicos, sino también para contender de manera más individualizada con aspectos fundamentales de la salud humana, entre los que señalaríamos el diagnóstico genético, la farmacogenómica y la terapia génica.

1.3. Herramientas utilizadas

1.3.1. Métodos estadísticos

- **Entropía Informacional de Shannon**

  Por medio del algoritmo de Shannon-Weaver para el estudio de mensajes codificados se intentará descubrir si el ADNnc posee o no información y en el caso positivo con qué proporción. Si se obtienen valores bajos, esto evidenciará un gran contenido informático, por el contrario, si son altos, una entropía creciente. Evidentemente, por medio de esta herramienta se refleja también el comportamiento aleatorio o dirigido del ejemplo experimental que se trate.

- **Procesos de Markov**

  Podemos hablar de la existencia de una cadena de Markov cuando tenemos una serie de datos en la cual, cada uno de ellos depende exclusivamente del anterior (estadísticamente) y de la probabilidad de transición entre éste y aquél. En tal caso, tenemos un sistema que posee *memoria a corto plazo*, es decir, los efectos estadísticos de los demás valores pierden peso en la determinación de la probabilidad de transición entre un dato y el siguiente.

  Cuando el valor del dato que nos interesa depende (en sentido estadístico) de datos muy alejados a él, se trata de un sistema no-Markoviano o de *memoria de largo alcance*. Esta memoria responde a una función del tipo Ley de Pontencia, la Markoviana

en cambio, produce un decaimiento exponencial (a lo largo de esta tesis podremos observar ambos tipos de comportamiento en las gráficas que se encuentran en el Cap. 5). Es a partir de estas funciones que podemos percibir cuando los elementos de un sistema están muy correlacionados; entre más lo estén, se puede hablar de dependencia estadística y viceversa. Por otro lado, es conocido que los sistemas estadísticamente dependientes son los que albergan la cantidad máxima de información, dado que se trata unidades estadísticamente estables, no así las distribuciones asociadas a un decaimiento exponencial que presentan una tendencia totalmente opuesta.

1.3.2. Métodos lingüísticos

- Distribución de Probabilidad (Análisis Lingüístico de Zipf)

  Los primeros trabajos sobre lenguajes naturales estudiados mediante distribuciones estadísticas de palabras fueron realizados por Georges Zipf, quien propuso un marco matemático para la dinámica poblacional de la evolución del lenguaje con particular énfasis en de qué manera las palabras son propagadas sobre generaciones [90] (en sus investigaciones, Zipf consideró lenguajes anglosajones -Inglés y Alemán-; en México, a principios de este siglo Torres lo hizo con el español y 3 lenguajes mesoamericanos [82]). Este es un método estadístico que puede ser comparado con un análisis algebraico (este último basado en los generadores gramaticales de Noam Chomsky [9]) sustentado en el conteo de las palabras.

  En el caso de este trabajo, se requería un texto lingüístico largo (con un número de *palabras* comparable al de los genomas estudiados) para realizar una comparación contra los cuatro *textos genéticos* [1]; esto con la finalidad de mostrar la coherencia existente en un texto ordenado de forma no aleatoria, de esta manera sería posible apreciar las diferencias -si es que existían- en los números obtenidos, así como en las gráficas. Si se notara una divergencia notable en los anteriores se tendría indicio de aleatoriedad; del mismo modo, si lograra comprobarse que existen exponentes de Zipf y pendientes similares en las gráficas resultantes de ambos grupos, podría concluirse que existe en la coherencia en la distribución de bases del ADN (mayormente constituido por ADNnc).

1.3.3. Métodos basados en propiedades de autosimilitud y fractales

- **Cálculo de la Dimensión Fractal de Haussdorf**

  Entre las cosas que se puede analizar en un texto (genético, lingüístico o artificial) está la complejidad que éste pueda tener. Esto es un indicativo de qué tan bien estructurado está, si lo es de manera sobresaliente será entonces un valor fraccionario y elevado de la dimensión de Hausdorff ; además, si una ley de Potencia tiene asociada aparte una

[1] además del *genoma* artificial del generador aleatorio utilizado a manera de control

dimensión fractal, la información que alberga es aún mayor [16]. Utilizando esta herramienta matemática, que básicamente cuantifica la densidad de puntos (nucleótidos) en radios progresivamente mayores dentro del espacio de fases, será posible reconocer esto tanto en valores numéricos como en gráficas.

- **Series de Tiempo**

  Para saber cuánta información se tiene en un genoma relativa a un nucleótido, se caracteriza su dinámica por medio de una serie de tiempo previamente renormalizada (para evitar las redundancias). Esta tira de números representa los niveles de concentración y distribución de la base. Es posible complementar esta información mediante la generación de gráficas y fractales que representen este mismo comportamiento de una forma diferente.

- **Fractales**

  En la última década las técnicas de análisis de escalamiento se han desarrollado con el fin de estudiar patrones estadísticamente invariantes y relacionarlos con las propiedades físicas de los fluídos complejos y otros sistemas. Puesto que las secuencias son largas cadenas poliméricas, algunas propiedades generales de invarianza de escala encontradas en la física de polímeros (alometría) [88] pueden aparecer en el ADN y las alteraciones de estas propiedades generales pueden servir para la caracterización de secuencias genómicas. Si le invarianza de escala implíca dimensiones de Hausdorff fraccionarias se habla de fractalidad o, mejor dicho en términos físicos de cuasifractalidad o multifractalidad [2]

  En los genomas de organismos no pueden existir fractales en el sentido matemático riguroso, sin embargo, en el esquema de descripción dado es posible generar fractales en el límite no biológico [3]. Estos cuasifractales pueden verse sugeridos en las gráficas de las distribuciones de frecuencias en los genomas de los organismos reales (así como en Generador Aleatorio) que se manejan en este trabajo.

  Por otro lado, es conocido que los fractales son objetos matemáticos que albergan gran cantidad de información, por lo cual se consideró apropiado construir algunos para los genomas manejados y hacer comparaciones posteriores.

1.3.4. Otros métodos (biología teórica)

- **Proporciones de Bases Complementarias**

[2] Técnicamente un multifractal es un conjunto fractal que tiene asociada una propiedad de campo en cada punto.

[3] este límite implicaría $k_s \rightarrow \infty$ donde k_s es el *factor de escala* que es una medida del detalle al que estudiamos un conjunto geométrico.

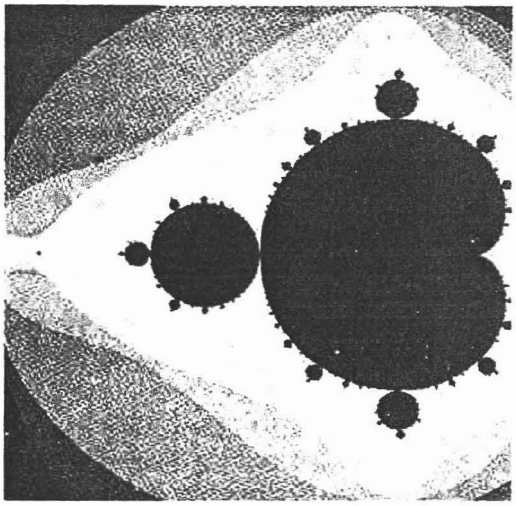


Figura 1.1: *Conjunto fractal de Mandelbrot*

Por medio de una ligera variación de las tasas descubiertas por Chargaff [5, 70] es posible observar si existe algún tipo de tendencia en la proporción de nucleótidos en el genoma o algún intervalo constante para sus tasas. A partir de estos datos puede inferirse la actividad bioquímica de las bases en la cadena de ADN y las probables implicaciones de esto en relación con su carácter no aleatorio.

Capítulo 2

Genética y Genómica

2.1. Conceptos básicos

Con la finalidad de tener un punto de partida apropiado para la consideración de los fenómenos lingüísticos, estadísticos y de orden relacionados con el genoma, punto medular de esta tesis, se procederá en este capítulo a establecer las nociones fundamentales que tienen que ver con la estructura, funcionamiento y desarrollo del código genético. Para lograr tal objetivo, se empezará enumerando los conceptos básicos de la Genética con un breve esbozo histórico que nos llevará hasta los recientes avances de la Ingeniería Molecular y la ciencia Genómica que han sentado las bases conceptuales para el desarrollo de este trabajo se comienza por definir conceptos como el de gene, cromosoma, ácidos nucleicos, secuencias genómicas y codificación genética [6]. Esto nos lleva a examinar de manera más detallada los mecanismos a través de los cuales la naturaleza almacena y utiliza la información funcional y estructural de los seres vivos en los genes. Una parte de esta información corresponde a mecanismos bioquímicos bien establecidos, como es el caso de la síntesis de proteínas en los segmentos de ADN conocidos como exones y más particularmente, en las triadas de nucleótidos conocidas como codones. Sin embargo, como ya se ha comentado, los mecanismos de codificación de la información en una basta mayoría del genoma de los seres vivos (intrones), son casi desconocidos; el primer paso para conocerlos sería averiguar la estructura subyacente al almacenamiento bioquímico de información y en último de los casos, el contenido informático presente en tales segmentos intrónicos, tema fundamental de esta tesis.

Todas las células poseen un material genético que alberga la información necesaria para regular todas sus actividades y que es transmitido de generación en generación.

La célula posee tres funciones biológicas fundamentales:

1. la replicación de su material genético y su transferencia a las siguientes generaciones
2. la expresión de los genes en los cromosomas.
3. la síntesis de proteínas a partir de la información genética que reside en el ADN

Los cromosomas son estructuras independientes y bien definidas, portan información hereditaria en los seres vivos que se emplea para construir y manejar el fenotipo del organismo en el que se encuentran las células que los albergan [23, 3]. La dotación entera del cromosoma actúa a modo de guía de supervivencia: describe lo que debe hacer cada parte del organismo y cómo deben reaccionar ante diversas circunstancias y estímulos procedentes del exterior. La propia imagen de unos genes ensartados a modo de las cuentas de un collar informa ya que ese mensaje vital codificado se ha escrito de forma lineal, como los renglones de las letras y palabras que constituyen el mensaje en un texto escrito. Las células siempre tienen un número par de cromosomas, aun cuando el número varía considerablemente entre organismos (46 en seres humanos, 4 en mosca, etc.).

Constituyendo a los cromosomas se encuentra el ADN, que es un polímero de tamaño variable según la especie a estudiar. Es el material que lleva la información genética en forma codificada de una célula a otra y de los progenitores a la descendencia. Toda la información necesaria para producir un nuevo organismo está contenida en su secuencia lineal y la fiel replicación de esta información se halla asegurada por la estructura de doble cadena que posee. En la bacteria el ADN es de doble hélice y circular. En las células eucariontes existe en las mitocondrias y cloroplastos.

La información celular reside en el ADN, que constituye el genoma (excepto en el caso de algunos virus en los que está en el ARN). Este ácido nucleico está constituido por cuatro nucleótidos, moléculas conformadas de la siguiente manera: un grupo fosfato, un azúcar desoxirribosa y una base nitrogenada (puede ser de dos tipos: purinas (adenina y guanina) y pirimidinas (timina, citosina y uracilo)). Estas macromoléculas están dispuestas en diversas combinaciones, tal como podemos imaginar cuentas de vidrio en cuatro colores diferentes formando un collar.

Erwin Chargaff descubrió en 1950 que las proporciones molares de las bases eran diferentes de un organismo a otro (relación de asimetría) y que la cantidad total de purinas que contenga una muestra de ADN (G+A) siempre es igual a la de pirimidinas (C+T); es más, hay tanta A como T, y tanta G como C. Son estas las denominadas proporciones de Chargaff. La suma de bases púricas entre pirimídicas es igual a 1 (relación de simetría).

En el siguiente nivel de complejidad, están las macromoléculas biológicas constituidas por estos nucleótidos. Las principales biomoléculas complejas encontradas en las células y tejidos de los animales superiores son el ADN, ARN, proteínas, polisacáridos y lípidos. De éstas, la familia de mayor trascendencia en lo que atañe a la estructura y funcionamiento de los seres vivos es la de las proteínas; su denominación viene a significar *principal* (entre las moléculas orgánicas).En el cuerpo humano se encuentran más de 50,000 distintas, todas las cuales participan en el correcto funcionamiento del organismo entero. Las proteínas se presentan en dos variedades principales: fibrosas, cuyas moléculas conservan en gran medida la estructura larga y delgada que se asocia automáticamente con una cadena, y globulares, cuyas

cadenas se enroscan en una bola.

La larga cadena polipeptídica que, por medio de abreviaturas químicas, podemos representar sobre el papel, constituye lo que se conoce por estructura primaria de la proteína, la secuencia que guardan los aminoácidos a lo largo de la cadena. En principio, el número de aminoácidos distintos que podrían existir es enorme, y muchos se han sintetizado artificialmente, pero, sólo 20 se encuentran en las proteínas y todos los seres vivos los emplean. Estos sillares no presentan propiedades biológicas intrínsecas, no se trata de moléculas vivas. Sin embargo, combinados en proteínas constituyen la materia prima de la vida.

Alrededor del 5 % del ADN está codificado, es decir, es traducido en proteínas mediante combinaciones variadas de tres nucleótidos (codones). A pesar de que el resto del ADN es no codificante en el sentido anterior, algunas regiones, se sabe, están involucradas en varios procesos regulatorios.

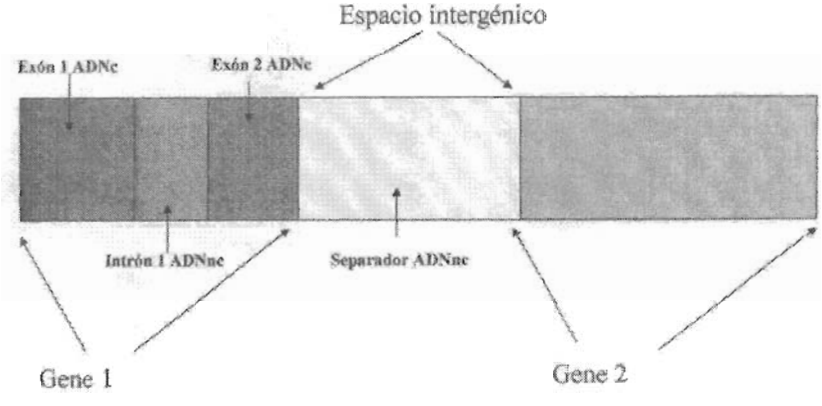


Figura 2.1: *Representación esquemática de las regiones de un segmento genómico*

El papel de material genético contenido en el ADN del genoma, consiste en señalar a la célula cómo y cuándo fabricar los diversos tipos de proteína. Sin lugar a dudas, tal información debe ser tan detallada y compleja como la codificada por los 20 caracteres del *alfabeto* aminoacídico. En las proteínas, el mensaje lo constituyen la secuencia de nucleótidos mientras que las palabras son éstos y las letras del código son los aminoácidos. Una larga ristra de *palabras* incorpora importantes instrucciones biológicas. Cualquier alteración estocástica del mensaje genético casi con seguridad le privará de sentido, en todo o en parte. Sin embargo, a veces esos cambios producen frases nuevas dotadas también de sentido, aunque distinto que el del mensaje original. Esto se traduce en un fenotipo que desempeña alguna labor con más

eficacia que sus rivales.

Por otro lado, la información genética contenida en el ADN es susceptible de sufrir variaciones, las cuales se denominan mutaciones. Las mutaciones se deben a un cambio en la secuencia de bases del ADN. Estas pueden deberse a errores de replicación, movimientos de reparación del ADN, etc. y ocurrir con una frecuencia baja para cada de las muchas divisiones celulares. Son varios los tipos de cambios que puede sufrir el ADN, y éstos pueden alterar desde un solo nucleótido, hasta cromosomas completos. Alteraciones en la secuencia de los nucleótidos de un gene, pueden ocasionar un cambio en la fase de lectura del gene, a nivel del ARNm, y por ello generarse una proteína alterada, la cual ya no será capaz de realizar su función original [6].

Algunos factores como virus, productos químicos, luz ultravioleta y radiación ionizante, incrementan la velocidad de las mutaciones. Estas a menudo afectan a las células somáticas y por tanto, en ocasiones son transmitidas a generaciones sucesivas de células dentro de un organismo. Sin embargo, puede haber cambios en la secuencia del ADN que no afecten al funcionamiento de la proteína codificada por el gene particular.

2.2. Fundamentos de Genética

Al estudio de los rasgos heredados biológicamente se le llama Genética. El término gene, fue introducido en 1909 por el botánico danés Wilhelm Johannsen para referirse a cada uno de los factores de información hereditaria que es transmitida por los progenitores a la descendencia en la reproducción.

La primera etapa de la genética se inicia con los experimentos del monje austriaco Gregorio Mendel quien trabajando con plantas de chícharo demostró en 1860 que las variaciones fenotípicas (las diferencias morfológicas entre los miembros de familias y especies) son el resultado de diferencias en elementos genéticos discretos (genes) y que se transmiten de padres a hijos conforme a ciertas reglas. Hoy sabemos que estos caracteres hereditarios se encuentran localizados en los cromosomas de todas las células de un organismo.

La teoría del monje era abstracta, se basaba en el razonamiento matemático. Dio nombre a *factores* invisibles, indetectables, que habrían de controlar la herencia. El momento y lugar adecuados para esa teoría se encontrarían después de que la microscopía alcanzara el desarrollo suficiente para estudiar el funcionamiento interno de la célula y para revelar la existencia de los componentes celulares denominados cromosomas [23].

Descubrió Mendel así, cinco puntos que están relacionados con la estructura íntima de la vida:

1. Todo carácter físico de un organismo se corresponde con un factor hereditario.

2. Los factores se presentan a pares.

3. Cada progenitor transmite un factor, y sólo uno, de cada par a todos y cada uno de los descendientes.

4. Los dos factores de cada par tienen igual probabilidad (en sentido estadístico escrito) de transmitirse de esa forma a cualquier descendiente.

5. Algunos factores son dominantes, mientras que otros son recesivos.

A principios del siglo XX, los trabajos de Morgan con la mosca *Drosophila melanogaster* fueron cruciales para el establecimiento de la genética como una disciplina biológica. Con este organismo fue posible establecer en unos pocos años el primer arreglo de algunos genes en los cromosomas, utilizando mutantes de este organismo y explicando el fenómeno fundamental del *entrecruzamiento genético* que implica el intercambio de información genética, es decir la recombinación molecular del material genético a nivel de los cromosomas, que se localizan en el núcleo de las células. Durante esta etapa de la genética, se establecieron los primeros mapas genéticos de algunos de los primeros organismos que hoy todavía se utilizan como modelos biológicos.

La siguiente etapa de la genética se centró en la identificación del tipo de sustancia química donde residía la información genética y en la identificación de la estructura a nivel molecular de este material. Esta fase se inicia en 1928 con los experimentos de Griffith quien logró transformar células no virulentas de la bacteria Pneumococo pneumoniae en células virulentas capaces de causar enfermedad mediante el tratamiento de éstas con material rico en ácidos nucleicos proveniente de células virulentas. Quince años después, en 1944 los investigadores ingleses Avery, McLeod y MacCarty siguiendo sus pasos demostraron que la capacidad para transformar bacterias inocuas en patógenas reside en un solo tipo de molécula el ácido desoxirribonucleico o ADN.

Posteriormente en 1950, los ingleses Franklin y Wilkins descubrieron características de simetría en la estructura de esta molécula y un año después también en Inglaterra, Dekker estableció la estructura covalente del esqueleto de la macromolécula del ADN que claramente definió la unión regular 3'-5' fosfodiéster [1] . Igualmente importante fue el trabajo de Chargaff, el cual permitió determinar las cantidades relativas de adenina, timina, guanina y citosina y demostrar que en cualquier ADN la cantidad molar de adenina es siempre igual a la de timina y también que la cantidad de citosina es la misma que de guanina [6, 3].

Considerando toda esta información, James Watson y Francis Crick realizaron en 1953, una de las contribuciones más fundamentales de la biología moderna: el desciframiento de la estructura molecular del ADN. Dicho descubrimiento ha venido a ser uno de los elementos unificadores en la biología moderna, ya que no solo es la misma en todos los seres vivos, sino

[1] Es decir los segmentos iniciales y terminales de un cadena genómica.

que además, la organización y regulación de los genes (fragmentos o segmentos específicos de esta hélice doble) también tiene en lo general, carácter universal en todos los organismos vivos. Esta característica es lo que posteriormente en 1973 permitió el nacimiento de la ingeniería genética, metodología mediante la cual es posible la edición a nivel molecular de este material.

La estructura general del ADN es exactamente la misma en todos los seres vivos, desde las bacterias hasta los seres humanos, es decir, el ADN es una doble hélice formada por dos polímeros antiparalelos y complementarios. Cada una de estas dos hélices o polímeros está, a su vez, integrada por monómeros que son como las cuentas de un collar (polímero). La diferencia fundamental entre todas las moléculas de ADN que forman los diferentes cromosomas de los seres vivos, es la secuencia de los millones de estos cuatro tipos de nucleótidos con sus bases, A, T, G, C en cada molécula de ADN, de la misma manera que sólo existen 28 letras en el alfabeto para formar todas las palabras y es la secuencia diferente de estas letras en las palabras lo que da un significado distinto para cada una de ellas. Además, en todo tipo de ADN, a un nucleótido con la base adenina le corresponde siempre, en el nucleótido de la hebra o hélice complementaria, uno con la base timina y a todo nucleótido con la base guanina corresponde un nucleótido con la base citosina en la hebra complementaria. Estas son reglas universales para todos los ADNs de todos los seres vivos.

Destaca en el modelo de Watson y Crick su sencillo planteamiento de la replicación. El ADN podía elaborar copias perfectas de sí mismo. Al desenrollarse las dos hebras de la hélice bicatenaria, se les añaden las bases pertinentes y van extendiéndose los esqueletos de azúcar y fosfato. Las dos nuevas hélices son idénticas entre sí y a la original, pues ambas poseen las mismas bases y en igual secuencia. Ambas llevan, por tanto, el mismo mensaje, redactado en código genético.

El mecanismo de replicación de los cromosomas cae a veces en los mismos errores que un mecanógrafo cansado: copia las mismas *palabras*, o repite u omite frases enteras. En una especie pluricelular como la nuestra, ese tipo de alteraciones de las células somáticas no suele tener repercusiones (para la progenie, aunque sí para el organismo). En cambio si se producen en la fase de copia que en la meiosis sigue a la recombinación, existe cierta probabilidad de que el cromosoma alterado se transmita a la descendencia.

Una copia de más de un gene, o de parte de él, no satisfará ninguna necesidad, pero aun así se replicará fielmente y se transmitirá de una generación a otra, cual si se tratara de un añadido al equipaje celular. Esas *piezas de repuesto* de material genético pueden también mutar, constituir sucesivamente varias o muchas versiones inútiles hasta que, por azar, aparezca un modelo que aporte algún beneficio a las células que lo contienen. Pueden crearse así genes enteramente nuevos. La copia, por tanto, puede introducir u omitir fragmentos de material en los cromosomas, puede alterar el mensaje codificado que determinen éstos. La regla empírica es que cualquier modificación de un mensaje lineal que podamos imaginar en una copia

probablemente se dé durante la transcripción y traducción del código genético.

De hecho, según se comprobó, es el ARN quien controla directamente la elaboración de proteína tras haberse sintetizado a partir de un molde de ADN. Ya desde un principio se concentraron los investigadores en una clave compuesta de tripletes, en la cual a cada carácter del alfabeto polipeptídico, esto es, a cada aminoácido, le correspondía una secuencia de tres bases. La razón de ello es bastante simple. Si se considera una sola base por vez tan sólo pueden *codificarse* cuatro objetos. Si se toman dos bases por vez (en un código compuesto por dobletes), se obtienen 16 parejas distintas de las cuatro bases (4X4), puesto que cada vez que una de ellas ocupa la primera posición, cualquiera de las cuatro puede aparecer en segundo lugar. Sin embargo, 16 combinaciones no cubren todavía los 20 aminoácidos que constituyen los sillares de todo organismo vivo. Si se consideran tres bases por vez, en tripletes, se alcanzan 64 combinaciones distintas (4X4X4), más que suficientes para codificar los 20 aminoácidos, más algunas combinaciones de *puntuación*, códigos que señalan el principio y el final de la cadena, indicaciones de *arranque* y de paro para las enzimas que ensamblan las cadenas.

En las células de los eucariontes hay varios sistemas genéticos en cargados de transcribir la información del ADN en copias del ARNm; es decir hay varios sistemas de ARN polimerasa. Además, el ADN se encuentra no sólo en el núcleo sino también en mitocondrias y cloroplastos.

Las proteínas son las moléculas biológicas que realizan la mayor parte de las funciones celulares y la función de cada proteína está determinada por su estructura. El ADN es una molécula que es el resultado de polimerizar (unir en forma de cadena), varios millones de los cuatro diferentes nucleótidos; las proteínas son también polímeros biológicos que están constituidos por decenas o centenas de veinte diferentes tipos de monómeros llamados aminoácidos. Por esta razón no puede haber correspondencia de un aminoácido por cada uno de los nucleótidos que integran los genes y la consecuencia es que cada uno de los aminoácidos de una proteína debe estar *codificado* por un grupo de nucleótidos. Al descifrar el código genético se comprobó que existen grupos conformados por tres nucleótidos, que se denominan tripletes o codones, son éstos los responsables de la información codificante del ADN. Se determinó también que en varios casos, más de un triplete codifica para un mismo aminoácido y que algunos tripletes codifican para señales de terminación o iniciación para la síntesis proteica.

La síntesis de proteínas, que de hecho es la traducción del mensaje del ARN, se lleva al cabo también en dirección 5' a 3', mediante la polimerización de aminoácidos en proteínas, a nivel de los ribosomas.

Los estudios más recientes sobre el material genético tratan de dilucidar precisamente la secuencia de nucleótidos del ADN (mapeo de secuencias, sondas de detección, técnicas de

clonación, etc.) ya que ello permitirá a su vez aumentar los conocimientos actuales sobre los genes.

Las partes no codificadas consisten de intrones y extremos. Se piensa que estos segmentos no-codificados tienen una importante labor en los procesos regulatorios y que además tienen cierta actividad promocional. Sin embargo, las palabras biológicamente significativas en estas regiones no han sido reconocidas. Un gene esta constituido comúnmente por un cierto número de exones separados por intrones.

Existen 64 combinaciones posibles de las 4 bases A C G T cuando éstas forman tripletes o codones. Tres de estas combinaciones, TAA, TAG y TGA son los llamados codones de terminación, de tal forma que el vocabulario significativo está formado por 61 tripletes. Observaciones sobre secuencias codificadas de DNA han revelado que muchos codones pueden tener la misma frecuencia, ha este problema se le ha llamado *degeneración en la frecuencia.*

Mucha de la investigación, actualmente se encuentra centrada en distinguir entre estas características estadísticamente universales y ciertas características privativas que identifican a cada gene.

2.3. Secuencias genéticas, Código y Genómica

Un libro, como el mensaje genético de la doble hélice, es lineal. Pero resulta imposible plasmar la historia de la doble hélice siguiendo un relato lineal. Como suele ocurrir en la ciencia, en la historia participa mucha gente, en diversos lugares, pero en intervalos temporales que se solapan; sólo puede observarse la tela entera tejiendo cada hilo de la historia por separado y retirándose a observar el tapiz acabado. A veces deben desandarse un par de pasos para recoger un hilo distinto y descubrir su relación con los que ya se han tejido. Igual ocurre con el relato de la resolución final del código genético.

Hoy podemos decir por lo que conocemos sobre el ADN que el descubrimiento de su estructura química ha venido a ser uno de los elementos unificadores en la biología moderna ya que no sólo la estructura del ADN es la misma en todos los seres vivos, sino que además, la organización y regulación de los genes (que son fragmentos o segmentos específicos de esta hélice doble) también tiene, en lo general, carácter universal en todos los organismos vivos. Esta característica es lo que posteriormente en 1973 permitió el nacimiento de la ingeniería genética, metodología mediante la cual es posible la edición, a nivel molecular, de este material. Podríamos decir que el material genético de todos los seres vivos tiene el mismo formato, y que por ello se pueden editar molecularmente en un tubo de ensayo ADNs de diferentes orígenes.

Una secuencia genética es un conjunto de símbolos puestos en orden (nucleótidos). Esa secuencia de símbolos contiene un mensaje. Las reglas gramaticales de ese mensaje están

dadas por las leyes de la genética y el estudio de la información que contiene ese mensaje es la Genómica, (Teoría de la información) hay también cierto vocabulario (diccionario) y ciertas palabras que tienen significado para ese ADN.

