ESTUDIOS DE ANTROPOLOGÍA BIOLÓGICA

VOLUMEN XIII

*

Editoras

Magalí Civera Cerecedo
Martha Rebeca Herrera Bautista

Instituto Nacional de Antropología e Historia
Consejo Nacional para la Cultura y las Artes
AMAB

Universidad Nacional Autónoma de México
Instituto de Investigaciones Antropológicas
Instituto Nacional de Antropología e Historia
Asociación Mexicana de Antropología Biológica
México 2007

Todos los artículos fueron dictaminados

Primera edición: 2007

ISSN 1405-5066

Impreso y hecho en México
Printed in Mexico

Antropología genética y molecular

El carácter lingüístico y la coherencia textual presentes en el ADN 'no codificante'

Brenda Cantú-Bolán
Enrique Hernández-Lemus*

Facultad de Ciencias, UNAM
**Departamento de Genómica Computacional, Instituto Nacional de Medicina Genómica*

Resumen

En este trabajo se teoriza que, así como es posible darse cuenta en un texto coherentemente ordenado que las palabras no están dispuestas al azar, el orden en la disposición de nucleótidos en un genoma es prueba de la coherencia y, por lo tanto, de la codificación (contenido informático) de éste en su totalidad. Así, se pretende mostrar por medio de métodos estadísticos que el ADN llamado "no codificante" (ADNnc) tiene una distribución no aleatoria similar a la de un texto lingüístico coherentemente ordenado. Se estudiaron las distribuciones de probabilidad asociadas con este ADNnc por medio de métodos estadísticos a fin de inferir algunas de sus propiedades lingüísticas. Se discuten brevemente las posibles implicaciones de los resultados obtenidos y sus aplicaciones. Con la finalidad de estudiar un espectro amplio de seres vivos, este estudio se efectuó sobre el genoma completo de una bacteria (*Mycoplasma pneumoniae*), así como en fragmentos del ADN de la mosca de la fruta (*Drosophila melanogaster*), de un tipo de gato doméstico (*Felis catus*), del pino negro japonés (*Pinus thunbergii*) y sobre un generador aleatorio como control.

Palabras clave: genómica, lingüística estadística, ADN intrónico

Abstract

In this work it is theoreticized that, much in the same manner that one can notice that words are not randomly distributed in a coherently ordered text, it is also

possible to see that the order in the nucleotide disposition in a genomic fragment is a proof of coherence. And it is also, a proof of codification (in the sense of information content) of the entire genome. It is intended to show through statistical methods that the so-called non-coding DNA (nc-DNA) has a non-random probability distribution in a similar fashion to a coherently ordered text. We studied the probability distributions of "words" associated with nc-DNA by means of statistical methods in order to dilucidate its linguistic properties. We briefly discuss the implications of the results we found, as well as some possible applications. In order to study a wide range of species, we analyze a complete bacterial genome (*Mycoplasma pneumoniae*), as well as fragments of the fruit fly (*Drosophila melanogaster*), a kind of domestic cat (*Felis catus*), a tree (*Pinus thunbergii*), and over a random generated sequence used as a control.

KEY WORDS: genomics, Statistical linguistics, intronicDNA.

ANTECEDENTES

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

Por otro lado, los estudios lingüísticos indican que si se agrupan letras en determinadas combinaciones, se obtienen palabras; si a su vez se unen palabras, se genera una frase; dichas letras, dichas palabras e incluso las frases no están dispuestas aleatoriamente, sino que tienen un orden tal que éstas entrañan un significado (es decir, poseen una cierta cantidad de información asociada). Si entonces se juntan frases, se produce un texto coherente que posee información ordenada. Igualmente es posible referirse a un genoma: si se agrupan nucleótidos

en combinaciones específicas, se obtiene un gene; si se juntan genes, se producirá una secuencia genómica, y si se unen varias de éstas, se produce un genoma. Los genes conllevan características que se expresan en el organismo a través de las proteínas y el genoma es comparable a un texto biológico donde se hace un recuento de cada rasgo del individuo. Siguiendo con la misma idea, se sabe que las proteínas constan de combinaciones de 20 aminoácidos alineados siguiendo diversas ordenaciones en cadenas de longitud arbitraria. Los escritos, por otro lado, constan de combinaciones de 28 letras (en el alfabeto para hispanohablantes), más una serie de signos de puntuación, alineados en distintas ordenaciones que suelen plegarse para que estén dispuestos a lo largo de las páginas. Del mismo modo que las letras de nuestra lengua pueden combinarse en una enorme cantidad de modos, también cabe combinar los 20 aminoácidos de que se vale la vida en una ingente variedad de proteínas distintas.

