11281

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

ANÁLISIS MATEMÁTICO DE SECUENCIAS DE AMINOÁCIDOS PROVENIENTES DE GENOMAS BACTERIANOS USANDO DIFERENTES CÓDIGOS GENÉTICOS

TESIS
QUE PARA OBTENER EL TÍTULO DE
DOCTOR EN CIENCIAS BIOMÉDICAS
PRESENTA

JOSÉ ANTONIO GARCÍA MACÍAS

TUTOR: DR. MARCO ANTONIO JOSÉ VALENZUELA

SEPTIEMBRE 2004

Índice general

Índice de figuras

Índice de cuadros

Prefacio

Presentación

Este trabajo se desarrolló bajo la dirección del Dr. Marco Antonio José Valenzuela y las cotutorías del Dr. Germinal Cocho Gil (Instituto de Física) y del Dr. Octavio Miramontes Vidal (Instituto de Física), en las instalaciones del Grupo de Biología Teórica del Instituto de Investigaciones Biomédicas de la UNAM, así como en las instalaciones de la Dirección de Posgrado e Investigación de la Universidad La Salle.

Los principales resultados de este trabajo, fueron aceptados para su publicación en la revista *Physica A*, artículo que se reproduce en el Apéndice A. Por otra parte, como parte de las actividades académicas desarrolladas en el programa de doctorado, se escribió un artículo de revisión para su posible publicación en la *Revista Mexicana de Física* (en proceso de revisión), el cual se reproduce en el Apéndice B.

Todas las abreviaturas empleadas en esta tesis siguen las convenciones internacionales de la *International Union of Biochemistry and Molecular Biology* y de la *International Union of Pure and Applied Physics*, con lo que se mantiene la consistencia de las abreviaturas en el texto en español e inglés.

En la versión electrónica de esta tesis, el texto coloreado (diferente al negro), indica ligas a las cuales se puede accesar mediante el *mouse*.

Sobre la naturaleza de la tesis

Esta tesis es fruto de un trabajo interdisciplinario que abarca las áreas de las ciencias biomédicas, la física y las matemáticas. En este sentido, se incluye una pequeña introducción al material biológico en el Apéndice C y, por otra parte, la revisión del Apéndice B sirve de introducción a las metodologías de la física aplicada.

Las preguntas de investigación que se abordan en este trabajo son fundamentalmente del área de la física aplicada a sistemas biológicos, motivo por el cual, aunque los resultados aquí presentados tengan probables implicaciones biológicas (discutidas en el Capítulo 6), las conclusiones obtenidas

sean particulares y de tipo técnico.

La naturaleza primordial de esta tesis es de tipo descriptiva. Los resultados aquí presentados han orientado al planteamiento de nuevos experimentos cuya naturaleza, se pretende, sea de tipo explicativa.

Agradecimientos

Agradezco a Marco V. José y Tzipe Govezensky, por sus aportaciones para el desarrollo de la presente tesis; a Alejandro Flores, por el desarrollo del programa *Code*; a Samantha Alvarez, Michelle Robles e Isabel Meza por su colaboración en la obtención de series de distancias; a Germinal Cocho y Octavio Miramontes, por su valiosa participación como miembros del comité tutoral; a los miembros del jurado: Carmen Gómez, Julio Collado, Lorenzo Segovia, Gabriel del Río, Mauricio Salcedo y Pedro Miramontes, por sus comentarios para el mejoramiento de la tesis; a los profesores que han contribuido en mi proceso de formación: Pedro Uriel, Bernardo Ayala, Araceli Sánchez, José Domingo Méndez, Armando Isibasi, Vianey Ortiz y Jorge X. Velasco; y a la Dirección de Posgrado e Investigación de la ULSA por haber financiado gran parte de este proyecto.

José A. García
México, D.F., septiembre 2004

Capítulo 1

Resumen

Nature gives to every time and season some beauties of its own;
and from morning to night,
as from the cradle to the grave,
is but a succession of changes so gentle and
easy that we can scarcely mark their progress.
- Charles Dickens -

There is something in the depths of our souls
which tells us that the world may be more than a mere combination of events
- Louis Pasteur -

Se han identificado una serie de características matemáticas relacionadas tanto a la fractalidad, como al contenido de información en secuencias codificantes del ácido desoxirribonucleico (DNA). En particular, sobresale la periodicidad módulo 3 (cada 3 posiciones) que se observa en series de distancias provenientes de secuencias codificantes. Esta periodicidad pudiera ser producto de la naturaleza del código genético, el cual se considera prácticamente universal.

Dado que el código genético es degenerado, los efectos de éste sólo pueden ser detectados en secuencias de aminoácidos (aa). En este sentido, se requiere saber, en primera instancia, si las características matemáticas detectadas en secuencias de DNA se observan también en secuencias de aa; y, en segunda instancia, si estas propiedades se deben a la naturaleza del código genético.

En el presente trabajo se analizaron las propiedades estadísticas de la distribución de aa en genomas cromosomales microbianos (MCGs), con el objetivo de probar la posible modificación de éstas, cuando se usan diferentes códigos genéticos para el proceso de traducción.

Se obtuvieron secuencias de aa de distintos MCGs usando ya sea el código genético universal, o bien otros cuatro códigos genéticos generados que incluyen a un código propuesto para el mundo del RNA. Como control negativo, se utilizaron secuencias de aa provenientes de la traducción (usando el código genético universal) de versiones de los MCGs previamente permutadas. A partir de estas secuencias, se obtuvieron series de distancia entre aa idénticos a lo largo de la secuencia. Sobre cada una de estas series, se aplicó el análisis sobre la función de autocorrelación (ACF), el análisis insesgado de fluctuaciones (DFA) y el método de máxima entropía (MEM) y se probó la hipótesis nula sobre la igualdad de las estadísticas analizadas en las series con los diferentes códigos, contra el código universal (control positivo).

Se encontraron mayor cantidad de diferencias significativas, en los análisis realizados sobre el genoma codificante de *Borrelia burgdorferi*, en comparación con los análisis realizados sobre el genoma completo de la misma bacteria. Inclusive, en este último caso, prácticamente se perdió la diferencia entre el control positivo y el control negativo. De los MCGs analizados, el de *Bacillus subtilis* presentó mayor robustez a las perturbaciones, mientras que el de *Haemophilus influenzae* fue el más frágil. En general los resultados del ACF y del MEM estuvieron correlacionados en los diferentes MCGs. Aunque el DFA fue de utilidad en el análisis del genoma codificante de *Borrelia burgdorferi*, en el caso de MCGs completos, esta técnica resultó ser menos sensible para detectar los cambios entre el código universal y los otros códigos.

Se concluye que el código genético universal contiene, para los MCGs estudiados y contra los códigos alternos propuestos, los valores más grandes referentes al ACF, al DFA y al MEM, aunque se encontró una amplia gama de respuestas.

Capítulo 2

Marco Teórico

DNA sequences shall encode biological instructions.
The nucleotides shall act as letters of Life's alphabet.
Three letters in a row shall be a codon.
A codon shall be like unto a word.
So DNA shall be like unto a text with
a thousand or ten thousand words.[1]

The history of life on earth is written in the cells and molecules of existing organisms.
- Christian de Duve -

[1] Book of Biology. The Bible according to Einstein. New York, USA. Jupiter Scientific Publishing Company (1997).

2.1. Análisis matemáticos sobre secuencias de DNA

El DNA puede considerarse como un alfabeto de cuatro letras, con el cual es posible escribir cierto número de palabras (genes). En este sentido, es posible aplicar ciertas técnicas empleadas en el estudio de los lenguajes al DNA. Asimismo, el DNA puede ser transformado en una serie de distancia y ser estudiado con técnicas propias del procesamiento de señales, con el objetivo de diferenciarla de secuencias aleatorias. En el Apéndice B se presenta una revisión sobre esta sección, donde pueden observarse gráficamente estas propiedades. A continuación se presentan brevemente algunos de estos análisis, que sirvieron como antecedente y motivación para el desarrollo de esta tesis.

2.1.1. Mapeo del DNA

Para poder aplicar técnicas de procesamiento de señales al análisis de secuencias de DNA, es necesario que estas últimas sean transformadas primero, en una serie consecutiva de números que, de alguna manera, represente a la cadena original.

Cuando los valores numéricos corresponden a observaciones sobre el mismo evento, pero a diferentes tiempos, se le denomina *serie de tiempo* (por ejemplo, datos de una cinética). En el caso del DNA pueden obtenerse datos numéricos al recorrer una de las hebras e ir registrando una determinada observación. A este procedimiento se le conoce como *mapeo* del DNA. En un mapeo del DNA, en lugar de considerar al tiempo como la variable que explica (variable independiente), se utiliza la distancia, por lo que, propiamente, se analizan *series de distancia.*

Una de las estrategias más simple para realizar el mapeo del DNA, consiste en la transformación binaria de una secuencia de nucleótidos usando alguna de las siguientes convenciones (73):

- Purinas R (A y G), o pirimidinas Y (T y C)
- Débiles W (A y T), o fuertes S (C y G)
- Aminas M (A y C), o cetonas K (T y G)

Esto es, se convierten todos los caracteres que comparten alguna de las características mencionadas por uno y el resto por cero. Otra alternativa, propuesta por el grupo de Stanley, es la *caminata aleatoria* a lo largo de una secuencia de nucleótidos (101). Este método asume la presencia de un caminador teórico que recorre una cadena de DNA. El caminador comienza en la posición $n = 0$ y da un paso hacia arriba $[u(n) = +1]$, cada vez que ve una pirimidina, y un paso hacia abajo $[u(n) = -1]$, cada vez que ve una purina. De esta manera, la cadena original de DNA se convierte en una secuencia de -1 y 1. Usualmente se grafica la caminata acumulada $y(n)$ contra la posición n en la cadena. Uno de los resultados más importantes de la caminata aleatoria, es el hecho de encontrar regiones relativamente largas, que son ricas en cierto tipo de nucléotidos (por

ejemplo, purinas), por que la distribución de las bases, a lo largo de una cadena de DNA no es uniforme (como sería el caso en una secuencia aleatoria) (101); dada esta característica, se dice que las secuencias de DNA presentan *parches* (55).

Algunas de las técnicas que se usan para el análisis de señales, dependen de estadísticas como la media o la varianza, las cuales no son constantes en las series obtenidas por la técnica de la caminata aleatoria (debido a la presencia de parches). Con el objetivo de evitar el efecto del parchado en el DNA y de preservar mayor información (la cual se pierde en las representaciones binarias), se ha propuesto la generación de series de distancia entre $n-$tuplas ($n =$ mono, du, tri, etc.) a lo largo de la secuencia (2, 87, 93). En esta estrategia, primero se localiza la posición de la $n-$tupla en la secuencia y, posteriormente, se calcula el número de bases nitrogenadas que se encuentran entre ellas. La serie de distancias se construye por la concatenación de todas las distancias calculadas.

2.1.2. Periodicidad

Partiendo de series de distancia de representaciones binarias del DNA, Shepherd encontró periodicidad módulo 3 (valores máximos en la posición 3 y múltiplos de 3) que podría ser atribuible al código genético (93). Shepherd utilizó la convención *RY* para transformar la cadena de DNA y, posteriormente, generó series de distancia entre diferentes combinaciones de $n-$tuplas. De esta manera, reportó que la secuencia *RNY* fue la más prevalente, hipotetizando que no sólo podría ser ancestro del código genético actual (28, 60), sino también que debieran existir vestigios de esta secuencia en los genomas actuales (94). De hecho, Jukes reportó más tarde, que efectivamente es posible detectar la prevalencia del patrón *RNY* en genomas actuales (51).

Posteriormente, Herzel & Große usando la función de información mutua sobre series de distancias, generadas por la probabilidad condicionada de encontrar al símbolo A_i y k caracteres después al símbolo A_j, concluyeron que la distribución no uniforme en el uso de codones del código genético en secuencias codificantes, era la responsable de las oscilaciones con período 3 (44).

Hay que hacer notar que la periodicidad detectada en la distribución de las series de distancia entre tripletes, se conserva al permutar la cadena original por tripletes y se pierde al permutar por monopletes.

Esta periodicidad módulo 3 se hace más evidente cuando se realiza el análisis sobre la función de autocorrelación (ACF) en secuencias codificantes de DNA (3, 4). En cambio, este patrón oscilatorio no ha sido observado ni en secuencias no codificantes, ni en secuencias aleatorias.

El ACF se explica, formalmente, en la sección 4.2.1. La autocorrelación es una operación matemática que se aplica sobre una serie de números (un vector), para determinar que tan relacionada se encuentra con sigo misma, pero desplazada n posiciones. A estos desplazamientos se les conoce como *retardos*. Cuando esta operación se aplica sobre varios retardos ($\tau = 1, 2, 3, \ldots, n$), entonces se habla de la función de autocorrelación. Al aplicar esta función sobre series de tiempo es posible, por ejemplo, detectar la presencia de ciclos en fenómenos biológicos como las epidemias.

2.1.3. Dinámica fractal y correlaciones a largo alcance

A principios de la década de los 80's, Benoit Mandelbrot (65) hizo la observación de que diversas estructuras encontradas en la naturaleza, como las nubes, los continentes o los árboles (por mencionar sólo algunos), presentan una geometría compleja a la que denominó *fractal*, por que se podría caracterizar por una dimensión no entera (por tanto fraccionaria). Hay que hacer mención que también es posible encontrar propiedades fractales en fenómenos naturales como el ritmo cardiaco y la temperatura.

Una característica importante de las estructuras con geometría fractal es que presentan *autosimilitud*, es decir, es posible identificar al mismo patrón en diferentes escalas (como en la ramificación de los árboles). Una consecuencia de este escalamiento es la presencia de correlaciones a largo alcance. En el caso concreto de los fenómenos biológicos de naturaleza fractal, existe un escalamiento anormal (es decir, no constante), que sigue una ley de potencias (ver más adelante) y que puede detectarse por al menos dos órdenes de magnitud (107), aunque algunos autores han reportado hasta cuatro órdenes (114, 115, 116).

Existen varios fenómenos tanto naturales, como de la acción humana, incluyendo el tamaño de las ciudades, los ingresos, las magnitudes de los terremotos, etc., que se encuentran distribuidos siguiendo una ley de potencias. En este tipo de fenómenos, los eventos pequeños son comunes, mientras que los eventos grandes son raros.

Una función, $f(x)$, sigue una ley de potencias cuando la variable independiente, x, tiene un exponente, es decir, se encuentra elevada a una potencia: $y = Ax^{\alpha}$ donde la potencia α puede ser cualquier número y A es la constante de normalización. Cuando $\alpha > 0$, la función crece, mientras que cuando $\alpha < 0$, la función decae. En este último caso se habla de una ley de potencias inversa. Usualmente las funciones que siguen leyes de potencia se grafican en escala logarítmica donde puede observarse una relación lineal entre las dos variables. La pendiente de la recta obtenida corresponde al exponente de la función graficada.

Los fenómenos que presentan una dinámica de leyes de potencias exhiben autosimilitud y, por tanto, una dinámica fractal con correlaciones a largo alcance (101). Dado que a partir de secuencias de DNA pueden generarse series de distancia, es factible determinar la presencia de correlaciones a largo alcance en estas secuencias.

En este sentido, partiendo del modelo de la caminata aleatoria, Peng, *et al.* han aplicado diversas técnicas relacionadas para analizar la presencia de correlaciones a largo alcance en secuencias de DNA (84, 85, 99, 100).

El primer acercamiento que utilizó este grupo, fue la determinación de la raíz de las fluctuaciones medias cuadradas, $F(l)$ alrededor del promedio del desplazamiento, definida como:

$$F(n) = \sqrt{\overline{[\Delta y(n) - \overline{\Delta y(n)}]^2}} \tag{2.1}$$

donde $\Delta y(n) = y(n_0 + n) - y(n_0)$ y las barras sobre las variables indican la media aritmética para todas las n posiciones en el gen.

Se entiende por fluctuaciones, a las variaciones a nivel microscópico que se presentan en los sistemas macroscópicos, por lo que las fluctuaciones en física son equivalentes al concepto de varianza en estadística. En la caso de $F(l)$, se pueden presentar dos posibles escenarios: (a) para procesos estocásticos o con sólo correlaciones locales, $F(n) \approx n^{1/2}$; y (b) para procesos con correlaciones libres de escala (correlaciones a largo alcance), las fluctuaciones se describen por la ley de potencias: $F(n) \approx n^{\alpha}$, con $\alpha \neq 1/2$. A través de este método (conocido como *min-max*), Peng, *et al.* encontraron correlaciones a largo alcance en regiones no codificantes, en contraste con las regiones codificantes, donde $\alpha \approx 0.5$ (84).

Estos resultados, han sido ampliamente discutidos e, inclusive, algunos autores señalan que no existe diferencia entre la dinámica de regiones codificantes y la de no codificantes (109). Una de las críticas principales a estos resultados, hace referencia al hecho de que se presenta heterogeneidad en las series de distancia obtenidas por la caminata aleatoria del DNA, es decir, se presentan secuencias relativamente grandes ricas en ya sea purinas o pirimidinas. Debido a este fenómeno de mosaico, métodos como el ACF y la $F(l)$ pierden validez ya que se basan en medias, las cuales, a su vez, cambian a lo largo de la serie de distancia (55).

Posteriormente, Peng, *et al.*, mejoraron su técnica para evitar el efecto mosaico del DNA, mediante la eliminación de la tendencia de las fluctuaciones sobre diferentes tamaños de ventana (85). A este técnica se le llama, en esta tesis, análisis insesgado de fluctuaciones (DFA). Esta técnica se explica formalmente en la sección 4.2.2.

En una serie de tiempo, las fluctuaciones crecen al aumentar la longitud de la ventana donde se calculan éstas. Este hecho nos permite asumir una relación de leyes de potencia entre los factores de escalamiento: $M_y = M_x^{\alpha}$, donde M_x y M_y son los factores de escalamiento y α es el exponente de escalamiento. El escalamiento en la variable independiente (usualmente, el tiempo) puede calcularse por:

$$M_x = \frac{n_2}{n_1} \tag{2.2}$$

donde n_1 y n_2 son dos longitudes diferentes de ventana. Una manera de calcular el escalamiento en la variable dependiente es mediante:

$$M_y = \frac{s_2}{s_1} \tag{2.3}$$

donde s_1 y s_2 son las desviaciones estándar de las distribuciones de los correspondientes histogramas. Con lo anterior, el exponente de escalamiento α, puede calcularse por:

$$\alpha = \frac{\ln M_y}{\ln M_x} = \frac{\ln s_2 - \ln s_1}{\ln n_2 - \ln n_1} \tag{2.4}$$

que es la pendiente obtenida al aplicar un modelo de regresión lineal sobre los logaritmos de los datos originales.

El DFA ha sido aplicado para estudiar diversos fenómenos, en el caso de biología, los trabajos más importantes han sido sobre secuencias de DNA (101) y sobre ritmos cardiacos (40). En el caso del DNA, los resultados de Peng, *et al.* fueron consistentes con sus resultados anteriores, es decir, las correlaciones a largo alcance sólo fueron detectadas para las secuencias no codificantes (85). En el caso de las secuencias codificantes, se detectó un cambio en la pendiente con una $\alpha = 0.51$ para la primera parte de la curva (equivalente a un proceso aleatorio).

En contraste con estos resultados, otros autores han encontrado correlaciones a largo alcance en secuencias codificantes. Esta diferencia se debe, fundamentalmente, a modificaciones en la metodología para generar la serie de distancia. En particular, Voss (109) y Sousa Vieira (108) calcularon el espectro de potencias proveniente de series no binarias, mientras que Mohanty & Narayana Rao obtuvieron los momentos factoriales de series que representaban el exceso de purinas sobre pirimidinas (74).

2.1.4. Entropía y contenido de información

La entropía es una medida de probabilidad estadística. Una situación que es altamente probable, genera mucha entropía, mientras que aquélla que es muy rara, tendrá poca entropía. En el caso de los gases, en los procesos espontáneos, las moléculas aumentan su velocidad por lo que puede asociarse la entropía al desorden molecular.

Uno de los métodos que se han utilizado para el análisis de lenguajes, ha sido la determinación del contenido de información, el cual puede calcularse por el logaritmo base 2 del número de mensajes posibles (86). Esta medida de información se conoce como entropía de Shannon (91) y está dada por la siguiente ecuación:

$$H_n = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{2.5}$$

donde p_i es la probabilidad (frecuencia relativa) del evento i.

El término de entropía se debe a que la Ec. 2.5 se encuentra relacionada con ciertas fórmulas de la mecánica estadística donde p_i es la probabilidad de que un sistema se encuentre en la celda i de su espacio de fase (91).

En el caso de eventos con dos posibles resultados, con probabilidades p y $q = 1 - p$, la entropía de Shannon alcanza su valor máximo cuando $p = q$. Este resultado puede generalizarse para cualquier n, número de probabilidades, con lo que H_n es máximo cuando todas las probabilidades p_i, son iguales, lo que representa la situación más impredecible (aleatoria).

Usando estos conceptos, varios autores han empleado la entropía de Shannon para determinar la redundancia en secuencias de DNA encontrando mayor contenido de información en secuencias codificantes, que en secuencias no codificantes o aleatorias (32, 45, 68).

Existen diversas formas de calcular la entropía, además de la de Shannon. En el caso de señales biológicas, como las relativas al clima, se ha utilizado el método de máxima entropía (MEM) (38). El MEM se explica formalmente en la sección 4.2.3.

El MEM se basa en el principio de parsimonia (también llamado *razuradora* de Occam[2]) que, en resumen, establece que, dados un conjunto de explicaciones igualmente posibles para un determinado fenómeno, la explicación más simple es la correcta. En modelación estadística el principio de parsimonia propone que (11):

- Los modelos deben tener el menor número de parámetros posibles.
- Se deben preferir los modelos lineales a los no lineales.
- Se deben preferir los modelos basados en pocas suposiciones, que los basados en muchas.
- Los modelos deben simplificarse hasta un mínimo adecuado.
- Se deben preferir las explicaciones sencillas a las complicadas.

En el trabajo de Shannon, la entropía es una función de una distribución de probabilidad determinada (91). Por el contrario, el principio de máxima entropía propone que una distribución de probabilidades puede determinarse a partir de la entropía, es decir, establece que la única distribución de probabilidades que representa o codifica el estado de información, es aquélla que maximiza el índice de incertidumbre (entropía), permaneciendo, a su vez, consistente con la información disponible (47).

Históricamente el MEM tiene su antecedente matemático en el principio de insuficiencia de Laplace, que establece que cuando no existe razón para hacer lo contrario, se deben asignar las mismas probabilidades a todos los eventos posibles, es decir, la distribución de probabilidades debe ser uniforme. Como se mencionó anteriormente, en este caso se obtiene el valor de máxima entropía. Si al aplicar este principio se encuentra experimentalmente que la distribución de probabilidades no es uniforme, se utiliza la nueva información para definir restricciones y se continúa con el mismo principio para el resto de la información.

2.1.5. Simetría bilateral inversa

Partiendo de series de distancias entre tripletes idénticos, Sánchez & José graficaron la posición acumulada a lo largo de un cromosoma bacteriano, encontrando que esta distribución muestra un

[2]Conocida así, porque se dice que *razuraba* las explicaciones hasta su mínima expresión posible.

cambio de pendiente justo a la mitad del cromosoma (correspondiente al punto de replicación) (87). Cuando se grafica la posición acumulada del triplete complementario (por ejemplo, ATG con CAT) se observa la formación de estructuras romboidales, lo que implica que la densidad de la distribución de un triplete (por ejemplo, ATG) en la primera mitad del cromosoma bacteriano, es idéntica a la de su complementario (en este caso, CAT) en la otra mitad del mismo cromosoma. A esta propiedad en los cromosomas bacterianos se le ha denominado simetría bilateral inversa (IBS). En la Fig. 2.1 se presentan las gráficas de posición acumulada (CPP) para los tripletes ATG y CAT en tres genomas. En el panel (a) se presenta la CPP para un genoma microbiano que contiene sólo regiones codificantes, mientras que en el panel (b) se presenta el mismo genoma, pero permutado (revuelto). Finalmente en el panel (c) se presenta la CPP para un genoma sintético obtenido por la concatenación aleatoria de un millón de bases nitrogenadas.

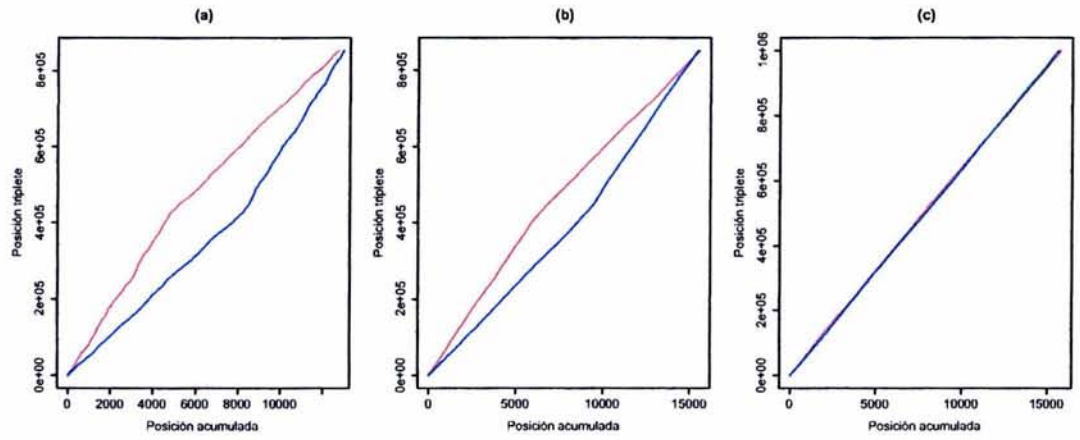

Figura 2.1. Gráficas de posición acumulada (CPP). En rojo se grafica la CPP para el triplete ATG mientras quen en azul se grafica la CPP para su complementario CAT. (a) Genoma codificante de *Borrelia burgdorferi*; (b) genoma permutado de *Borrelia burgdorferi*; (c) genoma sintético.

Como puede observarse en la Fig. 2.1 tanto el genoma original (codificante), como el permutado, producen estructuras romboidales que implican un cambio en la distribución de los tripletes en la segunda mitad del cromosoma. Este tipo de estructuras fueron descritas por Sánchez & José para varios tripletes y genomas microbianos (87). Hay que hacer notar que, en general, se observan trazos más finos para los genomas permutados que para los genomas originales. La distribución de los tripletes en el genoma sintético es uniforme, por lo que no se observan las estructuras romboidales.

Dada la degeneración en el código genético, no es posible detectar la IBS en secuencias de aa (carece de sentido, el concepto de complementaridad de bases). Sin embargo, es factible que las otras propiedades, anteriormente descritas, pudieran encontrarse en secuencias de aa, en cuyo caso, se podría evaluar el papel que juega el código genético en estas propiedades. En las siguientes secciones, se presentan las generalidades correspondientes al código genético, con el objetivo de completar el marco teórico de la presente tesis.

2.2. Antecedentes históricos del código genético

2.2.1. Código diamante

El primer código fue propuesto por el físico George Gamow en 1954[3] (36). Gamow propuso un código al que denominó "diamante", en el cual los aa se ensamblaban directamente usando al DNA como plantilla, de tal manera que cuatro bases que conformaban una cavidad tipo diamante entre ellas, determinaban un aa en particular. El diamante consideraba tres bases de una hebra del DNA y una cuarta que se encontraba apareada con la del centro; dada la restricción de la paridad, sólo 20 combinaciones eran posibles (ver Fig. 2.2).

El código diamante además tenía la propiedad de considerar que los tripletes se podrían traslapar, lo que podría maximizar la densidad de información contenida en el DNA (ver panel (a) de la Fig. 2.3). Una desventaja de los códigos con traslapes, es que una mutación puntual debería alterar no sólo a un aa, sino también a otros dos aa vecinos. Posteriormente, Sydney Benner analizaría las secuencias proteínicas reportadas en su tiempo para descartar la posibilidad de la existencia de códigos con traslapes (8).

2.2.2. Códigos libres de comas

A finales de los 50's ya existía evidencia experimental sobre la posible participación del mRNA en el proceso de traducción. Francis Crick propuso la hipótesis de la "molécula adaptadora", en la cual se consideraba que los aa interactuarían con el mRNA mediante un adaptador. Dado que para esos tiempos, se consideraba que el código no podría contener traslapes, Crick sugirió un código *libre de comas*, en el cual se tendrían que descartar todos los tripletes que se prestaran a traslapes (e.g. AAA), con lo que restarían, precisamente, 20 posibles combinaciones, correspondientes a las moléculas adaptadoras *con significado* (12). Es decir, un código libre de comas se construye de tal manera que cuando dos codones con significado (con traducción) se encuentran juntos, entonces los tripletes obtenidos por el traslape entre ellos, se descartan y se consideran sin significado. En el panel (b) de la Fig. 2.3 se presenta un ejemplo de la construcción de un código libre de comas.

2.2.3. El código genético

En 1961 Marshall Nirenberg y Heinrich Matthaei publicaron sus estudios en los cuales utilizaron un sistema libre de células donde mRNA artificial puede usarse para la síntesis de proteínas (78). El primer mRNA utilizado fue poli-U, obteniéndose un polipéptido de poli-fenilalanina. Hay que hacer notar, que el triplete UUU se consideraba sin sentido en los códigos libres de comas. En los años posteriores, se siguieron asignando codones, hasta que el código genético quedo totalmente establecido para 1965 (43). El código genético universal se presenta en la Fig. 2.4.

[3]Gamow propuso la teoría del Big Bang en cosmología.