Hoy en día, gracias al proyecto del Genoma Humano, y a otros proyectos que han permitido determinar la secuencia de todos los genes en varios genomas de diferentes organismos, y se tiene ya una gran colección de genes de diversos organismos, incluyendo a los de la raza humana, que hay permitido empezar a tener una idea más clara de la manera en que se regula la expresión de los genes en plantas, animales e incluso la especie humana.

A finales del siglo XIX los biólogos conocían la existencia de los cromosomas y sospechaban el papel que podía corresponderles en la herencia. Varias teorías pretendían explicar su mecanismo de acción.

Para 1950, a científicos como Alexander Dounce les era evidente ya que de algún modo, el ARN se copiaba del ADN del núcleo y se tenía la idea de que el orden de los aminoácidos de las proteínas depende del que sigan las bases a lo largo de las cadenas de ácido nucleico.

La prueba de que la anemia falciforme era lo que Pauling denominó una enfermedad molecular se publicó en 1949, el mismo año en que apareció un artículo de James Neel, de la Universidad de Michigan, donde se establecía de una vez por todas que la enfermedad la provoca un gene mutante recesivo que se transmite de una generación a la siguiente en exacto acuerdo con las leyes mendelianas de la herencia. Juntos, los dos trabajos establecieron que la hemoglobina de los pacientes de anemia falciforme presenta un cambio químico específico, producido por la alteración de un solo gene del juego cromosómico humano.

El ritmo de las investigaciones se aceleró en los años 70 gracias al desarrollo de las técnicas de recombinación de ADN, las técnicas de la ingeniería genética. A finales de esa década podía ya analizarse de forma rutinaria la secuencia de bases de segmentos de ADN de gran longitud e identificarse con exactitud su mensaje genético; se sintetizaban fragmentos de ADN artificial de varias docenas de pares de bases (en 1982 se fabricó el gene del interferón, que requirió el empalme de 514 pares de bases, ordenados en la secuencia adecuada); se había identificado, y se sabía utilizar las enzimas que cortan el ADN, así como las que reempalman los extremos libres tras la inserción en el hueco de un gene artificial. Las sorprendentes posibilidades de intervención humana en el genoma de los organismos vivos (hombre incluido) que procuran esos avances han sido objeto de enconados debates en la década de 1980. Por un lado prometen beneficios incalculables, por ejemplo la obtención de nuevos fármacos, como el interferón, de nuevas fuentes de insulina para los diabéticos, nuevas estirpes de vegetales y animales de interés alimentario y hasta la corrección de defectos genéticos congénitos como la anemia falciforme o la hemofilia.

Las estructuras de las cadenas no responden a ninguna sencilla regla química, como que

la glicina se halla siempre junto a la valina o a simples repeticiones, como, por ejemplo, seis leucinas seguidas de cuatro valinas y dos cisternas y repítase el bloque hasta el final de la cadena. En realidad, se describe mejor considerando que responde a un mensaje codificado. El descubrimiento de Sanger reveló una secuencia que carecía de reglas; sin embargo, contenía información y, para explicar la presencia de esa información en la proteína se necesita imprescindiblemente un código. En otros términos, toda molécula de proteína contiene un mensaje codificado que asegura que cualquier proteína globular posea una forma específica que la dota de manera sin igual para el desempeño de su papel de molécula biológica. Sanger proporcionó a los bioquímicos los medios para leer el mensaje, aunque no resolvió el código.

La especificidad de un segmento unitario de ácido nucleico sólo la expresa su secuencia de bases y cuando la información se ha transferido a la proteína no puede ya salir de ella. En ese sentido información es la determinación exacta de una secuencia, ya sea la de las bases del ácido nucleico como la de los aminoácidos de una proteína. Si bien el ADN puede controlar la elaboración de ARN y éste la de la proteína, el proceso no puede seguir el curso contrario. A ese concepto fundamental se denominó dogma central de la biología molecular. Ingram estableció finalmente que un gene determina verdaderamente una sola proteína.

Recuérdese que la adición o sustracción de una letra del mensaje mezclaba el contenido entero a partir de ese punto; desde la perspectiva del código genético, esa mutación inutilizaría el gene y provocaría la elaboración de una pieza de proteína sin sentido (o ni siquiera eso) en lugar de una enzima funcional. La adición o sustracción de dos caracteres no mejora el resultado. Pero supóngase que se añaden o se quitan tres letras en un breve segmento de mensaje. Por ejemplo, POR DOS MIL LES DIO PAN SIN FIN puede quedar en POR FDO SMI LGL ESH DIO PAN SIN FIN, o bien PRD OSI LLE SDO PAN SIN FIN. En ambos casos, aun cuando un breve fragmento del mensaje resulte un galimatías, el resto del mensaje sigue teniendo cierto sentido, pues el efecto global de las mutaciones es justo la extracción o la adición de una palabra de tres caracteres. Un gene que haya sufrido ese tipo de alteración bien pudiera seguir siendo funcional, al menos en parte, pues determinaría la producción de una proteína que sólo había sufrido variación en unos pocos aminoácidos.

Eso precisamente demostraron Crack y Brenner en 1961. Cada aminoácido viene lee a partir de un grupo de partida fijo y el código debe ser degenerado, puesto que existiendo 64 tripletes distintos y sólo 20 aminoácidos que codificar, algunos aminoácidos debían estar determinados por más de un triplete o codón.

Un físico, Erwin Schrödinger, publicó la primera interpretación convenientemente elaborada de código genético tal como hoy lo entendemos. Sus fundamentos bioquímicos eran erróneos (al redactar ¿Qué es la vida? Creía que la clave la portaban las proteínas), pero ha subsistido la noción de código análogo al de Morse. Y otro físico, George Gamow, llamó poderosamente la atención de los biólogos moleculares sobre esa idea a principios de la década de 1950, justo cuando acababan de publicar Watson y Crick sus dos primeros artículos

sobre la naturaleza del ADN.

El problema de la codificación es el enigma de cómo se traducía una retahíla de bases situadas a lo largo de la doble hélice en la ristra de aminoácidos de una proteína.

Como el propio Gamow lo expresaba en su artículo aparecido en Nature las propiedades hereditarias de cualquier organismo podrían caracterizarse por un largo número escrito en un sistema de cuatro dígitos. Por el contrario, las enzimas, cuya composición debe venir determinada completamente por la molécula de ácido desoxirribonucleico, son largas cadenas peptídicas constituidas por uno veinte tipos distintos de aminoácidos, y pueden considerarse palabras de gran longitud basadas en un alfabeto de 20 caracteres. La cuestión fue ¿cómo pueden traducirse números de cuatro dígitos en esas palabras?

En última instancia los teóricos se vieron forzados a retornar a la posibiliad más sencilla: la ristra de bases del ácido nucleico debe leerse de tres en tres, a partir de un punto concreto y acabando en otro. En el mensaje, a cada triplete le correspondería un aminoácido. Tal tipo de código presenta algunas características destacables, que de inmediato ilustran sobre el problema de las mutaciones, el mecanismo que aporta la variación y adaptación. Si se pierde una de las bases, el mensaje entero se embrolla irremediablemente a partir de ese punto, suponiendo que el mecanismo de lectura de la célula siga traduciendo el código por tripletes.

Un mensaje como HOY VAS CON MUY MAL PIE quedaría, sin más que quitarle una letra, en HYV ASC ONM UYM ALP IE. De igual modo, al añadirle una letra de más (una base más) se obtendría la frase tan carente de sentido como la anterior, HHO YVA SCO NMU YMA LPI E. Pero si se sustituye una letra, una base, por otra, sólo cambia el significado de una de las palabras del mensaje: uno de los aminoácidos de la cadena se atendrá a las nueva sin instrucciones. Puede así aparecer una palabra sin significado alguno en medio del mensaje, como HOY VAS CON MJY MAL PIE, o puede formarse una nueva palabra que sí tenga sentido: HOY VAN CON MUY MAL PIE. En este caso de ser correcta la teoría, la maquinaria celular se atendrá a la información y elaborará una proteína que diferirá en un aminoácido de la composición que le correspondería, quizá con valina donde debería ir alanina. Puede que la proteína desempeña igual de bien la función que tenga encomendada e, incluso, pudiera ser que la desempeñe mejor que la versión original, confiriendo al organismo que porta la versión mutante del mensaje genético correspondiente a esa proteína en particular una ligera ventaja en la lucha por la supervivencia. O puede que resulte fatal, que la nueva proteína no logre llevar a cabo la función que le corresponde. El descubrimiento de una de esas mutaciones causante de la anemia falciforme abrió nuevos caminos para la biología.

En 1977 se descubrió que la información del ADN se encontraba en segmentos no contiguos. Esto mostraba que los genes están interrumpidos por inserciones de ADN no codificantes, éstos son conocidos como intrones.

En cualquier célula, bastante más del 90 % de esa información no llega a emplearse nunca

para codificar proteínas. Permanece bloqueada en los filamentos de ADN, que se encuentran fuertemente enrollados en los cromosomas que encierra el núcleo celular. Tan sólo se transcribe a ARN, que a su vez se emplea en el control de la elaboración de proteínas, la minúscula cantidad de información genética que se refiere a la actuación de esa célula en particular; la propia de las células hepáticas, musculares, etc. Pese a ello, todas las células contienen un juego entero de cromosomas.

Se descubrió la existencia de largos fragmentos de ADN que no parecían contener ningún mensaje útil, pero cuya presencia en el genoma de muchas especies era tan abundante que el código de los genes útiles, funcionales, podía encontrarse escindido en porciones separadas por segmentos de ADN sin sentido, trozos de aparentes galimatías que debían extraerse del ARN transcrito a partir del ADN para que el ARN mensajero pudiera actuar en la síntesis de proteínas.

Otro descubrimiento espectacular efectuado por la biología molecular a finales de la década de 1970 es que los genes de los organismos superiores rara vez constan de una sola pieza sino que se encuentran diseminados a lo largo de un segmento de ADN, interrumpidos por fragmentos de ADN que a primera vista no parece portar mensaje alguno. Viene a ser algo parecido a como si este documento comenzara con unas pocas páginas redactadas en castellano, siguieran otras tantas, o más, sin sentido, luego otras pocas en castellano, que retomarían el hilo del relato exactamente donde se dejó, etc. Para leer y utilizar un mensaje genético de ese tipo, el mecanismo fundamental del que debe valerse la célula, que se remonta sin duda a la aparición de las primeras estructuras celulares, es la extracción de los intercalados que carecen de sentido y la reunión de las piezas del mensaje coherente para obtener una copia funcional del gene.

En 1977, Pierre Chambon estudiando el mecanismo de la síntesis de ovalbúmina, la proteína de la clara de huevo, trazó el mapa de la región cromosómica involucrada en ese proceso valiéndose de las, a la sazón, nuevas técnicas de recombinación genética. La longitud del ARN mensajero elaborado por la región cromosómica responsable de la síntesis de ovalbúmina parecía insuficiente para albergar toda la información contenida en el segmento de cromosoma. Casi simultáneamente, un equipo de virólogos del Cold Spring Harbor descubrió que uno de los genes víricos que analizaban presentaba dos porciones separadas por un fragmento que no parecía portar mensaje codificado.

No cabía otra respuesta: el ARNm no era en realidad una copia directa de la porción cromosómica entera, sino sólo de las porciones que portaban información genética de interés. Se había ignorado los intercalados sin sentido.

Según sabemos hoy, la síntesis de ARNm procede en dos etapas. En primer lugar se copia con entera fidelidad un segmento de ADN cromosómico en un largo filamento de ARN, incluidas las secuencias de ADN inútiles (secuencias intercaladas que se ha dado en denominar

intrones). Seguidamente, las enzimas celulares de corte y reempalme extraen justo los trozos inservibles de ARN y reúnen con precisión el resto, obteniéndose una molécula funcional de ARNm que la célula empleará de patrón en la síntesis de una proteína determinada. El mecanismo debe actuar con precisión absoluta, pues cualquier error inutilizaría el ARNm. De alguna forma, la célula reconoce en un nucleótido el punto donde debe cortar, identifica con exactitud comparable el otro extremo del intrón y escinde entonces el ARN inservible y empalma los exones. Recuérdese que la pérdida o adición de una sola base desbarataría por completo el mensaje genético, pues alteraría la fase de lectura según el código que traduce tres caracteres a una palabra. Esos errores prácticamente no se dan nunca. Sin embargo, en muchos casos los intrones abarcan la mayor parte de la longitud del ARN precursor.

Por ejemplo, en el ADN que porta el gene para la ovoalbúmina existen siete intrones; otro gene, el de la proteína conalbúmina, posee 17 secuencias intercaladas que carecen de sentido para codificar proteínas.

Se han avanzado varias explicaciones a la razón de ser de ese exceso de equipaje genético que acarreamos. Sostiene una escuela de pensamiento que la célula se vale del proceso de empalme para informar de los acontecimientos a otros genes. Otra, que los intrones desempeñan un papel regulador. Otra, que se trata de ADN primitivo y una más, que es un reservorio contra posibles mutaciones.

De esta manera, en 1973 Cohen y Boyer realizaron su experimento histórico en donde por primera vez se demostró que usando in vitro, es decir en un tubo de ensayo en el laboratorio, estas herramientas celulares, fue posible insertar el ADN de una rana en el ADN cromosomal de la bacteria Escherichia coli. Con este experimento se construye el primer organismo transgénico y se da inicio la era del manejo in vitro de la información genética, o la edición molecular del material genético mediante la metodología llamada ingeniería genética o de ADN recombinante.

Las secuencias genómicas contienen numerosas capas de información. Estas incluyen especificaciones para las secuencias de RNAm responsables de la estructura de las proteínas, la identificación de regiones codificadas y no codificadas de las secuencias, información necesaria para la identificación de secuencias reguladoras, promotoras, etc.; información que guiará las interacciones proteína-DNA e instrucciones para el empaquetamiento y despliegue del DNA. Mientras que algunos de los mecanismos de codificación de esta información son comprendidos, se sabe relativamente poco acerca de otras capas de información codificada y oculta en la molécula de DNA.

Las regiones que codifican para proteínas (también llamadas regiones codificadas o exones) tienen a las bases trabajando en grupos de tres para formar las proteínas. Estos tripletes son llamados codones.

2.4. Desarrollo histórico de la Genómica

A principios de la década de los años ochenta, en el siglo pasado, da inicio un esfuerzo muy importante encaminado al aislamiento y la caracterización de genes de organismos superiores incluyendo el humano, con el objetivo general de entender en detalle la organización y la expresión de los genes. Así, utilizando diferentes estrategias y metodologías, se aísla un primer conjunto de ADN complementarios a partir de copiar moléculas de ARN mensajeros específicos, los cuales fueron, a su vez, utilizados para aislar los genes correspondientes a partir directamente del ADN cromosomal. Sorpresivamente, muchos de los genes cromosomales de organismos superiores resultaron tener un mayor número de nucleótidos, es decir un mayor tamaño, que los ADN complementarios correspondientes utilizados para aislarlos.

La mayor parte de los genes en los organismos eucariontes, están constituidos por dos tipos de regiones de ADN. El primer tipo son regiones que codifican para la proteína, llamadas exones y el segundo tipo son regiones de ADN que no codifican para proteína, a las cuales se les denominó intrones. El ADN complementario, obtenido a partir de ARN mensajeros, sólo lleva, o está constituido por las regiones de ADN que son los exones, y por eso siempre es de menor tamaño que el gene a nivel de ADN cromosomal.

Con toda esta información y gracias al mejoramiento y sofisticación continuos de las técnicas de ADN recombinante, en particular la aparición de técnicas poderosas de amplificación de ADN, hoy es posible analizar inclusive sin clonar, genes de cualquier organismo, incluyendo al hombre. A través de ello, se ha iniciado la etapa o la era del genoma, en la cual el esfuerzo ya no se concentra solamente en genes aislados, sino en el análisis del conjunto de todos los genes presentes en un organismo. En el caso de las bacterias, organismos unicelulares, su genoma lo conforman sus 3000 a 5000 genes que se localizan en sus cromosomas, dependiendo de la especie, en el caso de los humanos tenemos cerca de 40000 genes.

En la actualidad pretendemos conocer inicialmente cómo están organizados y localizados todos los genes en los cromosomas; es decir, tenemos como objetivo global de todos los grupos que trabajamos en esta área, contribuir a determinar los mapas genéticos de los seres vivos. Asimismo, pretendemos conocer cómo se regulan y en qué tipo de procesos celulares participan los genes y sus productos proteicos.

En genética, un mapa es la posición que guarda un gene con respecto a otros en las cintas de ADN que forman los cromosomas de determinado organismo. La parte del genoma que no codifica para proteínas ocupa una fracción muy significativa de estos espacios. Este ADNnc puede ser de dos tipos:

1. **ADN espaciador**: Está formado por bases en una secuencia sencilla que está entre regiones codificantes del genoma;

2. **ADN repetitivo**: Formado por una secuencia que, al contrario que el espaciador, se

dispone por todo el genoma, debido a la existencia de múltiples copias. A este tipo de ADN pertenecen las RDFs.

En el caso de una cinta genética de ADN, sabemos también que se encuentra almacenada información para hacer varias y diferentes proteínas y que la información para cada una de estas proteínas o canciones biológicas, está almacenada en un segmento específico de la cinta para cada una de ellas, que se llama gene. Estos segmentos de la cinta genética o genes se encuentran organizados también de manera lineal, es decir uno después de otro y que al igual que los segmentos de la cinta musical que codifican cada uno de ellos para una canción diferente y diferente duración o tamaño, los genes codifican también cada uno de ellos para una proteína distinta y de diferente tamaño.

Hoy en día se ha logrado ya determinar la secuencia nucleotídica de todo el genoma de varias bacterias y también de varios organismos eucariontes y con esto nace la ciencia genómica, que nos permite el análisis comparativo de las secuencias de todos los genomas.

Únicamente a través del conocimiento de cuándo y qué proteína aparecerá en cierta etapa del desarrollo del organismo vivo, es que seremos capaces de entender en detalle la complejidad de toda interacción genética y proteica que subyace al más simple de los procesos de funcionamiento y desarrollo. Así pues, obtener el mapa genético así como la secuencia del genoma y del proteoma del organismo vivo es elemento primario para el entendimiento del fenómeno de la vida.

En los primeros meses del año 2000, se publicó la noticia de que se había ya determinado la secuencia de todos los fragmentos de ADN que conforman el genoma humano y que en un lapso de no más de dos años, se tendría una secuencia del genoma humano con más de un 95 % de certeza.

Además, también a principios del año 2000 se reportó la secuencia nucleotídica del cromosoma humano 22, en el cual se localizaron 545 genes. Asimismo, se logró también la determinación de la secuencia nucleotídica del cromosoma 21 del ser humano, el más pequeño de los 23 cromosomas de nuestra especie, y que tiene 33.5 millones de pares de bases y en el que se localizan 225 genes. Estos números de genes, relativamente reducidos, conforme a lo esperado, hizo pensar que el número final de genes humanos no llegaría a los 80 o 100 mil originalmente estimados, sino tal vez serían entre 35 y 40 mil los genes de nuestra especie humana. Finalmente, en febrero de 2001, dos grupos reportaron simultánea pero independientemente, en las revistas Nature y Science, la secuencia del genoma humano.

Dos seres humanos somos genéticamente 99.9 % idénticos. Lo anterior implica que, si bien entre dos individuos compartimos la mayor parte de nuestro material genético hay, sin embargo, aproximadamente en promedio, tres millones de nucleótidos diferentes entre dos miembros de la raza humana; encontrándose entonces en promedio entre dos individuos, un nucleótido diferente cada 1200. Estas diferencias son en lo general, el producto de muta-

ciones acumuladas en el genoma de la raza humana, a través de los años. Aquellas que en particular son el resultado de una mutación de un solo par de bases en un gene, reciben el nombre de polimorfismos genéticos de un solo nucleótido (SNP, por sus siglas en inglés).

Los polimorfismos genéticos son, en buena medida, la razón genética responsable de las diferencias físicas entre los seres humanos. Desde el punto de vista de la medicina, los SNP son marcadores genéticos importantes para entender no sólo las enfermedades, sino también las diferentes predisposiciones a las enfermedades que existen entre las razas y los individuos. Los polimorfismos genéticos son, en suma, los responsables de la identidad genética individual de cada ser humano. La presencia e identificación de los SNP en cada individuo, está dando lugar al desarrollo de una medicina molecular individualizada orientada al diagnósitco preventivo y al diseño de drogas específicas (farmaco-genómica) a nivel del individuo.

Los dos artículos que reportan la secuencia del genoma humano, señalan que no es posible fijar aún el número preciso de genes que conforman nuestro genoma ya que hay secuencias que pudieran ser realmente genes, pero hay todavía indefiniciones que tienen que precisarse, sobre todo por el problema de la presencia en el genoma humano de pseudogenes, material genético repetido que no funciona como un gene. Sin embargo, ambos grupos concuerdan que el genoma humano tendrá entre 30 y 40 mil genes.

Es relevante mencionar también que se han identificado ya más de mil genes involucrados con enfermedades humanas, y también se han determinado genes que codifican para proteínas que pudieran ser blanco del desarrollo de nuevas drogas para el tratamiento de problemas de salud.

Otra característica muy importante de nuestro genoma, es la presencia de lo que se conoce con el nombre de material genético repetido. Al menos el 50 % del total del material genético del genoma humano y probablemente más, son secuencias de bases que se repiten numerosas veces y de maneras diferentes. Este tipo de ADN en términos generales, se puede clasificar en cinco categorías, mencionadas por el Consorcio Internacional para la Secuencia del Genoma Humano: a)material repetido derivado de transposones; b) copias inactivas de genes llamadas pseudogenes; c) secuencias de tamaño corto, repetidas varias veces; d) duplicaciones de regiones del genoma de 10-300 kb, que han sido copiadas e incorporadas en otra región del genoma; y e) bloques de secuencias repetidas en fila, como los centrómeros y los telómeros de los cromosomas.

Las secuencias repetidas del tipo de los retrovirus, de las cuales existen en nuestro genoma del orden de 450 mil copias, provienen de infecciones de virus que tienen ARN como material genético, que insertaron copias de su genoma como ADN, en diferentes sitios del genoma humano, a lo largo del tiempo. Muchas de estas secuencias contienen todavía un gene activo que codifica para la transcriptasa reversa, que es una proteína con actividad de enzima que sintetiza ADN a partir de ARN, y es la enzima que usan el virus del ADN y otros

retrovirus para incorporar copias de su genoma en los cromosomas humanos.

El avance de la genética y de la ciencia genómica es vertiginoso. Lo anterior está contribuyendo al entendimiento profundo de los procesos finos involucrados en la organización, expresión y modificación de los genes, y a través de sus productos proteicos, del funcionamiento de la célula viva. Por otro lado, el comparar los procesos normales con procesos patológicos, ha facilitado substancialmente el entendimiento de enfermedades y problemáticas clínicas. Ciertamente, la determinación de la secuencia del genoma humano, nos permitirá abordar problemáticas clínicas más complejas y enfermedades multifactoriales, en donde participan varios genes, de manera mucho más certera e individualizada.

La revolución de la biología molecular -incluyendo la decodificación del genotipo humano, el proceso conocido para la comunidad científica internacional como genómica- está abriendo un amplio campo de investigación en medicina y farmacología.

La genómica es la piedra angular para el futuro éxito en el desarrollo de la farmacéutica innovadora.

En esta década, fue posible identificar solo entre 400 y 500 moléculas biológicas clave o simple blancos, que fueron adaptados para uso como instrumentos para medir y optimización de drogas sintéticas. Gracias a la genómica, esta figura puede presumiblemente ser multiplicada por un factor de diez dentro de un espacio de sólo unos pocos años.

Los científicos pueden seguir tres diferentes estrategias como medidas para utilizar la decodificación del genoma humano para sus propósitos:

1. La caza de blancos para genes causantes de enfermedades

   Este método ha causado el más fuerte eco en los medios, a pesar de que alcanza su éxito en el tratamiento de enfermedades raras debido a un defecto genético individual. Un ejemplo particularmente impresionante aquí es la enfermedad genética de Huntington, comúnmente conocida como el mal de San Vito. En tales desordenes monogenéticos, el uso de métodos analíticos de 'pedigree' genético hace posible rastrear variantes del gene defectuoso. Esto abre la oportunidad de reemplazar los productos genéticos defectuosos o faltantes en cuestión -proteínas- o más para compensar el defecto por medios farmacéuticos.

2. El análisis de la expresión génica en tejido saludable y enfermo.

3. La secuencia enteramente indirecta -en otras palabras no más orientada especialmente hacia condiciones 'saludables' o 'enfermas'- del genoma humano completo para el año 2005.

El sentido aquí es emplear métodos bioinformáticos para investigar la enorme cantidad de datos para secuencias potencialmente relevantes, por ejemplo por análisis de similitud,

por el cual es posible identificar miembros desconocidos de familias de genes. Una estrategia especial para la secuenciación del genoma humano entero está concentrada exclusivamente en aquellas secuencias genéticas que están transcritas de hecho -expresadas- y de las cuales segmentos individuales pueden ser caracterizados, los así llamados MSE (marcadores de secuencia expresada). Entre los métodos bioinformáticos están los métodos estadísticos y lingüísticos usados aquí.

No puede haber dudas respecto a que la genómica ha escrito una extremadamente plena fuente de información sobre proteínas biológicamente activas.

Se dice que la vida es un conjunto de datos investidos en una forma corpórea, codificando en billones de bloques de construcción que edifican las espiral del material hereditario ADN, y activadas a través de la interacción sutil de estructuras moleculares complejas inimaginables.

2.5. Alometría y escalamientos en Biología

La forma y el tamaño influyen todos los aspectos de la vida; definen funciones y estructuras a nivel orgánico y determinan la relación de los individuos o las poblaciones con el medio ambiente. En los organismos conforman las características anatómicas y definen las fisiológicas. A nivel colectivo, ser grande o pequeño, decide formas de interacción con los demás seres vivos y con el medio físico. La dinámica dentro y entre las especies depende esencialmente de la forma y el tamaño.

Los estudios morfométricos, básicos para comprender la mecánica del movimiento o la termodinámica de los procesos vitales, se hallan en el origen mismo de la biología: son el fundamento de la anatomía comparada, una disciplina capaz de establecer vínculos generales entre formas, dimensiones, mecanismos y funciones en los seres vivos, sin ella hubiera sido imposible el trabajo paleontológico de reconstrucción de distintos organismos ya extintos.

La vida es flexible y ubicua; de su capacidad de generar seres de tamaño tan variado depende su persistencia. Este hecho, que parece obedecer a una ley, es parte de la historia de la vida porque el tamaño influye determinantemente en la naturaleza de la mayoría de las formas de relacionarse los individuos inter e intraespecíficamente: por ejemplo, al extinguirse los dinosaurios -cuyas dimensiones imponentes señalaban el extremo superior para el tamaño de los animales terrestres de su época- los pequeños mamíferos ya estaban allí para sucederlos.

Además, la existencia de tales invariantes implica que, a pesar de su multitud, no todas las formas son posibles. En los organismos, los procesos ocurren y siempre influyen a la totalidad; de aquí derivan las restricciones estructurales. Así tanto en las dimensiones de los seres

vivos como en su fisiología hay pautas de interrelación que se cumplen en todas las escalas. Descubrir tales pautas es, desde nuestra perspectiva, equivalente a poder establecerlas en forma de ecuaciones. En todos los procesos de la vida y, particularmente en el crecimiento y la forma siempre es posible identificar patrones.

La alometría -del griego $\alpha\lambda o\varsigma$: variación, diferencia o cambio y $\mu\varepsilon\tau\rho o\nu$: medida- es el estudio de la variación de las magnitudes en los seres vivos. La alometría trata de las relaciones entre las diversas medidas corporales; analiza tanto las de tipo anatómco, propias de la arquitectura de los organismos, como las que puedan darse entre éstas y las variables con las que se cuantifican los procesos fisiológicos en un sentido amplio [7, 8]. Llama la atención el que las relaciones alométricas en diferentes niveles de organización sean de tipo pontencial. Se ha observado que las leyes de potencia tienen relación con criterios de optimización, como se menciona en la sección acerca de la ley de Zipf .

Resulta necesario hacer notar que, a pesar de la aparente similitud que existe entre las relaciones de escalamiento presentes en los estudios de alometría y la relación de escalamiento (a toda escala posible), que presentan los objetos fractales, ambos fenómenos describen facetas muy distintas entre sí, en el primer caso, se reflejan en características funcionales de entidades homólogas (hasta cierto grado) en especies diferentes, mientras que en el segundo caso, el escalamiento proviene de una propiedad intrínseca de un objeto matemático (que juega el rol de especie) con respecto a sí mismo. En este trabajo, más que centrarnos en los orígenes biológicos de la alometría, hacemos uso de los conceptos de autosimilaridad.

2.5.1. Fractalidad

El nombre viene del latín fractua, que significa irregular, pero a su descubridor, Benoit Mandelbrot también le gustaban las connotaciones de fraccional y fragmentario que hallaba en la palabra.

Los fractales abrazan no sólo los reinos del caos y el ruido, sino una amplia variedad de formas naturales que resultaban imposibles de describir mediante la geometría que se ha estudiado en los últimos dos mil quinientos años: formas tales como las líneas costeras, los árboles, las montañas, las galaxias, las nubes, los polímeros, los ríos, los patrones meteorológicos, los cerebros (la dimensión fractal en el caso del humano es de 1.79 a 2.73), los pulmones, etc. También hallamos estructuras fractales en las membranas de las células del hígado. Los huesos nasales del ciervo y el zorro ártico aumentan la sensibilidad olfativa de estos animales, concentrando la mayor superficie posible en un volumen pequeño. El resultado es una estructura fractal con una dimensión fraccional constante conocida como dimensión de Haussdorff, de la que hablaremos más adelante en extenso.

Tomemos por ejemplo nuestro sistema circulatorio. En un texto de anatomía, la repetida ramificación de las venas y arterias puede parecer caótica pero, si se la mira con mayor detalle, notamos que la misma y compleja ramificación se repite en vasos sanguíneos cada vez

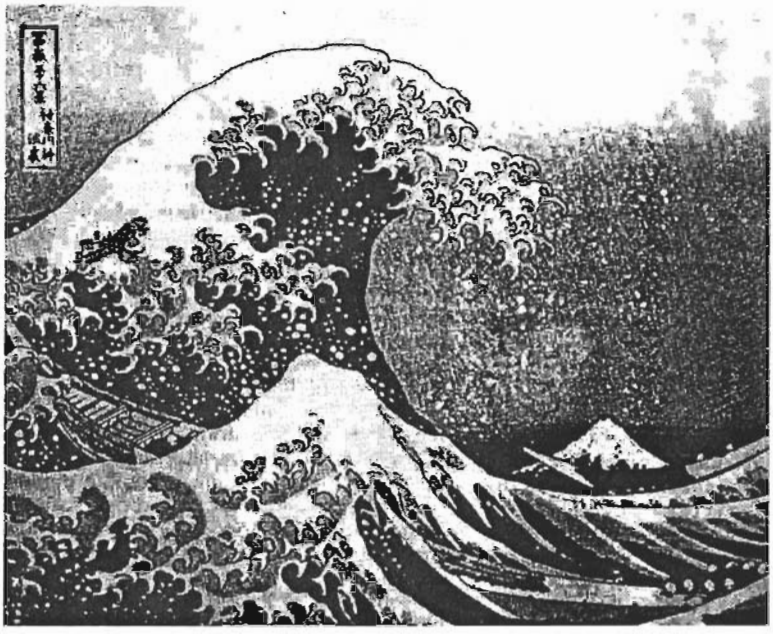


Figura 2.2: *La gran Ola* de Katsushika Hokusai

más pequeños, hasta llegar a los capilares. Lo mismo ocurre con una montaña. Visto desde kilómetros de distancia, el contorno de la montaña es muy reconocible, pero también es irregular. Cuanto más nos acercamos, más detalles apreciamos y cuando empezamos a escalar la montaña reparemos en el mismo patrón de irregularidad y detalle en cada roca individual. Los sistemas complejos de la naturaleza parecen preservar el aspecto de los detalles en escalas cada vez más finas. La cuestión de la escala surge nuevamente cuando observamos las maravillosas formas y estructuras de la naturaleza en un libro de fotografías tomadas a través de microscopios y telescopios. Las imágenes de escalas muy diferentes suscitan una sensación de similitud y reconocimiento de un mundo irregular y complejo pero al mismo tiempo ordenado.

Alrededor de 1890, Giuseppe Peano descubrió una curva que llena el espacio, es decir una curva que se torcía de modo tan complejo que llenaba el plano del papel donde se dibujaba. La línea curva de Peano incluía todos los punhtos del plano. Setenta años después, Mandelbrot tomó estas curvas en serio y siguiendo sus implicaciones, demostró conviencentemente que en dichas curvas residía el secreto del modo de medir la irregularidad del mundo real. El secreto de los fractales.

La figura 2 muestra la generación de un fractal originado en la curva copo de nieve elaborada por Helge von Koch en 1904. Esencialmente el fractal se crea mediante un proceso de

iteración en el cual cada paso se sigue en escala más pequeña. De este modo se produce una curva de considerable complejidad, la cual contiene un elevadísimo grado de detalle.

Al plantearse la longitud exacta de la línea costera de Gran Bretaña, Mandelbrot razonó que no era algo trivial. A primera vista parece un problema sencillo de solución simple: basta con medir. Pero las publicaciones y textos de geografía dan un kilometraje diferente para la misma línea costera o límite. Si fuera posible medir con una precisión cada detalle de la línea costera, incluidos rocas, guijarros, polvo y aún moléculas, la verdadera medida debe ser infinita.

Surgió ahora otra cuestión, como matemáticamente todas las líneas costeras con sus detalles reales deben tener una longitud infinita, ¿es posible comparar esas cifras? Mandelbrot descubrió que sí. Sin embargo ya no se trataba de medir la longitud cuantitativamente, sino de una nueva clase de medida cualitativa basada en escalas: la dimensión fractal.

Las dimensiones que encontramos en la vida cotidiana son números enteros: 0, 1, 2, 3. Sin embargo, ¿cuál es, por ejemplo, la dimensión de un ovillo de hilo?

Desde lejos, el ovillo luce como un punto y por tanto tiene dimensión cero. Pero a pocos metros de distancia el ovillo es tridimensional. Si nos acercamos mucho vemos un hilo curvado sobre sí mismo. El ovillo está compuesto por una línea curva y por lo tanteo es unidimensional. Desde más cerca, está línea se convierte en una columna de grosor finito y el hilo se vuelve tridimensional. Desde más cerca aun, dejamos de ver el hilo para ver las hilachas más finas que se tuercen unas alrededor de otras para formar el hilo: el hilo ha vuelto a ser unidimensional. En otras palabras, la dimensión efectiva del ovillo sigue cambiando de tres a uno una y otra vez. Su dimensión aparente depende de nuestra cercanía y determinarla no es tan simple como parece a primera vista.

Aunque al principio resulte desconcertante admitir que los objetos naturales tienen dimensiones efectivas, el concepto permite elaborar una dimensión fractal para una línea costera y descubrir que éste es un número fraccional mayor que uno. Si una curva o una dimensión fractal de la línea costera está cerca de uno, la costa es poco accidentada y no tiene detalles finos. Cuanto más se aleje este número de uno, más irregular o caótica será la línea costera, y esta irregularidad persistirá en escalas más pequeñas. Por otro lado, la dimensión fractal de una isla de Koch se encuentre entre 1 y 2 y se corresponde con una curva irregular que ocmparte algunas de las propiedades de una superficie bidimensional. Pero también hay una rica variedad de fractales cuyas dimensiones se encuentran entre la de un punto (0) y las de una línea (1).

Ahora comprendemos mejor la curva fractal creada por Peano. Esta curva se ha vuelto tan irregular en escalas infinitamente decrecientes que su dimensión fractal es dos. La línea tiene tantas sinuosidades que alcanza todos los puntos del plano. Sin embargo, a pesar de su

extrema complejidad, nunca se cruza consigo misma.

Los fractales se suelen caracterizar por los infinitos detalles, la infinita longitud, la carencia de inclinación (dimensión derivativa fraccional) y la autosimilitud y se pueden generar por iteración (como se hizo para producir la línea costera de Koch). Los fractales son muy complejos y muy simples al mismo tiempo. Son complejos en virtud de sus infinitos detalles y sus singulares propiedades matemáticas (no hay dos fractales iguales), pero son simples porque se pueden generar mediante sucesivas aplicaciones de la iteración simple.

Los fractales constituyen un sistema descriptivo y una nueva metodología para una investigación que sólo acaba de empezar. En las próximas décadas los fractales sin duda revelarán cada vez más acerca de los modos en que la estabilidad y el orden pueden nacer de la turbulencia y el azar subyacentes.

En su pintura La gran ola, el pintor japonés del siglo dieciocho Katsushika Hokusai capturó bellamente algunos de los interesantes aspectos del mundo fractal. En el siguiente capítulo veremos más a fondo la relación entre la alometría, la fractalidad y las leyes de escalamiento en las secuencias genómicas.

Capítulo 3

Lingüística, Estadística, Complejidad y Código Genético

Se considerarán a continuación, de manera más detallada, los fundamentos estadísticos y de Lingüística Matemática que constituyen la base de los métodos empleados en este trabajo con el fin de analizar las características esenciales comunes a los genomas de una variedad de especies. De particular importancia para este capítulo son los factores relativos al contenido de información (en el sentido formal del término), la coherencia y estructuración, así como las reglas de escritura para el ADN, particularmente para las secciones intrónicas de éste. Debido a que el presente estudio centra su atención en el carácter lingüístico de la codificación en el genoma, se estudiará la distribución de frecuencias de las palabras de este lenguaje genético con métodos usuales de la Lingüística Matemática. Se aplicaron métodos lingüísticos como el propuesto por Zipf, a fin de considerar coherencia y orden lingüístico tanto en segmentos del ADN de cuatro organismos, como en un generador aleatorio, así como en un texto en español contemporáneo. El fundamento lingüístico y matemático de la llamada Ley de Zipf y las adaptaciones necesarias describir su aplicación al código genético, se describen en este capítulo.

Así mismo, se describe detalladamente el algoritmo de Shannon-Weaver para la determinación del contenido informático de una cadena textual. Debido al carácter autosimilar de las distribuciones de frecuencia de palabras, se utilizaron métodos provenientes de la dinámica no lineal, el estudio de secuencias caóticas (en el sentido técnico) y la geometría fractal. Por otro lado, se caracterizaron las secuencias genómicas en términos de su dimensión fractal o de Hausdorff-Besicovitch calculada a través del algoritmo de Procaccia y Grassberger para la función integral de correlación. Además, se generaron cuasifractales a partir de las series de tiempo tras un procedimiento de renormalización. El estudio de las distribuciones de frecuencia consideradas como procesos estocásticos, se llevó a cabo estudiando las cadenas de Markov y procesos no Markovianos asociados con la persistencia estadística en la longitud de correlación a lo largo del genoma; adicionalmente se realizó un cálculo relacionado con la complementariedad de bases para evidenciar la presencia de proporciones específicas entre bases en las secciones intrónicas del genoma comparado con el comportamiento azaroso

presente en el generador aleatorio utilizado como control.

3.1. Modelos estadísticos

3.1.1. Entropía Informacional de Shannon

La teoría del lenguaje no es sólo un objeto formal, pues si es aplicada de manera correcta a problemas específicos puede proveer herramientas computacionales y construcciones útiles que dan resultados adecuados. Se utilizará una clase especial de lenguajes llamados lenguajes factorizables. A continuación se dará un breve resumen de la teoría del lenguaje en general.

Comenzamos con un alfabeto finito Σ ={ A,C,G,T } y colectamos todas las posibles cadenas de caracteres de estas letras en un conjunto infinito Σ^* , que incluye la cadena vacía, esto es la cadena que no contiene ninguna letra. Cualquier subconjunto L de Σ^* es llamado un lenguaje sobre el alfabeto Σ. A fin de definir qué clase de lenguaje estamos utilizando debemos dar la regla generadora de L, lo cual puede hacerse de varias maneras, por ejemplo:

1. Si el subconjunto L es finito, podemos simplemente enumerar sus elementos
2. Es posible desarrollar algunas reglas de producción y aplicarlas repetitivamente a algunas letras iniciales a fin de generar el lenguaje. Este es con mucho el modo más importante y bien estudiado de definir lenguajes. Si las reglas son aplicadas secuencialmente nos lleva a la gramática generativa de Chomsky. Si son aplicadas en paralelo nos llevan a los sistemas de Lindenmayer.
3. Para una clase especial de lenguajes, los lenguajes factorizables, es posible definir un lenguaje mediante indicar su conjunto de palabras prohibidas, este enfoque se sigue usualmente en los análisis de DNA.

Una clase especial de lenguajes factorizables se puede definir en un genoma completo: dado un genoma completo de un organismo G es posible cortar las secuencias del DNA en todas sus posibles subsecuencias y formar un lenguaje L=sub(G), mediante coleccionar estas subsecuencias incluida la cadena vacía. Este lenguaje es factorizable por definición por lo que es posible construir un lenguaje determinista a partir de él [25].

Otras medidas estadísticas de interés de las correlaciones de corto y mediano alcance en los lenguajes son la entropía y la redundancia. Esta última es una manifestación de la flexibilidad del código subyacente.

Tanto los lenguajes naturales como los genomas (un tipo de lenguaje biológico) son redundantes, esto quiere decir que repiten la información que contienen; sin embargo, no hacen esto de manera errática ni a lo largo de todo el texto (genético o lingüístico), más bien, recalcan la información que poseen en determinadas secciones para compensar la *corrupción* o pérdida de información (de lo que quiere expresar el emisor al texto, de ahí al primer receptor, de éste al segundo, etc; en el caso de los genomas podría hablarse de intercambio de

información entre genes) que se genera en este tipo de sistemas. Perder información es ganar entropía y viceversa, son complementarias, la suma de ambas es una constante. Que debería ser, aunque no siempre es, cero.

Todos los mensajes con un costo total menor son aquellos donde las palabras son usadas de acuerdo con la Ley de Zipf, esto es, que poseen la cantidad máxima posible de información. Zipf se dio cuenta de que la función $f_i = K_a R_i^{-\alpha}$ describía la información que podía transmitir un texto. Es posible calcular en este sentido la información contenida en una distribución. Esta información se mide con la fórmula de Shannon:

$$I = -K_s \sum_{\nu} p_\nu \log p_\nu \tag{3.1}$$

donde I es la información de Shannon-Weaver, K_s es una constante y p_ν es la frecuencia de aparición de la ν-ésima palabra.

3.1.2. Cadenas de Markov

Eventos aleatorios

Llamaremos *evento aleatorio* a alguno de los posibles resultados de un proceso probabilístico, sea que conozcamos o no las razones de su comportamiento aparentemente azaroso. Algunas veces lo llamamos también proceso puntual, conjunto aleatorio de puntos o función aleatoria. Comúnmente su distribución de probabilidad cambia en el tiempo, por lo que a veces lo llamaremos *evento estocástico*. El espacio muestral para éste, consiste en un conjunto *contable* de secuencias de funciones de distribución de probabilidad [4].

Por ejemplo, si nuestro evento consiste en hacer una torre con monedas y observar con que número de monedas se vendrá abajo la torre por inestabilidad mecánica tras un cierto tiempo, podemos asignar una función $Q_1(t_1)$ a la probabilidad de que la torre *de una moneda* se venga abajo tras el tiempo t_1 [1]. Luego tenemos la probabilidad de que la torre de dos monedas se venga abajo después de un tiempo $t = t_2$ que será $Q_2(t_2)$ y así sucesivamente tendremos para una torre de tamaño n la probabilidad $Q_n(t_n)$ de que se caiga al tiempo $t = t_n$. Sin embargo es notorio que si una torre tiene ya 18 monedas y es muy inestable, la torre de 19 monedas tiene una distribución de probabilidad de caerse **basada** en la de 18 monedas (y en principio mayor a esta última) y ésta a su vez en la de 17 monedas, etc. [2]. Entonces podemos escribir:

$$Q_n = \int Q_{n-1}(t_{n-1}) dt_{n-1} + Q_n^* \tag{3.2}$$

Donde el primer término $\int Q_{n-1}(t_{n-1}) dt_{n-1}$ representa la probabilidad de que la moneda n (digamos la 18) se caiga debido a una inestabilidad en la torre de 17 monedas (o sea en

[1] Obviamente por el contexto de este ejemplo $Q_1 = 0\ \forall t$.

[2] Si se cae la moneda 7, se vendrán abajo también la 8, la 9, etc.

el hecho de que se caiga *cualquiera* de las 17 monedas anteriores) y el término segundo Q_n^* representa al hecho de que se caiga la torre exclusivamente por la adición de la moneda 18. Ahora bien, podemos escribir una ecuación similar a 3.2 para la torre de tamaño $n-1$ y para la de tamaño $n-2$ y así sucesivamente. Si lo hacemos:

$$Q_n = \int\int\int \ldots \int\int Q_{n-1}(t_{n-1}) dt_{n-1} dt_{n-2} dt_{n-3} \ldots dt_3 dt_2 dt_1 + \sum_{j=2}^{n} Q_j^* \qquad (3.3)$$

Ahora pensemos en que en lugar de monedas se tiene una secuencia de bases nitrogenadas. Obviamente, la probabilidad de tener una T, por ejemplo en el octavo sitio dependerá fuertemente de qué nucleótido hay en el séptimo sitio, pero este depende de qué hay en el sexto sitio, por lo que la T en el octavo sitio depende de qué hay en el sexto sitio, ... y por lo tanto de lo que hay en el quinto, cuarto, tercero, segundo y primer sitio. Así pues, si no tenemos *ninguna otra información* acerca de cuál es la relación que existe entre las distribuciones de probabilidad en los demas sitios [3] tenemos un *gran* problema. La distribución de probabilidad para el sitio digamos 300 representaría CONOCER e INTEGRAR *300 veces* la distribución de probabilidad para cada sitio anterior en la cadena. Obviamente, no vamos a hacer esto. En la práctica resulta casi imposible para una cadena de este tamaño y es ABSOLUTAMENTE imposible para una cadena de tamaño real. Esta es una de las razones por las cuales un tratamiento estadístico de alto nivel se requiere. Posteriormente vamos a analizar dos casos en los que el horripilante problema planteado por la ecuación 3.3 se simplifica significativamente: los eventos independientes y las cadenas de Markov. Pero esos casos vendrán después, antes analizaremos el significado de la correlación estadística.

Funciones de correlación

Decimos que dos eventos estocásticos están *correlacionados* si el resultado de uno de ellos depende en alguna medida del resultado del otro, aunque sea en un corto intervalo de tiempo [11]. Así, por ejemplo, el tiempo empleado estudiando para un examen estará relacionado con la calificación obtenida. Si registramos el tiempo que cada alumno de una escuela en particular estudió para algún examen [4] y observamos el registro de calificaciones[5], notaremos que evidentemente hay una relación. Dado que la relación es mutua y en cierta medida para evitar incluir el posible prejuicio estadístico de achacar causalidad[6] llamaremos a este tipo de interrelación *correlación estadística* y diremos que ambos eventos están *correlacionados*. Por otro lado notaremos que este registro de calificaciones no guarda relación alguna con el registro alfabético de estudiantes, es decir ambos registros no están correlacionados.

La función matemática que indica qué tan relacionados están dos eventos aleatorios α y β se llama la *función de correlación* $C_{\alpha\beta}$ y es una función de la naturaleza de los eventos α y

[3] Recordemos que en nuestro caso la posición en la cadena reemplaza a la variable tiempo.

[4] Por ser muchos alumnos *distintos* este es un evento estocástico.

[5] Evento estocástico por las razones anteriores.

[6] Obviamente en este evento en particular si hay causalidad.

β, así como de los tiempos característicos t_α y t_β. Si $C_{\alpha\beta}$ tiene un valor muy grande entonces los eventos están muy correlacionados. Si $C_{\alpha\beta}$ vale cero los eventos no están correlacionados. Por simetría de las DPt's tenemos que $C_{\alpha\beta} = C_{\beta\alpha}$ [4, 11]

La función de correlación es una cantidad muy importante porque nos ayuda a saber cuando la relación entre dos eventos es tan pequeña que se puede despreciar y así nos permite simplificar los cálculos. Pensemos, por ejemplo en el precio de un paquete de jabón. Este precio depende tanto del tamaño y tipo del jabón que quieras comprar (lo que llamaremos calidad) como de dónde lo compras (digamos localidad). Sin embargo si calculáramos las funciones de correlación $C_{precio-calidad}$ y $C_{precio-localidad}$ probablemente nos daríamos cuenta de que la primera es mucho más grande y por lo tanto, de manera aproximada podemos afirmar que el precio de un paquete de jabón depende de su calidad y nos olvidaríamos de comparar su precio en diversas tiendas (es decir no calcularíamos la distribución de probabilidad precio-localidad).

Correlaciones cruzadas y autocorrelación

Hemos hablado ya de la correlación entre dos eventos distintos α y β, sea en el mismo tiempo t o a distintos tiempos t_α y t_β. A esta clase de relación la llamaremos *correlación cruzada*, por ejemplo ¿Qué tan probable es que llueva hoy, dado que *hace viento*? o ¿Qué tan probable es que llueva hoy si *hizo viento* ayer?. También podría interesarnos la relación que existe entre la probabilidad de que ocurra *el mismo evento* pero en dos tiempos distintos, por ejemplo ¿Cuál es la probabilidad de que llueva hoy dado que llovió ayer?. A este fenómeno lo llamamos *autocorrelación*. Para escribirlo taquigráficamente, tendríamos $C_{\alpha\beta}(t)$, $C_{\alpha\beta}(t_1, t_2)$ y $C_{\alpha\alpha}(t_1, t_2)$ respectivamente. Cuando estamos considerando una variedad de eventos aleatorios que pueden estar relacionados, suele ser conveniente definir una *matríz de correlación* C_{ij} donde $i, j = 1, 2, 3, \ldots$ representan a los diversos eventos estocásticos. Los términos no diagonales ($i \neq j$) se refieren a las correlaciones cruzadas y los terminos diagonales ($i = j$) a las autocorrelaciones.

3.1.3. Procesos estocásticos

Una vez que hemos definido una variable estocástica X, es posible definir un número infinito de otras posibles variables estocásticas relacionadas, es decir todas aquellas funciones Y que están relacionadas con X mediante una expresión del estilo:

$$Y_X(t) = f(X, t) \tag{3.4}$$

Llamamos a la expresión 3.4 una *función aleatoria*, o como el parámetro t casi siempre es el tiempo, un *proceso estocástico*. Así un proceso estocástico es de manera simple una función de dos parámetros, uno de los cuales es una variable aleatoria X y el otro es, generalmente el tiempo t.

Por ejemplo, la cantidad de personas que entra a un supermercado es, obviamente una variable aleatoria y podemos obtener su función de distribución si hacemos estadísticas (o sea datos X) sobre, por ejemplo las ventas de ese supermercado. Por otro lado, resulta obvio que en diferentes tiempos (diversas horas del día, diferentes días de la semana o épocas del año) la cantidad de personas que van al supermercado cambiará. Así esta variable de naturaleza aleatoria también depende del tiempo. Es decir, es un proceso estocástico. Si pudieramos determinar una función que en términos de las estadísticas medidas en la tienda (o sea las x's que forman la distribución X) y del tiempo t nos dijera cuántas personas entrarán a la tienda a una hora determinada, tendríamos una función de la distribución X y del tiempo. O sea, tendríamos una expresión análoga a la ecuación 3.4. Si de la variable X tenemos algún posible valor x^* en particular, obtenemos una *realización* o *función de muestra.*

$$Y_{x^*}(t) = f(x^*, t) \tag{3.5}$$

Es posible encontrar el valor promedio, en términos de la densidad de probabilidad $P_X(x)$ de la variable aleatoria:

$$\langle Y(t) \rangle = \int Y_x(t) P_X(x) dx \tag{3.6}$$

El símbolo $\langle Y \rangle$ representa el *promedio estocástico* de la variable Y a veces llamado también *promedio de ensamble* o, promedio simplemente. De manera similar podemos calcular los momentos estocásticos de la variable Y como lo hicimos con el evento aleatorio original X. Una cantidad muy importante es la *función de autocorrelación estocástica* $\kappa_Y(t_1, t_2)$ que definimos como sigue:

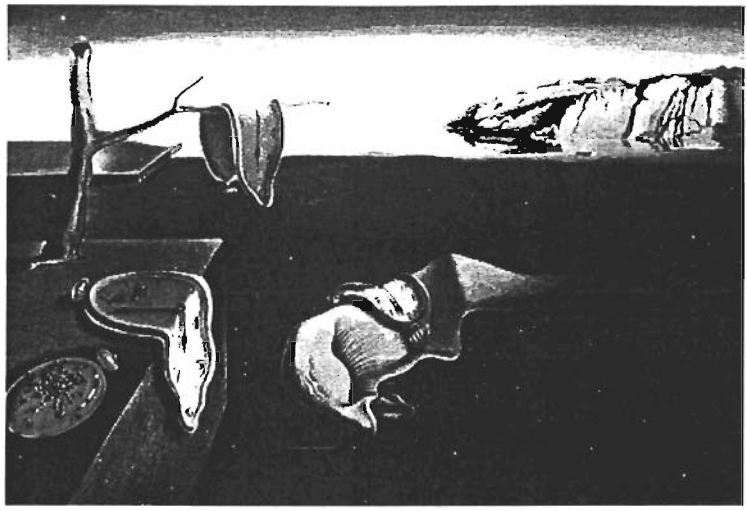
$$\kappa_Y(t_1, t_2) = \langle Y(t_1) - \langle Y(t_1) \rangle \rangle - \langle Y(t_2) - \langle Y(t_2) \rangle \rangle = \langle Y(t_1) Y(t_2) \rangle - \langle Y(t_1) \rangle \langle Y(t_2) \rangle \tag{3.7}$$

Esta expresión corresponde con el concepto que llamamos $C_{\alpha\beta}(t_1, t_2)$ para eventos aislados α y β, pero ahora para procesos completos. Si evaluamos la expresión 3.7 para $t_1 = t_2 = t$ esta se reduce a la varianza dependiente del tiempo.

$$\kappa_Y(t, t) = \langle \langle Y(t)^2 \rangle \rangle = \sigma_Y^2(t) \tag{3.8}$$

Procesos de Markov

Como seguramente recordamos hace algunas páginas prometimos encontrar algunos casos en los que la ecuación 3.3 se simplificaría enormemente. Pues bien, a continuación presentaremos al más *popular* de los procesos estocásticos, el *proceso de Markov.* Si tenemos un proceso estocástico definido para eventos sucesivos ξ_1, ξ_2, ξ_3, etc. que ocurren en tiempos t_1, t_2, ... también sucesivos, la probabilidad $p_n(t_n)$ de que ocurra el evento ξ_n dependerá en general de cada una de las probabilidades anteriores $p_{n-1}(t_{n-1})$, $p_{n-2}(t_{n-2})$, ..., $p_3(t_3)$, $p_2(t_2)$ y $p_1(t_1)$. Si, además, cada una de estas probabilidades tiene asociada una función de distribución Q_n nos hallamos frente a un caso formalmente idéntico a la ecuación 3.3. Al realizar la gran cantidad de integrales requeridas se reflejaría el hecho de que existe correlación



La persistencia de la memoria (1931)

estadística entre TODAS las Q_n's. Sin embargo, sabemos que la magnitud de estas correlaciones no es igual en todos los casos [11].

Por ejemplo, pensemos de nuevo en el ejemplo de la lluvia. La probabilidad de que llueva hoy, muy posiblemente se vea afectada por el hecho de que haya o no llovido ayer (con una cierta distribución de probabilidad), la cual también probablemente influencie el hecho de que llueva o no mañana (con alguna probabilidad). Lo cierto es que, difícilmente consideraríamos que el hecho de que haya llovido hace 8 meses o 5 años vaya a afectar el clima la semana entrante. Se presenta comúnmente el fenómeno de *memoria desvaneciente* [7] en mayor, o menor grado. Esto significa que después de un tiempo al sistema *se le olvida* su comportamiento del pasado lejano y sólo se acuerda del pasado reciente. En el caso extremo podríamos considerar un proceso estocástico en el que se cumple lo siguiente:

$$C_{Q_nQ_{n-1}} \neq 0\,; \quad C_{Q_nQ_{n-2}} = C_{Q_nQ_{n-3}} = C_{Q_nQ_{n-4}} = \ldots = C_{Q_nQ_2} = C_{Q_nQ_1} = 0 \quad (3.9)$$

Es decir, la probabilidad p_n depende formalmente de la probabilidad p_{n-1} pero no depende de ninguna de las demás probabilidades. La ecuación 3.9 define a un proceso de Markov. Observamos que la única correlación que importa es entre un evento n y su predecesor inmediato $n-1$ careciendo por completo de valor lo que haya ocurrido en eventos anteriores. Si hacemos esto, las integrales sobre las restantes variables en 3.3 son sobre distribuciones de eventos sin correlación. Ahora bien, recordamos que si el número de eventos es

[7] Aplicar el término *memoria* a entidades matemáticas puede parecer un antropomorfismo excesivo, pero es la mejor manera de relacionar el hecho matemático con la experiencia cotidiana.

grande *no-correlacionado* implica *estadísticamente independiente*. La ecuación 3.3 se transformará en:

$$Q_n = \int Q_{n-1}(t_{n-1})dt_{n-1} + \sum_{j=2}^{n} Q_j^* \tag{3.10}$$

Esta ecuación 3.10, salvo el último término es idéntica a la ecuación 3.2. Dado que el último término es una suma de constantes podemos reemplazarlo con una sola constante y las ecuaciones serían idénticas. Ejemplos de procesos de Markov son el dinero ganado o perdido en los *volados doble o nada*[8], el número de neutrones que están libres en un reactor nuclear [9], la velocidad (tanto su dirección como su magnitud) de una parícula browniana [10], las posiciones de los clasificados en la *Tour de France* [11], etc.