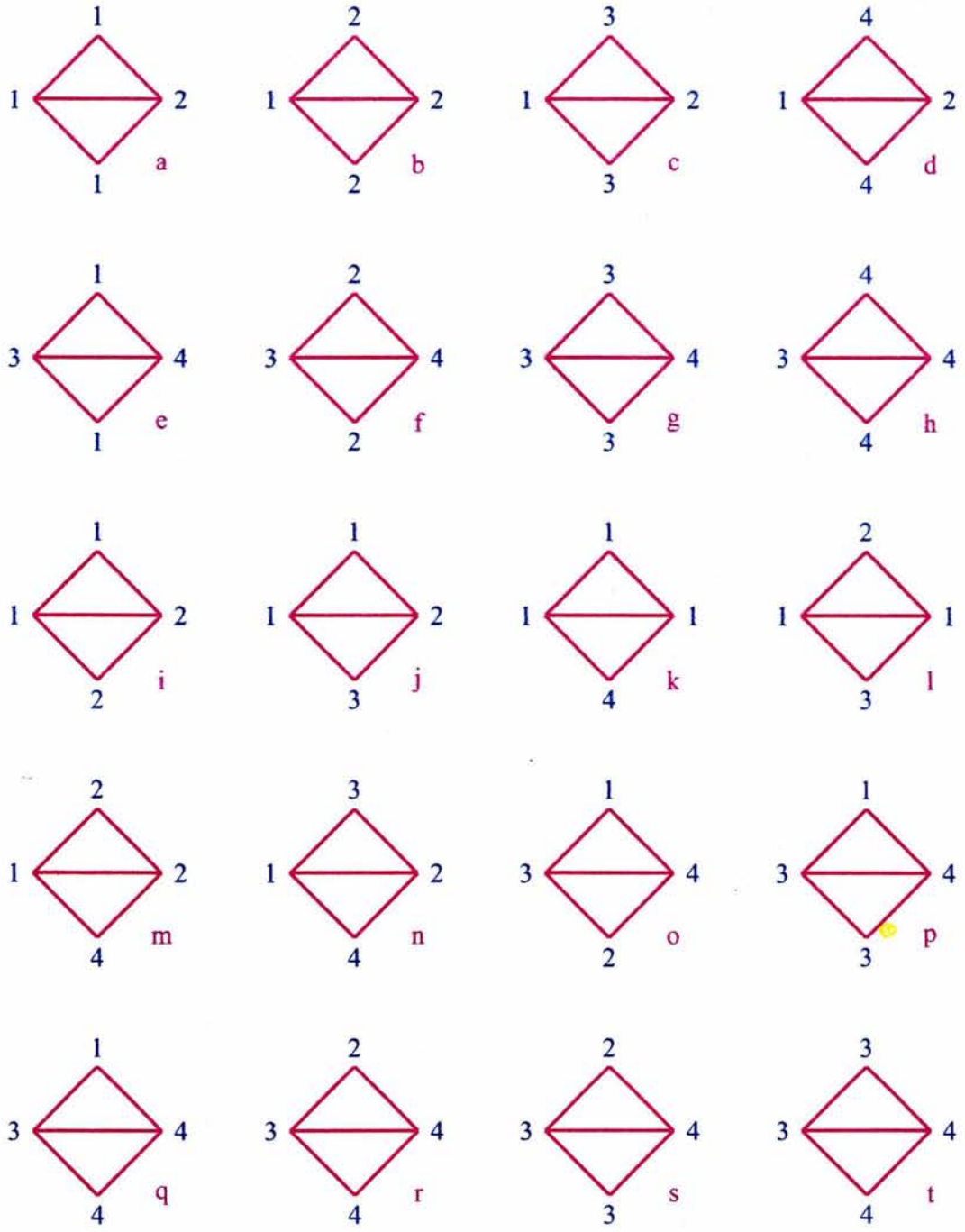

Figura 2.2. Código diamante propuesto por George Gamow (36). En este código las bases son designadas por números y los 20 codones por letras minúsculas. La línea horizontal indica que las bases se encuentran apareadas (son complementarias).

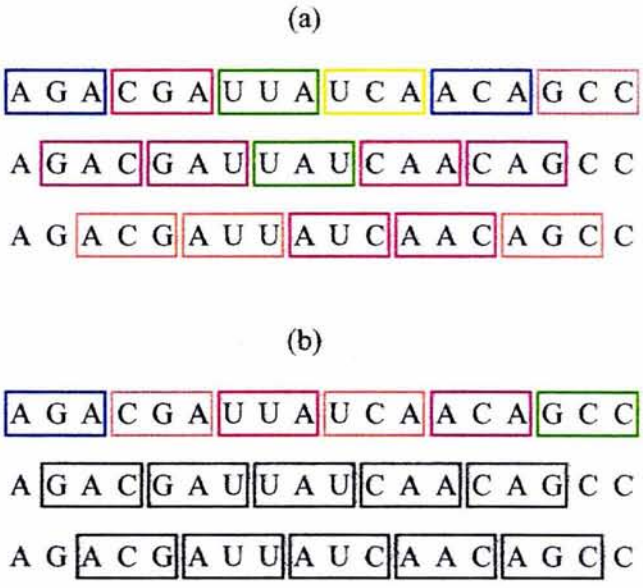

Figura 2.3. Comparación entre (a) códigos con traslapes y (b) códigos libres de comas. En una secuencia de 18 nucleótidos, un código con traslapes puede tener 16 codones, mientras que en uno libre de comas, sólo son posibles 6. Los cuadros con colores indican codones con significado, mientras que los cuadros negros indican codones sin significado. Esquema originalmente propuesto por Brian Hayes (43).

UUU UUC	Phe	UCU UCC	Ser	UAU UAC	Tyr	UGU UGC	Cys
UUA UUG	Leu	UCA UCG		UAA UAG	...	UGA UGG	... Trp
CUU CUC CUA CUG	Leu	CCU CCC CCA CCG	Pro	CAU CAC CAA CAG	His Gln	CGU CGC CGA CGG	Arg
AUU AUC AUA AUG	Ile Met	ACU ACC ACA ACG	Thr	AAU AAC AAA AAG	Asn Lys	AGU AGC AGA AGG	Ser Arg
GUU GUC GUA GUG	Val	GCU GCC GCA GCG	Ala	GAU GAC GAA GAG	Asp Glu	GGU GGC GGA GGG	Gly

1 codón
2 codones
3 codones
4 codones
5 codones
6 codones

Figura 2.4. Código genético universal. El codón de inicio es *AUG* y los codones de término se indican con puntos suspensivos.

2.3. Características del código genético

Pueden identificarse las siguientes características en el código genético:

1. Es universal. Aunque existen algunas excepciones, se ha encontrado que prácticamente el mismo código es usado por todas las especies. Este hecho hace pensar que el código genético debió establecerse temprano en la filogenia y que todos los organismos, provenimos de una célula ancestral común, que utilizó este código.
2. Hay convergencia. Cada codón codifica para un solo aminoácido.
3. Es degenerado. Un aminoácido puede ser codificado por más de un codón.
4. Presenta degeneración en la tercera base. El código está organizado de tal manera que, mutaciones en la tercera posición del triplete, usualmente no implican un cambio en el aa codificado, o bien, el cambio es por un aa con propiedades de polaridad semejante.
5. Contiene signos de puntuación. Para evitar los traslapes, el triplete AUG usualmente se emplea como codón de inicio (además de codificar para el aa metionina), mientras que los tripletes: UAG, UAA y UGA se emplean como codones de terminación.

2.4. El código *RNY* y el mundo del RNA

Una vez descifrado el código genético, se desecharon todos los modelos teóricos que previamente se habían formulado. Sin embargo, años más tarde, Crick propuso que probablemente los códigos libres de comas fueran antecedentes de un código primitivo, que posteriormente diera origen al código actual (15). Estas especulaciones llevaron a formular diferentes escenarios, que al ser enriquecidos con cierta evidencia experimental (10), llevaron a Gilbert a proponer al *mundo del RNA*, como un evento importante en el origen de la vida (ver Apéndice C) (39). A continuación se presenta una pequeña descripción de estas especulaciones.

Dentro de los patrones que garantizan una lectura sin traslapes (libre de comas), Crick propuso la secuencia de bases *RRY*, es decir, purina–purina–pirimidina para que fuera igual para todos los codones que especificaban un mensaje (15). Sin embargo, debido al exceso de purinas en esta secuencia, su estructura es termodinámicamente más inestable. Debido a esto, y a otros argumentos, Eigen & Schuster propusieron como alternativa de código primitivo, a uno que siguiera el patrón *RNY*, en donde *N*, representa cualquiera de los cuatro nucleótidos *A*, *U*, *G* o *C* (28).

Crick y cols. habían descartado este modelo porque presentaba como desventaja el que si la *N* fuese una pirimidina, el loop anticodón tenía que usar sus cinco nucleótidos centrales para formar pares de bases estables con el mensajero (15). Eigen & Schuster por su parte argumentan que este código da lugar a ocho aa por lo que uno puede excluir ciertas combinaciones que no satisfacen los

requisitos de estabilidad del complejo mensajero-peptidil-tRNA (28). Aparte de esto también mencionan que el código *RNY* también presenta las siguientes ventajas: carece de comas, es simétrico con respecto a las cadenas positiva y negativa, puede desarrollar simetría interna en una sola cadena y por lo tanto se permite la formación de estructuras secundarias (ver Fig. 2.5).

La propuesta de Eigen & Schuster es consistente con resultados experimentales, donde se observa la abundancia de 8 aa bajo condiciones que asemejan a la Tierra primitiva (72), así como por su amplia presencia en genes actuales (51, 93, 95). Asimismo, el patrón *RNY* ha sido considerado por Konecny y cols. como el código que pudo emplearse en el mundo del RNA (60). Las relaciones entre los diferentes codones en el patrón *RNY* se presentan en la Fig. 2.5.

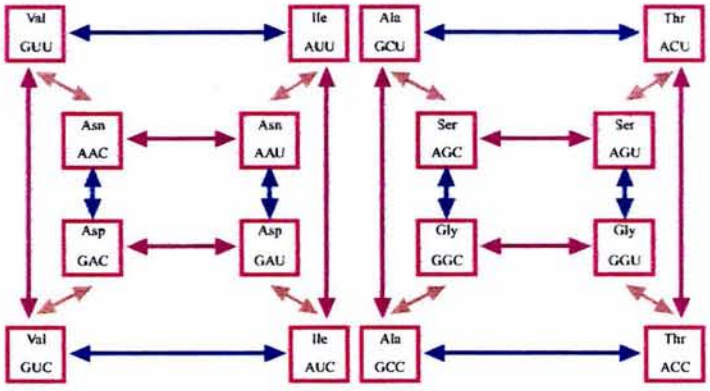

Figura 2.5. Relaciones entre codones en el patrón *RNY*. Esquema originalmente propuesto por Konecny, Schöniger y Hofacker (60). Las flechas de color azul indican mutaciones conservadoras; las flechas de color morado indican mutaciones silenciosas; las flechas de color café indican transformaciones sentido–antisentido (complementarios inversos).

2.5. Teorías sobre el origen del código genético

2.5.1. Teoría estereoquímica

Esta teoría sugiere que el código genético se originó a partir de las interacciones estereoquímicas entre codones o anticodones con los aa. La primera propuesta, dentro de esta corriente, fue la de Gamow con el código diamante (36), comentada en la sección 2.2.1.

Posteriormente, Melcher propuso modelos que establecían una correlación entre los aa y sus anticodones (69). La característica principal de estos modelos es la asociación entre los hidrógenos alifáticos de los aa con los electrones π de las bases nitrogenadas en el anticodón.

Recientemente, Yarus descubrió que hay interacción estereoselectiva entre el sitio de la guanosina, en el sitio catalítico de una molécula de RNA, y la arginina (122). Estos resultados podrían sugerir un papel importante de las interacciones estereoquímicas, en el origen del código genético (23).

2.5.2. Teorías fisicoquímicas y de reducción de la ambigüedad

La hipótesis fisicoquímica sugiere que el código genético se originó con base en las fuerzas que tienden a reducir las distancias fisicoquímicas entre los aa codificados por codones que difieren en sólo una base (23). Estas fuerzas han sido asociadas tanto con los efectos deletereos de las mutaciones (97), como en la presión selectiva que tiende a reducir los errores de traducción de los códigos genéticos primitivos (120).

Una hipótesis conceptualmente relacionada con la fisicoquímica, es la de la reducción de la ambigüedad. Esta hipótesis sugiere que grupos de codones relacionados fueron asignados a aa con estructura similar y que el código genético alcanzó su estructura actual a través de la reducción de la ambigüedad en la codificación entre grupos de aa (30, 119).

La hipótesis fisicoquímica está fundada en evidencia experimental que establece relaciones entre las propiedades fisicoquímicas de los aa y la organización del código genético, por lo que, se sugiere, que en ciertos estadios del origen del código genético, las propiedades fisicoquímicas de los aa pudieron haber jugado un papel muy importante en la estructuración y organización del código genético (23).

2.5.3. Teoría coevolutiva

Esta teoría, aunque tiene varios antecedentes (26, 77, 83), fue formalmente postulada por Wong (121) y sugiere que la estructura del sistema de codones es básicamente una impresión de las vías prebióticas que dan origen a los aa. En este sentido, el origen del código genético puede establecerse con base a las relaciones precursor–producto entre los aa disponibles en las vías metabólicas actuales. Es decir, en un principio en el código genético sólo se codificaron para aa precursores y, conforme fueron obteniéndose nuevos aa como productos, estos se fueron agregando gradualmente al código (23).

Esta teoría es la más favorecida actualmente, dado que además de contar con evidencia experimental (26, 83, 121), el mecanismo de la asignación de codones de aa precursores a aa producidos parece haber sido mediado por moléculas semejantes al tRNA (tRNA-*like*), en las cuales se llevaron a cabo las transformaciones correspondientes. En este sentido, esta hipótesis sugiere que fósiles moleculares, representativos de estas transformaciones podrían encontrarse actualmente y, de hecho, algunos de ellos han sido identificados (23).

Por otra parte, Jurka & Smith han sugerido que estructuras de vueltas β en las proteínas fueron

un objeto de selección en el medio prebiótico, e influenciaron tanto el origen del código genético como las vías metabólicas de los aa, dado que los aa precursores son los más abundantes en esas estructuras (52). Asimismo, Di Giulio ha encontrado que otro dominio importante en las proteínas primitivas fueron las hojas β (22), y dado que los aa precursores son también abundantes en estas estructuras, se sugiere que los mRNA primitivos codificaran para ellos (23). Finalmente, casi todos los aa reportados por Wong (121) siguen el patrón GNR o GNY que puede considerarse como un subconjunto del patrón RNY discutido en la sección 2.4.

Relacionada con estas posturas, Jukes (50) ha propuesto una teoría evolutiva del código genético. De acuerdo a esta teoría, el código primitivo consistía de 16 codones. Conforme aa adicionales fueron requeridos para la síntesis de proteínas, el código genético fue evolucionando hasta que el código universal actual emergió. Una vez presente, este código quedó congelado debido a que mutaciones en la asignación de codones hubieran provocado errores en muchas secuencias (13). Sin embargo, en el DNA mitocondrial, dado que el número de proteínas es considerablemente menor, las mutaciones pudieron presentarse con mayor frecuencia, dando origen a variaciones en el código original (75).

Sea cual fuere el origen del código, su universalidad sugiere que efectivamente quedó congelado. En este sentido, no se sabe si otros códigos alternos, no se observan en la naturaleza, porque no fueron seleccionados, o simplemente porque no fueron generados. En otras palabras, no se sabe si otros códigos pudieran haber sido tan exitosos como el que tenemos.

Con base en todos estos antecedentes, se decidió evaluar la posibilidad de que series de distancias provenientes de secuencias de aa heredaran algunas de las propiedades matemáticas observadas en cadenas de DNA, así como la posibilidad de que estas propiedades se modificaran al emplear diferentes códigos genéticos en el proceso de traducción.

Capítulo 3

Objetivos e hipótesis

How can the events in space and time
which take place within the spatial boundary of a living organism
be accounted for by physics and chemistry?
- Erwin Schrödinger -

Felix qui potuit rerum cognoscere causas
(Feliz aquél que conoce las causas de las cosas)
- Virgilio -

3.1. Objetivos

- Determinar si series de distancia entre aminoácidos idénticos, provenientes de la traducción de un genoma codificante, poseen propiedades matemáticas semejantes a las series que se obtienen de secuencias de nucleótidos.
- Seleccionar estadísticas representativas de estos análisis, que nos permitan distinguir entre secuencias codificantes y series de distancia obtenidas de la traducción de secuencias aleatorias.
- Comparar las estadísticas provenientes de la traducción de cromosomas microbianos usando el código genético universal, contra las estadísticas provenientes de la traducción de los mismos genomas usando códigos genéticos alternos.

3.2. Hipótesis

A continuación se presentan tanto las hipótesis de investigación (H_i) como las hipótesis nulas (H_0) correspondientes, que se prueban en esta tesis.

- H_i : Series de distancia entre aminoácidos idénticos provenientes de la traducción de genomas codificantes, pueden distinguirse matemáticamente de series provenientes de la traducción de secuencias aleatorias.
  - H_0 : Las estadísticas representativas de los análisis provenientes de las secuencias codificantes son indistinguibles de las estadísticas obtenidas de los análisis sobre secuencias aleatorias.
- H_i : Secuencias de aminoácidos provenientes de genomas microbianos usando el código genético universal, contienen estadísticas relativas al escalamiento (fractalidad) y al contenido de información, significativamente mayores a las obtenidas usando códigos genéticos alternos.
  - H_0 : Las estadísticas obtenidas usando el código genético universal son indistinguibles de las obtenidas usando códigos genéticos alternos.

Capítulo 4

Estrategia experimental

A mathematician is a blind man, in a dark room,
looking for a black cat, which isn't there
- Charles Darwin -

The purpose of models is not to fit the data
but to sharpen the questions
-Samuel Karlin -

4.1. Diseño experimental

El esquema del diseño experimental se presenta en la Fig. 4.1. En las subsecciones siguientes se describen con más detalle cada uno de los procesos. Brevemente, se obtuvieron secuencias teóricas de aa a partir de la traducción de genomas cromosomales microbianos (MCGs) usando diferentes códigos genéticos. A partir de estas secuencias, se generaron series de distancia entre caracteres idénticos. Se calcularon el ACF, DFA y MEM para cada una de las series y se obtuvieron estadísticas descriptivas de estos análisis. Finalmente se comparó la distribución de cada estadística obtenida con los diferentes códigos genéticos, contra la distribución de la estadística respectiva, obtenida usando el código genético universal.

4.1.1. Genomas microbianos

Con el objetivo de poder analizar con más detalle el efecto de diferentes códigos genéticos, sobre las propiedades matemáticas de secuencias de aa, se partió inicialmente de secuencias codificantes de *Borrelia burgdorferi*, concatenadas de acuerdo a su orden y orientación original en el genoma (87). A esta secuencia se le denomina, en este trabajo: *Borrelia burgdorferi*-OOO y contiene sólo secuencias codificantes (de ambas cadenas). Esta secuencia fue previamente generada y estudiada por Sánchez & José (87) debido a su tamaño (el genoma más pequeño reportado a la fecha de inicio de los estudios).

Por otra parte, se realizaron los mismos análisis matemáticos sobre series de distancias obtenidas de MCGs completos de 11 especies diferentes. A diferencia del caso anterior, estas secuencias contienen regiones intergénicas. Sólo se consideró la cadena líder del DNA para cada caso. Se seleccionaron estos genomas, debido a que provienen de especies representativas, bien estudiadas, de sus respectivos géneros. A continuación se en listan los genomas seleccionados, con una breve descripción y el motivo de su selección.

1. *Borrelia burgdorferi* NC 001318. Agente etiológico de la enfermedad Lyme, contiene un cromosoma linear de 910,725 bp que codifica para 853 genes (33). Este genoma se seleccionó como control, ya que era la única secuencia codificante con que se disponía.

2. *Bacillus halodurans* NC 002570. Bacteria alcalifílica con un genoma cromosomal de 4, 202, 353 bp y 4,066 secuencias potencialmente codificantes de proteínas (CDSs). La presencia de 112 transposones en su genoma sugieren un papel importante de la transferencia genética horizontal en el desarrollo evolutivo de esta bacteria (103). Esta bacteria se encuentra taxonómica y filogenéticamente muy relacionada con *Bacillus subtilis*, por lo que fue seleccionada como control de comparación, con el objetivo de verificar la consistencia de los análisis realizados en genomas similares.

3. *Bacillus subtilis* NC 000964. Contiene un genoma cromosomal de 4,214,810 bp con 4,100 genes codificantes de proteínas. Su genoma contiene al menos 10 profagos o remanentes de profagos que sugieren que la infección por estos virus jugó un papel importante en la

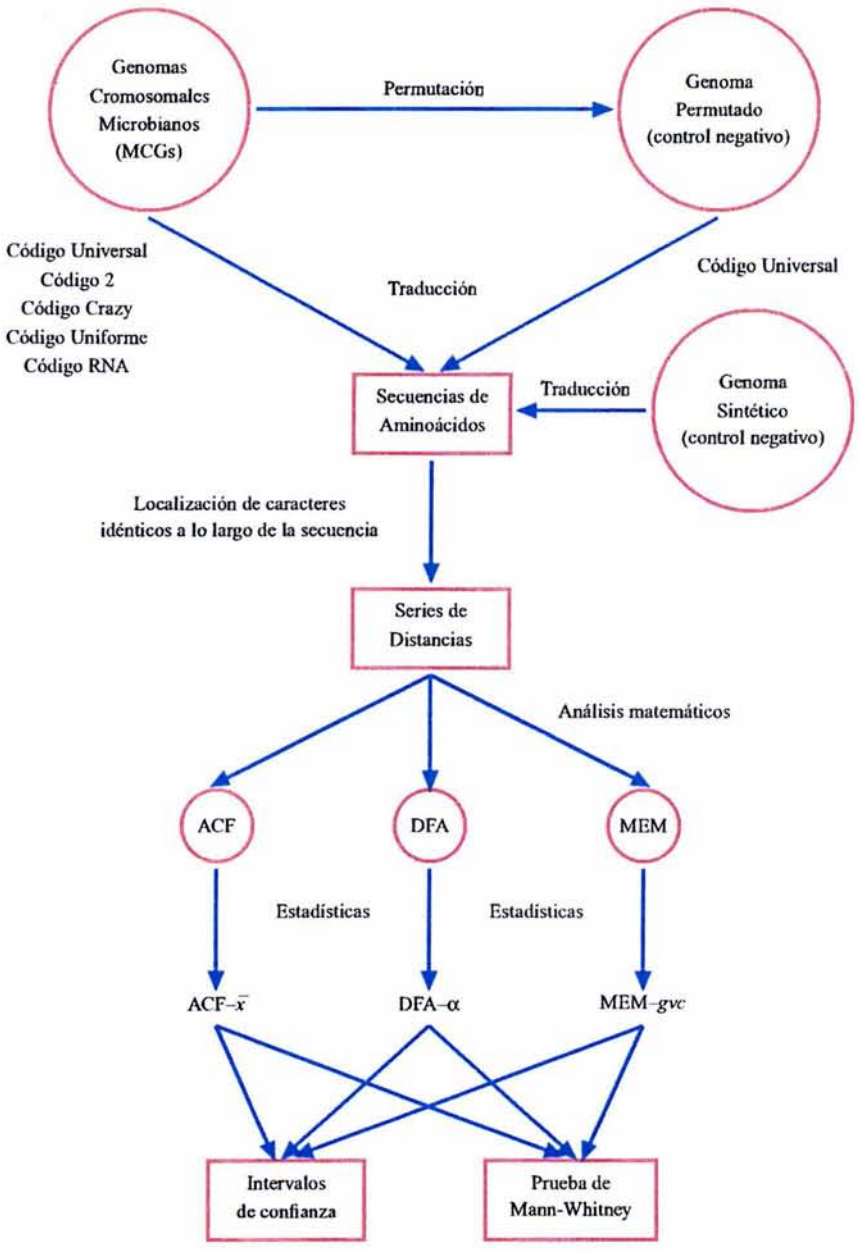

Figura 4.1. Esquema del diseño experimental.

transferencia genética horizontal (61). Se seleccionó este genoma debido a que corresponde a la bacteria Gram-positiva mejor caracterizada.

4. *Deinococcus radiodurans* R1 NC 001263. Es un organismo en el que todos los sistemas de reparación del DNA, exportación de DNA dañado, redundancia genética y recuperación a la desecación y hambruna se encuentran presentes en una célula. Su genoma completo está compuesto por dos cromosomas (2,648,638 y 412,348 bp), un megaplásmido (177,466 bp) y un plásmido pequeño (45,704) que dan un genoma total de 3,284, 156 bp (118). Para los análisis reportados en esta tesis, se utilizó solamente, la secuencia del cromosoma mayor. Esta bacteria fue seleccionada debido a su alta eficiencia en mecanismos de reparación que, en consecuencia, favorece la conservación de su genoma a lo largo del tiempo (salvo los eventos de transferencia horizontal).

5. *Escherichia coli* K-12 NC 000913. Contiene un genoma de 4,639,221 bp con 4288 genes codificantes de proteínas. Su genoma contiene secuencias de inserción, remanentes de fagos y otros tipos de parches que sugieren una alta plasticidad genómica adquirida a través de transferencia horizontal (6). Este genoma se seleccionó debido a que corresponde a la bacteria mejor caracterizada.

6. *Haemophilus influenzae* Rd NC 000907. Su genoma de 1,830,137 bp fue el primero en secuenciarse, iniciando, de esta manera, la era de la genómica (31). Esta bacteria Gram-negativa fue seleccionada, debido a que es un patógeno estricto donde se espera que los eventos de transferencia horizontal sean menos frecuentes.

7. *Methanococcus jannaschii* NC 000909. Arquea autotrófica con un genoma cromosomal de 1.66 Mbp y elementos extracromosomales de 58 y 16 kbp. Sus genes relacionados con la producción de energía, división celular y metabolismo son similares a los de las bacterias, mientras que aquéllos involucrados en la transcripción, traducción y replicación están más emparentados con los de eucariontes (9). Este genoma fue seleccionado, como control, dado que corresponde a la arquea mejor caracterizada.

8. *Streptococcus pneumoniae* NC 003028. Bacteria Gram-positiva causante de pneumonía, bacteremia, meningitis y otitis media. Contiene un genoma cromosomal de 2,160,837 bp con 2236 CDSs (104). Este patógeno fue seleccionado debido a su relativo acercamiento filogenético con los bacilos.

9. *Sulfolobus solfataricus* P2 NC 002754. Arquea con un genoma cromosomal de 2,992,245 bp que codifican para 2,977 proteínas. Su genoma muestra un alto grado de plasticidad debido a sus 200 secuencias de inserción, varios elementos móviles no-autónomos y evidencia de eventos de inserción mediados por integrasa (92). Este genoma fue seleccionado como otro representante de las arqueas, que tiene la particularidad de vivir en medios relativamente adversos, lo que podría favorecer su conservación genómica a lo largo del tiempo.

10. *Thermotoga maritima* NC 000853. Contiene un genoma cromosomal de 1,860,725 bp con 1,877 CDSs. Su genoma contiene genes con un alto porcentaje de homología con genes de arqueas (24 %). Su organización genómica sugiere transferencia genética lateral entre arqueas y eubacterias termofílicas (76). Esta bacteria fue seleccionada debido a su capacidad de vivir en medios relativamente adversos, además de encontrarse relacionada con las arqueas.

11. *Xylella fastidiosa* 9a5c NC 002488. Bacteria causante de enferemedades severas en naranjos. Contiene un genoma cromosomal de 2,679,305 bp y dos plásmidos de 51,158 y 1,285 bp. De sus genes, por lo menos 83 provienen de bacteriófagos e incluyen genes asociados a virulencia provenientes de otras bacterias lo que provee evidencia directa de transferencia genética horizontal mediada por fagos (96). Esta bacteria se seleccionó debido a que es un patógeno donde se han identificado eventos de transferencia horizontal.

En la Fig. 4.2 se presenta la relación taxonómica entre los genomas seleccionados, basada en los archivos del NCBI (117). Asimismo, en la Fig. 4.3 se presenta la relación filogenética, basada en la alineación múltiple del gen que codifica para el rRNA 16S.

- Bacteria
  - Gammaproteobacteria
    - Enterobacteriaceae ⇝ *Escherichia coli*
    - Pasteurellaceae ⇝ *Haemophilus influenzae*
    - Xanthomondaceae ⇝ *Xylella fastidiosa*
  - Bacilli
    - Bacillaceae
      - *Bacillus halodurans*
      - *Bacillus subtilis*
    - Streptococcaceae ⇝ *Streptococcus pneumoniae*
  - Deinococcus-Thermus
    - Deinococcaceae ⇝ *Deinococcus radiodurans*
  - Thermotogae
    - Thermotogaceae ⇝ *Thermotoga maritima*
  - Spirochaetes
    - Spirochaetaceae ⇝ *Borrelia burgdorferi*
- Archeae
  - Euryarchaceota
    - Methanococci
      - Methanocaldococcaceae ⇝ *Methanococcus jannaschii*
  - Crenarchaeota
    - Thermoprotei
      - Sulfolobaceae ⇝ *Sulfolobus solfataricus*

Figura 4.2. Relación taxonómica entre los genomas seleccionados.

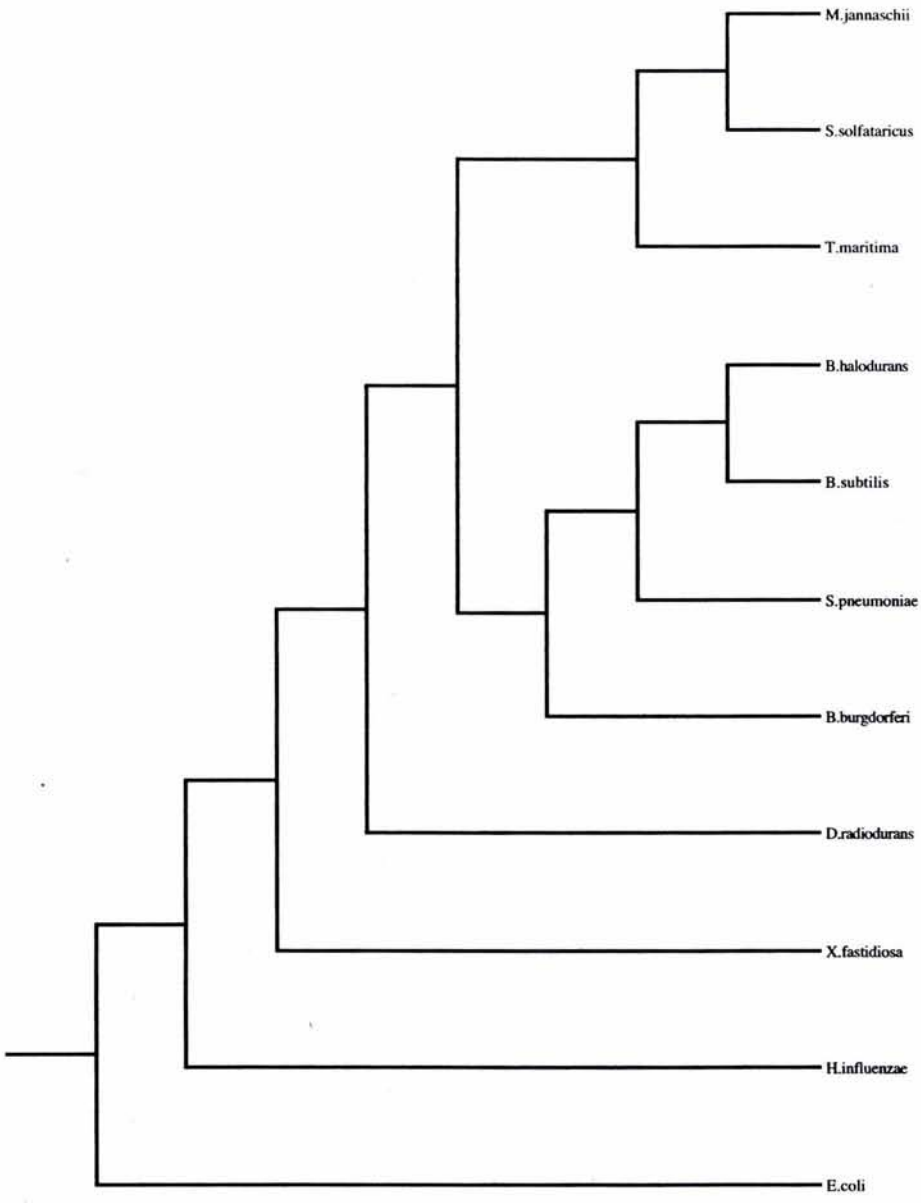

Figura 4.3. Relación filogenética entre los genomas seleccionados basada en la alineación múltiple del gen que codifica para el rRNA 16S. La alineación se realizó con el programa ClustalW, mientras que el dendrograma se obtuvo con los algoritmos del paquete Phylip y se graficó con el programa Treeview.