3.1.4. Cadenas de Markov

Una *cadena de Markov* es un conjunto discreto de eventos cuyas distribuciones estocásticas son procesos de Markov. Una serie de tiempo para eventos en la que la correlación no llega más allá de un sitio en la cadena [12]. Por ejemplo, en un hospital hay un cierto número de pacientes, estos pueden ser de 2 tipos leve (L) o grave (G). Si un enfermo leve mejora, sale del hospital y su lugar es ocupado por otro enfermo que puede ser leve o grave. Si el enfermo leve se pone grave continúa en el hospital. Si un enfermo grave muere su lugar es ocupado por otro enfermo grave o levemente enfermo. Si este paciente grave mejora hasta convertirse en leve continúa en el hospital. La proporción de enfermos de cada tipo [13] es una variable estocástica que obedece a una distribución de Markov y su registro en el tiempo es una cadena de Markov.

Veamos una *corrida* de tal cadena. Iniciamos con 50 L's y 50 G's. Vamos a evaluar el número de G's. Tras el primer día un enfermo leve mejora y sale del hospital, su lugar es ocupado por un enfermo grave [14], el segundo día no pasa nada, al tercer día otro enfermo leve mejora y se va del hospital, su lugar es ocupado por otro enfermo grave [15], al cuarto día un enfermo grave se muere y su lugar es ocupado por un enfermo leve [16], etc.

La lista $G(t) = \{50, 51, 51, 52, 51, \ldots\}$ diaria del número de enfermos graves es una cadena de Markov y junto con la distribución de probabilidad $p(t) = \{1, p_G(t_1), 1, p_G(t_3), p_L(t_4), \ldots\}$ constituyen un proceso de Markov.

[8] Si ganaste 100 veces antes, pero perdiste el último juego no tienes ya dinero o sólo te queda para un juego.

[9] Por la reacción en cadena los neutrones no quedan libres mucho tiempo.

[10] Si el último golpe aleatorio no fue suficiente para moverla no se mueve, si fue el golpe anterior ya se movió. Los demás golpes no cuentan.

[11] No importa si fuiste campeón hace 8 años, pero si ganaste el año pasado esta vez arrancas el primero.

[12] El tiempo de correlación vale una unidad.

[13] Recuerda que como sólo hay dos tipos conocer la proporción de un tipo nos da la información completa.

[14] con cierta probabilidad $p_G(t_1)$

[15] con cierta probabilidad $p_G(t_3)$

[16] con cierta probabilidad $p_L(t_4)$

3.1.5. Probabilidades de transición

Pensemos ahora en una cadena de Markov como una *secuencia de sucesos* que ocurren, cada uno como *consecuencia*[17] del anterior. Piensa en el ejemplo del hospital, el hecho de que en el tiempo $t = t_j$ un paciente leve pueda ser admitido depende de dos cosas:

- De que el *día anterior* ($t = t_{j-1}$) un enfermo grave un enfermo grave haya muerto o un enfermo leve haya sanado y por lo tanto tengamos un lugar en el hospital. Ambos eventos son, desde luego probabilistas.
- El hecho de en el día $t = t_j$ alguien este enfermo levemente y requiera ingresar al hospital.

Si hacemos esto nos daremos cuenta que la probabilidad de que un enfermo leve ingrese el día t_j al hospital en lugar de un enfermo grave que murió depende de dos eventos como los recientemente esbozados. Si a la probabilidad de que un enfermo grave muera el día t_{j-1} la llamamos $p_G^{\dagger}(t_{j-1})$ y a la probabilidad de que alguien enferme levemente el día t_j la llamamos $p_L^{\spadesuit}(t_j)$ y consideramos que *ambos eventos son estadísticamente independientes*, entonces la probabilidad de que ambas cosas ocurran es, desde luego, el producto de las probabilidades individuales. En los términos ya discutidos tenemos que:

$$p_L(t_j)^{\rightarrow G} = \left(p_G^{\dagger}(t_{j-1})\right)\left(p_L^{\spadesuit}(t_j)\right) \tag{3.11}$$

La ecuación anterior representa la probabilidad de que al tiempo t_j un enfermo leve reemplace a un enfermo grave que murió recientemente. Llamamos al primer factor en la ecuación 3.11 la probabilidad de transición del estado $j-1$ al estado j. Este término representa la probabilidad de que estén presentes las condiciones en sentido estadístico para que el evento j-ésimo pueda darse [18].

Obviamente la probabilidad *completa* de que alguien levemente enfermo ingrese al hospital depende tanto de que reemplaze a un enfermo grave que murió como si reemplaza a alguien levemente enfermo que sanó. Esto es:

$$p_L(t_j) = p_L(t_j)^{\rightarrow G} + p_L(t_j)^{\rightarrow L} \tag{3.12}$$

es decir:

$$p_L(t_j) = \left(p_G^{\dagger}(t_{j-1})\right)\left(p_L^{\spadesuit}(t_j)\right) + \left(p_L^{\heartsuit}(t_{j-1})\right)\left(p_L^{\spadesuit}(t_j)\right) \tag{3.13}$$

Donde, evidentemente los superíndices † y ♡ en el estado de transición de un paciente indican respectivamente muerte y mejoría.

[17]Consecuencia en términos estadísticos y no deterministas. No estamos aduciendo directamente a causalidad.

[18]Para ilustrarlo, la probabilidad de que yo cruze el periférico dependerá de dos probabilidades: una, la probabilidad de que yo esté situado a la orilla del periférico y dos, la probabilidad de que quiera cruzarlo. Aunque la decisión esencial es la segunda, ¡ no puedo cruzar el periférico si estoy en el Zócalo!.

Ecuaciones Maestras

En la ecuación 3.13 tenemos la suma de dos posibilidades que llevan al evento $p_L(t_j)$. Si tuviéramos muchas posibilidades distintas, todas ellas conduciendo al mismo evento al tiempo t_j y las sumáramos sobre todos los posibles tiempos t_j ($j = 1, 2, 3, \ldots n$) obtendríamos una ecuación para todas las posibles maneras en que tal evento ocurriría. Llamamos a tal expresión una *ecuación maestra.*

Las ecuaciones maestras suelen tener una cara como la siguiente:

$$P(x_j, t_j) = \int_{t_i<t_j} \sum_k Q_k(t_{i-1}) P_k(t_i) dt_i \tag{3.14}$$

Donde la suma va sobre las k posibilidades en que algún evento pueda ocurrir a un tiempo determinado y la integral[19] es sobre todos los tiempos t_i anteriores al tiempo t_j que nos interesa.

Cadenas de Markov y ADN

Los sistemas Markovianos corresponden a decaimientos exponenciales y los no-Markovianos poseen largas colas del tipo que corresponde a una ley de potencias. Analizando las gráficas de frecuencia de aparición vs. rango generadas a partir de las distribuciones de probabilidad (Análisis Lingüístico de Zipf) es posible notar a cuál de estas dos opciones pertenecen los genomas seleccionados y por tanto qué tipo de memoria manejan. En este trabajo se realizará este análisis tomando en cuenta el carácter persistente de las distribuciones de probabilidad para palabras en los genomas.

Las secuencias de ADN han sido analizadas utilizando una gran variedad de modelos que pueden ser considerados básicamente en dos categorías. Los del primer tipo son análisis locales que toman en cuenta el hecho de que las secuencias del ADN son producidas en orden progresivo, de manera tal que los pares de bases vecinas afectan el enlazamiento del siguiente par de bases. Este tipo de análisis tales como los modelos Markovianos de n pasos pueden describir algunas de las correlaciones de corto alcance observadas en las secuencias de ADN. La segunda categoría es de naturaleza más global y se concentra en la presencia de patrones repetidos (tales como las repeticiones periódicas y las repeticiones secuenciales de bases) que pueden ser encontrados en la mayoría de las secuencias genómicas. Un análisis típico de esta clase es el análisis de la transformada de Fourier [20] o análisis espectral que se basa en localizar los puntos de máxima frecuencia en una serie de tiempo.

[19] Que ahora es una sola y no como las miles de la ecuación 3.3.

[20] Podemos ver a la transformada de Fourier como un cambio de coordenadas o un cambio de variable que nos permite visualizar de manera más clara fenómenos en el *dominio de la frecuencia* o alternativamente en el del *tiempo*. De manera simple, si llueve una vez al mes el tiempo en el que se presenta lluvia es cada 30 días, es decir con período 30, en el dominio de la frecuencia diríamos que la frecuencia es de $\frac{1}{30}$. Aunque en este caso ambas descripciones alternativas son muy obvias y trivialmente equivalentes, este no es el caso para sistemas más complejos en los que la descripción recíproca es muchas veces indispensable para entender el fenómeno.

3.2. Modelos lingüísticos

3.2.1. Lingüística Matemática

En general, se entiende por lingüística matemática o lingüística cuantitativa todos aquellos métodos que utilizan cantidades exactas para proveernos información acerca del lenguaje.

Hay no menos de cuatro vertientes que pueden denominarse así:

1. La lingüística estadística, incluyendo la utilización de computadoras
2. La consideración del lenguaje como objeto de la teoría de la información
3. Las interrelaciones de la lógica matemática con el lenguaje, a nivel teórico y práctico
4. La traducción automática

Para fines del presente trabajo se procederá a analizar las primeras dos, que son las que tienen una relación más cercana con los fines que éste persigue.

La lingüística estadística

Hasta hace poco los recuentos de vocabulario de un autor, obra, época ... habían exigido la labor paciente de investigadores que leían, clarificaban y elaboraban datos. Hoy día, gracias al avance de las ciencias de la computación, existen ordenadores capaces de realizar estas arduas tareas matemáticas con un ingente ahorro de tiempo y esfuerzo.

Las computadoras, sin embargo, no trabajan de espaldas al hombre. Son unos instrumentos más, que proporcionan los datos buscados con mayor precisión y rapidez, pero que exigen la presencia del hombre: a) antes, para dirigir su trabajo; b) para recopilar e interpretar los datos que proporciona.

Supongamos que deseamos realizar un recuento exhaustivo de las palabras que aparecen en *el Quijote*. Para empezar, vertemos la obra a un código que la máquina comprenda. Por ejemplo, mediante una máquina de escribir con caracteres especiales o que produce perforaciones en la banda sobre la que escribe. Esta banda -traducción del texto a un código convencional- puede servir de texto de entrada a la máquina. La máquina puede leer -por su mecanismo- aquellos caracteres. El ordenador de acuerdo con los manejos de quien lo gobierna, ordenará aquella información que se le da, según cierto modo. Es posible así, por ejemplo realizar el recuento total de las palabras del Quijote. Con estos resultados, el lingüistas tendrá datos seguros sobre los que basar sus apreciaciones acerca de la obra. Nos hemos referido, y por encima, a un aspecto o un modo de trabajar de las computadoras. Las posibilidades de aplicación son muchísimas más.

En fonética se pueden estudiar de este modo las frecuencias relativas de los fonemas de una lengua, pormenorizando su aparición en sílabas, palabras, comienzos de frase, etc. Lo

mismo se puede hacer con respecdto a otros rasgos fonéticos: los acentos, la largura de la palabra, la entonación, etc. Con estos datos, se puede estudiar rigurosamente la configuración fonética de una lengua, confrontarla con otras, servirse de todos ellos para los laboratorios de idiomas y la enseñanza, etc.

En gramática, la estadística pone en claro los índices de frecuencia de los elementos gramaticales, su lugar en la oración, sus contornos léxicos, etc. Todo ello utilísimo para confeccionar la gramática de la lengua en cuestión [9].

Los índices de frecuencia en semántica pueden surtirnos de cantidad de datos que nos ayudarán tanto a la elaboración de un diccionario como a la confección de vocabularios básicos de un idioma. Sus aplicaciones en el campo de la estilística son innumerables: preferencias léxicas de un autor, repeticiones, ausencias de ciertos términos, combinaciones raras de palabras, etc. Todo ello puede deducirse de un índice de palabras bien elaborado. En el caso de los textos poéticos, la máquina puede recoger los datos estadísticos que se refieren al verso, sus tipos, rimas, acentos, etc.

El estudio estadístico de la lengua no acaba aquí, se prosigue con formulaciones matemáticas qu resultan de observar el trabajo bruto realizado por la máquina. Por ejemplo: se sostiene que las palabras se distribuyen a partir de unos esquemas matemáticos más o menos constantes y se enuncian reglas como la siguiente: la frecuencia de una palabra es función de su número de orden en la lista (listas de frecuencia), y el producto de este número por la frecuencia constante. Otras observaciones, sin ser tan rigurosas, no carecen de interés: la leyt del mínimo esfuerzo por la que tienden a repetirse las mismas palabras y se abusa de los sustitutos (algo, cosa, hacer, los pronombres . . .); la relación entre la frecuencia de una palabra y su longitud relativa a la longitud media de las palabras de aquella lengua o la relación entre su frecuencia y su número de acepciones, etc. Muchos de estos aspectos tan rápidamente enunciados aquí, han podido ser reducidos a leyes, cuando son se ha encontrado que obedecen incluso a fórmulas ya enteramente conocidas in abstracto o con respecto a otros objetos de estudio.

Caben aún experimentos más audaces. De hecho, un ordenador puede deducir modelos generalizados de uso, parámetros, modos de hacer que se repiten en una obra de cualquier tipo, o en una lengua. A partir de ellos cabe reconstruir el estilo del autor. Nos daremos una idea del alcance de estos estudios si meditamos sobre un experimento de este tipo: una computadora que, después de deducir los esquemas generales de las composiciones musicales de un autor (Bach), proporcionó el material suficiente y apropiado para componer lo que sería una obra de Bach, la obra ideal que nunca compuso como el resumen de todas ellas.

Teoría de la Información

El tratamiento cuantitativo del lenguaje se efectúa otras veces cuando se considera a este sub especie de información. La Teoría de la Información (una de cuyas ramas es la informáti-

ca) es una ciencia reciente que se basa en postulados muy sencillos. Ofrecemos la exposición muy clara de Umberto Eco:

Cuando entre dos acontecimientos, sabemos cuál se producirá, tenemos una información. Hemos de suponer que ambos acontecimientos tienen iguales probabilidades de producirse y que, por tanto, nuestra ignorancia respecto a la disyuntiva de probabilidades, es total. La probabilidad es la relación entre el número de casos favorables a la realización del acontecimiento y el número de casos posibles. Tirando una moneda al aire, para obtener cara o cruz, dispongo de una probabilidad de $\frac{1}{2}$ *para cada cara de la moneda.*

Tratándose de un dado con seis caras, tengo una probabilidad de $\frac{1}{6}$ para cada cara (en el caso de tirar dos dados, la probabilidad de que se produzcan *conjuntamente dos acontecimientos -de que se consiga sacar 6 y 5, por ejemplo es el producto de las probabilidades simples, es decir, de* $\frac{1}{36}$).

La relación entre una serie de acontecimientos y la serie de probabilidades correspondientes, es la relación entre una progresión aritmética y otra geométrica, y la segunda serie representa el logaritmo de la primera.

Eso quiere decir que, teniendo una eventualidad y 64 posibilidades de realización distintas (las de la posición de una figura en el tablero de ajedrez, por ejemplo), al saber cual de ellas se ha producido he obtenido una información equivalente a logaritmos 64 (que es 6). O sea que, para individualizar una eventualidad entre 64, hay sido precisas 6 disyuntivas o selecciones binarias.

Este mecanismo puede explicarse mejor mediante el esquema adjunto, reduciendo el número de elementos para facilitar la operación. Teniendo ocho eventualidades, de las que no podemos predecir cuál ocurrirá, la individualización de una de ellas se hace por medio de selecciones binarias e implica tres operaciones, tres opciones, tres alternativas.

Hemos indicado con letras alfabéticas los puntos de disyunción binaria. Y así, por ejemplo, para identificar la eventualidad número 5, se precisan tres selecciones binarias: 1) de A, selecciono entre B1 y B2; 2) de B2, escojo la dirección hacia C3; 3) de C3 escojo dirigirme hacia 5 en vez de hacia 6. Pues que se trata de individualizar una eventualidad entre ocho, la expresión logarítmica de la situación es: log2 B = 3

En la teoría de la información se le llama unidad de información o bit (del inglés binary digit o señal binaria), a la unidad de disyunción binaria que sirve para individualizar una alternativa. Si se trata de individualizar un elemento entero entre ocho, habremos recibido 3 bits de información; en el caso de los 64 elementos, habríamos recibido 6 bits.

Por el método de disyunción binaria es posible individualizar una eventualidad entre un

número infinito de posibilidades. Para ello bastará proceder con constancia en una serie de bifurcaciones sucesivas elininando progresivamente las alternativas que se presenten[2].

Esta teoría se ha llevado al lenguaje en estrecha relación con los sistemas de oposiciones fonológicas (para emplear un signo siempre realizamos una elección que excluye otros: en mano elegimos la n y rechazamos los demás fonemas). Se considera entonces a la lengua no como transmisora de significados -concepto muy difícil de medir-, sino de información, que se mide en bits. Para la teoría de la información sólo interesará la cantidad de información, los bits, y se desentenderá totalmente del significado de la expresit'pon que analiza. Lo difícil es, en cada caso, conocer el rumbo de alternativas necesarias para llegar a la expresión analizada; es decir: conocer el origen de aquella expresión y las sucesivas bifurcaciones que ha preferido seguir, rechazando otras [9]. En el caso de la lengua, estas bifurcaciones se pueden deducir gracias a las restricciones que impone el código lingüístico de cada comunidad. Un ejemplo fonológico: en español no podemos construir series como las siguientes porque las prohíbe el código: -pcesa, -pbesa, -pdesa, etc.

Detrás de p debemos elegir -nos dicen las reglas fonológicas españolas- o bien una vocal (peso, paso, puso, piso, poso) o bien las consonantes l (plano), r (prado) y quizá s (psique). El número de elecciones posibles no es, pues, infinito en este caso, el código lo reduce a ocho solamente.

Un ejemplo semántico. En la oración *El niño se ha fracturado la . . .*, el término que queda por elegir se escoge entre una serie determinada de palabras que pueden aparecer ahí (pierna, cabeza, naríz . . .). El resto de la oración prohíbe que la selección sea muy amplia, ya que entre los términos existentes no sería posible incluir ciertas palabras (luna, miércoles, Laura). Del mismo modo hay una restricción sintáctica: sólo puede aparecer un sustantivo o sintagma sustantivado.

La teoría de la información tiene su campo de aplicación en la vida comercial y publicitaria. La publicidad busca, por ejemplo, transmitir el máximo de información con el mínimo mensaje, para lo cual busca expresiones adecuadas (con muchos bits), resultantes de una larguísima preselección; este mismo argumento podría aplicarse al genoma de los organismos vivos.

El hecho de que la información sea susceptible de tratamiento cuantitativo (hemos hablado de logaritmos, frecuencias, proporciones, etc.) hace que se le considere a veces como lingüística matemática.

La teoría del lenguaje no es sólo un objeto formal, pues si es aplicada de manera correcta a problemas específicos puede proveer herramientas computacionales y construcciones útiles que dan resultados adecuados. Se utilizará una clase especial de lenguajes llamados *lenguajes factorizables*. A continuación se dará un breve resumen de la teoría del lenguaje en general.

Comenzamos con un alfabeto finito Σ=A,C,G,T y colectamos todas las posibles cadenas

de caracteres de estas letras en un conjunto infinito Σ^*, que incluye la cadena vacía, esto es la cadena que no contiene ninguna letra. Cualquier subconjunto L de Σ^* es llamado un lenguaje sobre el alfabeto Σ. A fin de definir qué clase de lenguaje estamos utilizando debemos dar la regla generadora de L, lo cual puede hacerse de varias maneras, por ejemplo:

1) si el subconjunto L es finito, podemos simplemente enumerar sus elementos.

2) es posible desarrollar algunas reglas de producción y aplicarlas repetitivamente a algunas letras iniciales a fin de generar el lenguaje. Este es con mucho el modo más importante y bien estudiado de definir lenguajes. Si las reglas son aplicadas secuencialmente nos lleva a la gramática generativa de Chomsky. Si son aplicadas en paralelo nos llevan a los sistemas de Lindenmayer.

3) para una clase especial de lenguajes, los lenguajes factorizables, es posible definir un lenguaje mediante indicar su conjunto de palabras prohibidas, este enfoque se sigue usualmente en los análisis de DNA.

Una clase especial de lenguajes factorizables se puede definir en un genoma completo: dado un genoma completo de un organismo G es posible cortar las secuencias del DNA en todas sus posibles subsecuencias y formar un lenguaje L=sub (G), mediante coleccionar estas subsecuencias incluida la cadena vacía. Este lenguaje es factorizable por definición por lo que es posible construir un lenguaje determinista a partir de él.

3.2.2. Distribución de Probabilidad (Análisis Lingüístico de Zipf)

El lenguaje trata sobre palabras y reglas. Mientras que hay una discusión en cuanto a si las reglas son aprendidas o innatas, está claro que las palabras tienen que ser aprendidas. Se define la tasa reproductiva básica de una palabra, R, y se muestra que si R mayor a 1 es requerido por las palabras para ser mantenidas en el lexicón de un lenguaje [90]. Suponiendo que la frecuencia de distribución de las palabras siga la ley de Zipf [21] , un límite superior es obtenido para el número de palabras en un lenguaje que recae exclusivamente en la transmisión oral.

Se ha descubierto que las distribuciones jerárquicas de palabras son bien aproximadas por Leyes de Potencia. Los resultados obtenidos en este tipo de investigación muestran el carácter altamente ordenado de los textos en el ADN.

Todos los sistemas tienden hacia un mínimo esfuerzo para lograr cierto objetivo; esto es verdadero también en el caso de los lenguajes humanos y el ADN. Desde la década de los

[21] Lo que implicaría una relación funcional específica entre la frecuencia de aparición de una palabra y su rango que representa cuan importante es en términos cuantitativos una palabra. Así la palabra más común en un texto tiene rango 1, la siguiente más común rango 2, etc. La relación funcional correspondiente al comportamiento de la ley de Zipf es una ley de potencias tal como la ecuación 3.15

1930's se han realizado investigaciones a partir de idiomas como el Inglés, el Alemán [90], el Español [82], algunas lenguas mesoamericanas, además del lenguaje de cómputo y el ADN [75, 76, 77, 78].

Los lenguajes naturales están caracterizados por estructuras determinadas por las reglas de la gramática. Las palabras enlazadas mediante el uso de estas reglas tienen significado, es decir expresan ideas, sentimientos y emociones de manera que es posible para los receptores del mensaje entenderlo. Las reglas gramáticales dan pues, coherencia y significado a los textos extensos; de esta manera los lenguajes tienen este *orden de largo alcance*. Los *espectros de frecuencia* de aparición de palabras muestran la presencia de periodos largos [22]. Estos son identificados por un comportamiento del tipo $1/f^{\beta}$ [23] en la región de baja frecuencia (tiempos largos) del espectro. Las palabras colocadas al azar tendrían un aspecto muy diferente sin orden de largo alcance.

La secuencia de las letras A C G T en el DNA tiene un espectro de frecuencias 1/fb. Es posible por lo tanto que estas secuencias presenten un orden de largo alcance que posea reglas gramaticales subyacentes. Las opiniones a este respecto continúan divididas, algunos han tomado el punto de vista de que el ADN tiene una estructura similar a la de un lenguaje. En las regiones codificadas, los periodos largos tienen una menor incidencia que en las partes no codificadas. El análisis de Zipf en las regiones diversas del ADN (exones, intrones, ADN separador, etc.) ha mostrado que el exponente a tiene un valor más alto en los segmentos no codificantes y este valor es más cercano al de los lenguajes naturales que en el ADNc.

Graficando el orden de la palabra (X) vs. Frecuencia de ésta (Y), se obtiene una línea de m = -1 aprox. Para un adulto y m = -0.6 aprox. Para niños. Este comportamiento es precisamente la Ley de Zipf ya mencionada. G. Zipf dedujo de esta línea recta una ley del mínimo esfuerzo, o sea un minimizar la cantidad de trabajo para lograr un determinado objetivo. La función de distribución se puede expresar como una gráfica de frecuencia vs. rango que se puede ajustar con una ley de potencias de la forma:

$$f_n = f_1 n^{-\alpha} \tag{3.15}$$

Donde f_1 es la frecuencia de la palabra de rango 1, f_n es la frecuencia de la palabra de rango n y α es un exponente característico. En el análisis original de Zipf se encontró que α tiene un valor aproximado de -1.

En muchas situaciones, algo más que una mínima cantidad de trabajo en el presente, resultará en una mayor ganancia en el futuro (redundancia y optimización). En este caso nos

[22] Un espectro de frecuencia es la representación en el espacio recíproco o de Fourier de una distribución de frecuencias. Tal construcción tiene por objetivo estudiar el comportamiento *dinámico* (periodicidades, persistencias, singularidades) de una distribución de probabilidad.

[23] Que es una ley de potencias estudiada comúnmente con relación a las distribuciones de frecuencia en muy diversas señales.

referimos a una comunicación social, antipleonasmos y que maneja elementos de redacción. Los métodos de la lingüística estadística han sido aplicados recientemente para las secuencias de ADN [64]. La idea es identificar características en las secuencias y correlacionarlas con las funciones biológicas. Los métodos de la lingüística estadística pueden proveernos con algorítmos adecuados [25]. Las secuencias están hechas de las bases nucleótidas A,C,G,T y el arreglo de estas bases sobre la cadena lineal determina el contenido de información que el ADN posee.

3.3. Métodos basados en objetos fractales

3.3.1. La dimensión fractal de Hausdorff

Dimensión topológica

De manera informal podriamos decir que la *dimensión* de un conjunto es la cantidad de información necesaria para especificar de manera precisa a los puntos en éste. Suele llamarse al número de parámetros necesarios para proporcionar esta cantidad de información la *dimensión topológica* de este conjunto.

La dimensión topológica que en lo sucesivo que denominaremos τ es la que nos resulta mas intuitiva y pragmática para comprender. Esta definición establece la dimensión de un punto $\tau_p = 0$, la de una curva $\tau_c = 1$, la de una superficie $\tau_s = 2$ etc., es decir, coincide con nuestra noción previa de dimensión.

Más formalmente escrito, un objeto Δ tiene dimensión topológica m cuando cualquier *recubrimiento* de ese objeto, tiene como minimo una dimensión topológica $\tau_{\Gamma_\Delta} = m + 1$. Donde Γ_Δ es un conjunto conocido como la *cobertura* de Δ. Para ilustrarlo de manera simple un circunferencia (que en lo sucesivo llamaremos T^1) [24] tiene por cobertura a un círculo (o sea la circunferencia con su *relleno*) $\Gamma(T^1)$. El círculo al ser una superficie tiene dimensión topológica $\tau_{\Gamma(T^1)} = \tau_s = 2$ mientras que la circunferencia tiene dimensión topológica $\tau_{T^1} = \tau_c = 1$ pues es una línea curva.

Aún más formalmente: la definición para conjuntos con dimensión topológica 0 queda como sigue: se dice que un conjunto F tiene dimensión topológica 0, $\tau(F) = 0$ si y solo si para todo x perteneciente a F y cualquier conjunto abierto U (para la topología relativa de F) que contenga a x, existe un abierto V tal que x pertenece a V que está incluido en U y la frontera de V con la intersección a F es vacía. Taquigráficamente:

$$\tau(F) = 0 \Leftrightarrow \forall x \in F, \forall U \in \mathcal{A}_F \;\; \exists V \in \mathcal{A}_F \backslash x \in V \subseteq U; \delta V \bigcap F = \emptyset \qquad (3.16)$$

[24] Utilizaremos la notación T^1 para una circunferencia, y en general T^n para el producto cartesiano de n de ellas por la conveniencia matemática de tratar con *toros n-dimensionales*.

La expresión anterior, a pesar de su aparente complejidad nos hace notar una casi obviedad: ¡¡ los puntos no tienen *orilla*!!.

En general describir el tamaño de un subconjunto de $\mathcal{R}^m$ implica naturalmente el uso de *medidas de Lebesgue* [25]. La medida de Lebesgue es, de manera informal una extensión de los conceptos comunes de medición en geometría como longitud, área, para conjuntos más generales, por ejemplo, los que involucran uno o varios conjuntos disjuntos.

Todo conjunto no vacío, acotado y abierto S puede representarse como la suma de un número finito o infinito pero contable de intervalos abiertos disjuntos, cuyos puntos terminales (orillas) no pertenezcan a S [26]. Entonces si $S = \sum_k (a_k, b_k)$ definimos la *medida de Lebesgue para el conjunto abierto* S como $\mu(S) = \sum_k (b_k - a_k)$. O sea, en términos informales: la suma de *longitudes* nos da la *longitud total.*

A pesar de que la medida de Lebesgue es una medida geométrica bastante general, en muchos casos los sistemas *complejos* [26] presentan *singularidades* o puntos con comportamiento *interesante*. Estos puntos al poseer dimensión topológica cero tienen, consecuentemente, medida de Lebesgue nula. O sea que en ciertos casos nuestras nociones usuales de dimensión, longitud, área, y en general cualquier noción de tamaño son inútiles. Debido a este hecho ha sido necesario definir *nuevas maneras de medir*, veamos algunas de estas.

Dimensión de Hausdorff o Hausdorff-Besicovitch

Sea A un conjunto no-vacío con una métrica [27]. Sea $N(r, A)$ el número mínimo de *bolas abiertas* de radio r necesario para cubrir (en el sentido de Γ) al conjunto A. En tal caso definimos la *capacidad* $dim_K(A)$ del conjunto A (en el sentido de Kolmogorov)[28] como:

$$dim_K(A) = \lim_{r \to 0} \sup \frac{\log N(r, A)}{\log(1/r)} \tag{3.17}$$

Aqui sup representa al mayor de estos números conforme tomamos el límite. Nos damos cuenta de que, por ejemplo para el conjunto de Cantor[29] tenemos que:

$$dim_K(A) = \lim_{r \to 0} \frac{\log N(r, A)}{\log(1/r)} = \lim_{n \to \infty} \frac{\log 2^n}{\log 3^n} = \frac{\log 2}{\log 3} < 1 \tag{3.18}$$

Notamos primero que reemplazamos el procedimiento de cobertura (con radios r) por el conteo de *rayitas* y *espacios* (dados por n) y así pudimos calcular una dimensión, o sea ya

[25] Considerando que la representación en series de tiempo que se realizó con las cadenas de ADN es isomorfa a un vector en $\mathcal{R}^m$ este es el caso que nos interesa

[26] Por ejemplo, en nuestro caso, la *dinámica fase* de las series de tiempo en el ADN

[27] que admita *alguna* manera de medir. Formalmente esto se da en el caso más común si existe un *producto interno* de valor real entre los elementos de este conjunto [19].

[28] $dim_K(A)$ es también llamada a veces, *dimensión de Kolmogorov*

[29] ver sección de antecedentes de fractales

podemos tener algunas ideas de *tamaño* o *magnitud* para esta clase de objetos [30]. Adicionalmente notamos que tal dimensión ¡ *no es un número entero* !.

Un calculo adicional involucra la medición de longitudes de una manera más general que la medida de Lebesgue ya mencionada. Para el conjunto A ya mencionado denotemos por σ una cobertura de A por una familia contable de subconjuntos σ_k de diámetro $r_k \leq r$ [31]. Dado un número real $\alpha \geq 0$ definimos:

$$m^{\alpha}(A) = \lim_{r \to 0^+} m_r^{\alpha}(A) \text{ donde } m_r^{\alpha}(A) = \inf_{\sigma} \left\{ \sum_{k=1}^{\infty} (r_k)^{\alpha} \right\} \tag{3.19}$$

Llamamos a $m^{\alpha}(A)$ la *medida de Hausdorff* para el conjunto A. Esta medida es equivalente al tamaño o longitud convencional pero para *fractales*.

Ahora bien, se puede mostrar que existe un valor α_0 del exponente α para el cual se cumple que $m^{\alpha}(A) = +\infty$ si $\alpha > \alpha_0$ y que $m^{\alpha}(A) = 0$ si $\alpha < \alpha_0$. A este valor α_0 lo llamaremos la dimensión de Hausdorff $dim_H(A)$.

Es posible mostrar [16] que existe relación entre la dimensión de Hausdorff y la capacidad de Kolmogorov. De hecho, siempre se cumplen las siguientes propiedades de la dimensión de Hausdorff:

$$dim_H(A) \leq dim_H(B) \text{ si } A \subset B \tag{3.20}$$

$$dim_H(A) \leq dim_K(A) \tag{3.21}$$

Si representamos a la dimensión de Hausdorff de manera similar a la representación de la capacidad de Kolmogorov [17] tenemos:

$$dim_H(A) = \lim_{\alpha \to \alpha_0} \frac{\log N(A, \alpha)}{\log(1/r)} \tag{3.22}$$

Es posible invertir esta ecuación para escribir:

$$dim_H(\log(1/r)) = \log N(A, \alpha) \tag{3.23}$$

Lo que implica que :

$$N(A, \alpha) = r^{dim_H} \tag{3.24}$$

La ecuación 3.24 refleja la estructura de ley de potencia que como vimos está implicada tanto en el fenómeno de autosimilitud como en la correlación estadística de largo alcance.

[30] Ya podemos medir y comparar estos objetos.

[31] Si conocemos un poco de álgebra sabremos que σ es la sigma-álgebra asociada al proceso de cubrir A [26]. Los detalles técnicos respecto al álgebra y topología de las sigma-álgebras no son, sin embargo, necesarios para nuestro argumento actual.

En el caso particular de este el trabajo método que se utilizó para calcular la dimensión fractal o de Hausdorff emplea una cota superior para esta dimensión dada por el procedimienro que Procaccia y Grassberger emplearon en secuencias caóticas [48].

En resumen, los objetos euclidianos diferenciables (las formas geométricas usuales) se ven con una correspondencia en su valor de dimensional topológica, de Kolmogorov y de Hausdorff-Besicovitch.Esto no resulta con los fractales, que son definidos por Benoit Mandelbrot como: *objetos tales que su dimensión de Hausdorff - Besicovitch excede estrictamente su dimensión topológica* [32].

El cálculo para determinar qué tan densa es una región dentro del espacio fase (el lugar geométrico donde se representan los procesos dinámicos (eventos) de tiempo, [33] se hace con la función integral de correlación (ICF por sus siglas en inglés) de Procaccia y Grassberger [48]:

$$C(r) = \frac{1}{N^2} \sum_{i=1}^{N} \Theta \left(r - \| X_i - X_j \| \right) \tag{3.25}$$

donde

r es el tamaño de la ventana por la que apreciamos qué tan cerca o qué tan lejos se encuentran los puntos entre sí. $\frac{1}{N^2}$ es un factor de normalización Θ es la función de Heavyside y X_i, X_j son puntos en el espacio de fases que pertenecen a la serie de tiempo.

Qué tan *densa* es la serie de tiempo en el espacio de fases según lo indica la dimensión de Hausdorff asociada implica qué tan fractal es el conjunto asociado.

Dentro de esta integral de correlación se encuentra una función conocida como *función de Heaviside* o *función escalón unitario* $\Theta(t)$, esta función normalmente se utiliza para presentar variables que se interrumpen en algún instante de tiempo y tiene múltiples usos dependiendo del campo en el que se aplique (en Ingeniería para teoría de conteo y análisis de señales, en Mecánica cuántica para describir cambios de potencial, cambios en niveles electrónicos o simplificar análisis de señal de un experimento. También puede formar funciones de pulsos o tipo puerta y existen muchas otras que se pueden expresar utilizando la suma o la multiplicación de funciones escalón unitario, es también probable que $\Theta(t)$ modele algún tipo de función que varíe en el tiempo, ya sea una expresión matemática, una variable estadística, etc.). Esta función matemática tiene como característica, el tener un valor de 0 para todos los valores negativos de su argumento y de 1 para todos los valores positivos de su argumento.

[32] Se han comprobado algunas excepciones a esta definición de Mandelbrot. El conjunto de Cantor es un ejemplo notable.

[33] Que para el presente trabajo, indica cómo cambian los nucleótidos cuando nos movemos en la serie de *tiempo* -que en este caso señala más bien la posición- que representa al genoma

Como hemos dicho los fractales cubren el espacio de fases de manera densa [16]. Qué tan fractal es un objeto es equivalente a qué tan denso es en su espacio de fases. Por otro lado entre más fractal sea un objeto su dimensión de Haussdorff es mayor[16]. Como ya se vió todo fractal tiene una dimensión de Haussdorff fraccionaria y menor a la del espacio de fases que lo contiene; para eso utilizamos la ICF.

Tomamos un valor de r = 1, por ejemplo
Si tengo muchos puntos Xi y Xj a distancia menor que 1 la integral de correlación tendrá un valor alto y positivo, por lo tanto sumaremos muchos 1's y el valor de C(r) será grande, lo cual indica que el conjunto de puntos es denso y por tanto muy fractal.

En el caso inverso, r = 0:

La distancia entre los puntos Xi y Xj es grande con lo cual se obtiene un valor negativo en el argumento de la función escalón unitario, entonces $\Theta = 0$ *a menudo y C(r) sería una cantidad pequeña lo cual indicaría una serie de tiempo poco densa y por lo tanto poco fractal.*

3.3.2. Series de Tiempo Renormalizadas Originales

Una serie de tiempo es una secuencia de datos cuyo comportamiento se analiza a través del tiempo. En este trabajo, se estudian cadenas genómicas cuya concentración de un nucleótido en especial cambia no según el tiempo, sino la posición en la tira de ADN.

Para construir la serie de tiempo de un nucleótido, se sustituye en el genoma con un 1 cada vez que aparezca la base de interés, de lo contrario se debe reemplazar con 0 el resto de la cadena de ADN. Posteriormente, es deseable llevar a cabo una renormalización de granulamiento grueso (se explicará más detalladamente el término en el Cap.4) para evitar repeticiones de información en el genoma (el ADNnc es particularmente abundante en redundancias).

Al graficar la serie de tiempo (concentración del nucleótido vs. posición en la cadena genómica) se obtienen los dominios de concentración correspondientes al nucleótido deseado. Cada pico que muestre la gráfica simboliza la presencia de una concentración alta de la base y la amplitud de éste, por cuántas y cuáles posiciones se conserva.

3.3.3. Conjuntos Fractales

Un fractal es un objeto geométrico de dimensión fraccionaria inmerso en un espacio de fases. En el caso específico de el trabajo que nos ocupa es posible determinar la densidad de puntos (nucleótidos) dentro de él y así darnos una idea de la concentración y proporción de

cierta base en el ADN, así como de la dinámica e información del genoma a analizar a través de la determinación de la dimensión de Hausdorff.

Los fractales obedecen a una fórmula del tipo Ley de Potencias del tipo $N = S^d$ donde al exponente $d = \frac{\ln N}{\ln S}$ [34] se la llama la *dimensión Fractal de Haussdorff* o dimensión de Hausdorff-Besicovitch.

Es frecuente encontrar fractales en la naturaleza, un ejemplo serían los copos de nieve o los perfiles de las costas. La principal característica de un fractal es que es autosimilar, esto quiere decir que tiene la misma forma incluso al cambiar de escala (de hecho, tiene un perímetro infinito). Un fractal puede ser simple o complejo, dependiendo de la forma que presente; esto es perceptible en su dimensión de Haussdorff, entre más alta sea, más denso o complejo es el fractal. Prácticamente, sólo se pueden obtener cuasifractales porque un fractal verdadero tiene un número de puntos infinito.

Al generar cuasifractales para los genomas aquí manejados es posible tener una idea de su complejidad y para eso se utiliza la Integral de Correlación de Procaccia y Grassberger recientemente mencionada.

3.4. Otros

3.4.1. Proporciones de Bases Complementarias

Se ha comprobado de manera experimental la actividad bioquímica entre los nucleótidos de una cadena de ADN. Sin dicha interacción sería imposible para éstos establecer enlaces químicos, puentes de Hidrógeno o participar en funciones tan importantes como la replicación, transcripción y traducción.

Analizando las proporciones de bases complementarias A-Ts y C-Gs (cada pareja integrada por una purina y una pirimidina respectivamente según la regla de Chargaff [5, 70]) en la cadena genómica será posible observar si existe cierta tendencia hacia alguna de las dos parejas que indique actividad bioquímica y por lo tanto, algún tipo de función biológica. Si se encuentran diferencias significativas, quizá esto evidencie un tipo de firma genética, por el contrario, si llega a existir un intervalo de valores constante, entonces probablemente puedan aplicarse estas proporciones a los seres vivos en general. El hecho de descubrir una proporción homogénea entre los nucleótidos mostraría un comportamiento azaroso y sin actividad bioquímica; dicho fenómeno sólo sería comprensible en el Generador Aleatorio, con lo cual se espera poder hacer una comparación fidedigna entre un comportamiento aleatorio y uno que no lo es.

[34] obtenido tras tomar logaritmos a ambos lados y despejar

3.5. Descripción individual de procedimientos

3.5.1. Entropía Informacional de Shannon

Con la finalidad de analizar el grado de complejidad en el texto (biológico, lingüístico o artificial), se procedió a sumar las frecuencias de aparición de cada palabra (dímeros, trímeros, tetrámeros, pentámeros, RDFs y palabras del libro de Isaías); posteriormente, se dividió cada frecuencia de aparición entre la sumatoria de todas ellas.

Se calcularon después los logaritmos naturales correspondientes a los resultados anteriores y fueron multiplicados por éstos mismos. Al final se sumaron los valores resultantes, la cifra negativa obtenida es la Entropía Informacional de Shannon.

Finalmente, se repitió el procedimiento para dímeros, trímeros, tetrámeros, pentámeros, la suma de todos los anteriores (global) y RDFs de cada organismo vivo y del Generador Aleatorio. En el caso del libro de Isaías, por obvias razones (no hay un parámetro establecido para obtener los cúmulos anteriores comparables a los del ADN en un texto escrito) solo se realizó una gráfica global.

3.5.2. Procesos de Markov

Según los conocimientos teóricos enunciados en el Cap.1 y 3 se procedió a analizar las gráficas de Distribución de Probabilidad (Análisis Lingüístico de Zipf) buscando principalmente un comportamiento del tipo Ley de Potencia que evidencie una memoria de largo plazo (alcance) y una gran correlación entre nucleótidos o una Markoviana que obedezca a un decaimiento exponencial y a independencia estadística.

3.5.3. Distribución de Probabilidad (Análisis Lingüístico de Zipf)

Con la finalidad de demostrar que en ningún caso el ADNnc se encuentra dispuesto aleatoriamente, se eligieron los siguientes fragmentos de genoma. Para los fines del presente trabajo fueron tomados como textos genéticos los siguientes segmentos *genómicos*:

1. *Mycoplasma pneumoniae* (en este caso, los datos obtenidos fueron del genoma completo)
2. *Drosophila melanogaster*
3. *Felis catus*
4. *Pinus thunbergii*
5. Libro de Isaías (utilizado como modelo)
6. Generador aleatorio (utilizado a manera de control)

Considerando que al analizar los genomas de una bacteria, un invertebrado, un vertebrado y una planta es posible abarcar a rasgos generales una amplia gama de formas de vida presentes en la naturaleza.

Así, se procedió con ayuda de un programa en el lenguaje C++ a realizar todas las combinaciones posibles de las cuatro bases nitrogenadas existentes en el ADN (A,T,C,G, las letras de nuestro alfabeto genético) para dímeros (16), trímeros (64), tetrámeros (256) y pentámeros (1024), así como RDFs (de tamaño 2 a 8). Obviamente, se podría trabajar con hexámeros, heptámeros, etc. pero se ha considerado que con los cálculos anteriores es posible llegar a una conclusión fidedigna.

Se realizó el conteo total de palabras genéticas, ya obtenidas las combinaciones para dímeros, trímeros, tetrámeros, pentámeros y RDFs; se utilizó un procedimiento computacional para determinar sus frecuencias de aparición.

Se les asignó rangos a las palabras, de forma tal que la palabra más frecuente tenía rango 1, la siguiente más frecuente rango 2, etc.

Zipf mostró que para los lenguajes naturales (especialmente los anglo-sajones que él estudió) la gráfica de rango contra frecuencia se puede ajustar con una ley de potencias de la forma $f_n = f_1 * n^{-\alpha}$

Para utilizar este método partimos de la fórmula para obtener probabilidades propuesta por Laplace (llamada también definición frecuentista de probabilidad): $P(A) = f(A)/P(T)$ Donde P(A) es la probabilidad de encontrar, por ejemplo, una Adenina en el genoma, f(A) es la frecuencia de aparición de Adeninas en el genoma P(T) es el número total de palabras genéticas.

Por ejemplo: cuando se dice que el tetrámero ACTT aparece 213 veces en determinado organismo, es 213 veces entre el número total de posibilidades existentes.

Si al graficar nuestros datos obtenemos una pendiente que muestre una distribución con largas colas, podemos observar una relación de largo alcance o que obedece a una ley de potencias. Al contrario, si la gráfica es exponencial, estamos ante un caso de orden de corto alcance o Markoviano.

3.5.4. Determinación de complejidad por dimensión fractal de Hausdorff

Se procedió a construir las series de tiempo para cada nucleótido. Ejemplificando con el caso de la adenina: se sustituye en el genoma cada A con un 1 y el resto de los nucleótidos con un 0, para hacer las series de tiempo; con los demás nucleótidos se emplea el mismo algoritmo.

Se obtuvo la primera derivada según la serie de derivadas fractales propuesto por Man-

delbrot [39] y se continuó de la misma manera hasta llegar a la derivada numero 10 [35]. Posteriormente, se construyó con tales derivadas la integral de correlación propuesta por Procaccia y Grassberger de la siguiente manera:

Se obtiene la distancia al cuadrado de la Serie de Tiempo Renormalizada Original y sus diez derivadas mediante la fórmula polinómica para conocer la *distancia* entre cada par de puntos.

$$dist = \sqrt{\left(\sum_{i \neq j}^{m} (x_i - x_j)^2\right)} \tag{3.26}$$

Utilizamos la norma euclidiana, considerando al espacio que contiene a las series de tiempo y sus derivadas isomorfo a $\Re^m$; de esta última columna de valores (las distancias euclideas entre puntos) se encontró el valor mayor: ese fue el radio máximo elegido (esto nos indica de qué *tamaño* es el fractal), después, continuaron proponiéndose radios con intervalos decrecientes constantes, hasta llegar a cero, que fue el radio mínimo.

A continuación se obtuvo la densidad de puntos (en este caso nucleótidos) para cada radio, restando los valores correspondientes de distancia a dichos radios, seguidamente, se hace un conteo: si la FIC es un número positivo, se suma un 1 si no, se suma un cero. Esta función es la ya conocida función de Heavyside. Finalmente, se sumó de manera individual cada *columna Heavyside*: ese valor es $C(r)$.

Se obtuvieron los logaritmos naturales correspondientes a cada r y $C(r)$ y se graficaron en dispersión X, Y tipo lineal (recordemos que la pendiente de una curva log-log es el exponente asociado a la ley de potencia).

Se repitió el procedimiento para cada ejemplo experimental y para cada uno de los cuatro nucleótidos.

3.5.5. Series de Tiempo Renormalizadas Originales

Dado que el genoma es redundante (ver Modelo de Entropía Informacional de Shannon) y hasta el más pequeño muy largo, se procedió a renormalizarlo; esto significa que se sustituyó con un 1 cada vez que aparecía en la cadena genómica el nucleótido deseado y al resto de ellos con ceros, después se tomaron tríos de números que fueron a su vez sustituidos con ceros o unos dependiendo de si su suma era mayor o menor a 1. Se repitió el procedimiento hasta lograr invarianza para cada ejemplo experimental (excepto libro de Isaías).

[35] acotamos en 10 el número de derivadas debido a razones de limitación computacional, de cualquier manera se sabe que el esquema de renormalización genera convergencia en el sentido de Cauchy, por lo que los primeros términos diferenciales son significativamente más importantes en la mayoría de los casos

Posteriormente se tomó la serie de tiempo renormalizada de cada nucleótido y se construyó una gráfica lineal que maneja concentraciones de la base deseada vs. t, es decir, posición dentro de la cadena genómica. Se repitió el procedimiento para los cuatro organismos vivos más el Generador Aleatorio.

3.5.6. Fractales de tipo Anillos y Radial

Se tomaron los valores de la Serie de Tiempo Renormalizada Original más los del conjunto de sus diez derivadas y se construyeron gráficas del tipo Anillos y Radial.

3.5.7. Proporciones de Bases Complementarias

Después de contar el número de cada uno de los cuatro nucleótidos en los genomas, se sumó el número de A-Ts y C-Gs; luego, por medio de una regla de tres se calculó el porcentaje correspondiente; para finalizar, se dividió el número de A-Ts entre el de C-Gs y viceversa para obtener sus tasas.

Capítulo 4

Resultados generales, discusión y conclusiones

A continuación se analizarán los resultados obtenidos en los procedimientos cuantitativos de análisis descritos en el capítulo 3. Primeramente, se enumeran de manera sintética estos resultados en las tablas 1 y 2. Estas tablas contienen cantidades utilizadas como referentes cuantitativos en la descripción supracitada de secuencias genómicas. La tabla 1 incluye los valores de los exponentes característicos de Zipf para las distribuciones de frecuencia de palabras constituidas por nucleótidos y combinaciones de hasta tamaño cinco de éstos.

Adicionalmente se incluyen d e coeficientes de correlación para tales ajustes de ley de potencias con la finalidad de establecer cotas de error. Además, se incluyen los valores de información de Shannon asociados a versiones renormalizadas de las series de tiempo pertenecientes a cada nucleótido en cada genoma. A estas mismas series de tiempo, se les asoció un conjunto cuasifractal representativo con el objetivo de discernir de manera más específica las características de fractalidad asociadas con la complejidad de la estructura informática en estas secuencias. A las series renormalizadas se les determinó una cantidad, la dimensión fractal o de Hausdorff, que especifica de manera cuantitativa tal nivel de complejidad, lo cual evidenciaría una estructuración específica y por tanto, probablemente una funcionalidad en la codificación de información subyacente a estos segmentos genómicos.

En todos los casos anteriores se incluyen gráficos semicuantitativos para expresar las relaciones matemáticas concurrentes. Finalmente, se discuten tales resultados en el marco de los paradigmas actuales de la Biología teórica, se esbozan conclusiones apropiadas bajo el mismo marco conceptual y se mencionan algunas perspectivas de desarrollo futuro.

4.1. Tabla 1

TABLA 1

	Mycoplasma pneumoniae	*Drosophila melanogaster*	*Felis Catus*	*Pinus thunbergii*	Texto de Isaías	Generador aleatorio
Zipf dímeros	R^2= 0.8061 y= 72142X^(-0.3156)	R^2=0.8323 y= 14426X^(-0.4171)	R^2=03865 y=12391X^(-0.4725)	R^2= 0.8234 y=13202X^(-0.4156)		R^2= 0.4456 y=6789.6X^(-0.0849)
Zipf trímeros	R^2= 0.8799 y= 39307X^(-0.4406)	R^2= 0.8729 y=7049.2X^(-0.4609)	R^2=0.4428 y=6813.3X^(-0.5861)	R^2= 07632 y= 5794.5X^(-0.4662)		R^2=0.4507 y= 1715.3X^(-0.0533)
Zipf tetrámeros	R^2= 0.8609 y= 25600X^(-0.5325)	R^2= 0.9000 y=4184.1X^(-0.5189)	R^2= 0.5044 y= 5183.7X^(-0.7024)	R2=0.7684 y=4139.9X-0.5285		R^2= 0.6296 y= 467.18x^(-0.0589)
Zipf pentámeros	R^2=0.8244 y= 20812X^(-0.6212)	R^2= 0.8953 y=2908.9X^(-0.5775)		R^2=0.7791 y=2028.9X^(-0.6044)		R^2= -0.7624 y= 157.15X^(-0.0995)
Zipf global	R^2=0.9519 y=724685X^(-1.0893)	R^2= 0.9708 y= 99613X^(-1.0468)	R^2= 0.793 y= 88740X^(-1.1572)	R2=0.9482 y=178273X-1.2068	R^2= 0.9685 y= 5028.8X^(-1.0328)	R^2= 0.9054 y= 25561X^(-0.8418)
Zipf RDFs	R^2=0.7703 y=221172X^(-2.7899)	R^2=0.8537 y=24396X^(-2.2736)	R^2=0.8682 y=16545X^(-1.9198)	R^2=0.8602 y=17555X^(-2.1617)		R^2=0.716 y=9578.6X^(-2.0819)
Shannon dímeros	-2.73855944	-2.71470778	-2.69845225	-2.71557437		-2.7682107
Shannon trímeros	-4.06800757	-4.058805744	-4.0256709	-4.0601716		-4.15659888
Shannon tetrámeros	-5.3902669	-5.38882369	-5.3416738	-5.3980608		-5.54280927
Shannon pentámeros	-6.71078407	-6.71456862	-6.64790101	-6.72826455		-6.92533465
Shannon global	-6.12614279	-6.143244837	-5.87160893	-5.9747581	-6.202783499	-6.272359234
Shannon RDFs	-2.7085	-2.8363	-3.2475	-2.8719		-2.9937
Contenido A/T	59.25%	63.55%	61.82%	61.50%		50.23%
Contenido G/C	40.75%	36.45%	38.18%	38.50%		49.77%

Figura 4.1: Resultados

4.2. Tabla 2

TABLA 2

	Mycoplasma pneumoniae	*Drosophila melanogaster*	*Felis catus*	*Pinus thunbergii*	Generador Aleatorio
Hausdorff A	d=0.45 R^2=0.93	d=0.37 R^2=0.90	d=0.37 R^2=92	d=0.35 R^2=0.91	d=0.17 R^2=0.91
Hausdorff C	d=0.04 R^2=0.89	d=0.04 R^2=0.87	d=0.07 R^2=0.89	d=0.03 R^2=0.85	d=0.17 R^2=0.90
Hausdorff T	d=0.24 R^2=0.91	d=0.43 R^2=0.92	d=0.41 R^2=0.92	d=0.37 R^2=0.91	d=0.19 R^2=0.91
Hausdorff G	d=0.04 R^2=0.89	d=0.05 R^2=0.89	d=0.06 R^2=0.88	d=0.04 R^2=0.88	d=0.16 R^2=0.90

Figura 4.2: Resultados adicionales

4.3. Discusión de Resultados Generales

4.3.1. Entropía Informacional de Shannon

Los resultados que se obtienen con la formula de la ecuación 3.1 puede sintetizarse a cuánta información posee un texto (ya sea lingüístico o biológico).

En las gráficas obtenidas es posible notar el significativo contraste existente entre el Generador Aleatorio y el resto de los elementos experimentales; ya que en éstos últimos se observa, como podía esperarse, que comienzan con una entropía baja, o dicho de otra manera: un alto contenido de información que van perdiendo poco a poco a lo largo de la cadena, en el texto la información se pierde del transmisor a la transmisión, luego al receptor, y así sucesivamente. Es perceptible en ellos una curva continua que muestra con claridad la estrecha interrelación entre los nucleótidos de la cadena de ADN o de las palabras en el texto lingüístico. Caso opuesto es el del Generador, el cual expone cuatro aglutinamientos sin continuidad alguna ocasionados por redundancias. Dado que no hay, como en el resto de los ejemplos, una regla gramatical ni genética que lo impida, este fenómeno se presenta por efecto estadístico, visto de esta forma, el sistema no es por completo carente de información, sin embargo ésta

no es útil para propósito alguno, como en el caso de una mantra.

Por otro lado, se aprecia que los aglutinamientos son periódicos en la secuencia de tamaño que poseen, esto es n^0, n^1, n^2yn^3, lo cual solo evidencia la presencia de dímeros, trímeros, tetrámeros y pentámeros respectivamente, dado que es un conteo global de la suma de las entropías para dichos cúmulos. No se nota aquí ninguna interrelación entre las bases generadas por computadora, de hecho, su independencia estadística es tan grande que es posible considerar que existe en éstas un comportamiento errático. De esta manera, puede decirse que se trata de un sistema dinámico disipativo (aquél que disipa energía y produce entropía).