4.1.2. Genomas control

Como controles negativos, particulares a cada genoma, se consideraron las versiones permutadas (a nivel de nucleótidos) de los genomas originales. Estas secuencias se obtuvieron mediante un muestreo, sin reemplazo, del genoma original. Además, se consideró como un control negativo general, una secuencia sintética (de un millón de bases), generada por la concatenación aleatoria de las cuatro bases. Ambos genomas fueron traducidos a secuencias de aa usando el código genético universal.

4.1.3. Códigos genéticos

El código genético actual presenta una distribución característica de codones a aa que se presenta en la Fig. 4.4 (a). En el código *univ*, usualmente la base que ocupa la tercera posición del triplete, es menos importante, dado que una mutación en esta posición no siempre implica un cambio en el aa codificado (degeneración de la tercera base). Para efectos comparativos, en el Cuadro 4.1 se muestran las equivalencias de los códigos usados en este trabajo, usando la abreviación de una letra para cada aa.

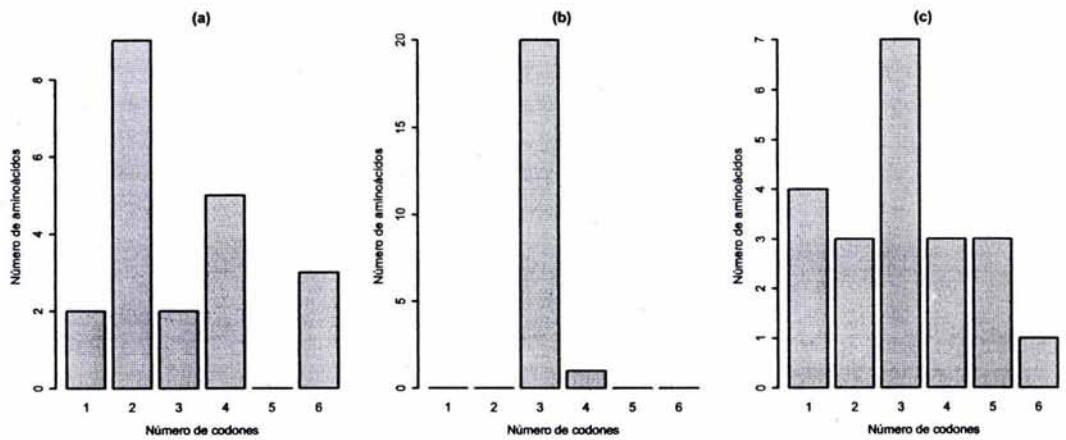

Figura 4.4. Gráficas de barras con distribución de codones-aa. (a) Distribución de codones-aa para los códigos *univ* y *cd2*. (b) Distribución de codones-aa para el código *unif*. (c) Distribución de codones-aa para el código *czy*.

Código 2 (*cd2*)

Para generar este código, los aa y señales de término se asignaron de manera aleatoria a cada triplete, conservando la distribución general del código universal (ver Fig. 4.4 (a)). Dado que la

Cuadro 4.1. Códigos genéticos*.

codón	*univ*	*cd2*	*czy*	*unif*	*rna*	codón	*univ*	*cd2*	*czy*	*unif*	*rna*
AAA	K	R	H	A	NA	GAA	E	Q	N	G	NA
AAC	N	P	A	A	N	GAC	D	V	F	D	D
AAG	K	K	N	P	NA	GAG	E	L	K	I	NA
AAT	N	K	M	F	N	GAT	D	Y	R	E	D
ACA	T	F	C	Y	NA	GCA	A	R	I	D	NA
ACC	T	C	T	R	T	GCC	A	L	S	L	A
ACG	T	M	C	G	NA	GCG	A	X	S	G	NA
ACT	T	D	X	P	T	GCT	A	A	M	T	A
AGA	R	M	G	I	NA	GGA	G	Q	Q	L	NA
AGC	S	N	H	N	S	GGC	G	C	I	X	G
AGG	R	D	Y	X	NA	GGG	G	T	C	I	NA
AGT	S	I	L	W	S	GGT	G	H	Y	K	G
ATA	I	T	D	Q	NA	GTA	V	M	A	W	NA
ATC	I	H	H	V	I	GTC	V	R	M	D	V
ATG	M	Q	S	E	NA	GTG	V	P	C	T	NA
ATT	I	R	T	V	I	GTT	V	R	Q	K	V
CAA	Q	K	E	H	NA	TAA	X	M	H	T	NA
CAC	H	F	N	Y	NA	TAC	Y	Y	L	H	NA
CAG	Q	F	L	Q	NA	TAG	X	E	P	X	NA
CAT	H	X	Q	F	NA	TAT	Y	M	F	X	NA
CCA	P	S	I	W	NA	TCA	S	D	K	R	NA
CCC	P	F	R	S	NA	TCC	S	A	R	C	NA
CCG	P	Q	P	E	NA	TCG	S	N	F	M	NA
CCT	P	Y	E	C	NA	TCT	S	R	M	K	NA
CGA	R	G	Y	Q	NA	TGA	X	M	A	A	NA
CGC	R	L	S	F	NA	TGC	C	E	V	C	NA
CGG	R	P	M	M	NA	TGG	W	K	X	P	NA
CGT	R	V	Q	H	NA	TGT	C	L	S	L	NA
CTA	L	K	P	Y	NA	TTA	L	P	W	M	NA
CTC	L	T	E	S	NA	TTC	F	W	R	N	NA
CTG	L	K	S	R	NA	TTG	L	I	Q	V	NA
CTT	L	T	C	S	NA	TTT	F	G	L	N	NA

*Para el codón de término se uso la letra "X". NA = no disponible.

asignación fue aleatoria, se pierde la degeneración en la tercera base. El código 2 se muestra en la Fig. 4.5.

Código Uniforme (*unif*)

En este código se asumió una distribución uniforme de los codones a aa, de tal manera, que todos los aa están codificados por 3 codones diferentes, y se consideran 4 codones de término (ver Fig. 4.4 (b)). El código uniforme se presenta en la Fig. 4.6.

Código Crazy (*czy*)

El código *czy* se generó mediante la asignación aleatoria de codones a una población de 64 caracteres obtenida mediante un muestreo con reemplazo de todos los aa y la señal de término (21 caracteres en total). De esta manera, se perdió la distribución original del código *univ* (ver Fig. 4.4 (c)). El código crazy se presenta en la Fig. 4.7.

UUU	Gly	UCU	Arg	UAU	Met	UGU	Leu
UUC	Trp	UCC	Ala	UAC	Tyr	UGC	Glu
UUA	Pro	UCA	Asp	UAA	Met	UGA	Met
UUG	Ile	UCG	Asn	UAG	Glu	UGG	Lys
CUU	Thr	CCU	Tyr	CAU		CGU	Val
CUC		CCC	Phe	CAC	Phe	CGC	Leu
CUA	Lys	CCA	Ser	CAA	Lys	CGA	Gly
CUG		CCG	Gln	CAG	Phe	CGG	Pro
AUU	Arg	ACU	Asp	AAU	Lys	AGU	Ile
AUC	His	ACC	Cys	AAC	Pro	AGC	Asn
AUA	Thr	ACA	Phe	AAA	Arg	AGA	Met
AUG	Gln	ACG	Met	AAG	Lys	AGG	Asp
GUU	Arg	GCU	Ala	GAU	Tyr	GGU	His
GUC		GCC	Leu	GAC	Val	GGC	Cys
GUA	Met	GCA	Arg	GAA	Gln	GGA	Gln
GUG	Pro	GCG		GAG	Leu	GGG	Thr

1 codón
2 codones
3 codones
4 codones
5 codones
6 codones

Figura 4.5. Código 2.

UUU	Asn	UCU	Lys	UAU		UGU	Leu
UUC		UCC	Cys	UAC	His	UGC	Cys
UUA	Met	UCA	Arg	UAA	Thr	UGA	Ala
UUG	Val	UCG	Met	UAG		UGG	Pro
CUU	Ser	CCU	Cys	CAU	Phe	CGU	His
CUC		CCC	Ser	CAC	Tyr	CGC	Phe
CUA	Tyr	CCA	Trp	CAA	His	CGA	Gln
CUG	Arg	CCG	Glu	CAG	Gln	CGG	Met
AUU	Val	ACU	Pro	AAU	Phe	AGU	Trp
AUC		ACC	Arg	AAC	Ala	AGC	Asn
AUA	Gln	ACA	Tyr	AAA		AGA	Ile
AUG	Glu	ACG	Gly	AAG	Pro	AGG	
GUU	Lys	GCU	Thr	GAU	Glu	GGU	Lys
GUC	Asp	GCC	Leu	GAC	Asp	GGC	
GUA	Trp	GCA	Asp	GAA	Gly	GGA	Leu
GUG	Thr	GCG	Gly	GAG	Ile	GGG	Ile

1 codón
2 codones
3 codones
4 codones
5 codones
6 codones

Figura 4.6. Código uniforme.

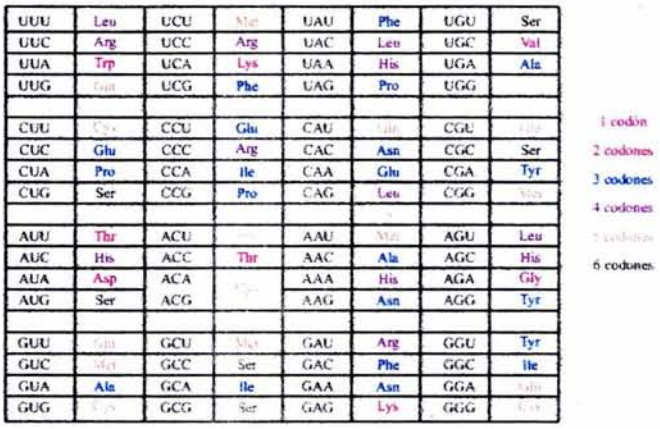

Figura 4.7. Código crazy.

Código del mundo del RNA (*rna*)

Este código se fundamenta en la hipótesis original de Gilbert sobre el mundo del RNA (39), usando el patrón *RNY* (purina – cualquier nucleótido – pirimidina) propuesto por Eigen & Schuster (28). Este código considera la presencia de sólo 16 tripletes que codifican para 8 aa diferentes y ha sido considerado como uno de los códigos más primitivos (60). El código del mundo del RNA que se uso en este trabajo, puede verse en la Fig. 2.5.

En el Cuadro 4.2 se presentan las relaciones entre los códigos utilizados, basada en la frecuencia de equivalencias entre ellos.

Cuadro 4.2. Frecuencia relativa de equivalencias entre códigos

	univ	*cd2*	*czy*	*unif*
cd2	3/64	1	2/64	3/64
czy	2/64	2/64	1	3/64
unif	4/64	3/64	3/64	1
rna	16/64	1/64	1/64	1/64

4.1.4. Series de distancias

Las secuencias de nucleótidos provenientes de cada genoma, fueron traducidas a secuencias de aa a partir del primer triplete ATG (codón de inicio en el código universal) en la secuencia original, y siguiendo ese marco de lectura. Para los codones de terminación se utilizó la letra "X". En el caso del código *rna*, se consideró sólo un marco de lectura, el cual se consideró que comenzaba a partir de la primera base en la secuencia; posteriormente se fue recorriendo la cadena por tripletes y se

conservaron solamente aquéllos considerados por la secuencia *RNY* (16 codones). Debido a esto, las series de distancia provenientes del código *rna* fueron más pequeñas que las provenientes de los otros códigos.

Para cada genoma se consideró como control positivo la traducción obtenida usando el código universal (*univ*) y como experimentos las secuencias de aa obtenidas usando los códigos alternos: código 2 (*cd*2), crazy (*czy*), uniforme (*unif*) y del mundo del RNA (*rna*). Como control negativo, se consideró la secuencia de aa proveniente de la traducción de una secuencia permutada del genoma original, usando el código universal. A esta secuencia se le llamó *ctl*. Asimismo, se consideró también como control negativo, una secuencia de aa proveniente de la traducción de una secuencia de nucleótidos teórica (de 1 millón de bases), generada por la permutación aleatoria de las cuatro bases del DNA, usando el código universal (genoma sintético). A esta secuencia se le denominó *rand*.

En lugar de usar el mapeo de las secuencias, siguiendo una caminata aleatoria (84), se calcularon las distancias en número de aa entre caracteres idénticos (21 caracteres en total). Esto es, primero se determinó la posición actual de cada caracter en la secuencia de aa y posteriormente, se calculó la distancia, en número de aa, entre caracteres idénticos.

El proceso de traducción así como la generación de las series de distancias se realizó con el programa *Code*. Este programa fue escrito *ex profeso* con estos objetivos en mente, se escribió en lenguaje C++, y se compiló en Mac OS X 10.2.8.

4.2. Análisis matemáticos

Con el objetivo de evaluar si una propiedad matemática determinada, era característica de secuencias codificantes, se calcularon diferentes funciones matemáticas sobre las series de distancia generadas por el código universal y por los controles negativos. De los análisis inicialmente seleccionados, se obtuvieron estadísticas representativas de la población, que nos permitieran comparar entre los dos tipos de secuencias (codificantes y aleatorias). Los tres análisis que tuvieron la suficiente sensibilidad para distinguir entre las secuencias, son los que se presentan a continuación.

4.2.1. Análisis de la función de autocorrelación (ACF)

El ACF permite probar la hipótesis nula sobre la independencia entre datos que forman parte de una serie de tiempo. Sea x una serie de tiempo y x_τ la misma serie, con un retardo de τ posiciones con respecto a x. El ACF calcula la correlación que existe entre x y x_τ mediante la fórmula de la Ec. 4.1, donde *SS* es la suma de cuadrados, definida en la misma fórmula, y las variables con barras

indican la media aritmética de la variable.

$$ACF = \frac{SSxx_\tau}{\sqrt{SSxSSx_\tau}} = \frac{\sum(x-\bar{x})(x_\tau-\bar{x_\tau})}{\sqrt{\sum(x-\bar{x})^2\sum(x_\tau-\bar{x_\tau})^2}} \tag{4.1}$$

En equivalencia con el coeficiente de correlación de Pearson (82), un valor de $ACF = 1$ indica una autocorrelación positiva del 100 %, un valor de $ACF = -1$ indica una autocorrelación negativa del 100 % y un valor de $ACF = 0$ indica independencia entre los datos de la serie, y por tanto se puede concluir que la serie sigue un comportamiento equivalente al del ruido blanco gaussiano (11, 54).

El ACF se calculó mediante el programa *R* version 1.8 para MacOSX (11, 46), que puede obtenerse de manera gratuita de la página web del CRAN (http://cran.r-project.org/). El ACF se aplicó sobre todas las series de distancias generadas con los diferentes códigos y los diferentes genomas para $1 \leq \tau \leq 38$ retardos.

4.2.2. Análisis insesgado de fluctuaciones (DFA)

El DFA se basa en un análisis modificado de las caminatas aleatorias para determinar las propiedades de correlación intrínseca de un sistema dinámico, al cual se le han eliminado posibles características tendenciales (sesgos) de origen externo a los datos (84, 85).

Brevemente, la serie de tiempo a ser analizada (con N muestras) se integra. Posteriormente, la serie integrada se divide en cajas equivalentes de tamaño n. En cada caja de longitud n, se ajusta una recta por mínimos cuadrados (que representa la tendencia de la caja). La coordenada y de los segmentos de la recta se denomina $y_n(k)$. Luego, se elimina la tendencia de la serie integrada $y(k)$ mediante la substracción de la tendencia local $y_n(k)$, en cada caja. La raíz cuadrada de las medias de la fluctuación, de esta serie, se calcula por la Ec. 4.2.

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k)-y_n(k)]^2} \tag{4.2}$$

El cómputo de la Ec. 4.2 se repite para todas las escalas de tiempo (tamaños de caja) para caracterizar la relación entre $F(n)$, el promedio de las fluctuaciones, como una función del tamaño de caja. Típicamente, $F(n)$ incrementa con el tamaño de caja n. Una relación lineal en una gráfica doble logarítmica (log-log) indica la presencia de escalamiento por leyes de potencias (fractales). Bajo estas circunstancias, las fluctuaciones pueden ser caracterizadas por el exponente de escalamiento α, que se obtiene de la pendiente que relaciona al log $F(n)$ con el log n.

Cuando el exponente de escalamiento, $\alpha = 0.5$, entonces se asume un comportamiento equivalente al del ruido blanco gaussiano. En cambio, cuando $0.5 < \alpha \leq 1$, se puede afirmar que la serie

de tiempo presenta correlaciones a largo alcance.

El DFA se calculó mediante el programa *dfa*, el cual se obtuvo de compilar el programa original *dfa.c* en Mac OS X 10.2.8. El programa *dfa.c* se puede obtener de manera gratuita de la página web: http://www.physionet.org/physiotools/dfa/ A los datos generados se les aplicó una regresión lineal usando mínimos cuadrados, para obtener el valor de la pendiente, que corresponde al exponente de escalamiento, α. El ajuste del modelo lineal se realizó con el programa *R* (11, 46).

4.2.3. Método de máxima entropía (MEM)

Procesos ya sea determinísticos o estocásticos pueden, en principio, caracterizarse por una función de la frecuencia f, en lugar del tiempo t. A esta función $S(f)$ se le conoce como espectro de potencias (37, 38). De tal manera que un proceso aleatorio, presenta un espectro suave y continuo, que indica que todas las frecuencias en una banda determinada, se excitan por dicho proceso. Por otra parte, un proceso periódico o cuasi-periódico presenta un espectro con número finito (usualmente pequeño) de lineas en el dominio de las frecuencias. Entre estos dos extremos podemos encontrar procesos con múltiples picos, característicos de procesos determinísticos pero caóticos.

El MEM se basa en aproximar una serie de tiempo con un proceso auto-regresivo (AR) lineal de orden M, AR(M). Brevemente, dada una serie de tiempo $X(t) : t = 1,\ldots,N$ que se asume es generada por un proceso estacionario con media cero y varianza σ^2, se obtienen los $M'+1$ estimados de los coeficientes de autocorrelación $\hat{\phi}_X(j) : j = 0,\ldots,M'$ mediante la Ec. 4.3.

$$\hat{\phi}_X(j) = \frac{1}{N+1-j}\sum_{t=1}^{N-j} X(t)X(t+j) \tag{4.3}$$

En ausencia de un conocimiento previo acerca del proceso que genera la serie $X(t)$, M' es arbitraria y tiene que ser optimizada. El objetivo de la Ec. 4.3 es estimar la densidad espectral $\hat{S}_X$ que está asociada con el proceso más aleatorio o menos predecible que contiene los mismos coeficientes de autocorrelación $\hat{\phi}$. En términos de la teoría de la información de Shannon (91), esto corresponde al concepto de máxima entropía.

La densidad espectral S_X de un proceso AR verdadero con coeficientes $\hat{a}_j : j = 0,\ldots,M$ está dada por:

$$\hat{S}_X(f) = \frac{a_0}{|1+\sum_{j=1}^{M}\hat{a}_j e^{2\pi i j f}|^2} \tag{4.4}$$

donde a_0 es la varianza del ruido residual de la ecuación del proceso AR (37, 38).

El MEM se calculó para todas las series de distancia, considerando un intervalo de muestreo igual a 1, un orden del MEM igual a 40 y 256 frecuencias muestreadas. El MEM se calculó usando

el programa *SSA-MTM Toolkit* (38), el cual se puede obtener de manera gratuita de la página web: http://www.atmos.ucla.edu/tcd/ssa/

4.3. Análisis estadísticos

Al aplicar los análisis matemáticos descritos en la sección anterior, se obtienen como resultado vectores que representan al fenómeno estudiado. Con el objetivo de tener un sólo valor (dato escalar), representativo del análisis aplicado para cada serie, se calcularon diferentes estadísticas, las cuales se mencionan a continuación.

1. Media aritmética de los primeros 38 retardos del ACF (ACF-$\bar{x}$). Se seleccionó esta estadística, debido a que los coeficientes de autocorrelación permanecieron relativamente constantes a lo largo de los primeros 38 retardos analizados (ver Fig. 5.4).

2. Exponente de escalamiento del DFA (DFA-α). Como se mencionó anteriormente, al realizar el DFA se obtienen dos vectores, cuya relación es lineal en la escala logarítmica (ver Fig. 5.6). La pendiente de esta recta corresponde al exponente de escalamiento y representa las posibles correlaciones a largo alcance para cada serie de distancia.

3. Coeficiente de variación geométrica (gvc) del MEM (MEM-gvc). Dadas las variaciones obtenidas en el espectro de potencias obtenido al aplicar el MEM (ver Fig. 5.7), se optó por una estadística de variación, en lugar de una de tendencia central. La ventaja de considerar al gvc en lugar del coeficiente de variación tradicional (aritmético), es que se reduce considerablemente el efecto de los valores extremos. El gvc se puede calcular mediante la siguiente ecuación:

$$gvc = \frac{\sigma(x)}{\tilde{x}} = \frac{\sigma(x)}{\sqrt[n]{\Pi x}} \tag{4.5}$$

donde $\tilde{x}$ es la media geométrica del vector x, definida en la misma ecuación y $\sigma(x)$ es la desviación estándar del vector x.

Debido a que existen diferencias en el uso de codones en cada organismo, algunos aa se encuentran más expresados que otros, dependiendo del código genético empleado en la traducción. Con el objetivo de poder determinar la posible diferencia entre los resultados obtenidos entre el genoma codificante y los genomas aleatorios, así como la posible diferencia entre el código universal y los códigos alternos, se determinaron los intervalos de confianza al 95 % (C.I.) para cada estadística y se calculó la probabilidad de que las distribuciones analizadas fueran de la misma población que la universal. En las subsecciones siguientes, se describen estos análisis.

4.3.1. Intervalos de confianza (C.I.)

Los C.I. nos permiten establecer con un 95 % de seguridad, la probabilidad de que una estadística determinada se encuentre entre el rango calculado. Cuando no existen traslapes entre los C.I. de dos muestras analizadas, se puede afirmar que pertenecen a poblaciones estadísticamente diferentes.

Una vez calculadas las estadísticas para cada serie de distancias, se determinaron los C.I. para todas las estadísticas calculadas. Dado que las series presentaron una distribución que no sigue un comportamiento gaussiano (ver Fig. 5.1), los C.I. se determinaron mediante una técnica no-paramétrica en la cual se realizaron 1000 muestreos con reemplazo de tamaño $n = 21$. Debido al teorema del límite central, la distribución de las medias de estas muestras es normal (11), por lo cual se obtuvieron de esta nueva población los cuantiles 0.025 y 0.975, que corresponden a los límites del 95 %. A esta técnica se le conoce como *bootstrap* (11, 27).

4.3.2. Prueba de Wilcoxon-Mann-Whitney (WMW)

Esta es una prueba que no requiere del empleo de estadísticas (como media y varianza) estimadas a partir de los datos (prueba no-paramétrica) y que nos permite determinar la probabilidad p, que dos muestras analizadas provengan de la misma población (66). Se consideraron diferencias estadísticamente significativas aquéllas que tuvieran valores de $p < 0.05$ (una diferencia significativa); $p < 0.01$ (dos diferencias significativas) o $p < 0.001$ (tres diferencias significativas).

Capítulo 5

Resultados

It is through science that we prove,
but through intuition that we discover
- Henri Poincare -

In mathematics you don't understand things,
you just get used to them
- John von Neumann -

5.1. *Borrelia burgdorferi-OOO*

Para el genoma codificante de *Borrelia burgdorferi*, la distribución de la serie de distancias para cada aa en particular, siguió un patrón similar con todos los códigos probados. La Fig. 5.1 muestra la función de densidad de probabilidad (PDF) para el caso del aspartato. Las PDFs para varios aa, obtenidos con diferentes códigos genéticos, presentaron un patrón con un decaimiento oscilatorio (e.g. Fig. 5.1e). La única excepción a este comportamiento se observó con el código RNA, en donde ningún aa presentó oscilaciones.

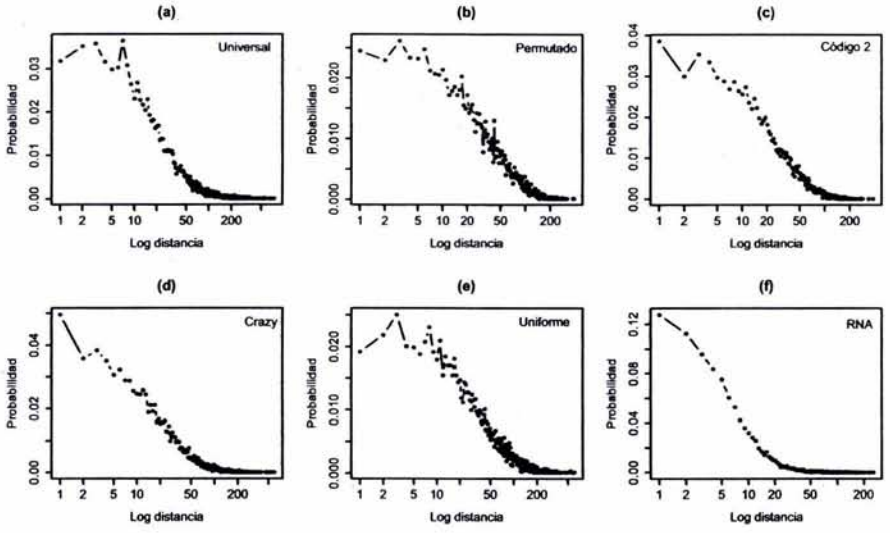

Figura 5.1. Función de densidad de probabilidad (PDF) para Asp en *Borrelia burgdorferi*-OOO. (a) PDF de Asp en código *univ.* (b) PDF de Asp en código *ctl.* (c) PDF de Asp en código *cd*2. (d) PDF de Asp en código *czy.* (e) PDF de Asp en código *unif.* (f) PDF de Asp en código *rna.*

Al graficar el logaritmo del número de veces que se detecta una determinada distancia contra el logaritmo de esa distancia, se pude detectar la presencia de leyes de potencia en las series generadas. En la Fig. 5.2 se muestran estas gráficas para las series de distancias de aspartato, obtenidas con los genomas control, mientras que en la Fig. 5.3 se muestran para las series obtenidas con los códigos alternos.

En el Cuadro 5.1 se presentan los valores estimados de la pendiente, y la ordenada, con sus respectivos errores estándar (s.e.), así como los de los coeficientes de regresión (r^2), para todos los casos analizados, aplicando un ajuste de regresión lineal por mínimos cuadrados. Como puede observarse, la relación lineal abarca mayor cantidad de datos en los casos del código universal y del mundo del RNA. Además, hay que hacer notar, que el error estándar del estimado de la ordenada para el código universal, es aproximadamente la mitad que el de los otros casos, lo cual puede interpretarse como un mejor ajuste del modelo lineal a los datos.

Existen diversos estudios donde se han reportado patrones periódicos en cadenas de DNA usando

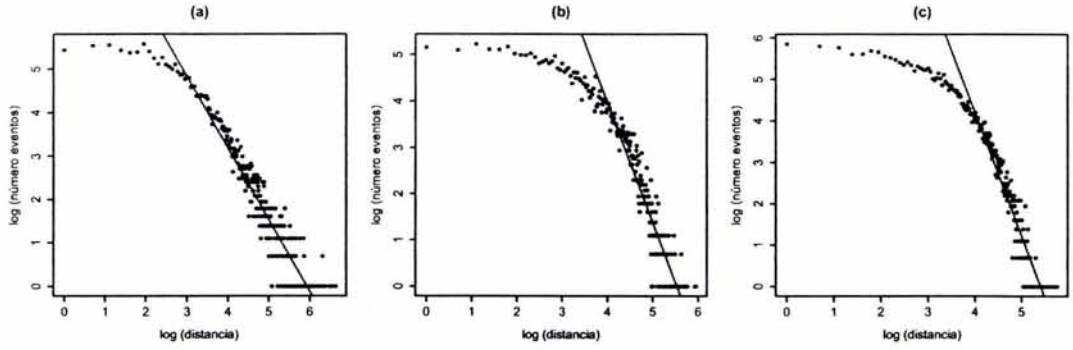

Figura 5.2. Ajustes lineales a las relaciones logarítmicas entre el número de eventos detectados, contra la distancia entre los caracteres, para las series de distancia de aspartato provenientes de la traducción de los genomas control de *Borrelia burgdorferi*-OOO. (a) Genoma original; (b) genoma permutado; (c) genoma sintético.

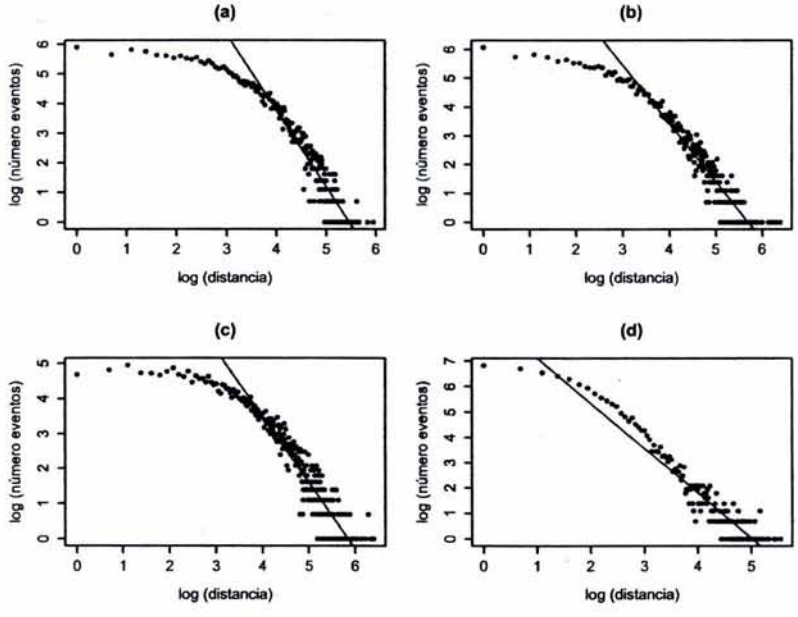

Figura 5.3. Ajustes lineales a las relaciones logarítmicas entre el número de eventos detectados, contra la distancia entre los caracteres, para las series de distancia de aspartato provenientes del genomas de *Borrelia burgdorferi*-OOO, usando diferentes códigos genéticos. (a) Código 2; (b) código crazy; (c) código uniforme; (d) código del mundo del RNA.

Cuadro 5.1. Estadísticas estimadas obtenidas de los modelos de regresión lineal aplicados en las Figs. 5.2 y 5.3.

Serie	Ordenada ± s.e.	Pendiente ± s.e.	r^2
univ	9.79 ± 0.19	-1.65 ± 0.04	0.87
ctl	14.33 ± 0.32	-2.59 ± 0.06	0.90
rand	16.13 ± 0.34	-2.98 ± 0.07	0.92
cd2	14.02 ± 0.33	-2.57 ± 0.007	0.89
czy	11.54 ± 0.25	-2.03 ± 0.05	0.88
unif	11.01 ± 0.24	-1.88 ± 0.05	0.86
rna	8.89 ± 0.31	-1.77 ± 0.07	0.84

el ACF (2, 4, 44, 55, 93). En este trabajo, también se utilizo el ACF aplicado a series de distancia entre aa, provenientes de la traducción del genoma codificante de *Borrelia burgdorferi*-OOO. En primera instancia se determinó la sensibilidad de la prueba, por su capacidad de discriminar entre el control positivo (código *univ*) y los controles negativos (*ctl* y *rand*). A diferencia del código universal, no se encontraron autocorrelaciones en los controles negativos, para todos los retardos probados (ver Fig. 5.4 y Cuadro 5.2) .