Es importante hacer hincapié en que a pesar de que, aparentemente, los valores del Generador Aleatorio y del libro de Isaías son muy similares (la variación es de 0.07 unidades de información), estas gráficas fueron hechas en escala logarítmica, por lo que, por poco que varíen, en términos globales implica una gran diferencia en los contenidos de información.

Dado que la fórmula para obtener la Entropía Informacional de Shannon es un logaritmo del tipo P log P (cf. ecuación 3.1, aplicamos el inverso de la operación, esto es un exponencial. Lo cual indica que la información que contiene el libro de Isaías es quince veces mayor que en el Generador Aleatorio.

En la Tabla 1 pueden observarse los valores para dímeros, trímeros, tetrámeros y pentámeros. Es interesante notar que, los valores que presenta el ADNc (dentro del conjunto de los trímeros se encuentran los codones) es mayor en todos los casos al de RDFs (ADNnc), mostrando así un menor contenido de información y consecuentemente un mayor desorden estadístico.

4.3.2. Distribución de Probabilidades (Análisis lingüístico) de Zipf

Este método de estudio denota la coherencia y estructuración de un escrito, es decir, cuando un conjunto de reglas -genéticas, gramaticales o semánticas- hacen que las palabras o los cúmulos de nucleótidos de un texto, lingüístico o biológico respectivamente, tengan un significado lógico por sí mismos y en interacción con otros elementos textuales. También muestra visualmente que los datos corresponden una Ley de Potencia que, como se explicó anteriormente (, es el tipo de función que implica una corelación de largo alcance entre los nucleótidos o palabras y por lo tanto, una memoria de largo plazo en la cadena (ya sea de bases o de letras). Se ampliará más esta noción al realizar el análisis de Markov.

Los resultados señalan la escasa coherencia del Generador, se trata de un texto con letras intercaladas al azar sin estructura, significado ni lógica alguna como puede suponerse en un texto producido artificialmente para un propósito como el de este trabajo. Por otro lado, son evidentes en la gráfica global los mismos cuatro *escalones* con aumento de tamaño periódico correspondientes a cúmulos de dos, tres, cuatro y cinco *nucleótidos* que ya se habían observado en su gráfica de Entropía Informacional de Shannon, evidenciando de nuevo su

independencia estadística e inexistente interrelación.

Aunado a que en la mayoría de los casos, el coeficiente de correlación R2 es más bajo, lo que implica el muy bajo carácter tipo Ley de Potencia su distribución de frecuencias), para absolutamente todos los cálculos (dímeros, trímeros, tetrámeros, pentámeros, global y RDFs) el Generador Aleatorio muestra los exponentes de Zipf menos negativos, dicho comportamiento es más evidente en las gráficas global y RDFs, donde en ningún modo se acerca a los valores -1 en global (GA= -0.84 Hay que recordar que, dado que este método de lingüística estadística obedece a una ley de potencia, un cambio pequeño en el exponente significa un cambio importante en el carácter coherente del texto) y -2 en RDFs (GA= -0.16), que el libro de Isaías y los organismos vivos sobrepasaron.

En contraste, todos los organismos generan exponentes cada vez más negativos conforme se aumenta el tamaño de los cúmulos; esto es razonable dada la mayor cantidad de *palabras* con significado biológico o lingüístico. Aunque existe también una variación de los exponentes de Zipf en el Generador Aleatorio, dicho cambio no es progresivo, sino confuso (GA= -0.08, -0.05, -0.05, -0.09, -0.84 y -0.16 respectivamente). Los valores máximos se encuentran en los resultados globales y de RDFs (alrededor de -1 y -2 en cada caso) resaltando de esta manera la coherencia existente en el ADNnc, que como sabemos, comprende aproximadamente el 95 % del genoma.

Es importante subrayar que los exponentes de Zipf para las RDFs fueron significativamente más grandes en todos los organismos, salvo en el Generador Aleatorio; esto resalta una poderosa estructura coherente. Se sabe por autores como Stanley [75, 76, 77, 78] y Dokholyan [41, 42, 43, 99] que la presencia de RDFs es dramáticamente superior en el ADNnc en comparación con el ADNc. En este caso fue posible comprobar que esa superestructura gramatical efectivamente presenta los valores máximos de coherencia (alrededor de -2); así como su presencia a gran escala en el ADNnc de los cuatro organismos vivos analizados, no así en ADNc (para trímeros el exponente de Zipf más alto se obtuvo en el caso de *Felis catus*, Fc= -0.5) ni en Generador Aleatorio.

4.3.3. Porcesos de Markov

En el ADNc las unidades de tamaño tres son muchas y existen numerosas posiciones posibles que pueden ocupar estos codones a lo largo de la cadena genómica (en genomas diferentes), de tal suerte que puede pensarse en ellas como moléculas distribuidas de manera aparentemente aleatoria e inestable porque, finalmente, cada unidad es estadísticamente independiente y sólo se afecta a sí misma, de ahí la memoria Markoviana que presenta el ADNc (como evidencia el carácter exponencial de su decaimiento en la correlación [75]), mientras que en el ADNnc hay regiones de hasta cientos de bases de longitud que están relacionadas estadísticamente, debido a una memoria no Markoviana o de largo alcance (como evidencia su comportamiento, que obedece a una ley de Potencia).

Así, grandes unidades constitutivas no pueden estar dispuestas de manera azarosa, sino que deben tomar el papel correspondiente dado por esta memoria de largo alcance. En el ADNnc como hay menos maneras de ordenar los necleótidos, el carácter aleatorio de este ordenamiento disminuye y podemos hablar de dependencia estadística y de estabilidad molecular.

Cadena de ADNc

ATA**GCG**GATCGG**GAG**TTC**CAT**GCCAATGC**CAG**GTTAGACTCAGG
Trímeros que funcionan como codones

A causa de su tamaño pequeño puede cambiar fácilmente de posición en el genomas (molécula inestable) sin afectar al resto del las bases (independiente estadísticamente). El desplazamiento de los codones puede ocurrir de tantas maneras en la cadena que tiene un comportamiento aparentemente aleatorio.

Cadena de ADNnc

ACTCG**ATTCAGCAGTA**AGGACATATATATATATATATATATAT
RDF de tamaño = 11

A causa de su enorme tamaño (algunos miden hasta cientos de bases) no puede cambiar fácilmente de posición en el genoma (molécula estable) sin afectar al resto de las bases (dependiente estadísticamente). El desplazamiento de la RDF puede ocurrir de poquísimas maneras y cuando ocurre es sumamente dirigido, lo cual implica orden en sentido estadístico.

Como hemos visto en las gráficas de Zipf, la presencia de un ajuste del tipo Ley de Potencia implica colas de largo alcance en la distribución de frecuencias y por lo tanto persistencia en la distribución de probabilidades. Este efecto de persistencia es visualizado como memoria de largo alcance correspondiente a un proceso no Markoviano.

Estudios realizados en fracciones exclusivamente de ADNc por autores como Stanley [75, 76, 77, 78] muestran que el decaimiento de la correlación en éstos es del tipo exponencial simple, lo que implica persistencia de corto plazo y memoria desvaneciente asociado a un proceso de Markov. Resulta significativo notar este efecto en los bajos coeficientes de correlación que para la Ley de Potencia tiene el Generador Aleatorio.

4.3.4. Dimensión Fractal de Haussdorff

Los valores d son indicadores de la magnitud de complejidad en un objeto matemático como puede ser un cuasifractal (estrictamente hablando, un fractal posee un número infinito

de puntos), entre mas grande sea *d*, la complejidad de éste es alta y viceversa. Por otro lado, cuando d no pertenece al conjunto de números enteros, sino al de las fracciones, se dice que el cuasifractal no es trivial. De este modo, si una Ley de Potencia tiene asociada aparte una dimensión fractal, la información que alberga es aún mayor, es decir, coherente, de largo alcance y compleja. En este caso lo que se busca conocer es la cantidad máxima de información que existe en un genoma relativa a un nucleótido (ver gráficas de Dimensión Fractal de Hausdorff y gráficas de Series de Tiempo Renormalizadas Originales) y qué tan compleja es (ver gráficas de Series de Tiempo Renormalizadas Originales, fractales y Tabla 2).

Se obtuvieron exclusivamente cifras fraccionarias, lo cual implica la cuasifractalidad (y consecuentemente un aumento en la cantidad de información) del conjunto de datos de los cinco sujetos experimentales. Al observar los resultados para la dimensión fractal en la Tabla 2, es patente que en todos los genomas hay 2 valores altos y dos bajos, coincidiendo los primeros para la proporción de Adeninas y Timinas (m= 0.7-0.8 para organismos vivos) y los segundos para la de Citosinas y Guaninas (m= 0.07-0.13 para organismos vivos). La periodicidad en todos es la misma porque en todos tenemos cuatro bases.

Sin embargo, en Generador Aleatorio se aprecian diferencias en cuanto al promedio de las sumas de A-Ts y C-Gs (0.36 y 0.33 respectivamente), esto es porque su proporción de concentraciones nucleotídicas difiere significativamente del resto de los casos (alrededor de 25 % para cualquiera de los cuatro, es significativo que todas las bases del Generador tuvieron la misma d, lo cual inidca que a pesar de existir complejidad, ésta es redundante en los 4 nucleótidos). También la R2, que es proporcional a la probabilidad, es prácticamente la misma para cada una de las bases, es decir, están equiprobablemente distribuidas en el genoma de manera homogénea. Lo anterior significa que, aunque el Generador posee *información* con cierto nivel de complejidad (en el caso de la proporción de G-Cs, supera la de todos los demás), no está bien estructurada pues no se observa ni la complementariedad en las bases ni una distribución de valores *d* que indiquen algún tipo de dinámica en el espacio de fases de la serie de tiempo. Su concentración de nucleótidos es la misma y por lo tanto, la probabilidad de encontrar alguna de ellos en el genoma, igual. Esto sería comparable a recorrer un largo y tortuoso camino para llegar al punto de partida, un tipo de complejidad redundante y azarosa.

En el resto de los organismos, por el contrario se aprecia una complejidad dirigida y persistente (como en el caso de un Caminante aleatorio dirigido) ; la probabilidad de encontrar cierto nucleótido no puede ser la misma en todos los casos, porque las concentraciones nucleotídicas difieren dependiendo de la sección genómica en la cadena de ADN y de la base que se trate (nótense las RDFs en ciertas partes del genoma). Se observa así en el ADN de cada ser vivo analizado, una misma estructura cuasifractal compuesta por dos conjuntos complementarios (de dos nucleótidos cada uno) que parecen entremezclarse: uno simple (G-Cs) y uno complejo (A-Ts).

Por otro lado, las gráficas de las Series de Tiempo Renormalizadas Originales muestran

los dominios de concentración de un nucleótido en particular donde un pico en un intervalo de posición X_o a X_n en la cadena de ADN representa abundancia de dicha base en esa región. Dado que la serie de tiempo se encuentra renormalizada, la concentración máxima que es posible alcanzar es de uno.

Para los cuatro organismos y el Generador Aleatorio se encontraron singuletes (picos sencillos) de tamaño dos (dos nucleótidos a lo largo del genoma). En los primeros casos, al menos en alguna de las cuatro gráficas que se hicieron para cada ser vivo se presentan dobletes (picos dobles y coalescientes), multipletes (picos truncados de tamaño variable) o duplicación de período entre singuletes; no así el Generador cuyo comportamiento más sobresaliente se encuentra en su gráfica para timinas, la cual exhibe cinco singuletes (número igualado por *Mycoplasma pneumoniae* en su gráfica para citosinas y superado por *Drosophila melanogaster* en su gráfica para citosinas) y con la que confirma su escasa complejidad.

Los fractales tipo anillos son una forma diferente de ver los dominios de concentración de un solo nucleótido. Como si se enrollaran individualmente las gráficas de las Series de Tiempo Renormalizadas Originales y se vieran de manera transversal, algunas veces los picos se sobrelaparían formando vetas, otras no, de modo que no sería tan evidente la presencia de dominios de concentración, no obstante, formarían zonas densas que permitirían reconocerlos, como es posible apreciar en cada ejemplo experimental.

En cuanto a los fractales de tipo radial es más perceptible la densidad en el espacio de fases, así como la estructura fractal de los genomas debido a las cuasiperiodicidades que se notan mientras se genera el fractal por computadora.

Es importante resaltar que aunque todos los cuasifractales son aparentemente muy parecidos entre sí (incluyendo los correspondientes a Generador Aleatorio), no todos exhiben las mismas dimensiones fractales de Hausdorff ni el mismo tipo de gráfica para Series de Tiempo Renormalizadas Originales, es sólo uniendo los tres tipos de evidencia que puede llegarse a una conclusión confiable en cuanto a cuáles genomas muestran los comportamientos más complejos y cuáles no.

4.3.5. Proporciones de Bases Complementarias

Chargaff, descubrió que la concentración de adeninas es, aproximadamente, la misma que la de timinas y que lo mismo ocurre con las citosinas y las guaninas [5, 70]. Una consecuencia directa de esta observación es que las tasas (G+A) / (C+T) y (G+T) / (A+C) tienen un valor cercano a uno. Se procedieron a hacer estas mismas operaciones para los organismos vivos y el Generador Aleatorio; los resultados para todos fueron los obtenidos por el investigador austriaco.

Conjunto Fractal Asociado para *Mycoplasma pneumoniae* Serie A Versión Radial

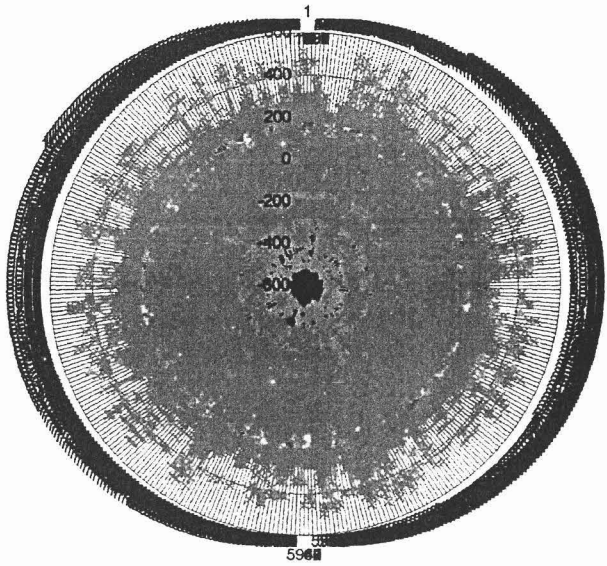


Figura 4.3: Conjunto fractal radial para *Mycoplasma pneumoniae* serie A

No obstante, al sumar la cantidad de adeninas y timinas y posteriormente la de citosinas y guaninas (bases complementarias en ambos casos) se observa una tendencia positiva en los porcentajes de A-Ts (59 a 63 %) y otra consecuentemente negativa en C-Gs (37-41 %). A causa de esta desigualdad compartida por los organismos vivos, al dividir el número de CGs entre el de ATs se obtiene un valor casi constante (0.61-0.68). Al dividir de manera inversa, es posible encontrar también valores muy semejantes. Se hace evidente que la actividad química entre los nucleótidos de la cadena genómica no es uniforme, lo cual es absolutamente indispensable para dar lugar a interacciones moleculares (puentes de Hidrógeno, fuerzas de Van der Waals, enlace químico, etc).

Como puede suponerse, en el caso del Generador Aleatorio se obtienen proporciones muy diferentes (51 % y 49 % en la proporción de A-Ts y C-Gs respectivamente y un valor aproximado de 1 para ambas divisiones), lo cual corrobora que sus *nucleótidos* están distribuidos de forma indistinta en el genoma, su actividad química es uniforme, podría incluso decirse que inerte, lo cual no permitiría establecer ningún tipo de actividad molecular entre sus *bases* en caso de que fuesen verdaderas.

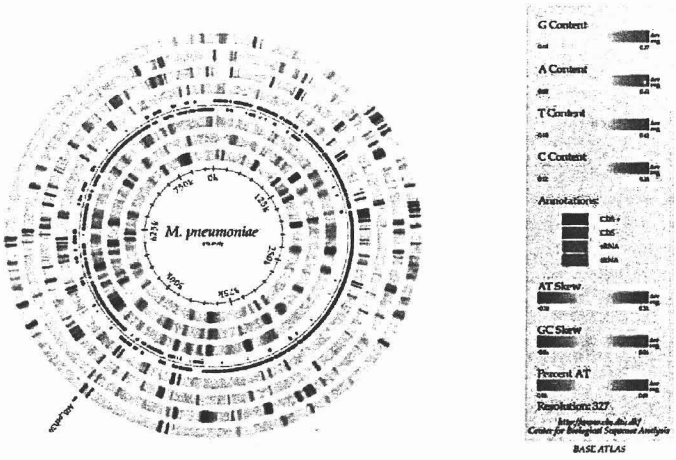


Figura 4.4: Gráfica de isocoras para *Mycoplasma pneumoniae*,

4.4. Conclusión

A lo largo del desarrollo de esta tesis se ha podido comprobar que el mal llamado ADN *no codificante*, que como hemos mencionado anteriormente constituye el 95 % del genoma de los seres vivos, no solamente no carece de información ni está dispuesto de manera aleatoria, sino que a través de los análisis realizados, fue posible constatar que:

1. Un análisis de la Teoría de la Información muestra una entropía menor que la del resto de los casos considerados (incluidas las secuencias de tamaño tres como las que constituyen al ADNc), puesto que entropía e información son cantidades complementarias, esto indica una mayor cantidad de información. El caso más sobresaliente son las Repeticiones Diméricas en Fila (RDFs).

2. No está dispuesto de manera aleatoria como lo demuestra la forma de su distribución de probabilidad (una Ley de Potencia). Lo anterior puede ser visto como consecuencia de una regla de carácter gramatical, semántico, o genético de manera similar a la propuesta por Zipf en sus estudios sobre el lenguaje humano. Es de suponerse una conclusión similar para los textos biológicos: coherencia y estructuración.

3. Posee un enorme grado de correlación estadística y por lo tanto, de coherencia interna, estabilidad ante mutaciones y estabilidad química debido a su memoria de largo alcance de carácter no Markoviano.

4. El análisis de los conjuntos fractales asociados nos muestra un enorme grado de complejidad en sentido estadístico.

Conjunto Fractal tipo anillo asociado al nucleótido A
en *Mycoplasma pneumoniae* (isocoras A)

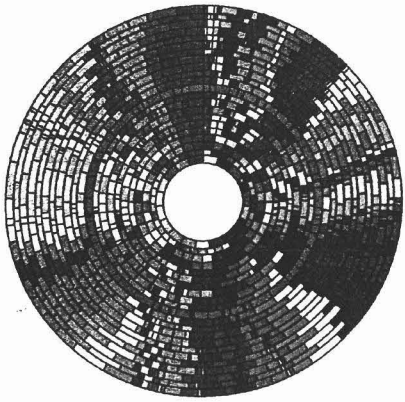


Figura 4.5: Conjunto fractal de anillos de isocoras para *Mycoplasma pneumoniae* serie A

5. Por otro lado, los conteos de nucleótidos, nos muestran la presencia de reglas específicas en su distribución similares a las propuestas por Chargaff [5, 70] en el contexto de cadenas complementarias. Esto resalta un principio subyacente en el texto genómico que indica que la proporción de nucleótidos posee un significado relacionado con su función .

Todo esto lleva a concluir que el ADN llamado *no codificante* contiene dentro de sí una gran cantidad de información dispuesta de manera coherente, estructurada y compleja cuyo carácter lingüístico (comunicativo) hace pensar en alguna clase de función biológica. Queda desde luego mucha investigación por hacer para dilucidar los complejos mecanismos bioquímicos y genéticos que se encuentran presentes en la información que nuestras pruebas estadísticas han revelado.

Se espera que este trabajo motive a biólogos moleculares y especialistas en genómica o ciencias del lenguaje a ahondar en este complejo y poco explorado fenómeno.

4.5. Perspectivas

El trabajo anterior es, desde luego, sólo un esbozo, por necesidad breve y limitado, del uso de herramientas lingüísticas, estadísticas y de la teoría de la información al estudio de secuencias genéticas. Mucho trabajo queda por hacerse: siguiendo estos esquemas han de

buscarse *reglas gramaticales* basadas en análisis estadísticos primero y funcionales más adelante,acreca de la presencia de palabras o grupos de éstas, en diversos segmentos del genoma. Asimismo deben generarse *diccionarios* que permitan conocer, más aún comprender, la manera en que se escribe cada rasgo presente en la materia viva a partir de sus *frases* u *oraciones* fundamentales.

Es posible intuir alguna relación, por ejemplo, entre la presencia de segmentos genéticos separadores [1] y las reglas de puntuación en lenguas humanas. El caracter reiterativo de los lenguajes (que, en muchas ocasiones llega a extremos en la repetición con fines enfáticos) puede verse representado por la aparición de RDFs. Hay mucho terreno que explorar para dilucidar que tan lejos puede llevarse la analogía, aunque, por lo visto, la relación entre el lenguaje y la genómica no es casual.

4.6. Gráficas

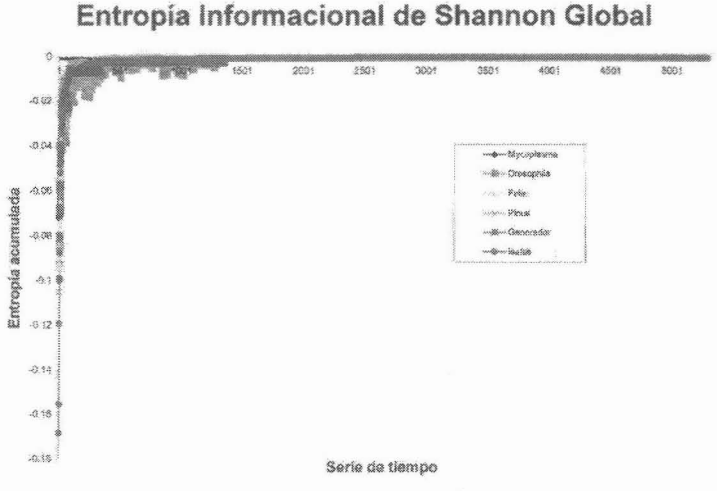


Figura 4.6: Gráfica de Entropía de Shannon-Weaver Global

[1] ver capítulo 2

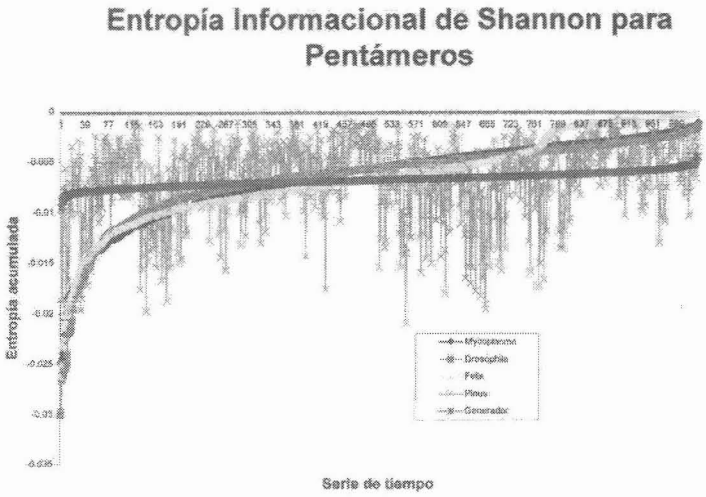


Figura 4.7: Gráfica de Entropía de Shannon-Weaver para Pentámeros

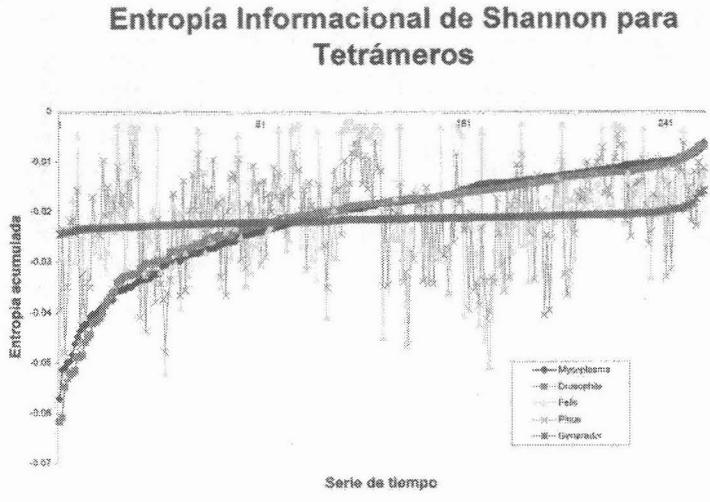


Figura 4.8: Gráfica de Entropía de Shannon-Weaver para Tetrámeros

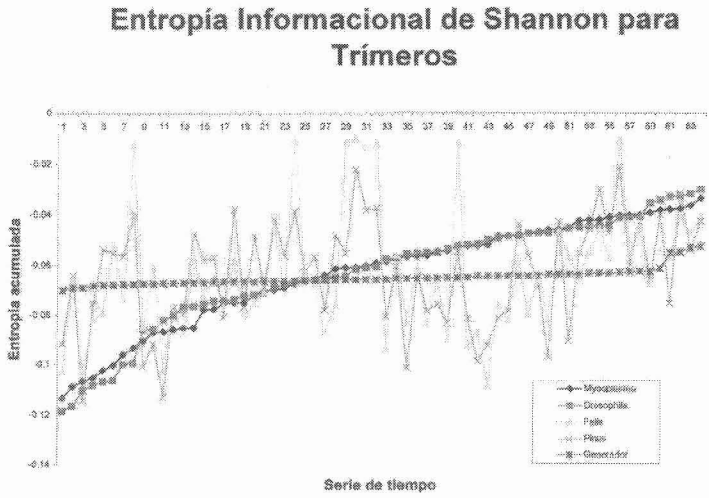


Figura 4.9: Gráfica de Entropía de Shannon-Weaver para trímeros

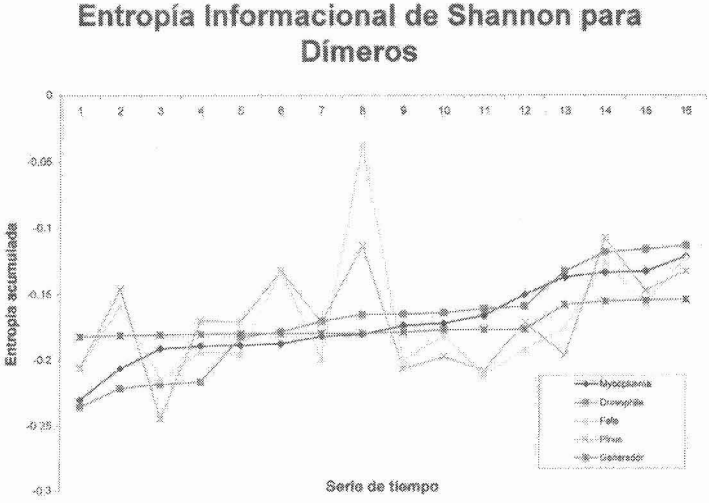


Figura 4.10: Gráfica de Entropía de Shannon-Weaver para dímeros

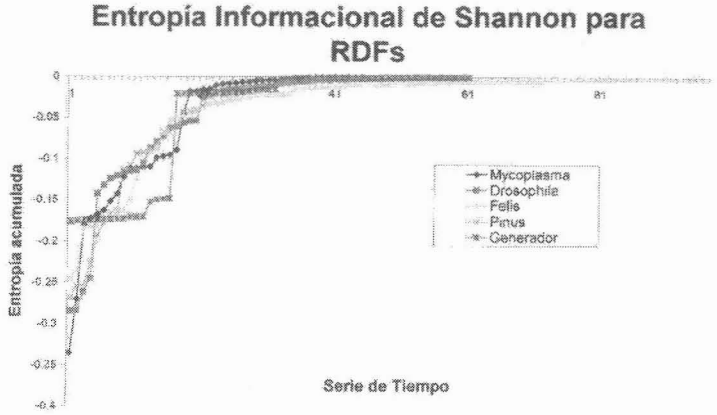


Figura 4.11: Gráfica de Entropía de Shannon-Weaver para RDFs

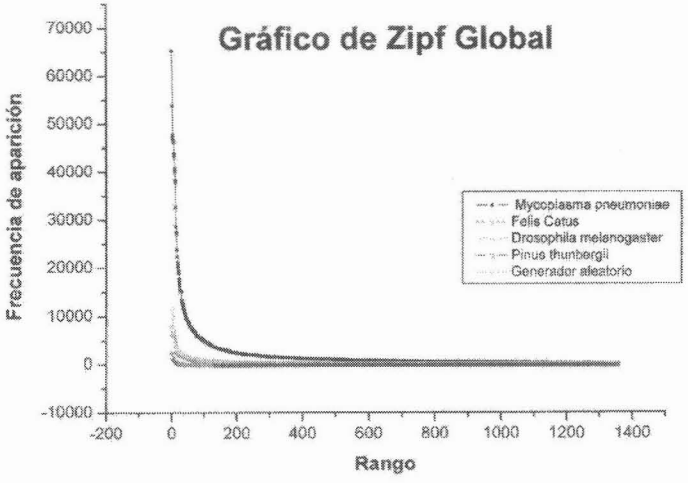


Figura 4.12: Gráfica de Zipf Global

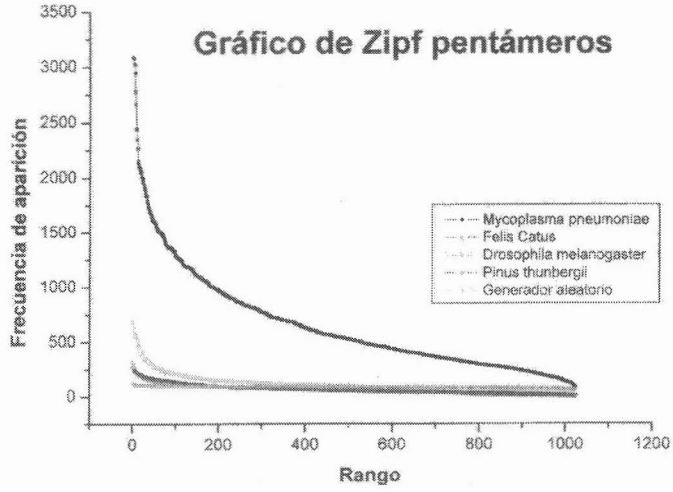


Figura 4.13: Gráfica de Zipf para Pentámeros

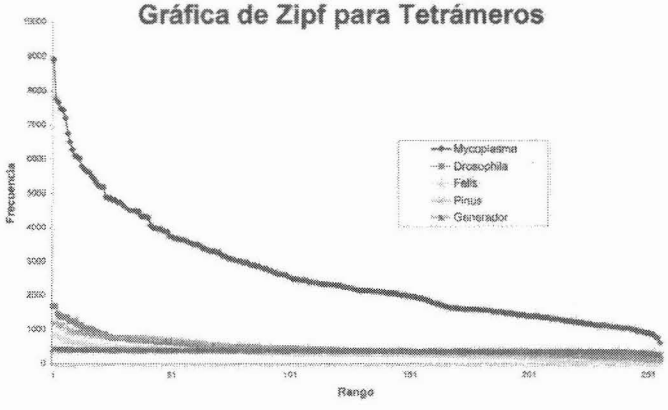


Figura 4.14: Gráfica de Zipf para Tetrámeros

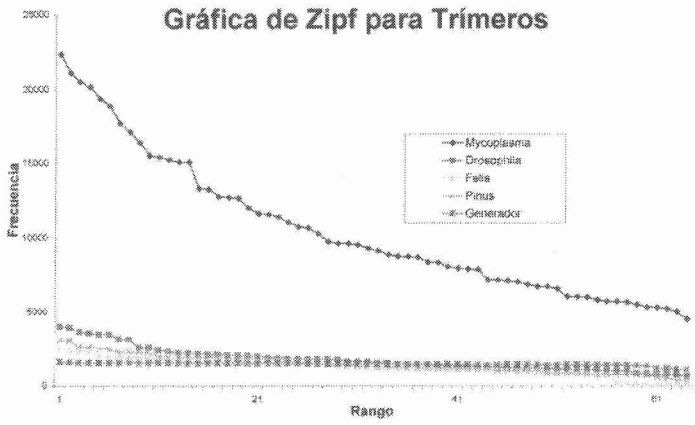


Figura 4.15: Gráfica de Zipf para Trímeros

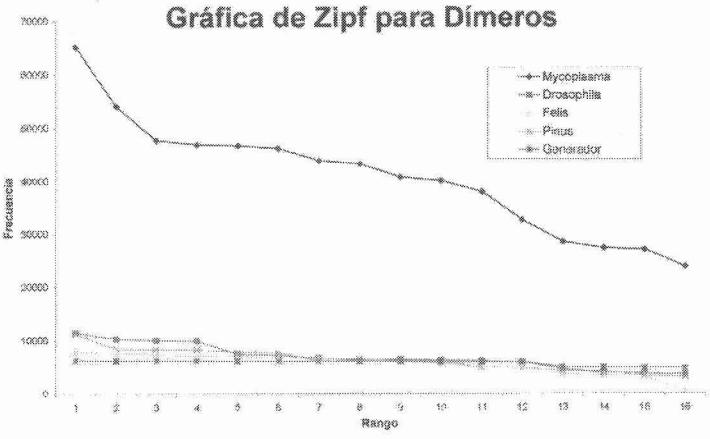


Figura 4.16: Gráfica de Zipf para Dímeros

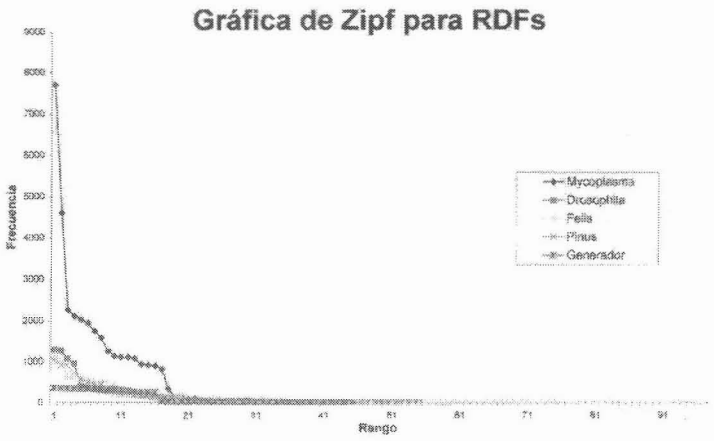


Figura 4.17: Gráfica de Zipf para RDFs

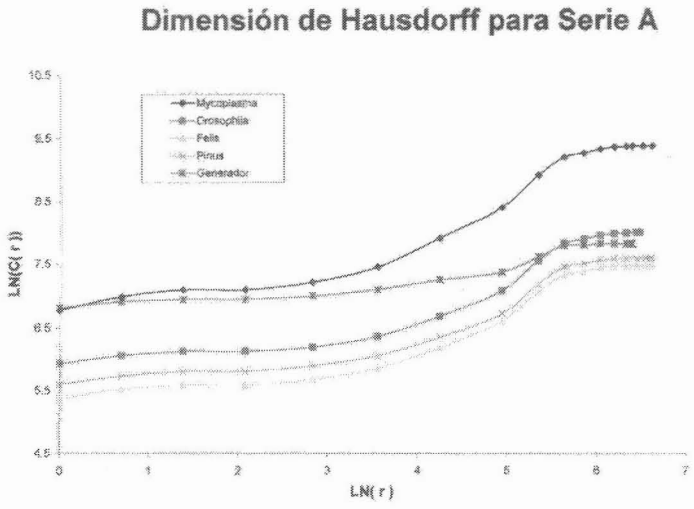


Figura 4.18: Gráfica de Dimensión de Hausdorff Serie A

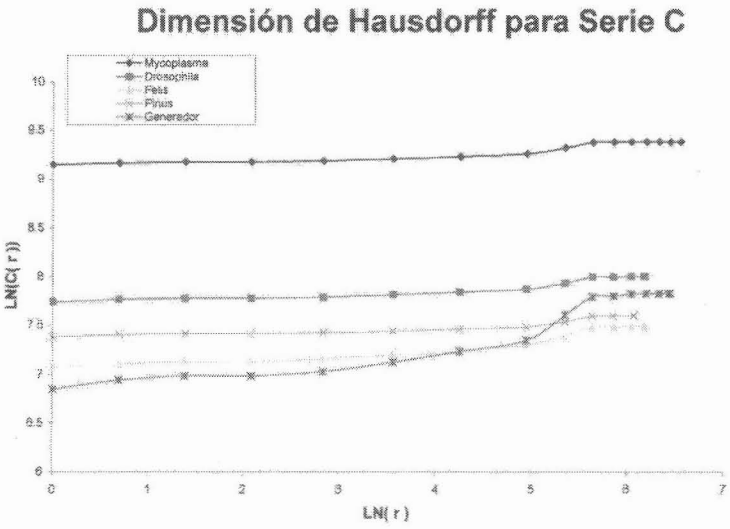


Figura 4.19: Gráfica de Dimensión de Hausdorff Serie C

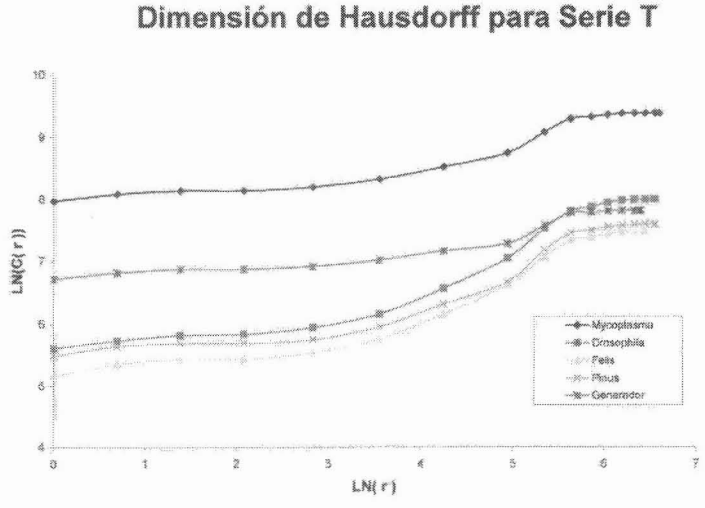


Figura 4.20: Gráfica de Dimensión de Hausdorff Serie T

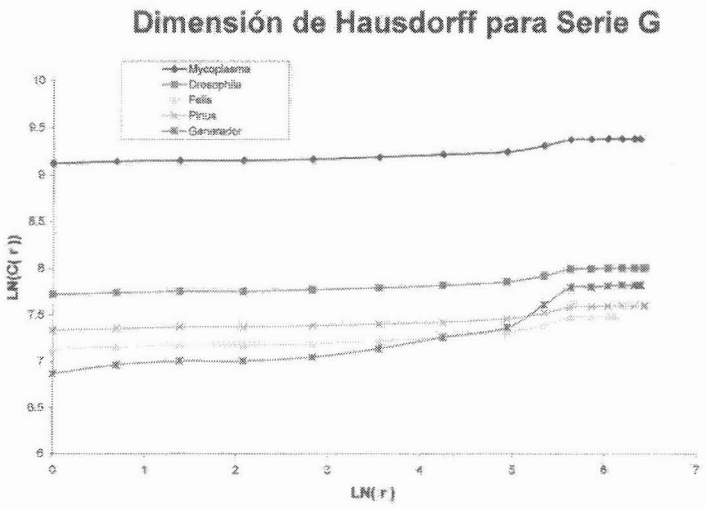


Figura 4.21: Gráfica de Dimensión de Hausdorff Serie G

Apéndice A

Glosario

1. **ADN**: Macromolécula biológica de gran tamaño conformada por ácido fosfórico, moléculas de azúcares de cinco carbonos y bases nitrogenadas. Está encargada de pasar la información hereditaria de una generación a otra. El ADN de eucariontes contiene dos tipos de secuencias denominadas intrones y exones, así como en segmentos intergénicos respectivamente.

2. **Aminoácido**: Compuestos derivados de los ácidos orgánicos que se obtienen de un grupo amino y de un carboxilo. Se reúnen formando cadenas de péptidos en las cadenas de proteínas. Aunque existen más de 20, son sólo éstos los que utilizan las moléculas relacionadas con la vida.

3. **ARN**: Molécula que se sintetiza a partir del ADN por transcripción, de forma que a partir de la molécula de ADN, se copia la secuencia complementaria de ARN. Existen tres tipos: los ARN mensajeros (ARNm), transferentes (ARNt) y ribosomales (ARNr). Todos ellos se sintetizan en el núcleo como precursores y posteriormente van sufriendo una serie de modificaciones enzimáticas postranscripcionales, produciéndose su maduración antes de adquirir su conformación definitiva.

4. **Base nitrogenada**: Molécula compuesta por () y son: Adenina, Citosina, Timina, Guanina y Uracilo (abreviadas A, C, T, G y U respectivamente). Dos de las bases, guanina y adenina guardan gran semejanza, ambas poseen una estructura de dos anillos y pertenecen a la familia de las purinas. Las restantes (timina, citosina y uracilo) poseen estructuras anulares en cierta forma más sencillas, también se parecen y pertenecen a la familia de las pirimidinas.

5. **Codón**: Trío de nucleótidos en una secuencia de ADN.

6. **Cromatina** Es un coloide incoloro, de carácter proteico con la capacidad para cambiar de sol a gel y que alberga las proteínas implicadas en el metabolismo del núcleo. Las proteínas disueltas en agua integrantes del carioplasma, constituyen una especie de red de carácter tridimensional en la que están inmersas el resto de las estructuras nucleares. Así por ejemplo, entre estas proteínas están todas las implicadas en los procesos de

replicación, reparación y transcripción del material genético. Dependiendo del grado de lasitud y la fase mitótica o meiótica por la que esté atravesando la célula, puede ser eucromatina o heterocromatina.

7. **Cromosoma** Unidades nucleares que aparecen individualizadas durante la división celular y que llevan en su interior organizado el material genético. Están constituidos por dos estructuras dispuestas longitudinalmente una con respecto a otra, denominadas cromátidas. También presentan un estrechamiento a cuyo nivel puede haber torsión, denominado constricción primaria. Las zonas del cromosoma que quedan a uno y otro lado de la constricción primaria son los brazos del cromosoma.

   El número de cromosomas varía entre diferentes especies, aunque en células somáticas y para una especie concreta permanece constante. En células somáticas, cada cromosoma está representado dos veces siendo ambos homólogos, es decir, hay 2n cromosomas que constituyen la dotación diploide. En las células reproductoras o gametos solo se encuentran la mitad de cromosomas, resultando un número n que constituye la dotación haploide.

   El tamaño de los cromosomas varía según la especie y oscila desde 0,2 hasta 50 m, excluyendo ciertos cromosomas de gran tamaño denominados gigantes, que pueden alcanzar dimensiones muy elevadas.

   Las características morfológicas de los cromosomas de una determinada especie constituyen su cariotipo.

8. **Entropía de Shannon**: una medida de la corrupción en los mensajes, tanto en un texto lingüístico como genético. A causa de ello, los lenguajes deben compensar ese hecho, por eso son redundantes (de lo que quiere expresar el emisor). Perder información es ganar entropía y viceversa, son complementarias, la suma de ambas es una constante.

9. **Enzima**: Familia de proteínas globulares que, en los organismos, favorecen o inhiben las reacciones químicas entre las otras moléculas. En la terminología especializada se diría que actúan de catalizadores: alteran la velocidad a la que proceden las reacciones químicas, sin verse ellas mismas alteradas por la reacción. El mejor modo de entender cómo proceden es imaginarse una gran molécula, más o menos esférica, (una proteína globular) en cuya superficie se aprecia una cavidad configurada de tal manera que se ajustan en ella otras dos moléculas, específicas y mucho menores. Al asentarse las dos moléculas en la cavidad tan adecuadamente proporcionada por la enzima, quedan alineadas de tal modo que de inmediato se establecen enlaces entre ellas. No existen ya dos moléculas, sino una sola; la enzima puede ya liberarla y proseguir sus tareas bioquímicas por el interior celular, tomando de nuevo dos pequeñas moléculas (exactamente iguales a las dos anteriores) del acervo de compuestos químicos que la rodean y repitiendo la operación cuanto sea necesario. Por un procedimiento semejante algunas

enzimas escinden otras moléculas.

Cada enzima está diseñada para un solo propósito, cada una encargada de una tarea concreta. Una enzima unirá un par de moléculas, quizá un enlace da alguna cadena polipeptídica, o las moléculas que intervienen en el aporte de energía a los músculos. Otra se dedicará por entero a escindir algún enlace concreto que uno un par de moléculas orgánicas. En muchos sentidos guardan parecido con las herramientas de las cadenas de producción industriales.

10. **Exones**: Región del ADN que codifica para una proteína.

11. **Fenotipo**: Conjunto de factores hereditarios que un individuo manifiesta.

12. **Gene**: Unidad determinante de factores hereditarios responsables de la manifestación de un carácter.

13. **Genotipo**: Conjunto de factores hereditarios que posee un individuo.

14. **Histona**: Proteínas de bajo peso molecular y gran cantidad de aminoácidos básicos tales como la arginina y la lisina. Junto con las protaminas, son proteínas que forman el esqueleto del cromosoma, a través de fuerzas electrostáticas que se establecen entre las zonas altamente positivas de las histonas y los restos negativos de los enlaces fosfodiéster del ADN. Estas proteínas pueden sufrir una serie de modificaciones reversibles, tales como acetilaciones-desacetilaciones o fosforilaciones-desfosforilaciones que pueden implicar cambios conformacionales de los edificios moleculares del cromosoma, quizá necesarios para que ocurran procesos tan importantes como la replicación y transcripción o los más complejos de la cariocinesis.

    Se han identificado cinco tipos de histonas, comunes a la mayor parte de las células eucariontes. La histona-1 (H-1) es la de mayor peso molecular e interviene de manera diferente a las otras en la arquitectura cromosómica. Las histonas 2A, 2B, 3 y 4 (H-2A, H-2B, H-3, H-4) forman unidades repetitivas asociadas al ADN llamadas nucleosomas. Cada nucleosoma está constituido por un par de cada tipo de éstas últimas histonas, asociadas entre sí por fuerzas hidrofóbicas.

15. **Interacción (de corto o largo alcance)**: Efecto que ejerce un nucleótido o conjunto de ellos sobre una sección del genoma. El ADNc revela una interacción de corto alcance o Markoviana. Por el contrario, el ADNnc muestra una interacción de largo alcance o apegada a una ley de potencias.

16. **Intrón**: Región del ADN que -se dice- no codifica para una proteína.

17. **Isocora**: Región del ADN de composición constante.

18. **Locus,loci: (Sing., Pl.)** Lugar o lugares que ocupan los genes dentro de un cromosoma.

19. **Meiosis**: Proceso celular que sirve para reducir el número cromosómico y, de esta manera, mantenerlo siempre estable. Está implicada en la formación de gametos y consta de dos mitosis: una reduccional y otra normal.

20. **Mitosis**: Proceso de división celular compuesto por cuatro fases principales: Profase, Metafase, Anafase y Telofase.

21. **Mutación**: Cambio que sufre el ADN por distintos factores que van desde radiaciones de diversos tipos hasta la dieta o la somatización del estrés. Las mutaciones en el material genético provocan enfermedades como el cáncer y son, por lo general, más bien dañinas que benéficas.

22. **Nucleosoma**: Unidad estructural que se repite, está formado por un par de cada una de las histonas H-2A, H-2B, H-3 y H-4. Los cuatro partes de histonas se asocian hidrofóbicamente entre ellas formando a modo de un tonel de unos 57 A de largo, sobre el que se espiraliza la doble hélice de ADN. Cada dos nucleosomas están unidos por el lazo internucleosómico, el cual representa la fibra de ADN que queda entre dos nucleosomas y a la que se une la histona H-1.

    Las histonas que integran los nucleosomas se encuentran en cantidades equimoleculares respecto al ADN. Así se cuantifican un par de cada tipo por cada 200 pares de bases. En el caso de la histona H-1, solo existe una molécula por cada 200 pares de bases. La histona 1 se une al ADN de forma laxa y juega un importante papel en el plegamiento de la fibra nucleosómica.

    El ADN unido a las histonas presenta un aspecto *con forma de rosario* (nucleosomas + fibra nucleohistónica).

    La fibra nucleosómica se pliega helicoidalmente para constituir los cromosomas. Cada vuelta contiene 6 nucleosomas.

23. **Nucleótido**: Componente más básico del ADN. Molécula biológica compuesta por una base nitrogenada, un grupo fosfato y un azúcar de cinco carbonos.

24. **Oligómero**: Polímero constituido por monómeros de diferente clase.

25. **Palabra**: En el sentido genético que se maneja en este trabajo, equivale a un nucleótido solo o en distintas combinaciones numéricas con otros nucleótidos o con sí mismo; cualquier expresión continua cuya ausencia o modificación cambiaría la funcionalidad del sistema (biológico o lingüístico).

26. **Persistencia**: Tamaño del intervalo en el cual hay independencia estadística en algún lado.

27. **Polímero**: La larga cadena resultante, que puede contener miles de unidades básicas en secuencia repetida.

28. **Procesos multiplicativos aleatorios**: Procesos estocásticos en los que las distribuciones de probabilidad multiplican el número posible de resultados, pudiendo generar de esta manera *ruido* en las descripciones.

29. **Proteína**: Las proteínas son las herramientas que tienen las células para llevar al cabo la mayor parte de sus funciones: en otras palabras, en ellas reside la información funcional de la célula. Existen en nuestro organismo más de cuarenta mil. Dichas proteínas son polímeros biológicos constituidos por veinte tipos de monómeros diferentes llamados aminoácidos. Cada proteína tiene una secuencia específica de varios aminoácidos dada por la secuencia de los tripletes de nucleótidos del gene que la codifica y esta secuencia es la que se conoce como la estructura primaria de la proteína. Gracias a esta secuencia primaria, la proteína puede adquirir una estructura secundaria que puede ser fundamentalmente de dos tipos: alfa hélice o beta plegada. Las estructuras secundarias permiten a su vez el doblamiento de las proteínas en estructuras terciarias y finalmente las estructuras terciarias permiten la asociación de varias moléculas de proteínas en lo que se conoce como estructura cuaternaria. La estructura o conformación espacial de cualquier proteína, es lo que le permite tener una función biológica particular a esa proteína: todas las proteínas tienen estructuras específicas y por ende funciones específicas en la célula viva.

    La función principal de las proteínas es actuar de enzimas, interviniendo así en la dinámica de las funciones nucleares, principalmente replicación y transcripción e inducción y represión de genes. Análogamente, la del material genético es controlar la síntesis de estas cadenas polipeptídicas.

30. **RDFs**: Abreviatura de Repeticiones diméricas en fila. Repeticiones pares de alguna palabra genética. e.g. Una RDF de tamaño 3 en la palabra genética GA equivale a: GAGAGA. La distribución de estos cúmulos en el ADN revela la presencia, ya sea de ADNc o de ADNnc.

31. **Secuencia**: Una serie de tamaño determinado en el genoma.

32. **Solenoide**: Conjunto de histonas.

33. **Tándem**: Serie repetitiva de letras, palabras o nucleótidos.

34. **Transposón**: Los genes transponibles en realidad no saltan sino que el original conserva su ubicación en el cromosoma: la maquinaria celular de replicación del ADN que actúa habitualmente elabora una copia y es ésta la que cambia de lugar, insertándose el genen el punto de escisión del cromosoma.

Además del gene activo, los transposones deben contener genes que trabajen en su intetés; genes que asegurenla elaboración de enzimas que se encarguen de practicar los cortes y empalmes necesarios para que el transposón se acomode en su nueva ubicación.

Además de la influencia directa que ejercen sobre sus vecinos, los genes transponibles provocan la gama entera de mutaciones habituales: deleciones, inserciones e inversiones. Quizá su papel principal sea el de controlar a sus vecinos, pero uno de sus efectos secundarios es el incremento de la tasa de mutación.

Todo esto resulta de importancia decisiva en la reproducción sexual pues permite la mezcla genética que se da por recombinación.

Apéndice B

Organismos estudiados en este trabajo

Firmicutes

B.0.1. *Mycoplasma pneumoniae*

División: Bacteria
Clase: *Mollicutes*
Familia: *Mycoplasmataceae*
Género: *Mycoplasma*
Especie: *Mycoplasma pneumoniae*

Uno de los seis géneros de eubacterias de la clase Mollicutes. Presenta fragilidad osmótica, forma colonial, filtrabilidad a través de membranas con poros de 450 nm. de diámetro y carece de pared celular bacterial.

Los géneros de la clase Mollicutes posee algunos de los tamaños genómicos más pequeños que se conocen hasta ahora; probablemente esto haya sido tolerado a causa del estilo de vida parasitario de la clase Mollicutes, ya que en la naturaleza depende de una célula huésped de vertebrado de la cual obtiene compuestos esenciales como ácidos grasos, aminoácidos, precursores para la síntesis de ácidos nucleicos y en algunos casos (como en el caso del género Mycoplasma), colesterol.

Mycoplasma pneumoniae es una bacteria patógena que ataca a los seres humanos causando traqueobronquitis y neumonía atípica primaria. Asociada a la célula huésped, la colonización superficial de las células epiteliales del tracto respiratorio tiene lugar. La carencia de la pared celular facilita el contacto estrecho entre M. pneumoniae y la célula huésped a la vez que garantiza el intercambio de compuestos necesarios para el crecimiento bacterial. Como consecuencia de este parasitismo, la célula huésped es gravemente dañada.

Entre los distintos géneros de la clase Mollicutes, Mycoplasma tiene el más alto contenido de bases nitrogenadas G+C (41 % mol para M. pneumoniae). El tamaño genómico de esta

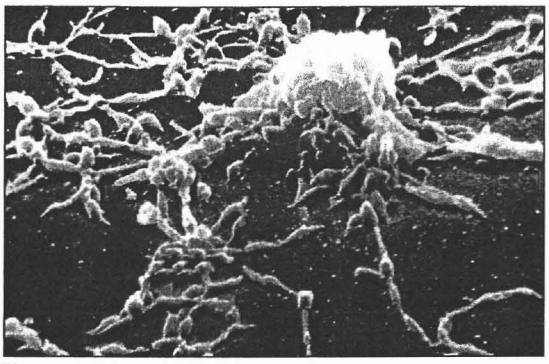


Figura B.1: *Mycoplasma pneumoniae*

especie es de 816 kbp. teniendo una capacidad para codificar 700 proteínas (687 genes) y un promedio de masa molecular de 40000 Da. Puesto que M. pneumoniae está entre las más pequeñas células auto-replicadoras conocidas hasta hoy (0.5 m. de diámetro) fue seleccionada como sistema modelo para determinar los requerimientos genéticos mínimos para una célula que se reproduce de manera autónoma.

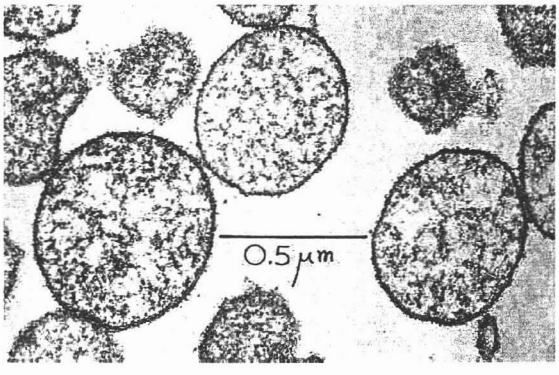


Figura B.2: Colonia de *Mycoplasma pneumoniae*

B.0.2. *Drosophila melanogaster*

División: *Animalia*
Clase: *Artropoda*
Familia: *Insectae*
Género: *Drosphila*
Especie: *Drosphila melanogaster*

Pequeño insecto de 3 mm. de largo, usualmente se le puede encontrar revoloteando alrededor de la fruta. Uno de los organismos más valiosos para la investigación biológica, particularmente en Genética y el Biología del desarrollo. D. melanogaster ha sido usada como organismo modelo para la investigación por casi un siglo y hoy, miles de científicos están trabajando en diferentes aspectos de esta mosca de la fruta. Su importancia en relación con la salud humana es ampliamente reconocida.

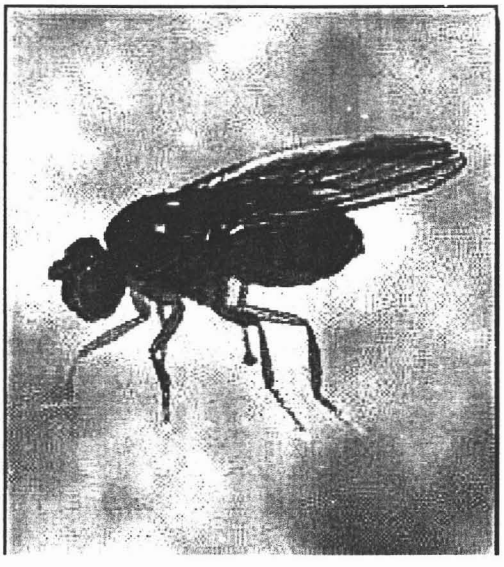


Figura B.3: *Drosophila melanogaster* -mosca de la fruta-

Parte de la razón por la cual la gente trabaja con este organismo es histórica (se sabe ya tanto de éste que es fácil manejarlo y entenderlo) y parte es práctica, es un animal pequeño, con un ciclo de vida corto de 2 semanas y de reproducción prolija, es barato y es fácil mantener poblaciones grandes de éste. Moscas mutantes con defectos en miles de genes están disponibles y se ha llegado a secuenciar su genoma entero.