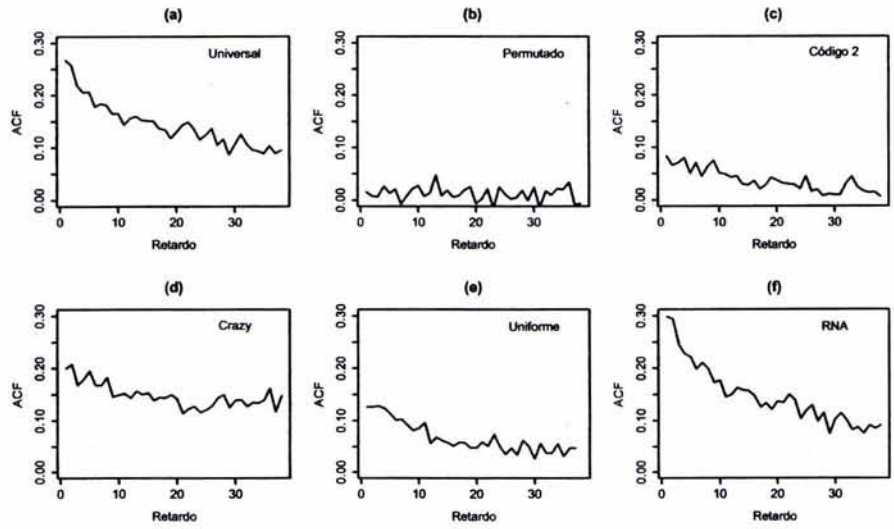

Figura 5.4. Función de autocorrelación para Asp en *Borrelia burgdorferi*-OOO. (a) ACF para Asp en código *univ*. (b) ACF para Asp en código *ctl*. (c) ACF para Asp en código *cd2*. (d) ACF para Asp en código *czy*. (e) ACF para Asp en código *unif*. (f) ACF para Asp en código *rna*.

Como puede observarse en el Cuadro 5.2, hay una clara diferencia entre el código universal contra los dos controles negativos. Una vez establecida la sensibilidad de la prueba, se aplicó el ACF sobre las series de distancia provenientes de la traducción con los códigos alternos. Se encontraron diferencias con los códigos 2, crazy y uniforme mediante la prueba de WMW. La distribución de

Cuadro 5.2. Intervalos de confianza (C.I.) y relación de aa hidrofílicos/aa hidrofóbicos (w/l) para *Borrelia burgdorferi*-OOO.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)	w/l
Universal	0.09 – 0.12	0.72 – 0.75	0.87 – 1.14	1.05
Permutado	0.03 – 0.06***	0.56 – 0.59***	0.23 – 0.46***	1.02
Sintético	-1.5e-3 – 1.2e-7***	0.49 – 0.51***	0.07 – 0.09***	0.97
Código 2	0.07 – 0.09**	0.69 – 0.73*	0.61 – 0.84**	1.27
Crazy	0.06 – 0.10*	0.69 – 0.73*	0.62 – 0.90**	0.83
Uniforme	0.07 – 0.10*	0.71 – 0.74	0.67 – 0.89*	0.82
RNA	0.06 – 0.13	0.65 – 0.73*	0.51 – 1.11	1.17

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

las estadísticas para los códigos alternos, presentó, en general, valores más pequeños que los del universal, aunque se presentó cierto traslape en los intervalos de confianza (ver Cuadro 5.2 y Fig. 5.5a).

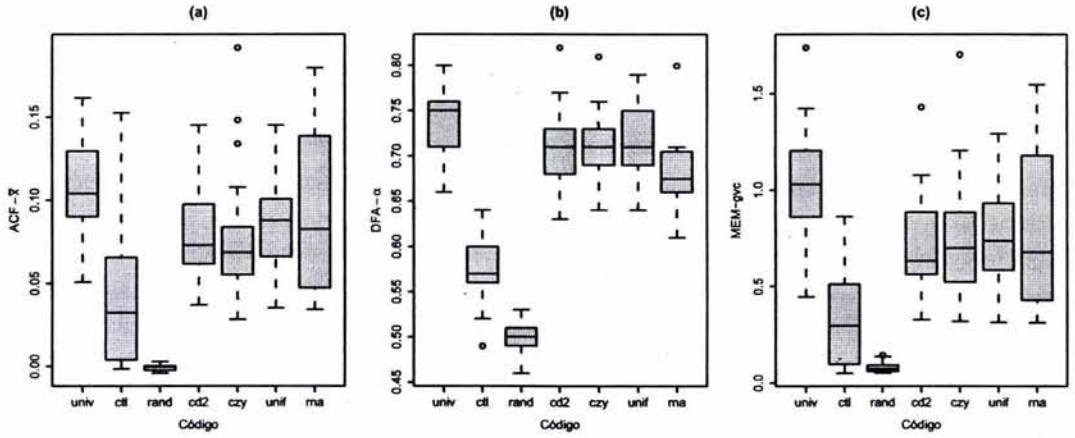

Figura 5.5. Distribución de estadísticas por códigos para *Borrelia burgdorferi*-OOO. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Por otra parte, también se han reportado correlaciones de largo alcance en el DNA (1, 74, 84, 85, 109). En este trabajo, con el objetivo de determinar correlaciones a largo alcance en las series de distancia de aa idénticos, se calculó el exponente de escalamiento del DFA para las series provenientes del código universal, y se comparó su distribución contra la de los códigos alternos. En la Fig. 5.6 se presentan, a manera de ejemplo, los resultados para el caso de la serie de distancia del aspartato. Las distribuciones por código, del exponente de escalamiento del DFA, se presentan en boxplots en la Fig. 5.5b. En los boxplots, el límite inferior de la caja corresponde al primer cuartil; el límite superior al tercer cuartil; la línea horizontal que cruza la caja a la mediana y los

círculos arriba o abajo de las líneas punteadas, a posibles valores extremos (*outliers*).

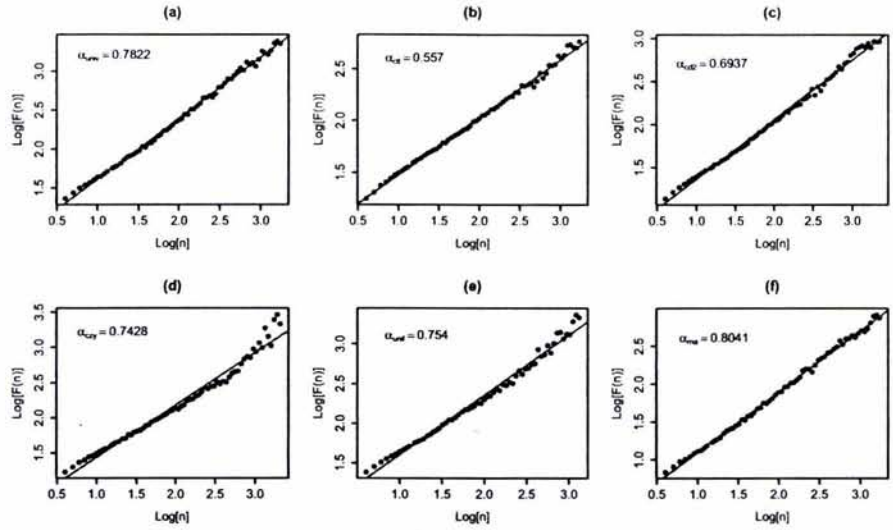

Figura 5.6. Análisis insesgado de fluctuaciones para Asp en *Borrelia burgdorferi*-OOO. a) DFA para Asp en código *univ.* (b) DFA para Asp en código *ctl.* (c) DFA para Asp en código *cd2.* (d) DFA para Asp en código *czy.* (e) DFA para Asp en código *unif.* (f) DFA para Asp en código *rna.* El valor del exponente de escalamiento α estimado, se indica en cada gráfica.

En el DFA se encontraron diferencias estadísticas entre el código universal y los códigos 2, crazy y RNA (ver Cuadro 5.2). Como era de esperarse, los exponentes de escalamiento del DFA para los controles negativos cayeron dentro de los valores que indican comportamientos estocásticos, como el movimiento browniano. Cabe hacer mención que los valores obtenidos para los controles negativos, se encuentran muy alejados de los otros códigos.

La entropía, como una medida de información, también ha sido utilizada para el análisis de cadenas de DNA en general y para comparar entre regiones codificantes y no codificantes, en lo particular (45, 67, 88). En este trabajo se calculó el MEM-gvc para todas las series de distancia generadas. Los espectros de potencia para el MEM, de las series de distancia de aspartato, se presentan a manera de ejemplo en la Fig. 5.7.

Como en los casos anteriores, hubo una clara diferencia en la distribución de la estadística medida en el código universal, contra los controles negativos (ver Fig. 5.5c y Cuadro 5.2). Asimismo, se encontraron diferencias estadísticas entre el código universal y los códigos 2, crazy y uniforme, con un cierto traslape en los C.I. (ver Cuadro 5.2).

El análisis de composición de un gran número de proteínas, ha revelado, que aproximadamente el 50 % de los aa presentes son de naturaleza hidrofílica (21, 90). Con el objetivo de determinar el efecto de los códigos alternos sobre la relación de aa por solubilidad, se cuantificaron todos los aa para cada secuencia, y se calculó la relación de aa hidrofílicos sobre hidrofóbicos (w/l). Estos resultados se muestran en el Cuadro 5.2. Como era de esperarse, la relación fue prácticamente de

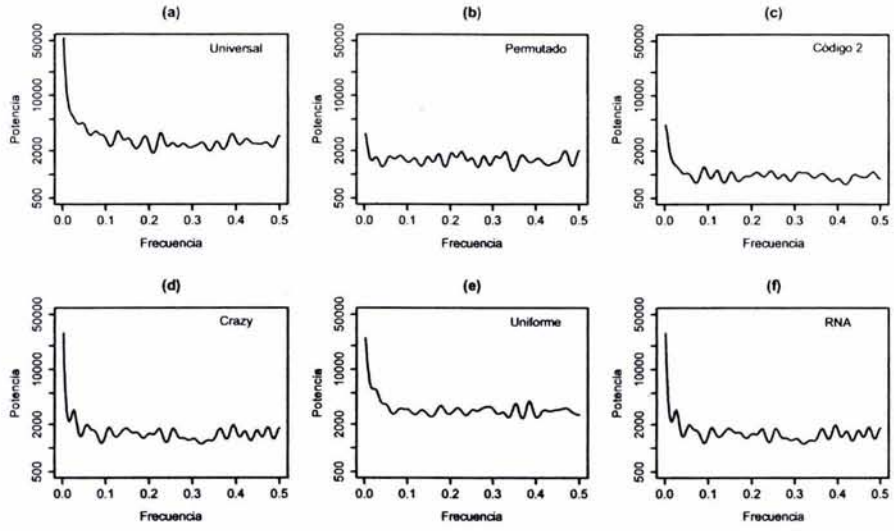

Figura 5.7. Espectro de potencias del MEM para Asp en *Borrelia burgdorferi*-OOO. (a) MEM para Asp en código *univ*. (b) MEM para Asp en código *ctl*. (c) MEM para Asp en código *cd*2. (d) MEM para Asp en código *czy*. (e) MEM para Asp en código *unif*. (f) MEM para Asp en código *rna*.

1 en las secuencias provenientes de los tres genomas control (universal, permutado y sintético) y diferente de 1 para las secuencias provenientes de las traducciones con los códigos alternos (código 2, crazy, uniforme y RNA).

5.2. *Borrelia burgdorferi*

En el caso del genoma completo de *Borrelia burgdorferi* se encontraron menor número de diferencias significativas, entre las estadísticas del código universal, contra las de los otros códigos. En la Fig. 5.8 se presentan las distribuciones de las estadísticas calculadas para este genoma, mientras que en el Cuadro 5.3 se presentan los C.I. correspondientes.

Como puede observarse en el Cuadro 5.3, la única estadística que pudo discriminar entre el código universal, contra el control negativo, fue el exponente de escalamiento del DFA. Para esta prueba se encontraron además, diferencias estadísticas con el código 2 y el código crazy.

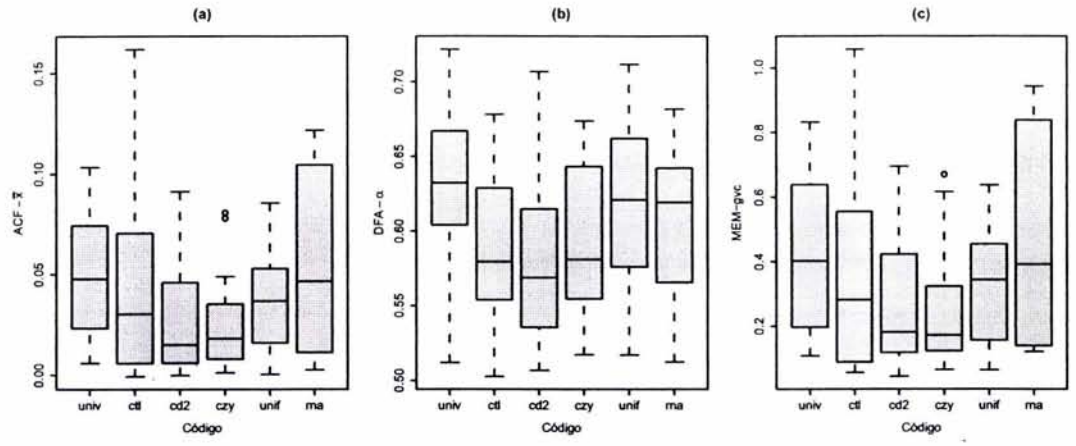

Figura 5.8. Distribución de estadísticas por códigos para *Borrelia burgdorferi*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.3. Intervalos de confianza para *Borrelia burgdorferi*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.04–0.06	0.62–0.66	0.33–0.55
Permutado	0.03–0.07	0.57–0.61*	0.24–0.48
Código 2	0.02–0.04*	0.56–0.60**	0.18–0.35*
Crazy	0.02–0.04*	0.57–0.61*	0.18–0.32*
Uniforme	0.03–0.05	0.60–0.65	0.25–0.40
RNA	0.02–0.09	0.56–0.64	0.25–0.73

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.3. *Bacillus halodurans*

Como puede observarse en la Fig. 5.9 y en el Cuadro 5.4, el genoma de *Bacillus halodurans* fue muy robusto[1] a las perturbaciones en el código genético, en las estadísticas calculadas. En particular, no se encontró ningún tipo de diferencia para el exponente de escalamiento del DFA (ver Fig. 5.9b).

Para este genoma, la estadística más sensible fue el ACF-$\bar{x}$, donde se encontraron diferencias significativas con el código 2 y el código crazy (además de con el control). Este último código, también presentó diferencia estadística para el MEM-gvc.

[1] Es decir, que sus propiedades matemáticas se modifican poco. En este caso que no hay diferencia en las estadísticas analizadas, al utilizar otros códigos en el proceso de traducción.

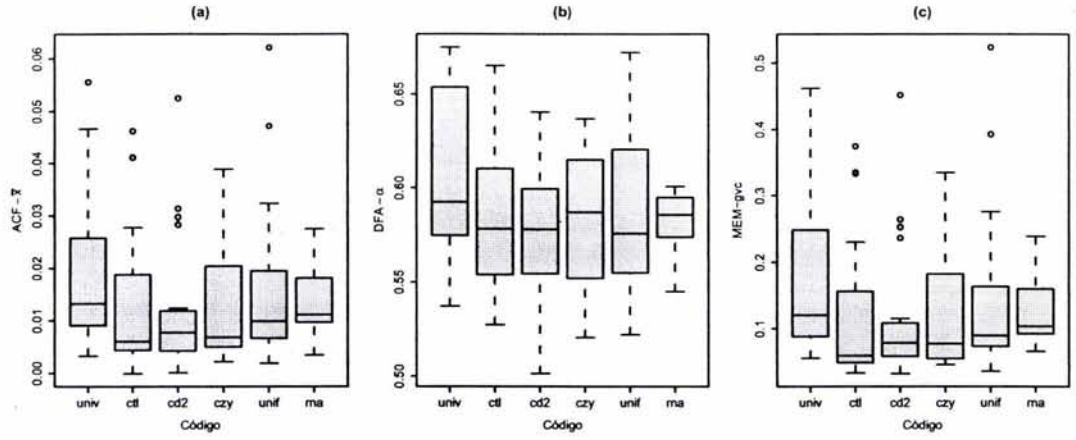

Figura 5.9. Distribución de estadísticas por códigos para *Bacillus halodurans*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.4. Intervalos de confianza para *Bacillus halodurans*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.01–0.03	0.59–0.62	0.13–0.23
Permutado	0.01–0.02*	0.57–0.61	0.08–0.17*
Código 2	0.01–0.02*	0.56–0.59	0.08–0.16
Crazy	0.01–0.02*	0.57–0.60	0.08–0.15*
Uniforme	0.01–0.02	0.57–0.60	0.10–0.20
RNA	0.01-0.02	0.57–0.59	0.09–0.17

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.4. *Bacillus subtilis*

Los resultados para el genoma de *Bacillus subtilis* fueron similares a los del genoma de *Bacillus halodurans*. En la Fig. 5.10 se presentan las distribuciones de las estadísticas calculadas y en el Cuadro 5.5 se presentan los respectivos C.I.

Cabe mencionar que el genoma de *Bacillus subtilis* fue el más robusto a las perturbaciones en el código genético, en las estadísticas calculadas. De hecho, la única diferencia se encontró entre el código universal y su control negativo, para el MEM-gvc, con un traslape en los C.I. (ver Cuadro 5.5). Es decir, las estadísticas provenientes del código universal de este genoma, fueron prácticamente indistinguibles de secuencias totalmente aleatorias. Es posible que estos resultados se deban al alto grado de transferencia horizontal que, se ha reportado, se ha presentado en esta bacteria (61).

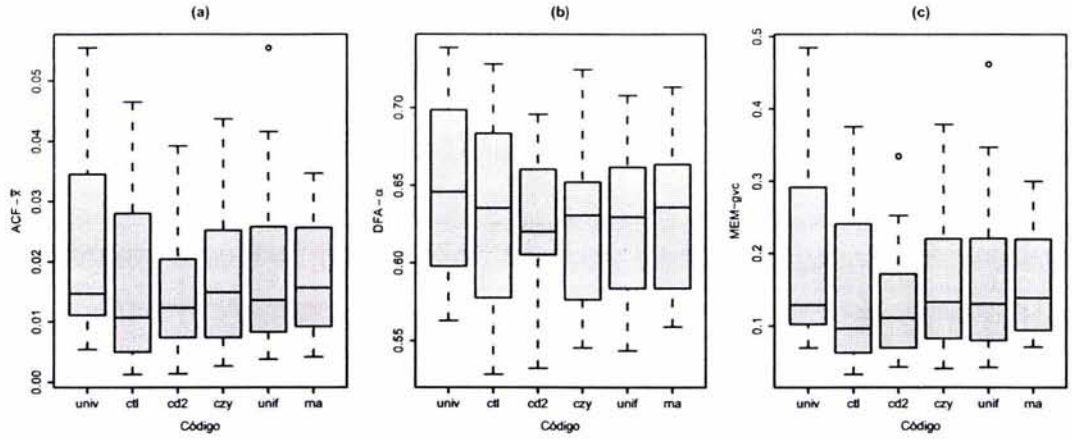

Figura 5.10. Distribución de estadísticas por códigos para *Bacillus subtilis*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.5. Intervalos de confianza para *Bacillus subtilis*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.02–0.03	0.62–0.67	0.16–0.27
Permutado	0.01–0.02	0.60–0.66	0.10–0.19*
Código 2	0.01–0.02	0.60–0.64	0.11–0.18
Crazy	0.01–0.02	0.60–0.65	0.11–0.19
Uniforme	0.01–0.02	0.61–0.65	0.12–0.22
RNA	0.01–0.03	0.59–0.67	0.10–0.22

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.5. *Deinococcus radiodurans*

En la Fig 5.11 se presentan las distribuciones de las estadísticas calculadas para el primer cromosoma de *Deinococcus radiodurans*, mientras que en el Cuadro 5.6 se presentan los C.I. correspondientes.

Como puede observarse en el Cuadro 5.6, existe una clara diferencia siginificativa, para todas las estadísticas calculadas, entre el control positivo (código universal) y el control negativo (permutado). Sin embargo, el código crazy fue el único código alterno en presentar diferencias siginificativas para dos de las tres estadísticas calculadas (ACF-$\bar{x}$ y MEM-gvc). Es decir, las estadísticas seleccionadas fueron capaces de distinguir entre ambos tipos de controles, pero no entre los posibles códigos empleados en el proceso de traducción.

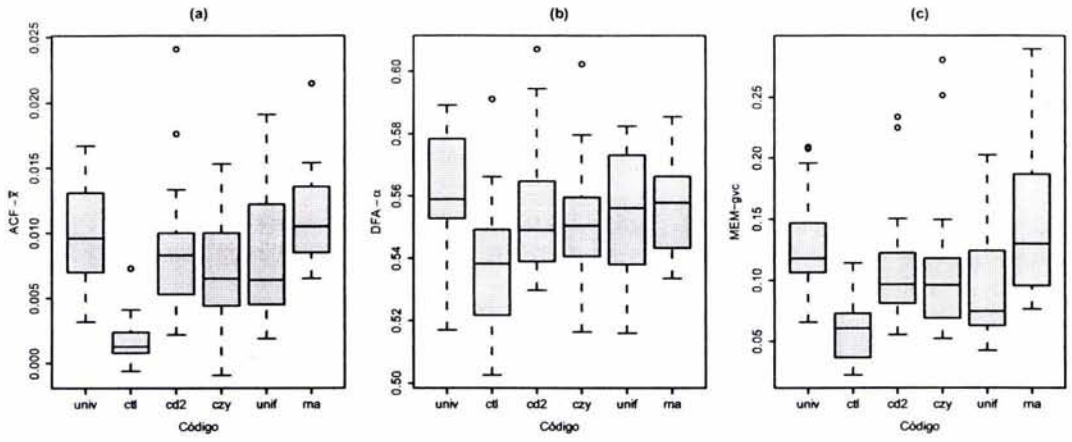

Figura 5.11. Distribución de estadísticas por códigos para *Deinococcus radiodurans*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.6. Intervalos de confianza para *Deinococcus radiodurans*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.008–0.011	0.55–0.57	0.11–0.14
Permutado	0.001–0.003***	0.53–0.55***	0.05–0.07***
Código 2	0.007–0.011	0.55–0.56	0.09–0.13
Crazy	0.005–0.008*	0.54–0.56	0.09–0.13*
Uniforme	0.006–0.010	0.54–0.56	0.07–0.11*
RNA	0.008–0.015	0.55–0.57	0.11–0.20

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.6. *Escherichia coli*

Para el genoma de *Escherichia coli* se encontraron resultados consistentes para las tres estadísticas calculadas (ver Cuadro 5.7), siendo el exponente de escalamiento del DFA, la estadística menos sensible, como puede observarse en la Fig. 5.12.

Con respecto a los códigos alternos, se encontraron diferencias significativas para el código 2 y el código uniforme, en las tres estadísticas calculadas (ver Cuadro 5.7), aunque en el caso del DFA-α se observó un traslape en los C.I.

Cuadro 5.7. Intervalos de confianza para *Escherichia coli*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.012–0.018	0.57–0.60	0.12–0.18
Permutado	0.002–0.004***	0.55–0.57**	0.04–0.05***
Código 2	0.008–0.013*	0.55–0.58*	0.09–0.13*
Crazy	0.008–0.014	0.56–0.59	0.08–0.14
Uniforme	0.007–0.012*	0.56–0.58*	0.08–0.11**
RNA	0.007–0.015	0.55–0.59	0.09–0.16

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

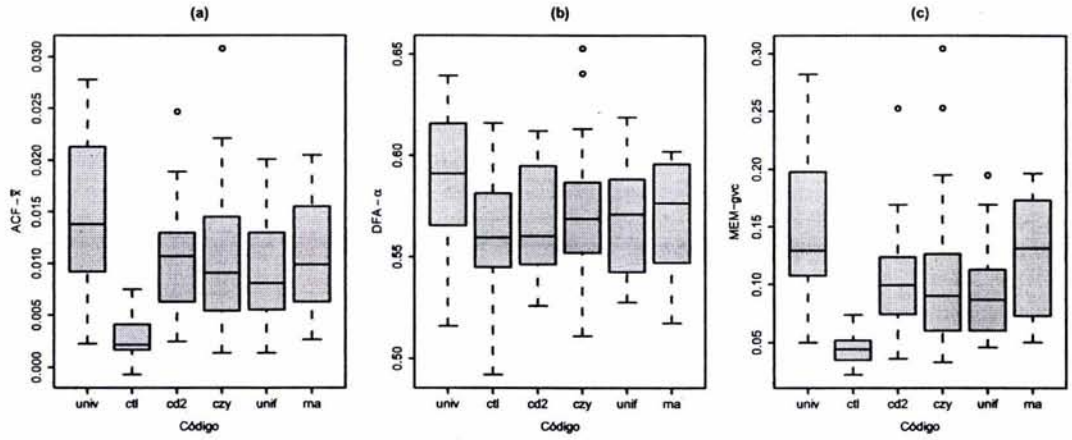

Figura 5.12. Distribución de estadísticas por códigos para *Escherichia coli*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

5.7. *Haemophilus influenzae*

El genoma de *Haemophilus influenzae* resultó ser el más sensible de los MCGs analizados en este trabajo, a las perturbaciones en el código genético. En la Fig. 5.13 puede observarse una clara discriminación entre los dos controles (positivo y negativo) para las estadísticas calculadas. Asimismo, se encontraron diferencias significativas con todos los códigos alternos (ver Cuadro 5.8).

El código alterno que presentó mayor número de diferencias estadísticas fue el crazy. El único caso donde no se encontró diferencia significativa con el código universal, fue para el DFA-α del código uniforme. Cabe hacer mención, que en todas las estadísticas medidas, el código universal obtuvo los C.I. más altos, lo que sugiere un posible proceso de optimización, en el origen de este código.

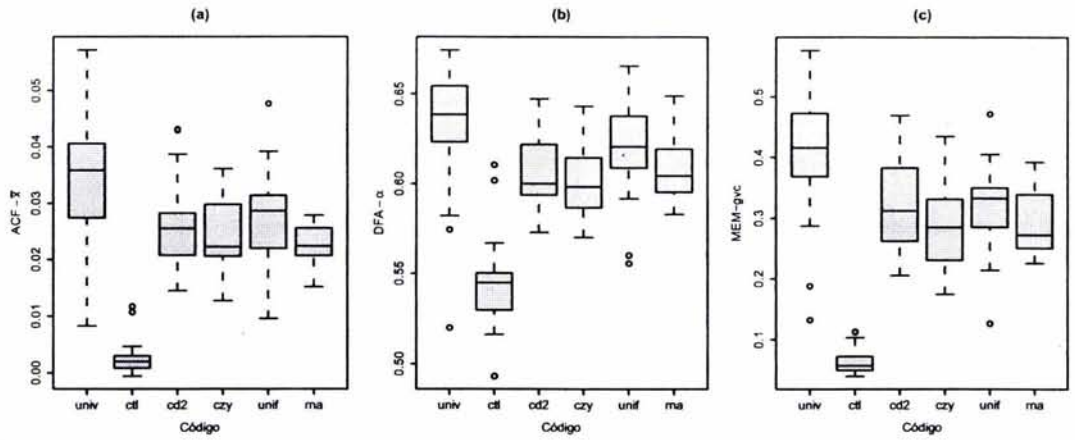

Figura 5.13. Distribución de estadísticas por códigos para *Haemophilus influenzae*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.8. Intervalos de confianza para *Haemophilus influenzae*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.029–0.039	0.61–0.64	0.36–0.45
Permutado	0.002–0.004***	0.53–0.56***	0.06–0.08***
Código 2	0.023–0.030*	0.60–0.62**	0.29–0.35**
Crazy	0.02–0.027**	0.59–0.61***	0.26–0.32***
Uniforme	0.024–0.031*	0.61–0.63	0.29–0.35***
RNA	0.020–0.025**	0.60–0.62*	0.26–0.33**

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.8. *Methanococcus jannaschii*

En el caso de esta arquea, las estadísticas seleccionadas pudieron discriminar entre el código universal y el control negativo (permutado), como puede observarse en la Fig. 5.14 y en el Cuadro 5.9, siendo el DFA-α la menos sensible, debido a un ligero traslape en los C.I.

El único código alterno que presentó diferencias significativas, con respecto al universal, fue el código crazy, para el caso del DFA-α y del MEM-gvc, presentándose, en ambos casos, un ligero traslape en los intervalos de confianza.

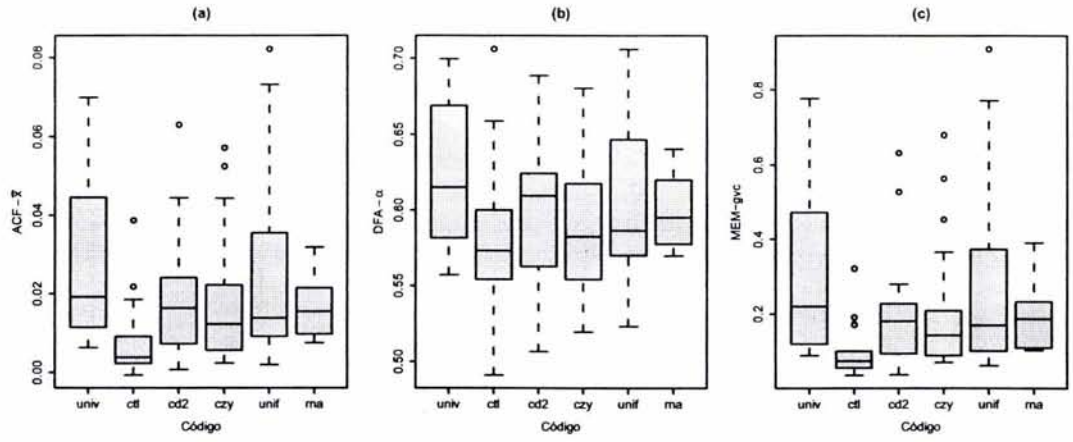

Figura 5.14. Distribución de estadísticas por códigos para *Methanococcus jannaschii*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.9. Intervalos de confianza para *Methanococcus jannashii*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.019–0.038	0.60–0.64	0.21–0.40
Permutado	0.004–0.011***	0.56–0.60**	0.07–0.12***
Código 2	0.012–0.024	0.57–0.61	0.14–0.26
Crazy	0.011–0.025	0.57–0.61*	0.13–0.28*
Uniforme	0.016–0.035	0.58–0.63	0.17–0.38
RNA	0.011–0.022	0.58–0.61	0.13–0.26

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.9. *Streptococcus pneumoniae*

En el caso de *Streptococcus pneumoniae*, se encontraron diferencias claras entre el código universal y el control negativo (permutado) para todas las estadísticas calculadas (ver Fig. 5.15 y Cuadro 5.10). Sin embargo, no se encontró ninguna diferencia significativa con los códigos alternos.

Cabe destacar que los resultados para el código RNA tuvieron una gran varianza, e inclusive, sus C.I. traslaparon con ambos controles (positivo y negativo).

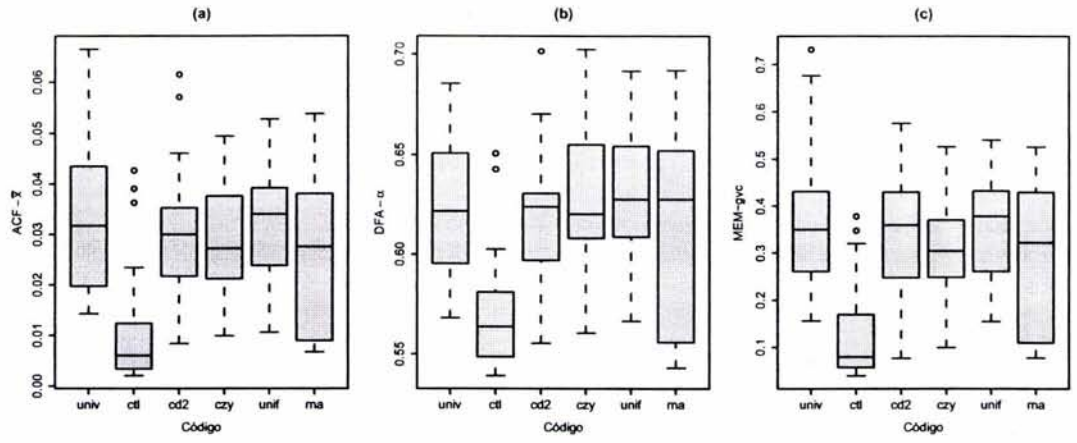

Figura 5.15. Distribución de estadísticas por códigos para *Streptococcus pneumoniae*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.10. Intervalos de confianza para *Streptococcus pneumoniae*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.028–0.041	0.61–0.63	0.30–0.44
Permutado	0.007-0.017***	0.55–0.58***	0.08–0.17***
Código 2	0.024–0.035	0.60–0.63	0.26–0.38
Crazy	0.024–0.033	0.61–0.64	0.26–0.36
Uniforme	0.027–0.037	0.61–0.64	0.29–0.39
RNA	0.014–0.037	0.57–0.65	0.17–0.40

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.10. *Sulfolobus solfataricus*

En el caso del genoma de *Sulfolobus solfataricus*, aunque las tres estadísticas determinadas pudieron discriminar entre los dos controles (ver Fig. 5.16 y Cuadro 5.11), el DFA mostró ser la menos sensible. En particular, la varianza del genoma permutado para esta prueba fue relativamente grande, y sus C.I. se traslaparon ligeramente, con los correspondientes del universal.

En la comparación con los códigos alternos, el código 2 presentó diferencias significativas para las tres estadísticas calculadas, mientras que los códigos crazy y RNA presentaron diferencias únicamente para el ACF-$\bar{x}$ y para el MEM-gvc.

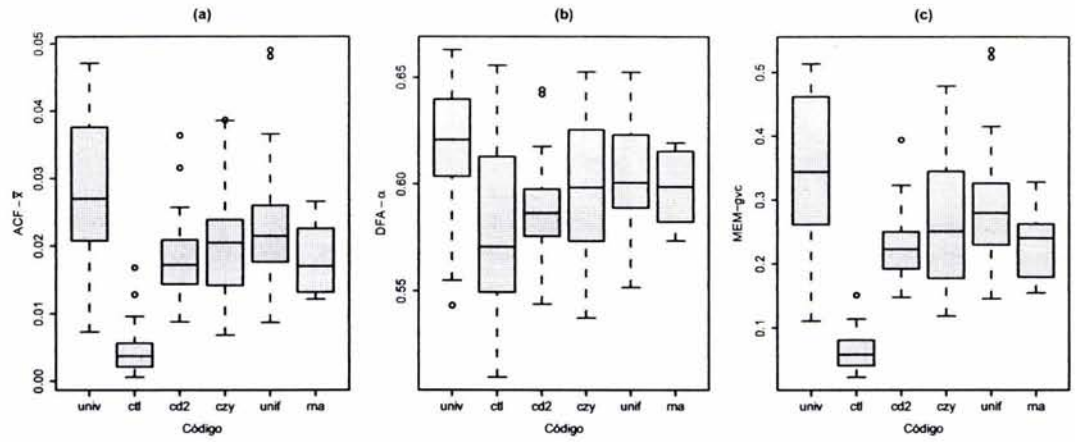

Figura 5.16. Distribución de estadísticas por códigos para *Sulfolobus solfataricus*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.11. Intervalos de confianza para *Sulfolobus solfataricus*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.024–0.033	0.60–0.62	0.30–0.39
Permutado	0.003–0.006***	0.56–0.59**	0.05–0.07***
Código 2	0.015–0.021**	0.57–0.60**	0.21–0.25***
Crazy	0.017–0.025*	0.58–0.60	0.21–0.30*
Uniforme	0.019–0.028	0.59-0.61	0.25–0.33
RNA	0.014–0.021*	0.58–0.61	0.19–0.26**

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.11. *Thermotoga maritima*

Como puede observarse en la Fig. 5.17, el MEM-gvc fue la estadística más sensible para analizar las perturbaciones en el código genético en *Thermotoga maritima*. En el caso del DFA-α, se encontró diferencia estadística entre los controles, sin traslape en los intervalos de confianza (ver Cuadro 5.12), aunque gráficamente la diferencia no fue clara.

En la comparación del código universal, contra los alternos, el código uniforme fue el único que no presentó ninguna diferencia significativa, para ninguna de las estadísticas estudiadas. En contraste, el código 2 presentó diferencias estadísticas en los tres casos. Nuevamente, el código universal presentó los valores más grandes para las tres estadísticas.

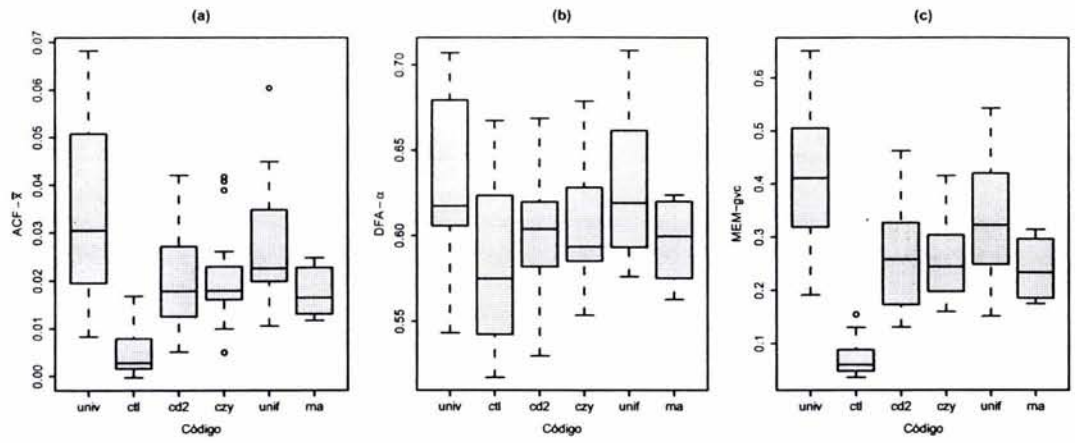

Figura 5.17. Distribución de estadísticas por códigos para *Thermotoga maritima*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.12. Intervalos de confianza para *Thermotoga maritima*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.026–0.041	0.61–0.64	0.35–0.46
Permutado	0.003–0.007***	0.56–0.60**	0.06–0.08***
Código 2	0.016–0.025*	0.58–0.61*	0.22–0.31**
Crazy	0.017–0.024*	0.59–0.62	0.22–0.28***
Uniforme	0.022–0.033	0.61–0.64	0.28–0.36
RNA	0.014–0.021*	0.58-0.61	0.20–0.27***

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.12. *Xylella fastidiosa*

En el caso del genoma de *Xylella fastidiosa*, el DFA-α no fue una estadística sensible para distinguir entre el código universal y el control negativo (ver Fig. 5.18 y Cuadro 5.13). Las otras dos estadísticas estuvieron correlacionadas, siendo el MEM-gvc la más sensible.

En cuanto a la comparación con los códigos alternos, el único que no presentó diferencias estadísticas, con respecto al código universal, fue el código RNA cuya varianza en las estadísticas medidas, fue muy grande, como en el caso de otros MCGs aquí reportados.

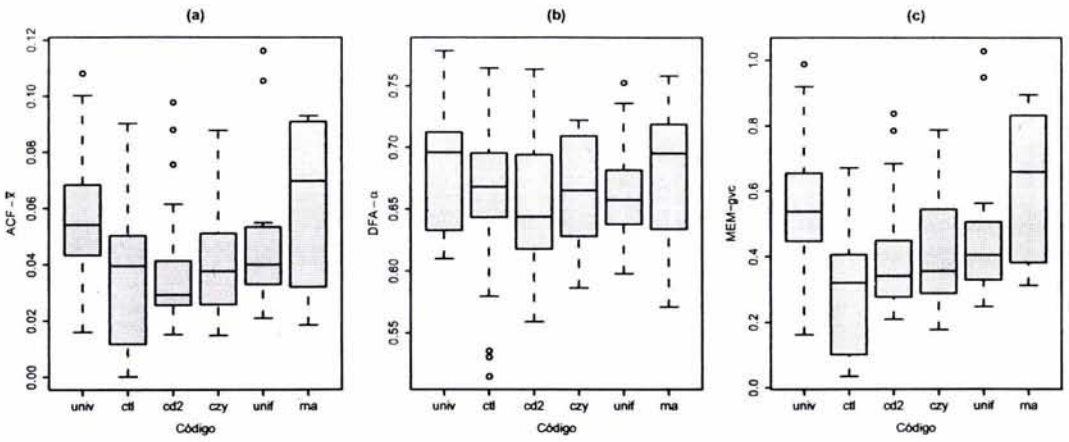

Figura 5.18. Distribución de estadísticas por códigos para *Xylella fastidiosa*. (a) Media de los primeros 38 retardos del ACF. (b) Exponente de escalamiento del DFA. (c) Coeficiente de variación geométrica del MEM.

Cuadro 5.13. Intervalos de confianza para *Xylella fastidiosa*.

Código	C.I. (ACF-$\bar{x}$)	C.I. (DFA-α)	C.I. (MEM-gvc)
Universal	0.049–0.069	0.66–0.70	0.49–0.66
Permutado	0.022–0.042**	0.62–0.68	0.19–0.34***
Código 2	0.029–0.048**	0.63–0.67	0.32–0.47**
Crazy	0.033–0.049*	0.64–0.68	0.34–0.47**
Uniforme	0.037–0.057*	0.64–0.68	0.37–0.54**
RNA	0.040–0.083	0.63–0.71	0.44–0.78

*p <0.05, **p <0.01, ***p <0.001 (prueba de WMW)

5.13. Relación estadística entre genomas

Con el objetivo de generar una referencia de relación entre los genomas estudiados, con base a las estadísticas analizadas, se concatenaron las estadísticas obtenidas de ACF$-\bar{x}$, DFA$-\alpha$ y MEM$-gvc$ de las series de distancias provenientes del código universal para cada genoma, y con estos vectores se realizó un análisis de agrupamiento jerárquico sobre la distancia entre medias. El resultado de este análisis se presenta en la Fig. 5.19.

Hay que hacer notar que la Fig. 5.19 no corresponde a la propuesta de un árbol filogenético, como el que se presenta en la Fig. 4.3, sino que, más bien, representa las relaciones entre los cromosomas estudiados, por las estadísticas obtenidas.

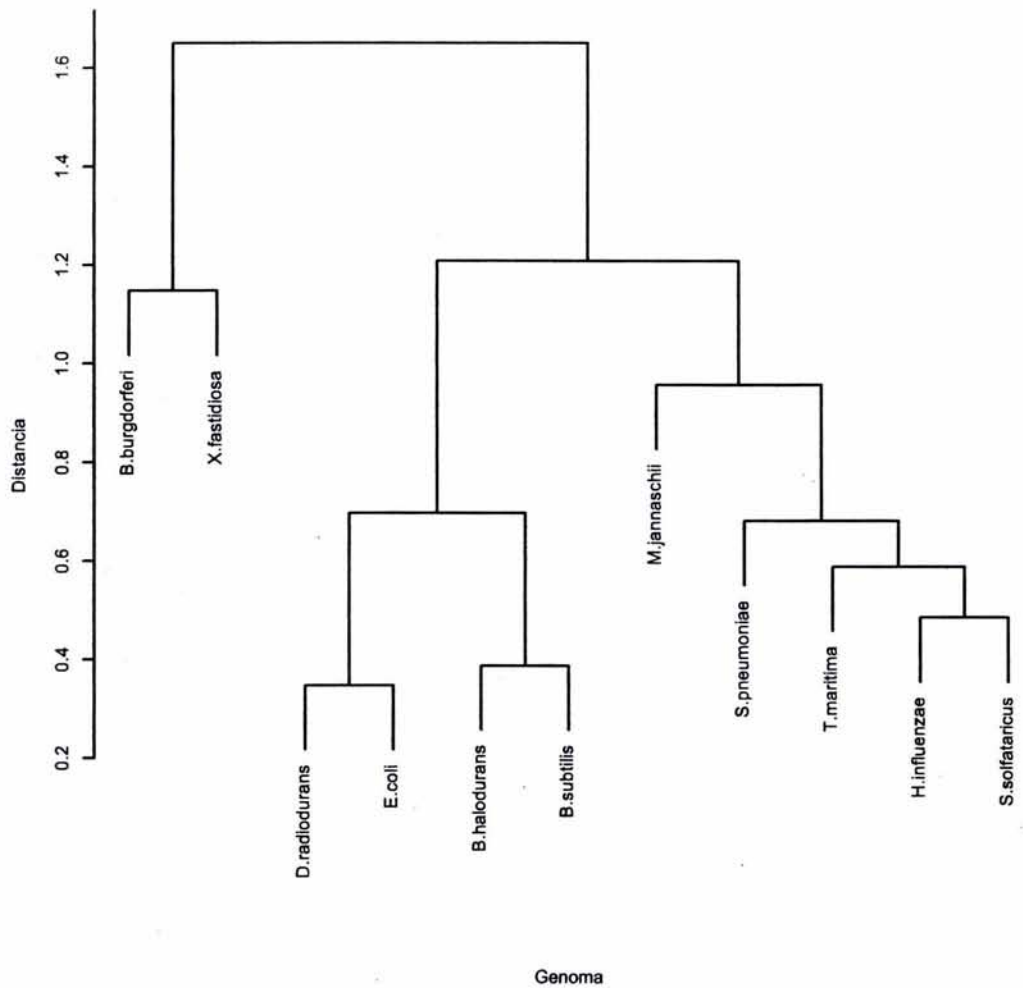

Figura 5.19. Dendrograma de agrupamiento jerárquico basado en la distancia entre medias aritméticas, para los MCGs analizados.

5.14. Relación estadística entre códigos

La relación en las estadísticas calculadas, entre los códigos propuestos, es diferente para cada estadística y para cada genoma. Con el objetivo de tener una referencia de comparación, en la Fig. 5.20 se presentan los dendrogramas correspondientes al análisis de agrupamiento jerárquico sobre la distancia entre las medias del MEM–*gvc* obtenidas para *Haemophilus influenzae*. Se tomó esta referencia, porque es donde se observó mayor sensibilidad (entendida como capacidad de distinción) en los genomas analizados. Esta figura, simula los niveles de expresión para cada aa. Como puede observase, la expresión es homogénea para los controles negativos. Asimismo, se observan claramente 3 grupos: por un lado están los controles negativos, luego tenemos al código universal y finalmente, a los códigos alternos agrupados.

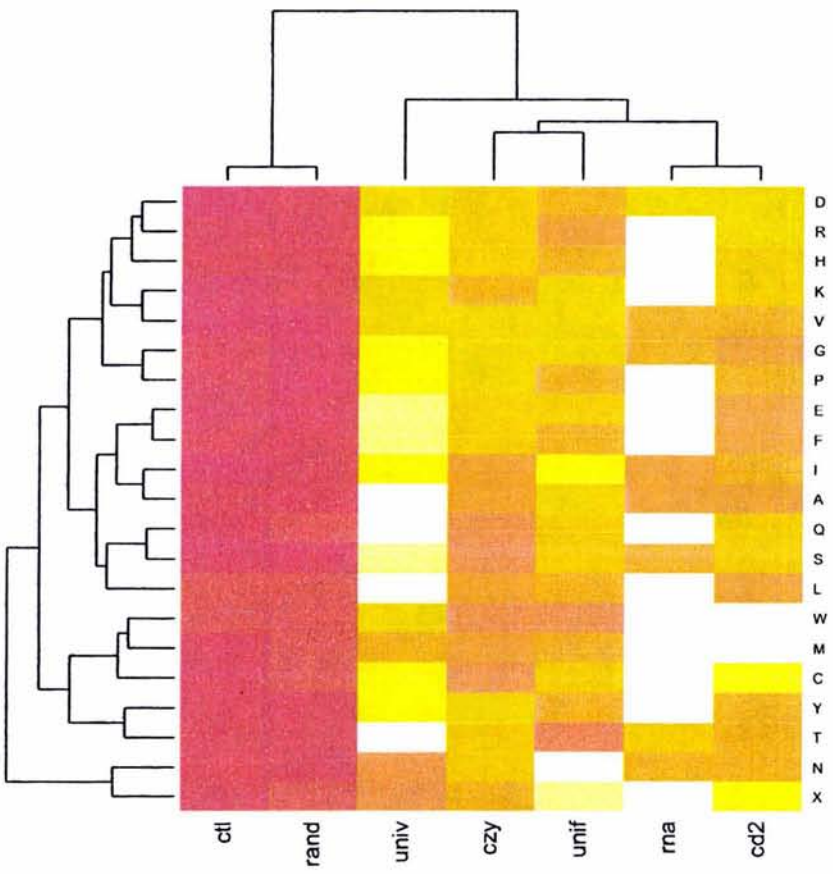

Figura 5.20. Dendrogramas de agrupamiento jerárquico basados en la distancia entre las medias aritméticas para los valores de MEM–*gvc* encontradas en *Haemophilus influenze*. Las letras a la derecha, corresponden a la convención de abreviación de una letra para los aa. El caracter *X* corresponde a los codones de terminación.

Capítulo 6

Discusión

O sweet spontaneous
earth how often have
the doting fingers of
prurient philosophers pinched
and poked thee,
has the naughty thumb
of science prodded thy beauty.
how often have religions taken thee
upon their scraggy knees
squeezing and buffeting thee
that thou mightest conceive gods
(but true to the incomparable
couch of death thy
rhythmic lover
thou answerest
them only with spring)
- e. e. cummings -

All our science,
measured against reality,
is primitive and childlike,
and yet it is the most precious thing we have
- Albert Einstein -

6.1. Propiedades matemáticas en series provenientes de secuencias codificantes

Se han publicado numerosos trabajos donde se aplican diversas funciones matemáticas sobre secuencias codificantes de DNA, que permiten distinguirlas claramente de secuencias aleatorias y de secuencias no codificantes (1, 4, 45, 68, 73, 79, 84, 93, 106). El primer objetivo de esta tesis, fue el determinar si algunas de estas propiedades podrían observarse también en secuencias de aa provenientes de la traducción de genomas completos. Con este objetivo en mente, se propusieron tres diferentes análisis, relacionados a la autosimilitud (ACF), la fractalidad (DFA) y al contenido de información (MEM), los cuales se aplicaron sobre series de distancia entre aa idénticos, provenientes de la traducción de un genoma que contiene sólo secuencias codificantes (*Borrelia burgdorferi*-OOO).

Para determinar si los resultados obtenidos aplicando estos análisis pudieran considerarse como realmente característicos de secuencias codificantes, se aplicó la misma metodología, pero sobre secuencias provenientes de la traducción de genomas aleatorios (uno producto de la permutación del genoma original, y otro sintético). Al comparar los resultados, se encontraron diferencias significativas entre ambos tipos de secuencias, lo cual sugirió que las estadísticas propuestas efectivamente podrían ser utilizadas para su diferenciación.

En las series obtenidas del genoma codificante, la fractalidad fue caracterizada tanto por el análisis de su distribución (la cual siguió una ley de potencias), como por sus correlaciones a largo alcance (DFA-$\alpha \neq 0.5$). Peng y cols. reportaron que secuencias no codificantes de DNA presentan correlaciones a largo alcance, mientras que las secuencias codificantes no las presentan (84, 85). Estos resultados son especialmente sorpresivos en el caso de los genomas bacterianos, donde casi todo el DNA es codificante, y no coinciden con los reportados en este trabajo. Estos resultados, aparentemente contradictorios, pueden deberse a diferencias en el diseño experimental, ya que mientras Peng y cols. obtuvieron las series de distancia mediante una caminata aleatoria sobre la cadena del DNA (84, 85), en este trabajo se obtuvieron las series mediante la identificación de la posición de cada símbolo a lo largo de toda la secuencia, y posterior cuantificación del número de caracteres entre símbolos idénticos (87). Además, en esta tesis se partió de un genoma que contiene únicamente secuencias codificantes.

Con el objetivo de determinar si las propiedades matemáticas identificadas fueran consecuencia de la naturaleza del código genético universal, se propusieron diferentes códigos alternos, con los cuales se llevó a cabo el proceso de traducción. En estos análisis, los C.I. para todas las estadísticas probadas se encontraron, por lo general, en los códigos alternos, por debajo de los valores respectivos en el código universal, indicando una disminución en el contenido de información, correlaciones a largo alcance más débiles y autocorrelaciones más pequeñas. En estos casos, parece ser que el código universal contiene a los valores óptimos, para las estadísticas analizadas.

Asimismo, en los análisis sobre el genoma codificante de *Borrelia burgdorferi*, independientemente de que se encontraron diferencias estadísticas con los códigos alternos, en todos los casos éstas tuvieron valores más cercanos a los del código universal (control positivo), que a los de

los controles negativos; lo cual sugiere que el código universal es muy robusto a perturbaciones, conservando su contenido de información y sus propiedades de autocorrelación a corto y largo alcance.

6.2. Propiedades matemáticas en series provenientes de genomas cromosomales microbianos (MCGs) completos

Una vez estandarizadas las técnicas, se procedieron a realizar los mismos análisis, pero ahora sobre series de distancias provenientes de la traducción de MCGs completos. Primero se comenzó, con el genoma completo de *Borrelia burgdorferi*, para tener una referencia de la diferencia entre el genoma que tiene sólo regiones codificantes del genoma completo. Hay que hacer notar que el MCG completo de esta bacteria contiene aproximadamente sólo un 8 % de secuencias intergénicas (no codificantes), sin embargo al aplicar los análisis matemáticos propuestos en este trabajo, se encontró una clara diferencia con respecto a su versión que contiene sólo secuencias codificantes (*Borrelia burgdorferi*–OOO); es decir, para el MCG completo el número de diferencias estadísticas disminuyó considerablemente e, inclusive, para dos estadísticas (ACF-$\bar{x}$ y MEM-gvc), no se encontraron diferencias significativas entre el control positivo y el control negativo. Es posible que estas diferencias se deban a los cambios en el marco de lectura que se producen por la inserción de regiones intergénicas, lo que provoca, a su vez, un cambio en las secuencias de aa obtenidas en el proceso de traducción.

Al aplicar los análisis propuestos sobre otros MCGs se encontró una amplia gama de respuestas. En el Cuadro 6.1 se presentan los MCGs estudiados en esta tesis, organizados de acuerdo al número de diferencias significativas que se encontraron entre el código universal y los códigos alternos incluyendo a los genomas permutados. Como puede observarse, los resultados variaron entre los MCGs analizados, siendo el MCG de *Bacillus subtilis* el más robusto a perturbaciones, por presentar el menor número de diferencias significativas en las estadísticas estudiadas; mientras que el de *Haemophilus influenzae* el más frágil, al presentar mayor número de diferencias significativas.

Frappat y cols. encontraron que existe una relación no-lineal entre la entropía de secuencias de nucleótidos y el contenido de GC (32), la cual se puede ajustar por una parábola invertida. Para probar la hipótesis nula sobre la independencia entre el contenido de GC y el número de diferencias significativas, se cuantificó la cantidad de cada una de las bases y se calculó el porcentaje de GC en cada uno de ellos (ver Cuadro 6.1). Como puede observarse en la Fig. 6.1, el número de diferencias significativas encontradas en las estadísticas estudiadas es independiente de su porcentaje de GC.

En aquéllos MCGs donde se encontraron pocas diferencias significativas, la tendencia fue la de acercar los valores de los C.I. respectivos hacia los obtenidos con los genomas permutados. Es decir, los valores de los C.I. para todos los genomas permutados estudiados, se encontraron dentro de los mismos rangos.

Cuadro 6.1. Organización de genomas cromosomales, por número de diferencias significativas entre el código *univ* con respecto a los códigos *ctl*, *cd2*, *czy*, *unif* y *rna*. Se incluye además, la relación de bases para cada genoma.

Genoma	Diferencias significativas	GC (%)
Bacillus subtilis	1	43.52
Bacillus halodurans	5	43.69
Borrelia burgdorferi	8	28.59
Streptococcus pneumoniae	9	39.70
Methanococcus jannaschii	10	31.43
Deinococcus radiodurans	12	67.01
Xylella fastidiosa	15	52.67
Escherichia coli	15	50.79
Sulfolobus solfataricus	20	35.79
Thermotoga maritima	20	46.25
**Borrelia burgdorferi*-OOO	21	28.82
Haemophilus influenzae	31	38.15

*Este genoma contiene sólo secuencias codificantes.

Con el objetivo de determinar si algún código lograba alejarse significativamente, mayor número de veces de las estadísticas obtenidas con el código universal, se concentraron los resultados, por código en el Cuadro 6.2. En este Cuadro se puede observar que los códigos 2 y crazy tuvieron el mayor distanciamiento, mientras que los códigos uniforme y RNA fueron los más cercanos. Cabe hacer notar que estos resultados son independientes del número de codones que comparten los códigos alternos con el universal ya que el código *unif* es tan diferente como el *cd2* y el *czy* (ver Cuadro 4.2) y, sin embargo, en las estadísticas estudiadas, su comportamiento es como el código *rna*.

Cuadro 6.2. Número total de diferencias significativas, contra el código universal.

Código	Núm. de diferencias significativas
Código 2	33
Código Crazy	30
Código Uniforme	14
Código RNA	13

Finalmente, para determinar si las estadísticas propuestas, fueron igual de sensibles en los MCGs estudiados, se concentraron gráficamente todas las diferencias significativas encontradas (incluidas las de los controles negativos) en la Fig. 6.2. Como puede observarse en esta figura, la estadística menos sensible fue el DFA-α, relacionada con propiedades de escalamiento, y la más sensible fue el MEM-gvc, relacionada con el contenido de información. De hecho, esta última, casi duplica el número total de diferencias estadísticas encontradas en la primera. Cabe hacer mención que,

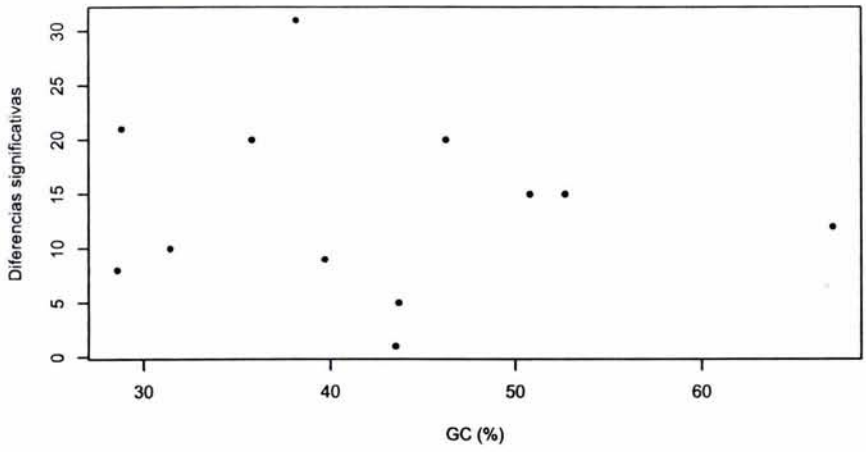

Figura 6.1. Número de diferencias significativas de acuerdo al porcentaje de GC en la cadena líder del cromosoma. Sólo se incluyen MCGs completos.

sin embargo, hubo casos, en donde la única estadística discriminante fue el DFA-α, de ahí la importancia de haber considerado más de una estadística para los presentes análisis.

6.3. Especulaciones sobre el origen del código genético

El código RNA que se utilizó en este trabajo, representa una propuesta original de Eigen & Schuster (28), como alternativa de código primitivo, al patrón *RRY* que hiciera Crick (15). El patrón *RNY* ha sido considerado como un código primitivo que pudo establecerse en el mundo del RNA, por Konecny y cols. (60), además que tiene cierta prevalencia en genes codificantes (51). Nuestros resultados parecen indicar que efectivamente, este código pudo ser el antecedente del universal, ya que es el que menos se aleja de éste, a pesar de considerar únicamente 8 aa codificados. Cabe destacar, además, que en las distribuciones de distancias, no se encontraron oscilaciones para ninguno de los 8 aa, lo cual sugiere que este tipo de comportamiento sea debido a que en este código no existen traslapes. Finalmente, con respecto a este código, cabe mencionar que frecuentemente se encontraron estadísticas con una varianza relativamente grande, en comparación con la de los otros códigos. Esta última observación, podría sugerir una mayor plasticidad de este código, al menos, en las estadísticas calculadas.

En los experimentos computacionales desarrollados en este trabajo, en ninguna ocasión los resultados obtenidos con los códigos alternos, lograron superar los valores obtenidos, para la misma estadística, con el código universal. En aquellos MCGs, donde se encontraron claras diferencias estadísticas (e.g. *Haemophilus influenzae*) los C.I. para las distintas estadísticas calculadas, estu-

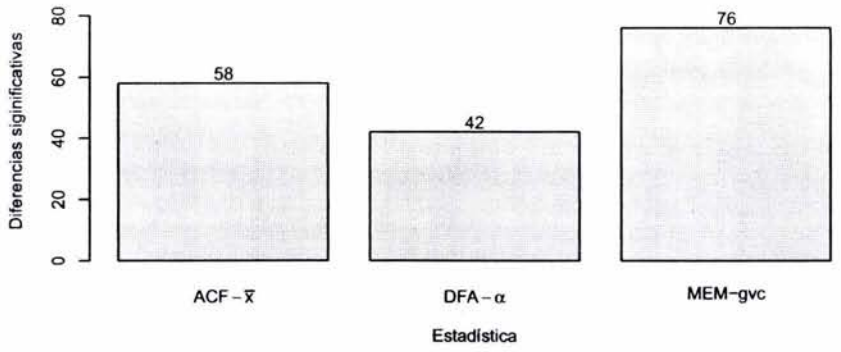

Figura 6.2. Número total de diferencias significativas por estadística calculada.

vieron por debajo de los C.I. del código universal y por arriba de los C.I. del control negativo. De hecho, para el caso de *Haemophilus influenzae*, el análisis de agrupamiento jerárquico mostró una organización de los códigos, que considera al código universal diferente a los demás (ver Fig. 5.20). Estos resultados sugieren que de los códigos propuestos y para los MCGs analizados, el código universal contiene valores óptimos en torno a propiedades de escalamiento, contenido de información y autocorrelación, que parecen contradecir la hipótesis del accidente congelado de Crick (13). Más bien, se podría hablar de una necesidad congelada, que pudo comenzar a establecerse desde el mundo del RNA.