Un huevo de D. melanogaster mide aproximadamente 1 mm. de largo. El embrión se desarrolla después de un día después de la fertilización, posteriormente se convierte en una larva vermiforme. La larva se alimenta y crece constantemente. Después de dos días (ya han pasado la primera, segunda y tercera etapa de desarrollo), forma una pupa inmóvil. Después de cuatro días más, el cuerpo ha cambiado totalmente hasta alcanzar la forma alada característica, así emerge del capullo y es fértil tras ocho horas.

D. melanogaster tiene cuatro pares de cromosomas: los cromosomas sexuales X/Y y los autosomas 2,3 y 4. El cuarto cromosoma es bastante pequeño y raramente mencionado. El tamaño del genoma es de aproximadamente 165 millones de bases y contiene unos 14 000 genes (el genoma humano tiene 3 300 millones de bases y tiene aproximadamente unos 70 000 genes; la levadura tiene unos 5800 genes en 13.5 millones de bases). El genoma fue (casi) completamente secuenciado en el 2000 y el análisis de los datos está en proceso hasta ahora.

Una característica peculiar en D. Melanogaster es la presencia de cromosomas politénicos. Conforme la larva crece mantiene el mismo número de células, pero necesita producir mucho más material genético. El resultado es que las células crecen y cada cromosoma se divide cientos de veces, pero todas las hebras se mantienen unidas unas con otras. El resultado es un cromosoma politénico ancho y masivo que puede ser visto fácilmente al microscopio. Mejor aún, estos cromosomas tienen un patrón de bandas oscuras y claras (102 en total, en promedio cada banda de letras contiene aproximadamente 300 kb de ADN y de 15 a 25 genes) como un código de barras que es único para cada sección cromosomal. Como resultado, leyendo en estas bandas politénicas es posible saber qué parte del cromosoma estamos analizando; cualquier delesión u otro rearreglo de partes del cromosoma puede ser identificado y usando pruebas de ácido nucleico, los genes clonados individuales pueden ser situados sobre el mapa politénico.

Para el presente trabajo se utilizó el genoma correspondiente al cuarto cromosoma de D. Melanogaster, que es el más pequeño que posee (comprende 3.5 % del genoma total y de 50-75 genes) y cuya característica singular es que posee regiones heterocromáticas (dominios R) y eucromáticas (dominios V) entremezcladas y asociadas estrechamente dentro de esta región del genoma cuyo tamaño es de 1.2 megabases.

B.0.3. *Felis catus*

División: *Animalia*
Clase: *Mammalia*

Orden: *Carnivora*
Familia: *Felidae*
Género: *Felis*
Especie: *Felis catus*

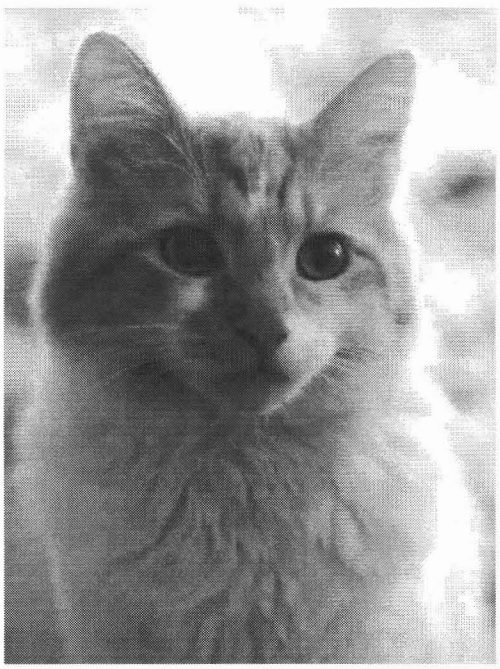


Figura B.4: *Felis catus*

También conocido como Felis domesticus, mamífero de cuerpo compacto, muscular y muy flexible. Su expectativa de vida es de aproximadamente 15 años.

El cuerpo de un gato doméstico es extremadamente flexible, su esqueleto contiene más de 230 huesos y su pelvis y hombros están fijos a la espina más laxamente que en otros cuadrúpedos. La habilidad trepadora del gato y su velocidad se deben en parte a su poderosa musculatura. Su cola provee balance cuando brinca o cae.

Las garras del gato están diseñadas para cazar, luchar y trepar; son filosas, ganchudas y retráctiles y están enfundadas en una cubierta de piel al final de cada dedo. El gato marca su territorio rascando y dejando su esencia en los objetos (sus garras dejan una marca visible y las glándulas secretoras en los cojinetes de sus patas dejan una marca aromática. Los dientes del gato están diseñados para morder no para masticar. Los poderosos músculos de su quijada

y filosos dientes le permiten dar una mordida mortal a la presa.

La visión del gato está bien adaptada para cazar, especialmente de noche. Tiene una excelente visión nocturna periférica extensiva y visión binocular, lo cual lo habilita para distinguir distancias acertadamente. La visión diurna del gato no es tan buena como en el caso de los humanos; los gatos perciben el movimiento más fácilmente que un detalle y se piensa que sólo ven un rango limitado de colores. El oído del gato es sensible y detecta una amplia cantidad de sonidos, incluyendo los de rango ultrasónico. Sus oídos son menos sensibles a frecuencias bajas.

El gato tiene desarrollado altamente el sentido del olfato, lo cual juega un papel vital en encontrar comida y en la reproducción. Muchas de las señales sociales de los gatos domésticos provienen de las esencias.

El sentido del gusto del gato está peculiarmente especialmente desarrollado: puede detectar las más pequeñas variaciones. Su lengua está cubierta con ásperas protuberancias o papilas que usa tanto para raspar carne de los huesos como para limpiarse a sí mismo. Las vibrisas del gato son muy sensibles al más leve roce y son utilizadas para sentir obstáculos o cambios en el ambiente.

El gato doméstico generalmente alcanza la pubertad a los 9 o 10 meses de edad. Las hembras se encuentran en estro varias veces al año, periodo durante el cual es receptiva a los machos. El periodo de gestación es de 65 días aproximadamente. El promedio de crías es de 4 y nacen ciegas, sordas e indefensas. Sus ojos se abren a los 8 o 10 días de edad y comienzan a caminar a las 6 semanas después del nacimiento.

El color de pelaje original en el gato doméstico fue probablemente el café grisáceo con rayas más oscuras, un color que provee un excelente camuflaje en una gran variedad de ambientes. Los demás colores y patrones de pelaje son el resultado de mutaciones genéticas; por ejemplo, el pelaje de color negro es resultado de un gene que suprime las rayas, un pelaje anaranjado es el resultado de un gene que transforma pigmento negro en anaranjado y el pelaje blanco es el resultado de un gene que suprime completamente la formación de pigmento.

Se conocen alrededor de 40 variedades o razas de gatos domésticos; a pesar de que todas ellas difieren notablemente, varían menos en tamaño (la raza más pequeña pesa de 2 a 3 Kg. Y la más grande de 7 a 9 Kg.

B.0.4. *Pinus thunbergii*

División: *Plantae*
Clase: *Pinopsida*
Orden: *Pinales*
Familia: *Pinaceae*

Género: *Pinus*
Especie: *Pinus thunbergii*

Figura B.5: Pino negro japonés - Pinus thumbergii

Árbol de hoja perenne (3 a 5 años), también conocido como pino negro japonés distintivo y pintoresco con una estructura abierta e irregular. Las ramas son largas y torcidas, extendidas horizontalmente y en algunas ocasiones formando péndulos. El follaje es abundante y tiende a concentrarse cerca de las puntas de las ramas más delgadas. Esta especie crece generalmente de 20' a 30' cuando es cultivada, aunque puede alcanzar 100' en su estado silvestre. Presenta hojas aciculares de color verde oscuro con márgenes ligeramente dentados y líneas estomáticas en cada superficie; miden de 3' a 5' de largo y 1/12' de ancho y están distribuidas en fascículos binarios. Generalmente posee brotes blancos, plateados o grises; vástagos verticales, yemas no resinosas de forma ovoide o cilíndrica, flores monoicas, corteza de color negro-grisáceo, fisurada en placas elongadas e irregulares, conos subterminales solitarios o distribuidos en cúmulos, simétricos que pueden ir de ovoides a cónicos, escamosos y aplanados de color café brillante, umbo pequeño, depreso, obtuso, con una pequeña espina. Existen numerosas variedades, muchas de ellas seleccionadas para bonsái.

Este pino es nativo de zonas como China, Corea y Japón. Es un árbol típico de los paisajes orientales aunque podemos encontrarlo en menor grado en los jardines costeros de la parte

norteña de Europa y el noreste de E.U.A. También es ampliamente empleado para estabilizar dunas de arena y en las últimas décadas como árbol de ornato en las ciudades, dada su tolerancia a las condiciones urbanas.

Apéndice C

Libro bíblico de Isaías

Escritor: *Isaías*
Lugar donde se escribió: *Jerusalén*
Fecha en que se completó: *Después de 732 a. E. C.*
Tiempo que abarca: *c. 778-después de 732 a. E. C.*

El escritor de este libro fue un agricultor israelita, hijo de Amoz (a quien no se debe confundir con Amós, otro profeta de Judá) (1:1). Las Escrituras no dicen nada de su nacimiento ni de su muerte, aunque según la tradición judía fue aserrado en dos partes por el rey Manasés (Heb.11:37). Sus escritos [no solo el libro bíblico que lleva su nombre, sino muy probablemente por lo menos un texto histórico: Los asuntos del rey Uzías que debía formar parte de los registros oficiales de la nación (2 Cró. 26: 22] lo sitúan en Jerusalén.

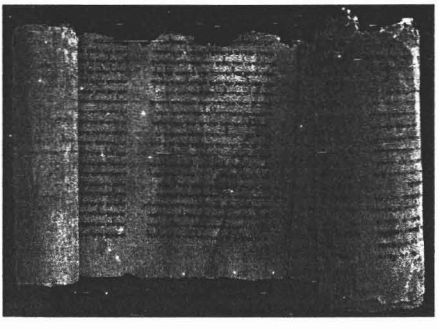


Figura C.1: Fragmento del libro de Isaías, perteneciente a los rollos hallados en la cueva de Qumrán cerca del mar muerto.

Isaías sirvió durante el tiempo de por lo menos cuatro reyes de Judá: Uzías, Jotán, Acaz y Ezequías; y parece que su servicio empezó alrededor del año 778 a. E. C. (cuando murió Uzías, o posiblemente antes) y continuó por lo menos hasta después de 732 a. E. C. (el dec-

imocuarto año de Ezequías), o no menos de 46 años. Otros profetas de su tiempo fueron: en Judá, Miqueas; y al norte, Oseas y Oded. (Miq. 1:1; Ose. 1:1; 2Cró. 28:6-9.)

A principios de 1947, de unas cavernas no lejos de Qumrán, cerca de la costa noroeste del Mar Muerto, se sacaron unos documentos antiguos. Estos fueron los rollos del Mar Muerto, y entre estos estuvo el libro de Isaías. Este documento está escrito en hebreo premasorético y tiene unos 2000 años de antigüedad.

Los primeros seis capítulos del libro de Isaías dan el marco de las circunstancias en Judá y Jerusalén, relatan la culpa de Judá ante Jehová y describen cómo se comisiona a Isaías. Los capítulos 7 a 12 tratan de amenazas de invasiones enemigas y de la promesa de alivio por medio del Mesías. En los capítulos 36 a 39 se describen sucesos históricos del reinado de Ezequías. Los capítulos restantes, 40 a 66, tratan el tema de soltar de Babilonia a los cautivos, el regreso del resto judío y la restauración de Sión.

Bibliografía

[1] Bolívar, Francisco **Obra Científica**. Tomos III y IV. Colegio Nacional. México. 1998. 261pp./120pp.

[2] Eco, Humberto; **La Estructura Ausente**, Lumen, Barcelona, (1972)

[3] Espinosa, Francisco et al. **Biología**. Alambra. 6ł ed. España. 1998. 370 pp.

[4] Feller, W., **An Introduction to Probability Theory and its Applications**, Wiley, N. York, (1957)

[5] Gribbin, John. **En busca de la doble hélice**. Biblioteca Científica Salvat. 2ł ed. Barcelona. 1986. 287 pp.

[6] Griffits, Anthony; **Modern genetic analysis**, W. H. Freeman, New York, 4a. ed., 2001, 675 pp.

[7] Gutiérrez, José Luis & Sánchez, Faustino; **Matemáticas para las ciencias naturales**, Sociedad Matemática Mexicana, México, 1998, 595 pp.

[8] Gutiérrez, José Luis & Sánchez, Faustino; **Matemáticas del crecimiento orgánico. De la alometría y los fractales al crecimiento estacional**, (en prensa)- Comunicación personal-.

[9] Jauralde, Pablo **Introducción al conocimiento de la lengua española**. Everest. España. 1982. 444pp.

[10] Jerman, I. & Stern, A. **The Gene in Waves.** ***The Forming of New Biology***. Znanstveno Publicisticno Sredisce. 1996.

[11] van Kampen, N.; **Stochastic processes in physics and chemistry**, North Holland, (1992)

[12] Kull, Kalevi & Tiivel, T. **Lectures in Theoretical Biology: The Second Stage**. Tallinn: Estonian. Academy of Sciences. 1993.

[13] Fernández Lagunilla, Marina; **Sintaxis y Cognición**, Sintesis, Buenos Aires, (1999)

[14] Fernández Pérez, M. **Introducción a la Lingüística. Dimensiones del lenguaje y vías de estudio**, Barcelona, Ariel, (1999).

[15] Peng, C.K. et al. **Fractals in Biology and Medicine**. Birkhauser Verlag. Boston. 1994. 397pp.

[16] Ruelle, D., **Chaotic Evolution and Strange Attractors**, Lezioni Lincee, Accademia Nazionale dei Lincei, Cambridge University Press, Cambrige (UK) (1989)

[17] Ruelle, D., **Statistical Mechanics: Rigurous Results**, Advanced Book Classics, Addison Wesley, Massachusetts, (1989).

[18] Rozanov, Y.A. **Probability: A Concise Course**. Dover Publications. New York. 1977. 148pp.

[19] Hirsch, M. W. y Smale, S., **Differential Equations, Dynamical Systems and Linear Algebra**, Academic Press, London, (1994).

[20] Schöedinger, E. **What is life?**. Canto series, Cambridge University Press, Great Britain, 1992.

[21] Sveshnikov, A.A. **Problems in Probability. Therory, Mathematical Statistics and Theory of Random Functions**. Dover Publications. New York. 1978. 471pp.

[22] Swokowski, Earl **Matrices y Determinantes**. Iberoamericana. México. 1986. 32pp.

[23] Tamarin, Robert. **Genética**. Reverté. Barcelona. 1996. 607 pp.

[24] **Traducción del nuevo mundo de las Santas Escrituras**. Watchtower Bible and Tract Society of Pennsylvania. 3a ed. E. U. A. 1987. 1595 pp.

[25] Welsh, Dominic **Codes and Criptography**. Oxford University Press. Oxford. 1988. 257pp.

[26] Williams, D., **Probability with Martingales** Cambridge Mathematical Textbooks, Cambridge University Press, Cambrige (UK) (1998)

[27] Zipf, Georges **Human Behavior and the Principle of Least Effort**. Addison-Wesley. Cambridge. 1949. 573pp.

ARTÍCULOS

[28] Anthony, T. G. et al. *Identification of domains within the e-Subunit of the Translation Initiation Factor eIF2B that are Necessary for Guanine Nucleotide Exchange Activity and eIF2B Holoprotein Formation. Biochimica et Biophysica Acta* 1492 (2000) 56-62.

[29] Ball, R.C. *Protein Design Depends on the Size of the Amino Acid Alphabet. Phys. Rev. E.* 66 031902 (2002).

[30] Bernaola-Galván, P. et al. *Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method. Phys. Rev. Lett.* 85 (6) 1342 (2000).

[31] Berthelsen, C.L. et al. *Global Fractal dimension of Human DNA Sequences Treated as Pseudorandom Walks. Phys. Rev. A.* 45 (12) pp.1992.

[32] Bonhoeffer, S. et al. *No Signs of Hidden Language in Noncoding DNA. Phys. Rev. Lett.* 76 (11) 1977 (1996).

[33] Buldyrev, S. V. et al. *Analysis of DNA Sequences Using Methods of Statistical Physics. Physica A* 249 (1998) 430-438.

[34] Buldyrev, S. V. et al. *Expansion of tandem repeats and oligomer clustering in coding and noncoding DNA sequences. Physica A* 273 (1999) 19-32

[35] Campos, P.R.A. et al. *On the Structure of Genealogical Trees in the Presence of Selection. Physica A* 283 (2000) 11-16.

[36] Czirók, A. et al. *Correlations in Binary Sequences and a Generalized Zipf Analysis. Phys. Rev. E.* 52 (1) 446 (1995).

[37] Czirók, A. et al. *Possible Origin of Power-Law Behavior in n-tuple Zipf Analysis. Phys. Rev. E.* 53 (6) 6371 (1996).

[38] Dasgupta, S. *Why sexual reproduction? Why four bases?. Physica A* 298 (2001) 465-470.

[39] Derrida, B. et al. *Distribution of Repetitions of Ancestors in Genealogical trees. Physica A* 281 (2000) 1-16.

[40] Diness, V. et al. Recombinant human factor VIIa (rFVIIa) in a rabbit stasis model. Thromb. Res. 67 (2) 233 (1992).

[41] Dokholyan, N.V. et al. *Distribution of Base Pair Repeats in Coding and Noncoding DNA Sequences. Phys. Rev. Lett.* 79 (25) 5182 (1997).

[42] Dokholyan, N. V. et al. *Model of Unequal Chromosomal Crossing Over in DNA Sequences. Physica A* 249 (1998) 594-599.

[43] Dokholyan, N. V. et al. *Distributions of Dimeric Tandem Repeats in Non-Coding and Coding DNA Sequences. J. Theor. Biol.* 202 (2000) 273-282.

[44] Domany, E. *Cluster Analysis of DNA-chip and Antigen-chip Data.* Personal communication. (preprint)

[45] Dreyer, O. & Puzio R. *Allomeric Scalling in Animals and Plants. J. Math. Biol.* 43 144-156 (2001).

[46] Duga, S., et al.; *Biochim. Biophys. Acta*, 29, 1490, 225 (2000)

[47] Filippini, D, et al.; *Biochem. Biophys. Research Comm.* 288, 1, p. 16 (2001)

[48] Grassberger, P. y Procaccia, I., *Phys. Rev. Lett.* **50**, 346, (1983)

bibitemgold Goldberg, B. & Stricker, R..B. *Bridging the Gap: Human Diploid Cell Strains and the Origin of AIDS. J. theor. Biol.* 204 (2000) 497-503.

[49] Gorban A.N. et al. *Statistical Approaches to Automated Gene Identification without Teacher*. Personal communication. (preprint)

[50] Gorban A.N. et al *Classification of Symbol Sequences over their Frequency Dictionaries: Towards the Connection between Structure and Natural Taxonomy*. Personal communication. (preprint)

[51] Gorban A.N. et al. *Self-Organizing Approach for Automated Gene Identification in Whole Genomes*. Personal communication (preprint).

[52] Gutiérrez, J.M. et al. *Multifractal Analysis of DNA Sequences Using a Novel Chaos-game Representation. Physica A* 300 (2001) 271-284.

[53] Hagedorn, T. R. and Landweber, L. F. *Phylogenetic Invariants and Geometry. J. Theor. Biol.* 205 (2000) 365-376.

[54] Halibard, M. & Kanter, I. *Markov Processes and Linguistics. Physica A.* 249 (1998) 525-535.

[55] Hao, B. L. *Fractals from Genomes: Exact Solutions of a Biology-Inspired Problem. Physica A* 282 (2000) 225-246.

[56] Hattori, M. et al. *The DNA Sequence of Human Chromosome 21. Nature.* 405 (6784) 311 (2000).

[57] Havlin, S. *The distance between Zipf plots. Physica A* 216 (1995) 148-150.

[58] Herrick, J. et al. *Kinetic Model of DNA Replication in Eukaryotic Organisms*. Arxiv: Physics 2001- 1-9 HRC.

[59] Ishii, H. *A Statistical-Mechanical Model for Regulation of Long-Range Chromatin Structure and Gene Expression. J. Theor. Biol.* 203 (2000) 215-228.

[60] Israeloff, N. E. et al. *Can Zipf Distinguish Language from Noise in Noncoding DNA?. Phys. Rev. Lett.* 76 (11) 1976 (1996).

[61] Kirillova, O. V. *Comparative Statistical Analysis of Bacteria Genomes in Word Context*. Physica A 290 (2001) 453-463.

[62] Kurzynski, M. *Internal dynamics of biomolecules and statistical theory of biochemical processes*. Physica A. 285 (2000) 29-47.

[63] Li, W. *Zipf's Law in Importance of Genes for Cancer Classification using Microarray Data*. arxiv:physics/0104028 v1 6/Apr/2001

[64] Mantegna, R.N. et. al., *Phys. Rev. Lett.* 73, (1994) pág. 3169.

[65] Mariño, I. P. et al. *Exploiting the Natural Redundancy of Chaotic Signals in Communications Systems*. Phys. Rev. Lett. 85 (12) 2629 (2000).

[66] Montemurro, Marcelo A. *Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics*. Physica A 300 (2001) 567-578.

[67] Nowak, M. A. *The Basic Reproductive Ratio of a Word, the Maximum Size of a Lexicon*. J. Theor. Biol. (2000) 179-189.

[68] Pinchuk, A. O. and Vysotskii, V. I. *Long-Range Intermolecular Interaction between Broken DNA Fragments*. Phys. Rev. E. 63 (3) 031904-1 (2001).

[69] Ramdsen, J. J. and Vohradsky, J. *Zipf-like Behavior in Procaryotic Protein Expression*. Phys. Rev. E. 58 (6) 7777 (1998).

[70] Resendis, O. & García-Colín, L. S. *Application of the Theory of Stochastic Processes to the Configuration of Biological Systems*. Physica A 290 (2001) 203-210.

[71] Román, R.R. et al. *Complejidad de Secuencias Simbólicas. Aplicación a Secuencias de ADN*. Revista Española de Física. 13 (2) 33 (1999).

[72] Scala, A. et al. *Thermodynamically Important Contacts in Folding of Model Proteins*. Phys. Rev. E. 63 (3) 032901-1 (2001).

[73] Seybold, Paul G. *Why Are There Four Bases in DNA?*. Int. J. Quantum Chem. Quantum Biology Symp. (3) 39-43 (1976).

[74] Som, A. et al. *Codon distributions in DNA*. Phys. Rev. E. 63 (5) 051908-1 (2001).

[75] Stanley, H.E. et al. *Scaling Features of Noncoding DNA*. Physica A 273 (1999) 1-18.

[76] Stanley, H.E. et al. *Scale Invariance and Universality: Organizing Principles in Complex Systems*. Physica A 281 (2000) 60-68.

[77] Stanley, H.E. *Exotic Statistical Physics: Aplications to Biology, Medicine, and Economics*. Physica A 285 (2000) 1-17.

[78] Stanley, H.R.R. et al. *Clustering of Identical Oligomers in Coding and Noncoding DNA Sequences*. Journal of Biomolecular Structure & Dynamics. 17 (1) 79-87 (1999).

[79] Stauffer, D. *Grand Unification of Exotic Statistical Physics. Physica A* 285 (2000) 121-126.

[80] Stella, A.L. *Scaling in DNA Denaturation Models: Excluded Volume, Stiffness, and a Block Copolymer Network Picture.* Personal communication.

[81] Thoumine, O. and Meister, J. J. *A Probabilistic Model for Ligand-Cytoskeleton Transmembrane Adhesion: Predicting the Behavior of Microspheres on the Surface of Migrating Cells. J. Theor. Biol.* (2000) 381-392.

[82] Torres, J.L. *Optimality in Human and Genetic Messages. Lectures on thermodynamics and Statistical Mechanics. Proceedings of the XXIII Winter meeting on statistical physics.* Costas, M., R. Rodríguez, A.L. Benavides (ed.) p.250-265. World Scientific, Singapore (1995)

[83] Voss, R. F. *Comment on Linguistic Features of Noncoding DNA sequences. Phys. Rev. Lett.* 76 (11) 1978 (1996).

[84] Wanant, S. and Quon, M. J. *Insulin Receptor binding Kinetics: Modeling and Simulation Studies. J. Theor. Biol.* (2000) 355-364.

[85] Wolfsberg, T.G. et al. *Guide to the Draft Human Genome. Nature.* Vol 409/15 Feb/ (2001) 824-859.

[86] Yanai, I. et al. Predictions of Gene Family Distributions in Microbial Genomes: Evolution by Gene Duplication and Modification. Phys. Rev. Lett. 85 (12) 2641 (2000).

[87] Zu-Guo, Yu et al. *Multifractal Characterisation of Length Sequences of Coding and Noncoding Segments in a Complete Genome. Physica A* 301 (2001) 351-361.

PÁGINAS WEB

[88] Alometría:

1. http://www.km0.com/catedra/fisi52.htm
2. http://www.km0.com/catedra/fisi53.htm
3. http://www.km0.com/catedra/fisi55.htm
4. http://www.km0.com/catedra/fisi56.htm
5. http://alexia.lis.uiuc.edu/ standrfr/zipf.html

[89] Genética:

1. http://www.microbiologia.com.ar/genetica/transformacion.html
2. http://www.microbiologia.com.ar/antimicrobianos/anti-proteínas.html
3. http://www.arrakis.es/ ibrabida/viginsercion.html

[90] Ley de Zipf: http://alexia_lis_uiuc_edu/standrfr/zipf.html

[91] *Mycoplasma pneumoniae*:

1. http://www.zmbh.uni-heidelberg.de/M_pneumoniae
2. http://www.zmbh.uni-heidelberg.de/M_pneumoniae/genome/Introduction.html: Introduction to *Mycoplasma pneumoniae* .
3. http://www.zmbh.uni-heidelberg.de/M_pneumoniae/: Understanding the biology of a 'minimal cell' .
4. Cebrat, S. and Dudek, M. *Symmetry in DNA Domains of Yeast Chromosomes*. http://smorfland_microb.uni.wroc.pl/dnasymmetry/dnasymmetry.html

[92] *Drosophila melanogaster*:

1. http://www.euchromatin.org/E02.htm : The Fourth Chromosome of *Drosophila melanogaster*: Interspersed Euchromatic and Heterochromatic Domains . Published in: Proc. Natl. Acad. Sci. USA, vol. 97, no. 10, pp 5345 (May 9, 2000).
2. http://ceolas.org/VL/fly/intro.html: A quick and simple introduction to *Drosophila melanogaster* .

[93] *Felis catus* : http://ww2.isys.ca/drazen/cats.htm:

[94] *Pinus thunbergii* :

1. http://www.nobleplants.com/classnotes/fall/fallprofiles/pinusthunbergii.htm: *Pinus thunbergii* - Japanese black pine
2. http://www.floridata.com/ref/P/pinu_thu.cfm:*Pinus thunbergii*

CONGRESOS Y CONFERENCIAS

[95] *The Linguistics of Biology and the Biology of Language*. Conferencistas Varios Centro Internacional de Ciencias Centro de Investigación sobre Fijación de Nitrógeno. UNAM Cuernavaca, Morelos 23-27 de Marzo 1998

[96] *De la Genética a la Genómica* Francisco Bolívar Zapata Universidad Autónoma del Estado de Morelos El Colegio Nacional 13 de Junio de 2001

[97] *México en el Umbral de la Era Genómica* Guillermo Soberón Academia Nacional de Medicina El Colegio Nacional Fundación Mexicana para la Salud 20 de Abril de 2001

[98] *Maquinaria Molecular para la degradación de Proteínas* Robert Huber, Premio Nobel de Química 1988 Colegio Nacional 28 de Marzo 2002

TESIS

[99] Nikolay V. Dokholyan, *Applications of Statistical Mechanics to Biological Macromolecules* Boston University Graduate School of Arts and Sciences Doctoral These 1999