Di Giulio ha cuestionado el origen universal del código genético, con base al proceso de formilación de la metionina en el tRNA correspondiente, dado que éste se presenta sólo en el dominio de las bacterias y en los organelos celulares (24). En el presente trabajo, además de analizar genomas cromosomales provenientes de bacterias, se consideraron dos genomas cromosomales provenientes de arqueas. Los resultados para *Methanococcus jannaschii* y para *Sulfolobus solfataricus* no difirieron significativamente de los resultados de los otros MCGs, lo que hace suponer, que nuestros resultados podrían generalizarse para el caso de las arqueas.

Los resultados obtenidos con el genoma sintético indican que es muy poco probable, que el origen de los genomas primitivos fuera totalmente aleatorio. En conjunto, estos resultados, con los de las estadísticas analizadas, sugiere que más bien, se hayan generado secuencias relativamente cortas, con cierta información codificada que, dada su estabilidad y probable ventaja selectiva, se fueron duplicando a lo largo de la misma secuencia (lo cual explicaría el comportamiento fractal detectado en el DFA), incorporando mutaciones en el proceso. El acumulo de mutaciones dio origen a nuevas proteínas, que, a su vez, confirieron de una mayor ventaja con respecto a otras cadenas. Una vez establecido el patrón *RNY* del mundo del RNA, debieron incorporarse nuevos codones, hasta quedar establecidos los 64 actuales. El código actual, al contener mayor densidad de información (MEM) y mayor plasticidad a las mutaciones (degeneración de la tercera base), debió establecerse sobre otros posibles códigos y congelarse hasta la actualidad.

Capítulo 7

Conclusiones

¿Dónde estabas cuando cimenté la tierra?
Dímelo, si es que sabes tanto.
¿Quién señaló sus dimensiones? –si lo sabes–,
¿o quién le aplicó la cinta de medir?
¿Dónde encaja su basamento
o quién asentó su piedra angular,
entre la aclamación unánime de los astros de la mañana
y los vítores de todos los ángeles?
- Job 38: 4–7 -

Con base a los resultados obtenidos en el genoma codificante de *Borrelia burgdorferi* y siguiendo los objetivos e hipótesis propuestos en la tesis, se concluye que, por lo menos para este genoma:

1. Es posible diferenciar matemáticamente secuencias de aminoácidos provenientes de secuencias codificantes de DNA, de aquéllas provenientes de secuencias aleatorias.
2. Series de distancia entre caracteres idénticos de secuencias de aminoácidos, provenientes de la traducción completa del genoma codificante de *Borrelia burgdorferi*, contienen propiedades de fractalidad como la autocorrelación y correlaciones a largo alcance así como mayor contenido de información con respecto a sus contrapartes provenientes de secuencias aleatorias.
3. De los códigos genéticos propuestos, el correspondiente al universal contiene los valores más grandes para el ACF-$\bar{x}$, DFA-α y MEM-*gvc*.

Asimismo, se puede concluir que para los MCGs estudiados y con las estadísticas propuestas:

1. La estadística más sensible para discriminar entre diferentes códigos fue el MEM-gvc.
2. Los resultados provenientes del genoma cromosomal de *Bacillus subtilis* presentaron el menor número de diferencias significativas con respecto al código universal, mientras que los resultados provenientes del genoma cromosomal de *Haemophilus influenzae* presentaron el mayor número.
3. El número de diferencias significativas con respecto al código universal es independiente de su proporción de bases en el genoma.
4. Las perturbaciones más severas al código universal se reflejaron en los códigos 2 y crazy.

Apéndice A

Statistical analysis of the distribution of amino acids in *Borrelia burgdorferi* genome under different genetic codes

Artículo aceptado para su publicación en *Physica A*, desarollado por:

José A. García, Samantha Alvarez, Alejandro Flores
Tzipe Govezensky, Juan R. Bobadilla y Marco V. José

Statistical analysis of the distribution of amino acids in *Borrelia burgdorferi* genome under different genetic codes

José A. García[1,2], Samantha Alvarez[1], Alejandro Flores[1]
Tzipe Govezensky[2], Juan R. Bobadilla[2] y Marco V. José[2]

[1]Research Department, La Salle University, México.
[2]Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, México.

Abstract

The genetic code is considered to be universal. In order to test if some statistical properties of the coding bacterial genome were due to inherent properties of the genetic code, we compared the autocorrelation function, the scaling properties and the maximum entropy of the distribution of distances of amino acids in sequences obtained by translating protein-coding regions from the genome of *Borrelia burgdorferi*, under different genetic codes. Overall our results indicate that these properties are very stable to perturbations made by altering the genetic code. We also discuss the evolutionary likely implications of the present results.

Keywords: Genomics, genetic codes, statistical analysis

PACS: 87.10.+e, 05.40.+j

A.1. Introduction

Organisms use the genetic code to translate the information stored in DNA or RNA nucleotide sequences to synthesize amino acids sequences called proteins. The same code is used in all living organisms (there is an exception in the mitochondrial genome), so it is nearly universal (1).

The universality of the genetic code suggests that it should have been established early in evolution, so once it appeared in nature it was "frozen" (2). An unanswered question is whether other genetic codes could accomplish the same function or be as efficient as the actual universal genetic code.

In order to test the properties of different genetic codes, we analyzed some discriminating statistics of the distance distribution of amino acids (aa) derived from protein-coding regions from the genome of *Borrelia burgdorferi*, through numerical experiments in which the actual genetic code was perturbed. The chosen statistics are related to the content of information, the scaling properties of the distances series, and the autocorrelation properties.

A.2. Experimental design

We started with a sequence of protein-coding regions of the genome of *Borrelia burgdorferi* (see 3). This sequence was translated to a sequence of aa using different genetic codes (see below). For the three stop codons, we assigned character X. Then, we generated distance series between identical aa along the chromosome for each character (either aa or an stop codon). Our master control was the universal genetic code itself. As negative controls, we considered both a shuffled version of the original sequence (shuffled code), and a synthetic DNA sequence 10^6 nucleotides long, obtained by sampling the four DNA nucleotides with replacement (random code).

Instead of the classic random walk mapping of a DNA sequence (4,5), we followed a different approach for studying the statistical properties of aa sequences. In particular, for a given aa, we determined its actual position along the whole sequence, and from this we measured, as distance, the number of aa which lies between two identical characters. Hence, we obtained the actual distance series for each character.

A.2.1. Genetic codes

The universal code presents a characteristic distribution of codons to aa. In this distribution, there are several aa which are encoded by more than one codon, so it is degenerated (1). Often the base in the third position is less significant, as a mutation in this position does not imply a change in the encoded aa (third-base degeneracy).

The first perturbation was called code 2. Amino acids were randomly assigned to codons, preserving the universal distribution of codons to aa i.e. the degeneracy of the code is the same as the universal code. In this way, a mutation in the third base altered the coded aa, thus the third-base degeneracy is not longer sustained.

In the uniform code we assumed a uniform distribution of codons to aa, so each aa is coded by three different randomly chosen codons. As there are 20 aa, the uniform code has four stop codons.

For generating what we called the crazy code, we built a population, in which the 21 characters (representing either an aa or an stop codon) were sampled with replacement, and each of them was randomly assigned to one out of the 64 codons, so the distribution of codons to aa is both not universal-like, and not uniform, e.g. three different aa can be translated with five different codons.

Finally, we also tested a perturbation following the RNA world hypothesis (6) using the RNY (purine-any nucleotide-pyrimidine) pattern proposed by Eigen & Schuster (7). In the RNA world code aspartic acid is coded by GAC and GAU; asparagine is coded by AAC and AAU; alanine is coded by GCU and GCC; isoleucine is coded by AUU and AUC; glycine is coded by GGC and GGU; serine is coded by AGC and AGU; threonine is coded by ACU and ACC; and, valine is coded by GUU and GUC. This code has been proposed as the primeval genetic code (see also 8).

A.2.2. Statistical analysis

In order to test for statistical differences among the master, the negative and the synthetic codes (code 2, uniform code, crazy code and RNA world code), both the p-value of the Wilcoxon-Mann-Withney test (9) and the bootstrap (10) 95 % confidence intervals (C.I. 95) were calculated for three different statistics. The first technique is used to test differences of means regardless of the particular distribution of the random variable and the latter estimates non-parametric confidence intervals by sampling few data several times (e.g. 1000 times) with replacement. The bootstrap C.I. 95 from the random code gives an estimation of the white noise bandwidths.

The chosen statistics were: a) The average of the mean of the first 38 lags of the autocorrelation function (ACF); b) Mean of the detrended fluctuation analysis scaling exponent (DFA)(); c) Average of the geometric variation coefficient maximum entropy ($ME - gvc$), where gvc is the ratio of $sd(x)/\tilde{x}$, where $sd(x)$ is the standard deviation of the maximum entropy of the series x, and $\tilde{x}$ is the geometric mean of the maximum entropy of the series x. Maximum entropy of the series was calculated with the *SSA-MTM Toolkit*, v4.2 (12). All means were calculated for the 21 characters within each code.

Both the ACF and the DFA look for autocorrelations within the series. The DFA technique is based on a modified root mean square analysis of a random walk, to assess the intrinsic correlation properties of a dynamic system separated from external trends in the data, and is intended to determine the scaling properties of a time series (11). When the DFA calculated scaling exponent is equal to 0.5 it is indicative of white noise; if the value lies betweeen 0.5 and 1 then the time series exhibits long-range correlations.

A.3. Results

The distance distributions for each particular aa changed for all the tested codes. Fig. A.1 shows the probability density function (pdf) of the distance distribution for aspartic acid as an example. Interestingly, the pdfs of various aa obtained with different codes presented an oscillatory decaying pattern. The only exception to this pattern was observed in the RNA world code, in which none of the aa presented oscillations.

Several works have reported periodical patterns in DNA sequences by means of autocorrelation function (ACF) analysis (13, 14, 15, 16, 17). We also utilized the ACF, but applied to the aa distance series, coming from both the universal code and the negative controls (shuffled and random codes). We found autocorrelations in the universal code, and no autocorrelation for all lags in the negative controls (as indicated by the bandwidth of white noise). As can be seen from Table A.1, there is a clear difference, as measured by the bootstrap C.I. 95 between the universal code and the negative controls. Further, we tested if the average of the mean of the first 38 lags of the ACF of the universal code, was statistically different to any of the other codes. We found differences with code 2, crazy code, and uniform code with a non-parametric test. These synthetic codes distributions

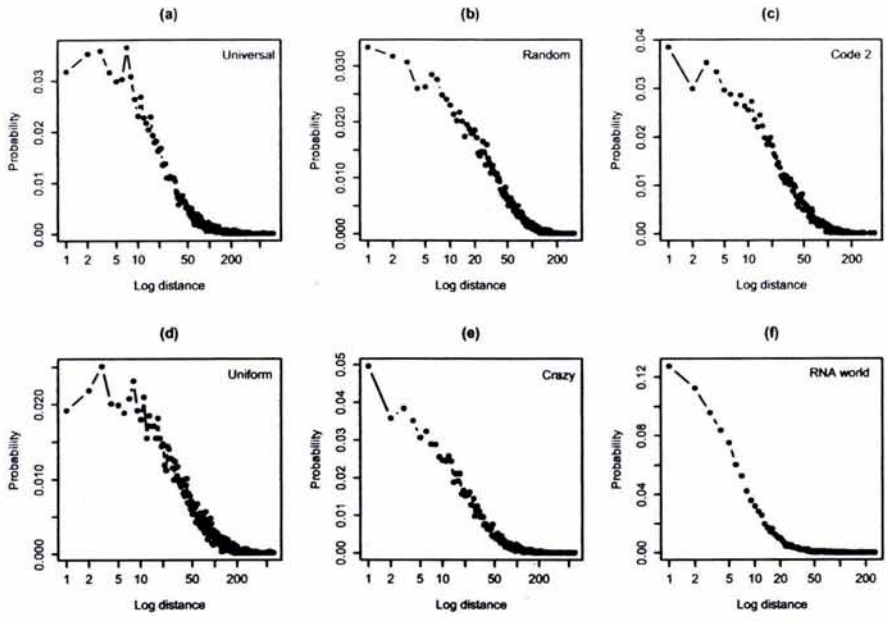

Figura A.1. Probability density functions of the distance distribution of aspartic acid. Distance is measured as the number of aa that are between two identical aa. a) Universal code; b) Random code; c) Code 2; d) Uniform code; e) Crazy code; f) RNA world code.

were displaced to lower values to respect of the universal code distribution, although with an slight overlapping (see Table A.1).

Several authors have reported long-range correlations in DNA (5, 18, 19, 20, 21). Here, in order to look for long-range correlations between each aa, we calculated the DFA scaling exponent (11) for each distance series, and then tested for differences in the mean value for each code against the corresponding value obtained with the universal code. We found statistical differences with code 2, crazy code and the RNA world code, although with slight overlaps in the bootstrap C.I. 95 in all cases (see Table A.1). As expected, the DFA scaling exponents of the negative controls lie within (random) or very close (shuffled) to values that indicate brownian motion. It is worth to mention, that both negative controls are strikingly different when compared with all other codes.

Entropy, as a measure of information, has also been used to analyze DNA and to make comparisons between coding and non-coding regions (22, 23, 24). In the current study, we calculated the average of ME-gvc for each distance series. Again there was a clear difference between the statistics of the negative controls, and all the other codes. Statistical differences, were also found with code 2, crazy code and uniform code against the universal code with slight overlap in the bootstrap C.I. 95 (see Table A.1).

Cuadro A.1. Bootstrap confidence intervals (C.I.) for: autocorreletion function analysis (ACF); detrended fluctuation analysis scaling exponent (DFA); and geometric variation coefficient of maximum entropy (ME-gvc).

Code	C.I. (ACF)	C.I. (DFA)	C.I. (ME-gcv)
Universal	0.09 – 0.12	0.72 – 0.75	0.87 – 1.14
Shuffled	0.03 – 0.06‡	0.56 – 0.59‡	0.23 – 0.46‡
Random	-1.5e-3 – 1.2e-7‡	0.49 – 0.51‡	0.07 – 0.09‡
Code 2	0.07 – 0.09†	0.69 – 0.73*	0.61 – 0.84†
Crazy	0.06 – 0.10*	0.69 – 0.73*	0.62 – 0.90†
Uniform	0.07 – 0.10*	0.71 – 0.74	0.67 – 0.89*
RNA World	0.06 – 0.13	0.65 – 0.73*	0.51 – 1.11

$^{*}p < 0.05$, $^{\dagger}p < 0.01$, $^{\ddagger}p < 0.001$ (Wilcoxon-Mann-Withney test)

A.4. Discusion

There have been several papers, some of them considered as classics, which have contributed to our understanding of the origin of the genetic code (2, 6, 7, 8, 25, 26, 27). However, none of them addressed the statistical properties of the translated sequence of aa. Here we carried out numerical experiments with different genetic codes in which some statistical properties of the translated products are analyzed. In order to study the coding DNA, we based our analysis on sequences of aa obtained by translating the protein coding sequence from *Borrelia burgdorferi* genome.

Peng, *et al.* (5) have found that noncoding DNA sequences show long-range autocorrelations whereas coding sequences do not. This is remarkable in the case of bacterial chromosomes since most of the DNA content is coding. Indeed, they showed that in bacteria there is a lack of autocorrelation. We found autocorrelation in both DNA coding sequences (3) and in aa sequences coming from translating bacterial coding DNA. These apparently contradictory results are presumably due to differences in the experimental design, as we looked for the distance series between characters (either aa or $n-$tuples of DNA).

In general the bootstrap C.I. 95 for all the tested statistics of the synthetic codes, showed a diminution of information content, weaker long-range correlations, and smaller values of the scaling exponent, when compared with the master code. Then, the universal code seems to contain optimum values for those statistics.

Regardless of finding statistical differences with alternative codes, in all cases the statistics have values closer to the universal code than to the negative controls. This suggests that the genetic code is very robust to perturbations, as information measures, as well as long and short correlations are maintained. Thus, once the universal code was established, it became fixed and resistant to evolutionary changes. The question of what makes unique the universal code remains an unanswered problem.

A.5. References

1. B. Lewin, Genes VII (2000) New York, Oxford University Press.
2. F.H.C. Crick, J. Mol. Biol. 38 (1968) 367.
3. J. Sánchez, M.V. José, Biochem. Biophys. Res. Comm. 299 (2002) 126.
4. W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655.
5. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature 356 (1992) 168.
6. W. Gilbert, Nature 319 (1986) 618.
7. M. Eigen, P. Schuster, The hypercycle: A principle of self-organization (1979) Berlin, Springer-Verlag.
8. J. Konecny, M. Schöniger, G.L. Hofacker, J. theor. Biol. 173 (1995) 263.
9. H.B. Mann, D.R. Whitney, Ann. Math. Statist. 18 (1947) 50.
10. B. Efron, R.J. Tibshirani, An introduction to the bootstrap (1998) San Francisco, Chapman & Hall.
11. C-K. Peng, S. Havlin, H.E. Stanley, A.L. Goldberger, Chaos 5 (1995) 82.
12. M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, D. Kondrashow, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi, P. Yiou, Rev. Geophys. 40 (2002) 1.
13. J.C.W. Shepherd, J. Mol. Evol. 17 (1981) 94.
14. D. G. Arquès, C.J. Michel, J. theor. Biol. 143 (1990) 307.
15. S. Karlin, V. Brendel, Science 259 (1993) 677.
16. D.G. Arquès, C.J. Michel, BioSystems 44 (1997) 107.
17. H. Herzel, I.Groβe, Phys. Rev. E 55 (1997) 800.
18. R.F. Voss, Phys. Rev. Lett. 68 (1992) 3805.
19. C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simmons, H.E. Stanley, A.L. Goldberger, Phys. Rev. E 49 (1994) 1685.
20. A. Arneodo, E. Bacry, P.V. Graves, J.F. Muzy, Phys. Rev. Lett. 74 (1995) 3293.
21. A.K. Mohanty, A.V.S.S.N. Rao, Phys. Rev. Lett. 84 (2000) 1832.
22. R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. Lett. 73 (1994) 3169.

23. S. Havlin, S.V. Buldyrev, A.L. Goldberger, R.N. Mantegna, C.-K. Peng, M. Simons, H.E. Stanley, Fractals 3 (1995) 269.

24. A.O. Schmitt, H. Herzel, J. theor. Biol. 188 (1997) 369.

25. J. Konecny, M. Eckert, M. Schöniger, G.L. Hofacker, J. Mol. Evol. 36 (1993) 407.

26. P. Béland, T.F.H. Allen, J. theor. Biol. 170 (1994) 359.

27. H. Hartman, J. Mol. Evol. 40 (1995) 541.

Apéndice B

Mathematical properties of DNA sequences from coding and noncoding regions

Artículo de revisión sometido a: *Rev. Mex. Fís.*, por:

José A. García y Marco V. José

Mathematical properties of DNA sequences from coding and noncoding regions

José A. García[1,2] and Marco V. José[2]

[1]Research Department, La Salle University, México.
[2]Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, México.

Abstract

Several nonlinear techniques have been applied to analyze DNA sequences. As a result, some mathematical properties of both coding and noncoding regions had emerged. We review and apply, some of these techniques selecting some examples and comparing our results with previously published data. We also discuss the main controversies that have been raised in terms of the different taken approaches, particularly the presence or absence of long-range correlations in coding regions.

Keywords: long-range correlations; DNA mathematical analysis; fractal dynamics

PACS: 87.10.+e, 05.40.+j

B.1. Introduction

DNA is the molecule in which life organisms store information for their biological processes. In a DNA strand, it is possible to find sequences which can be transcribed to complementary RNAs, such as tRNAs, rRNAs and mRNAs. These sequences, also known as genes, are the coding regions in DNA. Between genes (intergenic regions), we can find regulatory sequences for transcription control. In the case of eukaryotic cells, besides, the vast majority of genes are not continuous: not expressed sequences, known as introns, lie between expression-coding sequences, known as exons. Thus, both intergenic regions and introns are noncoding DNA.

As a replicating information unit, DNA has fascinated not just biologist, but also, other scientists, like physicists, chemists, mathematicians and astrobiologists. The former, have made a lot of contributions to DNA understanding and, recently, they have applied several mathematical techniques for analyzing coding and noncoding regions.

In this paper, we apply, compare and review some mathematical methods, which have been commonly used in order to reveal signature properties between coding and noncoding DNA sequences. As there are a lot of controversies among some of the results, we include numerical experiments and discuss the different interpretations among them.

B.2. Biochemistry of DNA

DNA is a double anti-parallel helix builded by concatenating nucleotide blocks. Each nucleotide has a nitrogenous base, a deoxyribose and a phosphate group. The bases are inside the molecule, while the phosphates are in contact with the hydrophilic medium.

DNA has four nitrogenous bases: adenine (A), thymine (T), cytosine (C) and guanine (G). There is complementarity between both DNA strands, as an A on one strand, always binds with a T on the other, and a C always binds with a G. The binding between the bases is through hydrogen bonds: two between A and T and three between C and G. Thus, following some physicochemical properties, DNA bases have been classified using three different dichotomies: (a) Purines (R), A and G; or Pyrimidines (Y), T and C; (b) Weak (W), A and T; or Strong (S), C and G; and (c) Amines (M), A and C; or Ketones (K), T and G.

As the DNA backbone is constant (i.e. a chain of deoxyriboses bound by phosphodiester bonds), its biological properties reside in the sequence of bases along one strand (as the other is complementary). In this sense, DNA can be seen as a four letter alphabet, or could be transformed to a distance series. In the current paper, we review the main mathematical techniques used to study DNA sequences, considering the latter case.

B.3. DNA mapping

In order to apply signal processing techniques to DNA analysis, a DNA sequence must be transformed to a distance series. There have been several approaches to accomplish this. Herein, we present the main three.

B.3.1. Binary representation

The easier approach is to transform a DNA sequence to a binary sequence using one of the three conventions mentioned before (e.g. all weak bases -A and T- are changed by 0, and all strong bases -C and G- are changed by 1). The obtained series could now be subject to further mathematical analysis (22).

B.3.2. DNA random walk

This technique could be seen as a particular case of a binary representation. Consider a conventional one-dimensional random walk model, in which a theoretical walker crosses a DNA strand (37). The walker starts at position $n = 0$ and gives one step up $[u(n) = +1]$ with each pyrimidine,

and one step down $[u(n) = -1]$ with each purine. To graphically represent the walking, one usually plots the cumulative walk $y(n)$, against the position n as shown on Fig B.1a for the first 50,000 nitrogenous bases of the coding genome of *Borrelia burgderfori*.

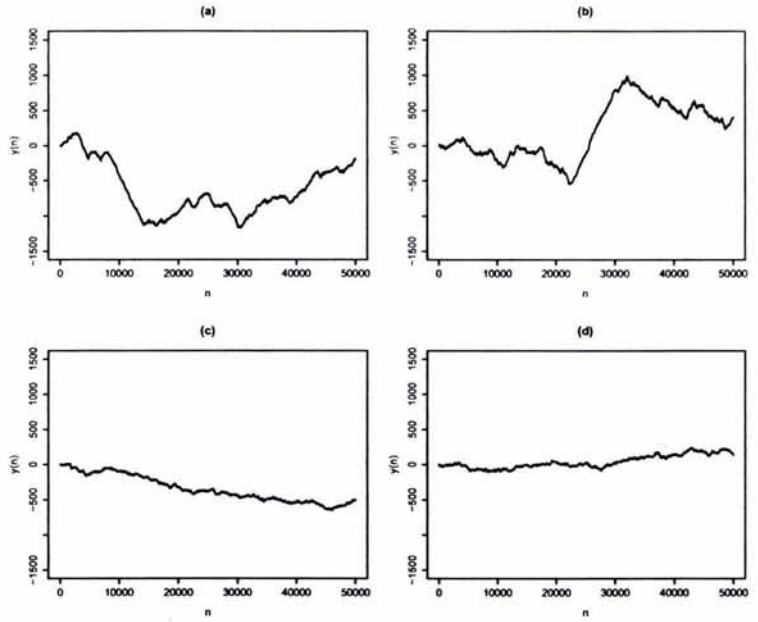

Figure B.1. DNA walk dislplacement $y(n)$ against nucleotide distance n for the first 50000 nitrogenous bases in: (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB (human beta globin chromosomal region); (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

The mathematical techniques reviewed here have been used in order to differentiate between coding and noncoding regions. In the current paper we apply these techniques to: a whole coding bacterial genome, obtained by concatenating all the coding genes of *Borrelia burgderfori* in its original order and orientation (31); the human beta globin chromosomal region (HUMHBB), a mainly non-coding sequence; and two control sequences: a shuffled version of the original coding genome of *Borrelia burgderfori*, and a synthetic DNA of one million bases, obtained by randomly sampling with replacement the four bases (9). These control sequences do not represent intergenic regions; they are just representations of pure stochastic processes, in order to be able to distinguish between sequences with information (for protein synthesis), and corresponding random sequences. Thus, on Fig. B.1 the four DNA walk displacements are shown. Note that both the coding and noncoding sequences presents jagged contours with local regions rich in either purines or pyrimidines (see Fig. B.1a and b), while the control sequences presents less variable displacements (see Fig. B.1c and d). This kind of pattern on DNA could be related to biological structure.

B.3.3. Actual distance series

An alternative to binary representations, is the production of actual distance series obtained by calculating the number of characters between identical $n-$tuples (n = mono, du, tri, etc.) along the whole DNA sequence (31). For example, to generate the distance series for the ATG triplet, the actual positions of this sequence is first identified using the three different reading frames, and then the number of bases which lie between consecutive ATGs is computed. By using this approach, less information is lost in comparison with the binary representations, due to oversimplification process in the latter.

B.4. DNA mathematical properties

B.4.1. Periodicities

Shepherd found purine-pyrimidine rhythms on viral genomes using actual distance series from binary DNA representations (33). He used the *RY* convention to transform the original sequence, and then looked for the actual distance series between different combinations. As an example, the results for the triplet *YRY* in the coding genome of *Borrelia burgdorferi* are shown on Fig. B.2a.

As shown on Fig. B.2a maxima occurs regularly every three bases. This rhythm was preserved for all the studied combinations with an exception of an irregularity in $n = 13$ for *Y.R* counts (33). In contrast, there is no pattern found in both the noncoding sequence (see Fig. B.2b), and the control sequences (see Figs. B.2c and d).

It is worth to mention that by looking at such periodicities, Shepherd found the sequence *RNY* as the most prevalent than other sequences, thus he hypothesized that not only this sequence was an ancestor of the actual universal code (7,20), but also that vestiges of this pattern are still detectable on current genomes (34). Although this hypothesis was challenged by Wong & Cedergren (40), and also by Jukes (18), several cases have been found with *RNY* prevalence in actual genomes (18).

Arquès and Michel used a similar approach to look for periodicities in coding and noncoding regions (3). They studied sequences from virus, prokaryotes and eukaryotes, and looked for the $i-$motiff $m_i = YRY(N)_iYRY$, with i in the range [0,99], i.e. two triplets YRY separated by any i bases. They found that the motif $YRY(N)_6YRY$ had preferential occurence on the vast majority of the studied sequences (3). Also two kinds of periodicities were found, the previously mentioned periodicity three, $P3$ (in both coding regions and also in noncoding regions from virus and mithocondria); and a periodicity two, $P2$ was identified in eukaryotic introns. The last periodicity was attributed to regulatory functions (2).

Although, using a mutual information function to distance series (joint probabilities of finding the symbol A_i and k characters downstream the symbol A_j), Herzel and Große concluded that the

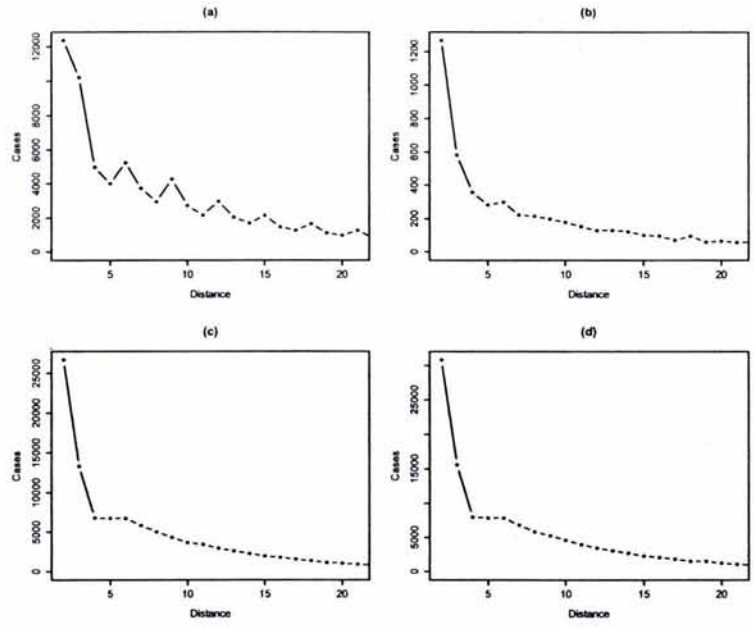

Figure B.2. Distribution of distances. The number of cases of triplet YRY for the first 21 distances are plotted for (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

nonuniform codon usage in protein coding sequences, is responsible of the period-three oscillations (12), this periodicity has also been found in whole chromosomal bacterial genomes, with no relation with protein translation process (16). Thus, we believe that the periodicity three is an intrinsic property of coding sequences, independent of the codon usage.

B.4.2. Autocorrelation function (ACF)

ACF allow us to prove the null hypothesis over individual data independence in a time series. Let $x(a)$ be a time series, and $x(a-\tau)$ the same series with a τ positions delay. The general ACF computes the correlation between $x(a)$ and $x(a-\tau)$ using the following equation:

$$ACF = \langle x(a)x(a-\tau)\rangle - \langle x(a)\rangle\langle x(a-\tau)\rangle \tag{B.1}$$

In equivalence with the correlation coefficient of Pearson, a value of $ACF = 1$ is indicative of a complete positive autocorrelation between the series; an $ACF = -1$ is indicative of a complete negative autocorrelation between the series; and an $ACF = 0$ is indicative of independence between the series, i.e. it is related with Gaussian white noise.

ACF has been used by Arquès and Michel to study both the $YRY(N)_6YRY$ preference in different

kinds of genomes with random mutations (4), and to identify subsets of triplets having a preferential occurrence frame (5). Following a similar approach, we obtained actual distance series for the ATG triplet, calculating the number of cases for each distance, and then computing the ACF. The results are shown on Fig. B.3.

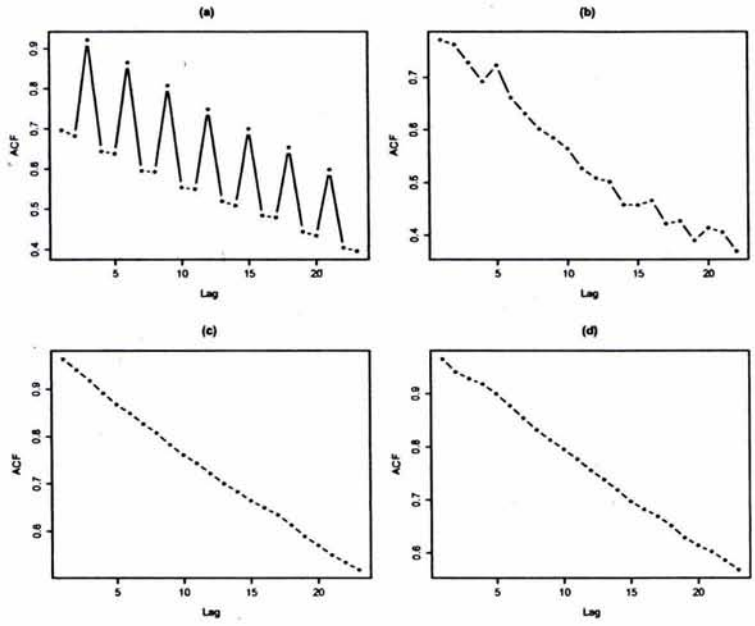

Figure B.3. Autocorrelation function analysis (ACF). The ACF is computed from the distribution of ATG (number of cases vs. position) for (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

As shown on Fig. B.3, while a coding sequence presents an oscillatory decaying pattern with a clear-cut rhythmical alternation of points, which are at distances of multiples of three (Fig. B.3a), a noncoding sequence has no apparent periodicity (Fig. B.3b), with a pattern similar with stochastic processes (see Fig. B.3c and d). The dynamics observed in the coding DNA sequence, is typical of an scale-invariant power-law behavior.

B.4.3. Nearest neighbor nucleotide patterns

Several physicochemical properties of DNA depend on the interactions between consecutive bases, thus, the identification of patterns from nearest neighbor bases could help in the characterization of nucleotide sequences (22).

Although, the group of Korenberg was the first to measure nearest neighbor frequencies on DNA (17), it was not until recently, when some patterns were identified from the analysis of whole genomes (24,25).

Nussinov counted the number of different dinucleotides, and found two kinds of patterns: (a) unequal frequencies of appearance of some asymmetric pairs, and (b) preferences of certain nucleotides with specific nearest neighbors over equivalent dinucleotides (24). In the first case, she found that the asymmetries AT >TA; CT >TC; TG >GT; and, GC >CG occurs in all the examined genomes, including both prokarytes and eukaryotes. On Table B.1 the counts differences, for these duplets are shown.

Table B.1. Differences in nearest neighbor counts.

Duplet	Original genome	HUMHBB	Shuffled genome	Synthetic genome
AT – TA	13557	589	62	-72
CT – TC	2016	616	34	208
TG – GT	15572	1205	94	135
GC – CG	17577	1648	24	-393
average	12180.5	1014.5	53.5	-30.5

As shown on Table B.1, the highest differences in the counts were detected on the coding genome (original) of *Borrelia burgdorferi*. There is one order of magnitude in the difference between a coding and a noncoding sequence (HUMHBB), and three orders of magnitude in the difference between a coding sequence and its corresponding shuffled version. Furthermore, in the case of the pure random control (synthetic genome), a switch in the relative counts were detected.

B.4.4. Long-range correlations

A power-law behavior of the form $y = f(x) = Ax^{\alpha}$, where α is the scaling exponent and A is the normalization constant, is related with processes exhibiting self-similar properties (fractal dynamics), such as time series with long-range correlations (37). As DNA sequences can be transformed to distance series, it is feasible to characterize long-range correlations in both coding and noncoding regions.

Using the one-dimensional random walk model, discussed before, Peng, *et al.* applied different, but related, techniques to study long-range correlations in DNA (27,28,35,36). Their first approach was the use of the root-mean square fluctuation, $F(l)$ about the average of the displacement, defined as:

$$F(n) = \sqrt{\overline{[\Delta y(n) - \overline{\Delta y(n)}]^2}} \tag{B.2}$$

where $\Delta y(n) = y(n_0 + n) - y(n_0)$, and the bars indicate the arithmetic mean over all positions n in the gene. There are two possible scenarios: (a) for both pure random process, and for local correlations, $F(n) \sim n^{1/2}$; and (b) for correlations with no characteristic length (long-range correlations), the fluctuations are described by the power law, $F(n) \sim n^{\alpha}$, with $\alpha \neq 1/2$. Using this

method (known as "min-max"), Peng, *et al.* found long-range correlations in noncoding regions, in contrast with coding regions where $\alpha \approx 0.5$ (27).

The above results have been widely discussed, as some authors argue that there is no difference between coding and noncoding regions (39). One of the main critics, was the finding that there is heterogeneity in the random walk series obtained from DNA, thus, it has been claimed that DNA presents "patchiness" (19). Patchiness can be clearly detected on Figs. B.1a and b, in which the "walker" moves far away from the origin (compare with Figs. B.1d which resembles a pure random process). Due to DNA patchiness, methods like ACF and the root-mean square of fluctuations are not valid, as they depend on averages, which in turn, change over the DNA sequence.

In order to avoid the effect of DNA patchiness, Peng, *et al.*, improved their method by detrending the fluctuations over different windows or boxes (28). This technique was called detrended fluctuation analysis (DFA).

In DFA, first a sequence of length N is divided into N/l nonoverlapping boxes, each containing l nucleotides, and the local trend for each box is calculated. Then, the detrended walk, $y_l(n)$ is obtained with the difference between the original random walk, $y(n)$ and the local trend. Next, the variances about the detrended walks are computed; and, finally, the averages of these variances over all the boxes $(F_d^2(l))$ are calculated (28).

The reported results of Peng, *et al.*, were essentially the same as before (28), i.e. a long-range correlations were detected on noncoding regions. In the case of the analyzed coding sequence, a crossover in the slope was detected, with an $\alpha = 0.51$ for the first part of the curve (in equivalence with pure random sequences).

There have been other approaches in order to eliminate local patchiness in DNA. Arneodo, *et al.*, introduced the use of the wavelet transform modulus maxima (WTMM) to study long-range correlations in DNA sequences (1). In the WTMM the scaling properties of a time series is investigated in terms of their wavelet coefficients. By applying the WTMM to a DNA random walk series, from both coding and noncoding sequences, they also found long-range correlations in noncoding sequences, and uncorrelated steps, undistinguishable from Brownian motion, in coding sequences (1).

In contrast with the above results, other authors have found long-range correlations in coding sequences (6,30). In particular, instead of starting with the random walk series, Voss (39) and Sousa Vieira (38) calculated the power spectrum into equal-symbol correlation series, whereas Mohanty and Narayana Rao applied factorial moments to series representing the excess or deficit of purines over pyrimidines (23). In these cases, long-range correlations were identified in large coding sequences.

In order to illustrate the long-range correlations in DNA sequences, here, we applied DFA to both, a time series obtained from a one-dimension random walk method, as well as a time series obtained from the actual distance series of triplet ATG. The results for the latter, are shown on Fig. B.4.

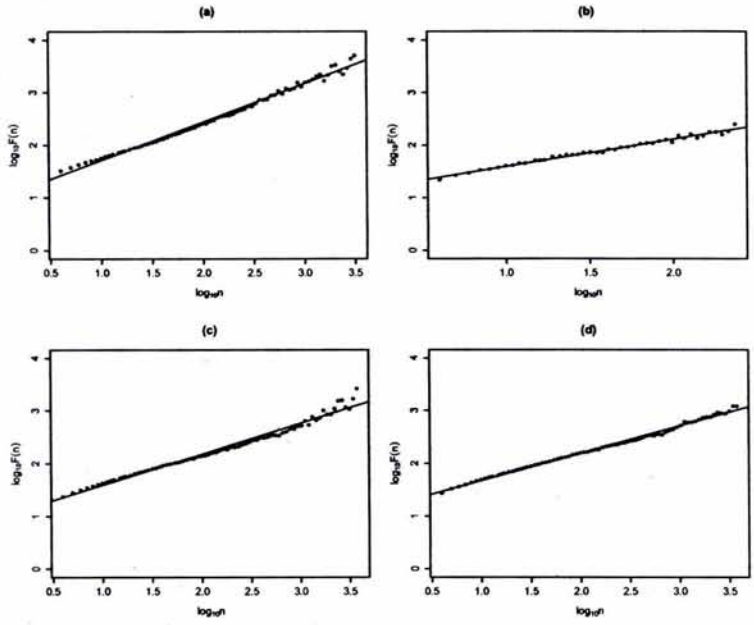

Figure B.4. Detrended fluctuation analysis (DFA) for the actual distance series of triplet ATG in (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome. The corresponding scaling exponents are shown on Table B.2.

As shown on Fig. B.4a, there was not a crossover in the slope of the curve. On Table B.2 we present the comparison of the obtained scaling exponents α, for both time series.

Table B.2. Calculated scaling exponents α, for the studied cases.

Genome	Random walk α	ATG distance α
Original (coding)	0.62*	0.73
HUMHBB (noncoding)	0.67	0.52
Shuffled	0.50	0.59
Synthetic	0.50	0.51

*Overall α, see text for explanation.

In accordance with Peng, *et al.*, (28), we detected crossovers on the coding sequence. In our case, two crossovers were identified (not shown), thus three different scaling exponents could be obtained: $\alpha_1 = 0.68$, with $n = 11$, number of points; $\alpha_2 = 0.52$, with $n = 29$; and, $\alpha_3 = 0.80$, with $n = 36$. On Table B.2 we present the overall scaling exponent ($n = 76$).

On the other hand, in contradiction with Peng, *et al.*, (28), we detected long-range correlations in a coding sequence using both kinds of time series as input. In fact, the highest value of α was

obtained from the ATG distance series from the coding genome of *Borrelia burgdorferi*. It is worth to mention, that this sequence, does not have any intergenic regions, thus is pure coding. We believe that using actual distance series of triplets are a better option than using the one-dimension random walk method, as no information is lost (due to binary representations) (31), and is more biological related (due to translation) (9).

Another approach to look for fractal dynamics on DNA sequences has been the use of the chaos game representation (CGR) of gene structure (11,13,15,26). CGR is a scatter plot derived from a DNA sequence. First a CGR image is divided into squares in which each corner represents one base. Starting from a random point (e.g. (0,0)) the next point is plotted in the mid point from the straight line which connects the current point with one of the corners, determined by the DNA sequence. On Fig. B.5, we illustrate CGRs from the DNA sequences discussed in the current paper. As expected, there is no pattern on the synthetic genome (Fig. B.5d). A fractal dynamics was observed on both the coding (Fig. B.5a) and noncoding sequences (Fig. B.5b), although it was more clear on the latter. CGRs were applied for the first 50,000 bases in the corresponding sequences.

B.4.5. Information content

Information can be measured in terms of the number of binary digits, i.e. by the logarithm of the number of possible messages (29). This measure of information is called the Shannon entropy (32):

$$H_n = -\sum_{i=1}^{n} p_i log_2 p_i \tag{B.3}$$

The term entropy is due to its relation with certain formulations of statistical mechanics where p_i is the probability of a system being in the cell i of its phase space (32).

In the case of two possible outcomes, with probabilities p and $q = 1 - p$, the Shannon entropy reaches its maximum value when $p = q$. This result can be generalized for any n, number of probabilities, thus, H_n is a maximum when all the p_i are equal, which is the most uncertain situation.

Several authors have used the Shannon entropy to analyze DNA information content (8,14,21). Here, in order to illustrate the information content quantification in different sequences, we computed the Shannon entropy for the frequencies of all possible triplets (64) in the previously mentioned genomes. The results are show on Table B.3.

Note that the entropy value from the synthetic genome had the expected value for a pure random process ($H_n \approx 6$). On the other hand, although the coding genome of *Borrelia burgdorferi* had the minimum entropy (i.e. more information content), its value was very closed to its shuffled version; this was unexpected, as the shuffling was made by nucleotides and not by triplets.

Another alternative to calculate entropies from nonlinear time series, is the maximum entropy

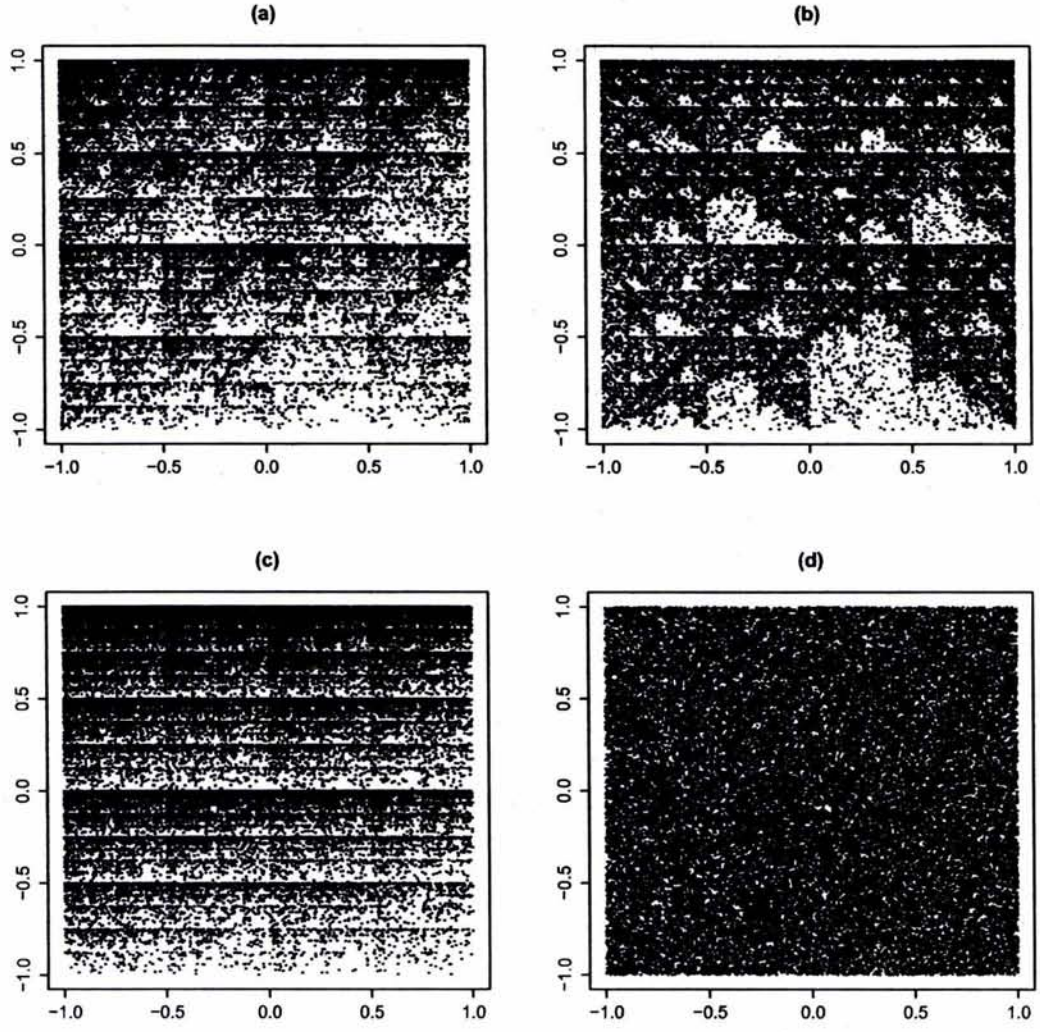

Figure B.5. Chaos game representation (CGR) applied to DNA sequences. *A* shifts to the left upper corner; *T* shifts to the upper right corner; *C* shifts to the lower left corner; *G* shifts to the lower right corner. CGR was applied to: (a) coding genome of *Borrelia burgdorferi*; (b) HUMHBB; (c) shuffled genome of *Borrelia burgdorferi*; and (d) synthetic genome.

Table B.3. Shannon entropy for triplets frequencies.

Genome	H_n
Original (coding)	5.60
HUMHBB (noncoding)	5.82
Shuffled	5.63
Synthetic	5.99

method (MEM), which is based upon the power spectrum of autocorrelation coefficients (10). The MEM has been recently used to study information content in series of amino acids obtained from translating whole bacterial chromosomes (9). In this case, the information content was proportionally related with the maximum entropy.

B.5. Concluding remarks

It is important to mathematically distinguish coding DNA sequences from non-coding ones, because through these kind of tools it is possible to identify quickly potential genes in the genome data bases, saving valuable time for a better experimental design. Furthermore, mathematical characterization of DNA sequences could help in the understanding of structural relationships among different genes along the chromosomes.

Although there has been controversies among the presence of long-range correlations in coding DNA, we have shown that the use of actual distance series between triplets is a better approach than the random walker DNA representation, as less information is lost, and a better characterization is made. When the analysis are carried out based upon the actual distance series, the presence of long-range correlations in coding sequences is clear and its in accordance with the CGR of the same sequence.

The presence of periodical rhythms in the ACF, long-range correlations and more information content in coding DNA sequences suggests that, although spontaneous mutations and horizontal genetic transfer occurs at random, there should be some kind of structural rules which favor the natural selection of sequences in which these properties are maintained.

B.6. Acknowledgments

We thank Julio Collado and Imelda López for valuable interactions and reading of the manuscript.

B.7. References

1. A. Arneodo, E. Bacry, P.V. Graves and J.F. Muzy *Phys. Rev. Lett.* **74** (1995) 3293–3296.
2. D.G. Arquès and C.J. Michel. *Nucleic Acids Res.* **15** (1987) 7581–7592.
3. D.G. Arquès and C.J. Michel. *J. theor. Biol.* **143** (1990) 307–318.
4. D.G. Arquès and C.J. Michel. *Math. Biosci.* **123** (1994) 103–125.
5. D.G. Arquès and C.J. Michel. *Biosystems* **44** (1997) 107–134.

6. C.A. Chatzidimitriou-Dreismann and D. Larhammar. *Nature* **361** (1993) 212-213.

7. M. Eigen and P. Schuster. *The hypercycle. A principle of natural self-organization.* (1979) Berlin, Alemania. Springer-Verlag.

8. L. Frappat, C. Minichini, A. Sciarrino and P. Sorba. *Phys. Rev. E* (2004) In press.

9. J.A. García, S. Alvarez, A. Flores, T. Govezensky, J.R. Bobadilla and M.V. José *Physica A* (2004) In press.

10. M. Ghil, *et al. Rev. Geophys.* **40** (2002) 1–41.

11. N. Goldman. *Nucleic Acids Res.* **21** (1993) 2487–2491.

12. H. Herzel and I. Große. *Phys. Rev. E* **55** (1997) 800–810.

13. K.A. Hill, N.J. Schisler and S.M. Singh. *J. Mol. Evol.* **35** (1992) 261–269.

14. D. Holste, I. Große and H. Herzel. *Phys. Rev. E* **64** (2001) 1–9.

15. H.J. Jeffrey *Nucleic Acids Res.* **18** (1990) 2163–2170.

16. M.V. José, J.A. García, J.R. Bobadilla, and T. Govezensky. International Conference on Biological Physics, Gothenburg (2004).

17. J. Josse, A.D. Kaiser and A. Korenberg. *J. Biol. Chem.* **236** (1961) 864–875.

18. T.H. Jukes. *J. Mol. Evol.* **42** (1996) 377–381.

19. S. Karlin and V. Brendel. *Science* **259** (1993) 677–680.

20. J. Konecny, M. Schöniger and L. Hofacker. *J. theor. Biol.* **173** (1995) 263–270.

21. R.N. Mantegna, *et al. Phys. Rev. E* **52** (1995) 2939–2950.

22. P. Miramontes, *et al. J. Mol. Evol.* **40** (1995) 698–704.

23. A.K. Mohanty and A.V.S.S. Narayana Rao. *Phys. Rev. Lett.* **84** (2000) 1832–1835.

24. R. Nussinov. *Nucleic Acids Res.* **8** (1980) 4545–4562.

25. R. Nussinov. *J. Biol. Chem.* **256** (1981) 8458–8462.

26. J.L. Oliver, P. Bernaola-Galván, J. Guerrero-García, and R. Román-Roldán. *J. theor. Biol.* **160** (1993) 457–470.

27. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley. *Nature* **356** (1992) 168–170.

28. C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L Goldberger. *Phys. Rev. E* **49** (1994) 1685–1689.

29. J.R. Pierce *An introduction to information theory. Symbols, signals and noise.* (1980) New York, USA, Dover Publications, Inc.

30. V.V. Prabhu and J.M. Claverie. *Nature* **359** (1992) 782.

31. J. Sánchez and M.V. José. *Biochem. Biophys. Res. Comm.* **299** (2002) 126–134.

32. C.E. Shannon. *Bell Syst. Tech. J.* **27** (1948) 379–423.

33. J.C.W. Shepherd. J. Mol. Evol. **17** (1981) 94–102.

34. J.C.W. Shepherd. *Proc. Natl. Acad. Sci. USA* **78** (1981) 1596–1600.

35. H.E. Stanley, *et al. Physica A* **191** (1992) 1–12.

36. H.E. Stanley, *et al. Physica A* **205** (1994) 214–253.

37. H.E. Stanley, *et al.* In: *Fractal geometry in biological systems. An analytical approach.* P.M. Iannacone and M. Khokha, eds. (1996) New York, USA, CRC Press, pp 15–30.

38. M.S. Vieira. *Phys. Rev. E* **60** (1999) 5932–5937.

39. R.F. Voss. *Phys. Rev. Lett.* **68** (1992) 3805–3808.

40. J.T.F. Wong and R. Cedergren *Eur. J. Biochem.* **159** (1986) 175–180.

Apéndice C

Introducción a conceptos biológicos

And the winds of Earth continued blowing gently.
And the waters of the Earth were continually stirred,
while the mud of shallow lakes and pools was continually mixed.
And nucleotides, amino acids and minerals were intermingled.
And new molecules were made.
And among them were certain organic complex molecules.
And these complex molecules were proto-proto-cells.
And they were continually being formed.[1]

[1]Biogenesis I. The Bible according to Einstein. New York. USA. Jupiter Scientific Publishing Company (1997).

C.1. Los secretos de la vida

Uno de los problemas conceptuales más antiguos en la historia de la biología, es la definición de lo que se considera como vida. El acercamiento más usado ha sido a través del análisis comparativo con la materia inerte. En este sentido, las características particulares descritas para los seres vivos llevaron a los físicos del siglo XIX a pensar que sería posible encontrar nuevas leyes físicas en los fenómenos biológicos; esto llevó, por ejemplo, a Schrödinger dar una serie de conferencias en Dublín en 1943, que culminaron con la publicación de su obra clásica: ¿Qué es la vida? (89).

Históricamente se han manejado dos versiones en torno al problema de la vida, el *vitalismo* y el *mecanismo*. Dado que la comunidad científica, prácticamente ha descartado la primera, se considera solamente la última, cuyas propuestas pueden clasificarse en dos vertientes: fisicoquimicalismo y maquinismo (64). De acuerdo a los primeros, los organismos representan simplemente sistemas físicos o fisicoquímicos altamente complejos, sin características propias. En cambio, para los últimos, los organismos además de representar estos sistemas físicos, son sistemas tipo máquina, o bien, propiamente, máquinas (64). Recientemente, Mahner & Bunge han propuesto una tercera opción, denominada "biosistemismo", que sostiene que: (a) los sistemas vivos, aunque compuestos por subsistemas fisicoquímicos, poseen propiedades emergentes, que sus componentes carecen, y (b) las unidades de la ciencia biológica son los organismos en su medio, así como sus varios subsistemas (moléculas, células, órganos) y supersistemas (poblaciones, comunidades, ecosistemas) (64). De esta manera, Mahner & Bunge proponen que un ser vivo es un sistema material, tal que:

1. Su composición incluye ácidos nucleicos y proteínas.
2. Su medio ambiente incluye algunos de los precursores de sus componentes (y por tanto permite el auto-ensamblaje de las biomoléculas).
3. Su estructura incluye la capacidad de metabolizar así como de auto-mantenimiento y auto-reparación (dentro de ciertos límites).

Con el objetivo de presentar las principales características de los seres vivos, a continuación se agrupan éstas en tres apartados: los secretos genéticos, los de complejidad y los físico-matemáticos.

C.1.1. Los secretos genéticos

La característica más notable captada originalmente fue la capacidad de los seres vivos de reproducirse. Claramente, el encontrar los misterios de la herencia fue el objetivo biológico primordial de principios del siglo XX. Así, comenzó una carrera científica entre diferentes grupos de investigación que culminó con los trabajos de Watson & Crick (111).

La aportación del modelo del ácido desoxirribonucleico (DNA) a la biológía fue fundamental al avance del entendimiento de la vida. Esta molécula se convirtió en el centro de estudio biológico. Se dio "el octavo día de la creación" (49).

Las características estructurales y funcionales del DNA pueden encontrarse en libros de texto de Bioquímica (20) y Genética (63), por lo que sólo se hará mención de algunos puntos de interés.

Ahora sabemos que los procesos biológicos de las células están controladas fundamentalmente por proteínas. Estas biomoléculas poseen, al menos, tres características esenciales que les permiten facilitar numerosos procesos en los organismos: son específicas; son regulables y son dinámicas (su conformación tridimensional puede modificarse dependiendo de los solutos con los que interactúa).

La información para la síntesis de proteínas está codificada en la molécula del DNA. Al replicarse esta molécula se asegura que la información será transmitida -con cierta fidelidad- a la siguiente generación. Parte de la "magia" del DNA reside en la especificidad de la interacción entre las bases púricas y pirimídicas: la adenina (A) se asocia con timina (T) y la guanina (G) con citosina (C). Se tienen entonces dos cadenas complementarias que pueden replicarse, transmitiendo su información a generaciones posteriores.

El llamado "dogma central" de la genética establece la replicación del DNA, la transcripción del ácido ribonucleico (RNA) y la traducción en una molécula ejecutora: la proteína (14). Un esquema no exhaustivo de la relación entre estas moléculas se presenta en la Fig. C.1.

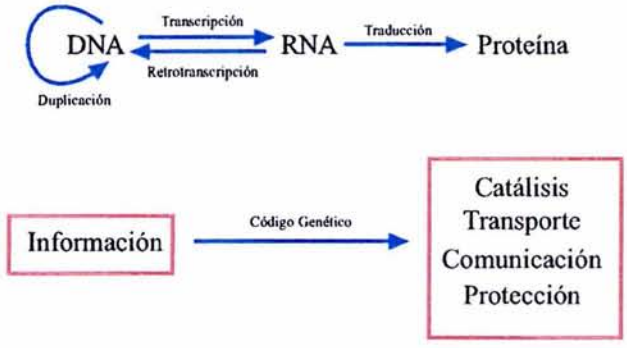

Figura C.1. Esquema no exhaustivo de la relación entre DNA, RNA y proteínas. En este esquema la información fluye de manera unidireccional de la molécula del DNA a las proteínas, las cuales son responsables de controlar varios procesos en la célula (se indican los principales). Algunos virus tienen la capacidad de hacer el proceso de retrotranscripción indicado en la figura.

Cuando un fenómeno biológico es entendido a nivel molecular, entonces es posible regularlo y manipularlo. La biología molecular, con esto, se ha convertido en una ciencia autónoma y auxiliar fundamental para el análisis profundo de los procesos biológicos.

Con respecto al papel que juega el DNA en los organismos, algunos autores han llegado al extremo reduccionista de explicar todos los fenómenos biológicos como una consecuencia de la acción de una molécula *egoísta* (19). Como se menciona a continuación, la vida es más compleja y requiere, para su entendimiento, de otros factores además del genético.

C.1.2. Los secretos de complejidad

Para entender la vida podemos comenzar analizando las características fundamentales de las mínimas entidades posibles que aún consideramos materia vida: la célula. Las células son entidades autónomas delimitadas por una membrana, y contienen todos los elementos para que en un medio favorable puedan dividirse.

Las células conocidas más simples siguen representando sistemas muy complejos con múltiples interacciones entre sus diferentes componentes. Resulta difícil explicar la aparición de la primera célula por los mecanismos clásicos de selección natural. Esta complejidad puede ser responsable, en cambio, del orden encontrado en los seres vivos.

Kauffman (57) ha analizado el comportamiento de sistemas complejos que se van desarrollando a partir de reglas de interacción sencilla. Por ejemplo, puede pensarse en una serie de nodos inicialmente no conectados, e ir aumentando de manera sucesiva las interacciones, de manera aleatoria. Cuando uno lleva a cabo este procedimiento, se obtienen redes que *per se* dan origen a sistemas más conectados como se muestra en la Fig. C.2.

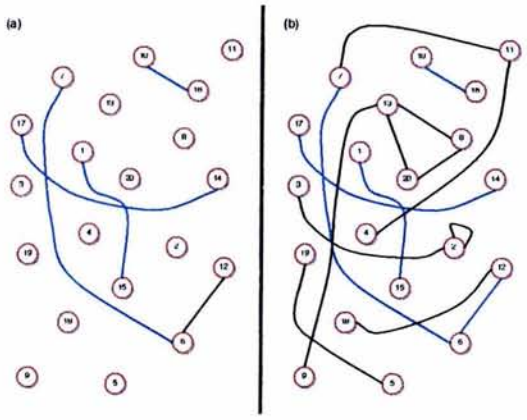

Figura C.2. Redes de conexión propuestas por Kauffman. En estas redes, un número de nodos son conectados aleatoriamente (adaptado de (57)). (a) Caso para 20 nodos con 5 aristas. (b) Caso con 20 nodos y 15 aristas.

Una de las propiedades más notables de estos sistemas autónomos (que se mantienen por sí mismos), es que siguen cierto tipo de patrones deterministas (es decir, no aleatorios). No se puede asegurar que el modelo de Kauffman sea trasladable a la vida real, sin embargo, sí es un hecho la alta complejidad de los sistemas biológicos.

En este ámbito de sistemas complejos surgen los conceptos de auto-organización y orden. Los sistemas biológicos parecieran oponerse a la segunda ley de la termodinámica que establece la tendencia al desorden (entropía). La explicación es relativamente sencilla: los seres vivos mantienen su orden interno a costa de un aumento de la entropía del universo que, efectivamente, aumenta.

Debido fundamentalmente a las características de las membranas biológicas, los seres vivos mantienen un desequilibrio con el medio que los rodea. Aunque el costo energético es alto, este procedimiento permite la supervivencia de la célula. De hecho podemos considerar que la muerte celular se da cuando se alcanza el equilibrio con el medio. Bajo este marco, surge la definición de Weiner, el padre de la cibernética, de los seres vivos: "demonios de Maxwell metaestables cuyo punto de equilibrio es la muerte". Los demonios de Maxwell se explican en la sección C.3.3.

C.1.3. Los secretos físico–matemáticos

La vida se desarrolla dentro de los marcos del universo físico con sus leyes, patrones, formas, estructuras, procesos y sistemas. Aunque los genes son los responsables de orientar la dirección que deben tomar los diferentes organismos, son las leyes de la física y la química las que controlan la respuesta de éstos a las instrucciones genéticas. En este sentido, estas leyes representan el otro secreto de la vida (102).

Uno de los primeros investigadores en establecer un posible vínculo entre el mundo biológico y las matemáticas fue D'Arcy Thompson, quien se concentró en encontrar reglas matemáticas que pudieran explicar ciertos patrones encontrados en los organismos vivos (105). Es así como Ian Stewart rescató y actualizó los trabajos de D'Arcy Thompson bajo el marco de los paradigmas genéticos actuales (102).

Para Stewart el DNA es necesario, pero no suficiente. Lo primero que se tendría que trabajar es un concepto de vida que no involucrara al DNA. Esto es en el sentido de que la vida es un proceso y no una sustancia. La diferencia entre lo que está vivo y lo que no es vida no se encuentra en las partes que los constituyen, sino en la manera en que están organizadas éstas.

No sabemos si hay otros tipos de vida en el universo, pero dado que la posibilidad existe, debe concebirse que estas formas pudieron haber surgido de maneras de organización muy diferentes, que no incluyeran al DNA, con lo cual se podría considerar a esta molécula como un "accidente local" para la presencia de la vida en nuestro planeta (102).

A pesar de que los organismos vivos están constituidos por elementos que encontramos en la química inorgánica, parecen trascender la rigidez de estos orígenes. A esta clase de trascendencia se le ha llamado "comportamiento emergente", el cual representa una red de casualidades tan intrincada que la mente humana no puede captarlo (102). Estos comportamientos emergentes no pueden ser explicados sólo con la molécula del DNA.

El DNA nos ha permitido comprender muchos procesos a nivel molecular incluyendo la replicación y la transmisión de la información. Sin embargo, la autonomía es una característica que tendría que ser explicada de otra manera. La alternativa es a través del entendimiento de las leyes de la física y matemáticas que rigen los fenómenos a nivel molecular.

Existen muchos ejemplos en la naturaleza donde podemos encontrar a las matemáticas. Los más perceptibles son a través de la geometría que resulta de la morfogénesis de los organismos vivos.

Así, la concha del caracol tiene una espiral logarítmica; la disposición de las flores de girasol contienen a la serie de Fibonacci; muchos vegetales tienen estructuras fractales -que se repiten a sí mismos-; etcétera.

Asimismo, las matemáticas han servido de herramienta para el entendimiento de muchos fenómenos biológicos. Por ejemplo, varios de los avances en la relación hospedero-parásito del virus de la inmunodeficiencia humana (HIV) se han establecido con la ayuda de modelos matemáticos. Tan sólo en el área de la biología molecular, los algoritmos matemáticos han sido empleados para la predicción de antígenos vacunales, la predicción funcional de proteínas, el análisis del proyecto del genoma humano, la relación filogenética de las proteínas, etc. En el área de la ecología son clásicas las ecuaciones de Lotka-Volterra que estudian la relación depredador-presa.

C.2. El origen de la vida

Una de las interrogantes que más ha preocupado al hombre y que sigue sin resolverse satisfactoriamente, es aquélla sobre el origen de la vida en la Tierra. A ciencia cierta no sabemos mucho sobre esto y quizá nunca lo sepamos. Las teorías creacionistas no son consideradas científicas y por otra parte, las teorías evolucionistas aún se encuentran incompletas. De ahí que en esta sección se debe tomar la precaución de que los conceptos aquí vertidos son el resultado de conceptualizaciones teóricas, apoyadas experimentalmente que, sin embargo, no pueden comprobarse aún.

Los estudios de los científicos han llevado a la conclusión de que nuestro universo actual tiene aproximadamente 14 mil millones de años, mientras que nuestro sistema solar (incluida la Tierra) tiene 4 600 millones de años (17, 35). Las sustancias más simples que pudieron existir dentro de lo que llamamos "vida" fueron probablemente agregados moleculares que incluyeron ácidos nucleicos y proteínas.

Existen evidencias que sugieren que todos los organismos que actualmente conocemos provienen de un ancestro común (35), aunque no se puede saber si esto es producto de un origen monofilético (que la vida haya aparecido sólo una ocasión), o bien de un proceso de selección.

Blomberg (7) ha propuesto tres estadios principales para el estudio del origen de la vida:

1. Fase química: producción espontánea de moléculas orgánicas sencillas y posteriormente de macromoléculas. Las interrogantes incluyen el lugar donde comenzaron a darse estas reacciones, así como la aparición de la primera molécula autorreplicante.
2. Fase de organización: caracterizada por la síntesis de macromoléculas y la aparición de otras actividades catalíticas. Esta fase abarca desde el desarrollo de la primera macromolécula autorreplicante, hasta la aparición de la primera célula. Corresponde al llamado mundo de RNA.
3. Fase del desarrollo de la vida: caracterizada por el desarrollo de las primeras células autónomas, y que se convirtieron en el ancestro común de las actuales.

Uno de los puntos más importantes es, sin lugar a dudas, la aparición de las primeras macromoléculas. Este ha sido un problema clásico de qué fue primero, ¿el huevo o la gallina? Se sabe de la importancia que tienen los ácidos nucleicos como portadores de información con la capacidad de replicarse, sin embargo, también se sabe de la necesidad de una molécula como la proteína para regular el proceso. Dado el descubrimiento de las ribozimas (10), se ha postulado que la primera molécula pudo haber sido el RNA, ya que ésta lleva información genética y podría ser capaz de catalizar su propia replicación. En este sentido se habla de un mundo de RNA (39).

La teoría del mundo del RNA aún está muy incompleta, a pesar de ser la más aceptada en la comunidad científica. Una de sus problemáticas principales se encuentra en tratar de explicar la alta estabilidad celular, a partir de un mundo teóricamente inestable que, además, no sigue los patrones de selección natural establecidos por la Síntesis neodarwiniana. Para explicar la aparición de la primera célula, se han elaborado tesis alternativas que incluyen un mundo asistemático, genéticamente controlado, de producción de proteínas (7), así como el desarrollo de redes proteínicas interconectadas que propiciaron el orden de manera espontánea en la primera célula (57).

La última tesis es resultado de un grupo de científicos norteamericanos que desde mediados de los 80's han estudiado los fenómenos complejos en "la era del orden y el caos" en el Instituto Santa Fe (110). Este grupo considera que la aparición del orden es espontánea y que se puede explicar bajo la teoría de la "complejidad", que es el paradigma central del mencionado instituto. En el caso de la aparición del orden en los procesos biológicos, caben destacar los trabajos de Kauffman (56, 57) quien propone que más que pensar en una sola macromolécula precursora, se puede concebir a un conjunto de macromoléculas interrelacionadas con actividades catalíticas, que de manera espontánea, dieron origen a sistemas más complejos, organizados en sistemas limitados como los "coascervativos" de Oparin (81). Existen evidencias recientes que muestran como sistemas de péptidos autocatalíticos pueden dar origen a moléculas más complejas (62).

C.3. La Síntesis neodarwiniana

C.3.1. Desarrollo de la Síntesis

A finales del siglo XVIII y principios del siglo IXX en la biología dominaba el concepto de las especies fijas que no cambian. Si los organismos literalmente no evolucionan en otras especies, como actualmente creemos, si están fijas por siempre, ¿a qué se deben las profundas similitudes entre los organismos? Aún atribuyendo esas similitudes al trabajo de Dios, uno desearía tener un creador racional. Para los científicos de aquella época, estas similitudes se interpretaron como reglas de forma establecidas por el estudio de las anatomías comparadas.

Además de la búsqueda de estas normas biológicas, existía una tradición filosófica coherente fundamentada sobre todo en Kant (112) quien distinguió a los organismos de los aparatos mecánicos. Para Kant, los organismos eran fundamentalmente auto-replicativos, por tanto, entes auto-organizados. En los aparatos mecánicos, las partes existen sólo para las otras partes, en el sentido

que el conjunto de todas ellas llevará a cabo una función. En contaste, en los organismos las "partes" existen tanto para las otras como por las otras. Para Kant un organismo "es aquello en lo que todo es a la vez fin y medio". En la "Crítica del Juicio", Kant argumenta la necesidad de usar principios teleológicos para hacer inteligible a la organización biológica (53).

Estos conceptos cambiaron radicalmente con el desarrollo de la Síntesis neodarwiniana, la cual se formó a partir de las siguientes aportaciones:

1. La teoría de la evolución de Darwin (16).
2. Las leyes de la transmisión genética de Mendel (70).
3. El concepto de desarrollo de las células germinales de Weismann (113).
4. La genética de poblaciones de Fisher (29).

C.3.2. La Síntesis neodarwiniana y los neutralistas

La Síntesis ha buscado entender las condiciones bajo las cuales alelos mutantes con ventajas escasas en uno o varios genes pueden esparcirse en una población. El análisis ha tratado la influencia del tamaño de la población, la ventaja selectiva del alelo, y las propiedades relativas de homocigotos y heterocigotos para diferentes alelos de un mismo gen. En la corriente de la genética de poblaciones, la Síntesis ha examinado los efectos de la relación de varios genes en un cromosoma y los efectos de la recombinación homóloga.

Por otra parte, la teoría neutral de Kimura (58) representa la argumentación más consistente a favor del papel asignado a los procesos estocásticos (*random drift*). Esta teoría afirma que toda o casi toda la evolución a nivel molecular se debe a procesos estocásticos entre variantes genéticas neutrales seleccionadas, inclusive en poblaciones grandes. Trata de entender hasta que punto la substitución evolutiva de un alelo por otro es o no es debido a diferencias selectivas, o bien a fluctuaciones azarosas.

C.3.3. El demonio de Maxwell y la selección natural

Para el análisis de la selección natural, Kauffman hace la analogía de ésta con el demonio de Maxwell (56). Este concepto se presenta en forma esquemática en la Fig. C.3.

Considérese una caja cerrada con dos cuartos separados por una pared con una válvula. En el interior de la caja existen moléculas de gas con una temperatura promedio constante. El demonio es capaz de distinguir la rapidez de cada molécula, de modo que acumula a las moléculas más rápidas en uno de los cuartos. El resultado es que al paso del tiempo, el demonio consigue aumentar la temperatura de uno de los cuartos, con respecto al otro, sin emplear energía; el demonio rompe entonces con la segunda ley de la termodinámica.

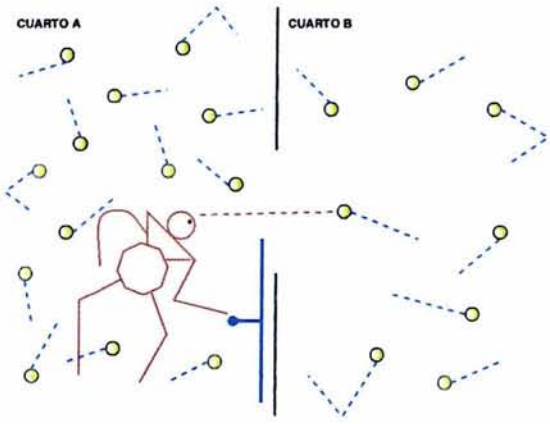

Figura C.3. Demonio de Maxwell. El demonio puede detectar la energía cinética de las partículas contenidas en los 2 compartimentos y decidir abrir la puerta para permitir el paso de alguna de ellas.

Cuando el demonio actúa, la presión en una de los cuartos aumenta y se opone al esfuerzo del demonio. Si el demonio es lo suficientemente fuerte, él tendrá éxito en separar a las moléculas más rápidas de las lentas. Sin embargo, si el demonio es finito, y más bien débil, entonces será capaz de movilizar sólo algunas moléculas rápidas en uno de los cuartos, antes que éste se le oponga con la presión del gas.

Para Kauffman, la selección es análoga al demonio, en el sentido que ésta tratará de llevar a la población a adquirir propiedades raras en el ensamble, mientras que la "presión de retorno" de las mutaciones hacia las propiedades estadísticamente típicas del ensamble aumentan (56).

C.4. Nuestro lugar en el universo

Recientemente se celebraron los 50 años de las conferencias de Schrödinger en Dublín sobre las características de los seres vivos. El comentario de estas conferencias fue publicado en Science y contiene las tres preguntas filosóficas claves de los biólogos (80).

El comentario comienza planteando un escenario en el cual uno podría obtener la respuesta a cualquier pregunta que uno hiciera -"la pregunta al ángel"-, ¿Qué pregunta se haría? Los físicos, químicos y biólogos coincidieron en que las preguntas serían ¿Quiénes somos? ¿De dónde provenimos? ¿Hacia dónde vamos?

C.4.1. ¿Quiénes somos?

En términos biológicos, la respuesta puede no agradar a mucha gente, ya que nuestro DNA tiene una homología del 99 % con el de los chimpancés. De hecho para algunos investigadores, la clasificación de nuestra especie debería corresponder como una tercera clase de chimpancés y no como una especie separada, como actualmente se encuentra (80).

Lo que marcaría la diferencia de nuestro éxito como especie podría ser nuestra capacidad inventiva fundada en la aparición del lenguaje. Cabe hacer mención de que estas propiedades también han sido identificadas por otras áreas del conocimiento. Por ejemplo, para Fullat una de las características distintivas del educando (el ser humano), que lo separa de la φυσις o *natura* es el lenguaje (34).

Es tal la semejanza con los primates, que si no fuéramos capaces de transmitir nuestra cultura (paideia) seríamos considerados animales, de ahí surge la necesidad del hombre de ser educado (34). En este sentido, no podemos encontrar todas nuestras respuestas en el DNA; como Ortega y Gasset dijo: "Yo soy Yo y mis circunstancias".

C.4.2. ¿De dónde provenimos?

Esta pregunta fue tratada con más detalle en la sección del Origen de la Vida en la Tierra. Brevemente se puede considerar, desde un punto de vista evolucionista, que provenimos de una célula ancestral común, que a su vez provino de las múltiples interacciones entre diferentes moléculas.

También es posible entender nuestro origen bajo el marco del orden espontáneo generado por las interacciones moleculares entre ácidos nucleicos y proteínas, como se mencionó anteriormente (57).

C.4.3. ¿Hacia dónde vamos?

En el mencionado artículo de Science (80), se comenta la participación de Stephen Jay Gould en el sentido que a pesar de los cambios fuertes que se están provocando en el clima, no se considera que nuestra especie vaya "hacia algún lado", es decir, que ocurra la especiación.

La teoría de la Síntesis neodarwiniana considera la evolución a dos niveles: la microevolución y la macroevolución. La primera representa un fenómeno demostrado, en la que se va favoreciendo la reproducción de cierto grupo de individuos dentro de la especie. La segunda, en cambio, implica cambios drásticos que provoquen el surgimiento de una nueva especie.

Es precisamente la macroevolución otro de los puntos de debate entre los científicos. Para algunos intelectuales, este proceso simplemente no ha ocurrido. Como fundamento se tienen los reportes experimentales en los cuales se ha sometido a numerosas mutaciones a la mosca de la

fruta (*Drosophila melanogaster*) sin obtener aún una nueva especie. Sin embargo, recientemente se reportó un evento de macroevolución en una especie de salmón (48), lo cual va en contra de la estasis fenotípica observada en las especies conocidas. Conforme se publique más evidencia experimental, será posible entender mejor este fenómeno de especiación.

Bibliografía

[1] A. Arneodo, E. Bacry, P.V. Graves y J.F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Phys. Rev. Lett.* **74:** 3293–3296.

[2] D.G. Arquès y C.J. Michel. 1990. Periodicities in coding and noncoding regions of the genes. *J. theor. Biol.* **143:** 307–318.

[3] D.G. Arquès y C.J. Michel. 1994. Analytical expression of the purine/pyrimidine autocorrelation function after and before random mutations. *Math. Biosci.* **123:** 103–125.

[4] D.G. Arquès y C.J. Michel. 1997. A code in the protein coding genes. *BioSystems* **44:** 107–134.

[5] P. Béland y T.F.H. Allen. 1994. The origin and evolution of the genetic code. *J. theor. Biol.* **170:** 359–365.

[6] F.R. Blattner *et al.* 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1474.

[7] C. Blomberg. 1997. On the appearance of function and organisation in the origin of life. *J. theor. Biol.* **187:** 541–554.

[8] S. Brenner. 1957. On the impossibility of all overlapping triplet codes in information transfer from nucleic acid to proteins. *Proc. Natl. Acad. Sci. USA* **43:** 687–694.

[9] C.J. Bult *et al.* 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273:** 1058–1073.

[10] T.R. Cech, A.J. Zaug y P.J. Grabowski.1981. In vitro splicing of the ribosomal RNA precursor of *Tetrahymena*: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell* **27:** 487–496.

[11] M.J. Crawley. 2002. *Statistical computing. An introduction to data analysis using S-Plus.* New York, USA. John Wiley & Sons, Ltd.

[12] F.H.C. Crick, J.S. Griffith y L.E. Orgel. 1957. Codes without commas. *Proc. Natl. Acad. Sci. USA* **43:** 416–421.

[13] F.H.C. Crick. 1968. The origin of the genetic code. *J. Mol. Biol.* **38:** 367–379.

[14] F.H.C. Crick. 1970. Central dogma of molecular biology. *Nature* **227:** 561–563. Este artículo se puede obtener en: http://www.euchromatin.org/Crick01.htm

[15] F.H.C. Crick, S. Brenner, A. Klug y G. Pieczenik. 1976. A speculation on the origin of protein synthesis. *Origins of Life* **7:** 389–397.

[16] C. Darwin. 1966. *On the origin of species.* Cambridge, USA. Harvard University Press. Se puede obtener el texto en: http://www.esp.org/books/darwin/origin/facsimile/

[17] P.L. Davies. 1999. *The fifth miracle. The search for the origin and meaning of life.* New York, USA. Touchstone.

[18] R. Dawkins. 1987. *The blind watchmaker: Why the evidence of evolution reveals a universe without design.* New York, USA. Norton.

[19] R. Dawkins. 1989. *The selfish gene.* Oxford, UK. Oxford University Press.

[20] T.M. Devlin. 1997. *Textbook of biochemistry with clinical correlations.* New York, USA. Wiley-Liss.

[21] T.G. Dewey y J.G. Bann. 1992. Protein dynamics and 1/f noise. *Biophys. J.* **63:** 594–598.

[22] M. Di Giulio. 1996. The β-sheets of proteins, the biosynthetic relationships between amino acids, and the origin of the genetic code. *Origins Life Evol. Biosph.* **26:** 589–609.

[23] M. Di Giulio. 1997. On the origin of the genetic code. *J. theor. Biol.* **187:** 573–581.

[24] M. Di Giulio. 2001. The non-universality of the genetic code: The universal ancestor was a progenote. *J. theor. Biol.* **209:** 345–349.

[25] M. Di Giulio. 2001. The landscape of optimization in the origin of the genetic code. *J. Mol. Evol.* **52:** 372–382.

[26] L.S. Dillon. 1973. The origins of the genetic code. *Bot. Rev.* **39:** 301–345.

[27] B. Efron y R.J. Tibshirani. 1998. *An introduction to bootstrap.* San Francisco, USA. Chapman & Hall.

[28] M. Eigen y P. Schuster. 1979. *The hypercycle. A principle of natural self-organization.* Berlin, Alemania. Springer-Verlag.

[29] R.A. Fisher. 1930. *The genetical theory of natural selection.* Oxford, UK. Oxford University Press.

[30] W.M. Fitch y K. Upper. 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 759–767.

[31] R.D. Fleischmann *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512.

[32] L. Frappat, C. Minichini, A. Sciarrino y P. Sorba. 2004. Universality and Shannon entropy of codon usage. *Phys. Rev. E*. En prensa.

[33] C.M. Fraser *et al.* 1997. Genomic sequence of a Lyme disease spirocheate, *Borrelia burgdorferi. Nature* **390:** 580–586.

[34] O. Fullat. 1995. *El pasmo de ser hombre.* Barcelona, España. Ariel Filosofía.

[35] D.J. Futuyma. 1998. *Evolutionary biology.* Third edition. Massachusetts, USA. Sinauer Associates, Inc.

[36] G. Gamow. 1954. Possible relation between deoxyribonucleic acid and protein structures. *Nature* **173:** *318.*

[37] N. Gershenfeld. 1999. *The nature of mathematical modeling.* Cambridge, UK. Cambridge University Press.

[38] M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, F. Varadi y P. Yiou. 2002. Advanced spectral methods for climatic time series. *Rev. Geophys.* **40:** 1–41

[39] W. Gilbert. 1986. The RNA world. *Nature* **319:** 618.

[40] A.L. Goldberger, C.-K. Peng, J. Hausdorff, J. Mietus, S. Havlin y H.E. Stanley. 1996. Fractals and the heart. En: *Fractal geometry in biological systems. An analytical approach.* P.M. Iannacone y M. Khokha, eds. New York, USA, CRC Press, pp 249-266.

[41] D. Haig y D.L. Hurst. 1991. A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.* **33:** 412–417.

[42] H. Hartman. 1995. Speculations on the origin of the genetic code. *J. Mol. Evol.* **40:** 541–544.

[43] B. Hayes. 1998. The invention of the genetic code. *American Scientist* **86:** 8–14.

[44] H. Herzel y I. Große. 1997. Correlations in DNA sequences: The role of protein coding segments. *Phys. Rev. E* **55:** 800–810.

[45] D. Holste, I. Große y H. Herzel. 2001. Statistical analysis of the DNA sequence of human chromosome 22. *Phys. Rev. E* **64:** 1–9.

[46] R. Ihaka y R. Gentleman. 1996. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **5:** 299–314.

[47] E.T. Jaynes. 1957. Information theory and statistical mechanics. *Phys. Rev.* **106:** 620–630. Se puede obtener este artículo en la página: http://bayes.wustl.edu/etj/node1.html

[48] B. Jónsson y S. Skúlason. 2000. Polymorphic segregation in Arctic charr *Salvelinus alpinus* (L.) from Vatnshlídarvatn, a shallow Icelandic lake. *Biol. J. Linnean Soc.***69:** 55–74.

[49] H.F. Judson. 1979. *The eighth day of creation.* New York, USA. Knopf.

[50] T.H. Jukes. 1983. Evolution of the amino acid code. En: *Evolution of genes and proteins.* M. Nei y R. Koehn, eds. Mass., USA. Sinauer Associates.

[51] T.H. Jukes. 1996. On the prevalence of certain codons ("RNY") in genes for proteins. *J. Mol. Evol.* **42:** 377–381.

[52] J. Jurka y T.F. Smith. 1987. β-turn-driven early evolution: The genetic code and biosynthetic pathways. *J. Mol. Evol.* **25:** 15–19.

[53] I. Kant. 1993. *Crítica del juicio. Buenos Aires, Argentina. Losada.*

[54] D. Kaplan y L. Glass. 1995. *Understanding nonlinear dynamics. New York, USA. Springer-Verlag.*

[55] S. Karlin y V. Brendel. 1993. Patchiness and correlations in DNA sequences. *Science* **259:** 677–680.

[56] S. Kauffman. 1993. *The origins of order.* New York, USA. Oxford University Press.

[57] S. Kauffman. 1995. *At home in the universe.* New York, USA. Oxford University Press.

[58] M. Kimura. 1983. *The neutral theory of molecular evolution.* Cambridge, UK. Cambridge University Press.

[59] J. Konecny, M. Eckert, M. Schöniger y G.L. Hofacker. 1993. Neutral adaptation of the genetic code to double-strand coding. *J. Mol. Evol.* **36:** 407–416.

[60] J. Konecny, M. Schöniger y L. Hofacker. 1995. Complementary coding conforms to the primeval comma-less code. *J. theor. Biol.* **173:** 263–270.

[61] F. Kunst *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256.

[62] D.H. Lee, K. Severin, Y. Yokobayashi y M.R. Ghadiri. 1997. Emergence of symbiosis in peptide self-replication through a hypercyclic network. *Nature* **390:** 591–594.

[63] B. Lewin. 2000. *Genes VII.* New York, USA. Oxford University Press.

[64] M. Mahner y M. Bunge. 1997. *Foundations of biophilosophy.* Berlin, Alemania. Springer.

[65] B.B. Mandelbrot. 1982. *The fractal geometry of nature.* W.H. Freeman. San Francisco, USA.

[66] H.B. Mann y D.R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.* **18:** 50–60.

[67] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons y H.E. Stanley. 1994. Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* **73:** 3169–3172.

[68] R.N. Mantegna, S.V. Buldyrev, A.L. Goldberge, S. Havlin, C.-K. Peng, M. Simons y H.E. Stanley. 1995. Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics. *Phys. Rev. E* **52:** 2939–2950.

[69] G. Melcher. 1974. Stereospecificity of the genetic code. *J. Mol. Evol.* **3:** 121–141.

[70] G. Mendel. 1866. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn 4:* 3–47. La traducción al inglés de este artículo, se puede accesar en la página web: http://www.mendelweb.org/MWpaptoc.html

[71] S.L. Miller. 1953. A production of amino acids under possible primitive earth conditions. *Science* **117:** 528–529.

[72] S.L. Miller y L.E. Orgel. 1974. *The origins of life on Earth.* New Jersey, USA. Prentice Hall.

[73] P. Miramontes, L. Medrano, C. Cerpa, R. Cedergren, G. Ferbeyre y G. Cocho. 1995. Structural and thermodynamic properties of DNA uncover different evolutionary histories. *J. Mol. Evol.* **40:** 698–704.

[74] A.K. Mohanty y A.V.S.S. Narayana Rao. 2000. Factorial moments analysis show characteristic length scale in DNA sequences. *Phys. Rev. Lett.* **84:** 1832–1835.

[75] M. Nei. 1987. *Molecular evolutionary genetics.* New York, USA. Columbia University Press.

[76] K.E. Nelson *et al.* 1999. Evidence for lateral gene transfer between Archea and bacteria from genome sequence of *Thermotoga maritima. Nature* **399:** 323–329.

[77] M.W. Nirenberg, O.W. Jones, P. Leder, B.F.C. Clark, W.S. Sly y S. Pestka. 1963. On the coding of genetic information. *Cold Spring Harbor Symp. Quant. Biol.* **28:** 549–557.

[78] M.W. Nirenberg y J.H. Matthaei. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proc. Natl. Acad. Sci. USA* **47:** 1588–1602.

[79] R. Nussinov. 1980. Some rules in the ordering of nucleotides in the DNA. *Nucleic Acids Res.* **8:** 4545–4562.

[80] L. O'Neill, M. Murphy y R.B. Gallagher. 1994. What are we? Where did we come from? Where are we going? *Science* **263:** 181–183.

[81] A.I. Oparin. 1957. *The origin of life on Earth.* New York, USA. Academic Press.

[82] K. Pearson. 1900. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is shuch that it can be reasonably supposed to have arisen in random sampling. *Phil. Mag. Ser. 5* **50:** 157–175.

[83] S.R. Pelc. 1965. Correlation between coding-triplets and amino-acids. *Nature* **207:** 597–599.

[84] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons y H.E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature* **356:** 168–170.

[85] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley y A.L Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49:** 1685–1689.

[86] J.R. Pierce *An introduction to information theory. Symbols, signals and noise.* (1980) New York, USA, Dover Publications, Inc.

[87] J. Sánchez y M.V. José. 2002. Analysis of bilateral inverse symmetry in whole bacterial chromosomes. *Biochem. Biophys. Res. Comm.* **299:** 126–134.

[88] A.O. Schmitt y H. Herzel. 1997. Estimating the entropy of DNA sequences. *J. theor. Biol.* **188:** 369–377.

[89] E. Schrödinger. 1944. *What is life? The physical aspect of the living cell.* Cambridge, UK. Cambridge University Press. Se puede obtener este texto en: http://home.att.net/ p.caimi/schrodinger.html

[90] E.I. Shakhnovich y A.M. Gutin. 1990. Implications of thermodynamics of proteins folding for evolution of primary sequences. *Nature* **346:**773–775.

[91] C.E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* **27:** 379–423. Se puede obtener este artículo en: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[92] Q. She *et al.* 2001. The complete genome of the crenarcheon *Sulfolobus solfataricus* P2. *Proc. Natl. Acad. Sci. USA* **98:** 7835–7840.

[93] J.C.W. Shepherd. 1981. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. *J. Mol. Evol.* **17:** 94–102.

[94] J.C.W. Shepherd. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. USA* **78:** 1596–1600.

[95] J.C.W. Shepherd. 1990. Ancient patterns in nucleic acid sequences. *Methods. Enzymol.* **183:** 180–192.

[96] A.J.G. Simpson *et al.* 2000. The genome sequence of the plant pathogen *Xylella fastidiosa.* The *Xylella fastidiosa* Consortium of the Organization for Nucleotide Sequencing and Analysis. *Nature* **406:** 151–157.

[97] T.M. Sonneborn. 1965. Degeneracy of the genetic code, extent, nature, and genetic implications. En: *Evolving genes and proteins.* V. Bryson y H.J. Vogel, edits. New York, USA. Academic Press, pp. 377–397.

[98] S. Spiegelman. 1970. Extracellular evolution of replicating molecules. En: *The neuro sciences: A second study program.* F.O. Schmitt, edit. New York, USA. Rockefeller University Press, pp 927–945.

[99] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, J.M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino y M. Simons. 1992. Fractal landscapes in biologycal systems: Long-range correlations in DNA and interbeat heart intervals. *Physica A* **191:** 1–12.

[100] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, Z.D. Doldberger, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.-K. Peng y M. Simons. 1994. Statistitical mechanics in biology: how ubiquitous are long-range correlations? *Physica A* **205: 214–253.**

[101] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, C.-K. Peng y M. Simons. 1996. Scale invariant features of coding and noncoding DNA sequences. En: *Fractal geometry in biological systems. An analytical approach.* P.M. Iannacone y M. Khokha, eds. New York, USA, CRC Press, pp 15–30.

[102] I. Stewart. 1997. *Life's other secret: The new mathematics of the living world.* New York, USA. John Wiley & Sons.

[103] H. Takami *et al.* 2000. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis. Nucleic Acids Res.* **28:** 4317–4331.

[104] H. Tettelin *et al.* 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae. Science* **293:** 498–506.

[105] D. Thompson. 1966. *On growth and form.* Cambridge, UK. Cambridge University Press.

[106] E. N. Trifonov. 1998. 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A* **249:** 511–516.

[107] T. Vicsek. 1996. Fractal geometry. En: *Fractal geometry in biological systems. An analytical approach.* P.M. Iannacone y M. Khokha, eds. New York, USA, CRC Press, pp 317–343.

[108] M.S. Vieira. 1999. Statistics of DNA sequences: A low-frequency analysis. *Phys. Rev. E* **60:** 5932–5937.

[109] R.F. Voss. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* **68:** 3805–3808.

[110] M. Waldrop. 1992. *Complexity: The emerging science at the edge of order and chaos.* New York, USA. Touchstone Books.

[111] J.D. Watson y F.H.C. Crick. 1953. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171:** 737–738. Se puede obtener este artículo en: http://www.nature.com/genomics/human/watson-crick/

[112] G. Webster y B. Goodwin. 1981. History and structure in biology. *Perspect. Biol. Med.* **25:** 39–62.

[113] A. Weismann. 1885. The continuity of the germ-plasm as the foundation of a theory of heredity. En: *Readings in heredity and development.* J.A. Moore, edit. New York, USA. Oxford University Press. Un texto equivalente puede obtenerse en: http://www.esp.org/books/weismann/germ-plasm/facsimile/

[114] G.B. West, J.H. Brown y B.J. Enquist. 2001. A general model for ontogenetic growth. *Nature* **413:** 628–631.

[115] G.B. West, W.H. Woodruff y J.H. Brown. 2002. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. *Proc. Natl. Acad. Sci. USA.* **99:** 2473–2478.

[116] G.B. West, V.M. Savage, J. Gillooly, B.J. Enquist, W.H. Woodruff, J.H. Brown. 2003. Why does metabolic rate scale with body size? *Nature* **241:** 713.

[117] D.L. Wheeler *et al.* 2000. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **28:** 10–14.

[118] O. White *et al.* 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286:** 1571–1577.

[119] C.R. Woese. 1965. On the origin of the genetic code. *Proc. Natl. Acad. Sci. USA* **54:** 1546–1552.

[120] C.R. Woese, D.H. Dugre, S.A. Dugre, M. Kondo y W.C. Saxinger. 1966. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symp. Quant. Biol.* **31:** 723–736.

[121] J.T. Wong. 1975. A co-evolution theory of the genetic code. *Proc. Natl. Acad. Sci. USA.* **72:** 1909–1912.

[122] M. Yarus. 1988. A specific amino acid binding site composed of RNA. *Science* **240:** 1751–1758.