Hipótesis de trabajo

El desarrollo de este trabajo parte de las siguientes hipótesis:

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

2. Por otro lado, si es posible mostrar que la concentración de Repeticiones Diméricas en Fila (RDFs) localizadas en el ADNnc tiene un comportamiento del tipo Ley de Potencia, entonces podemos concluir que posee orden de largo alcance en sus nucleótidos. Este hecho apoyaría la hipótesis 1 de no aleatoriedad del ADNnc.

Importancia técnica

1. Actualmente algunos cálculos genómicos realizados por computadora, como el proyecto Genoma Humano (M. Hattori *et al.* 2000), consumen muchos recursos computacionales y tiempo. Si encontráramos peculiaridades *lingüísticas* y/o comunicativas (en sentido estadístico) en los genomas por estudiar, reduciríamos significativamente el tiempo de cómputo al optimizar los algoritmos de búsqueda

(un ejemplo real: dado que los cúmulos de tamaño 5 en adelante de AT/TA no están presentes en el genoma del gato, desde el principio omitiremos buscarlos y no los tomaremos en cuenta para cálculos posteriores).

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

3. Las características generales del ADNnc lo hacen especialmente útil para su aplicación en la identificación forense (A. Griffits 2001). Como se puede deducir de su trascendente función, el ADN esencial está formado por secuencias altamente conservadas con muy pocas variaciones interindividuales e intergeneracionales, ya que de lo contrario se podrían ver afectadas funciones básicas para la vida de las personas. Los mínimos cambios que tienen lugar, cuando son viables, aumentan el polimorfismo de proteínas y enzimas, aunque también pueden tener efectos negativos. Por el contrario, el ADNnc presenta una gran variabilidad de unos individuos a otros, ya que estas secuencias no son conservadoras al no afectar sus cambios a la fisiología normal del individuo. Las variaciones debidas a cambios de bases sencillos, procesos de inserción-deleción o de intercambio de ADN (recombinación) durante la formación de las células germinales (meiosis) hacen que se modifique el número de repeticiones o el orden de las bases de un determinado fragmento repetitivo, pudiendo producirse en un locus sencillo o en múltiples loci, y éste es el origen de la variación que hace que no haya dos personas, a excepción de los gemelos univitelinos, que tengan la misma secuencia del ADN.

Importancia práctica

1. Una célula no utiliza la expresión de características propias de otro tipo de tejido para dividirse y dispersarse en el medio sin control (W. Li 2001; I. Yanai *et al.* 2000). No cabe duda de que todo ello depende del proceso de transcripción a ácido ribonucleico (ARN) y luego a

proteína de ciertas porciones del mensaje genético codificado en los cromosomas. La cuestión es descifrar cómo se conectan y desconectan las diversas porciones del mensaje genético. Esa línea de investigación suma la posibilidad obvia de comprender primero y quizá más delante de controlar esta enfermedad en la cual las células dejan de obedecer las órdenes genéticas que dictan su comportamiento.

2. Se ha determinado que la presencia de cierta enfermedad (un tipo específico de cáncer) está asociada ($p < 0.05$) con la existencia de RDFs de tamaño 11 de TAs. Sería posible ver en qué cromosoma pasa eso (cuál tiene más de estas repeticiones y usar terapia génica para curar esa enfermedad), atacar el gene enfermo, o cambiarlo de alguna manera para ver si la asociación persiste (W. Li 2001).

3. Se sabe que las Repeticiones Diméricas en Fila (RDFs) en el genoma constituyen una importante fracción del ADNnc y son relativamente escasas en las secuencias que codifican para proteínas. Las RDFs son de considerable interés teórico y práctico debido a su alto polimorfismo. Por ejemplo: se sabe que las RDFs del tipo (CA)n se expanden debido al plegamiento en el proceso de replicación; estos errores frecuentemente son eliminados por la enzima reparadora MSH2. Sin embargo, una mutación en el gene MSH2 lleva a una expansión descontrolada de repeticiones que es una causa común del cáncer de ovario (I. Yanai *et al.* 2000), y mecanismos similares se han atribuido para otros tipos de cáncer. Estas investigaciones sobre las RDFs en ADNnc podrían conducir a avances en el diagnóstico oportuno y posible prevención de enfermedades de esta clase. Las RDFs se denominan también ADN-satélite, porque con frecuencia pueden segregarse de la masa del ADN por centrifugación en cloruro de cesio. El número de RDFs de un tipo determinado puede variar entre distintos individuos, dando lugar a una huella única de ADN (H. E. Stanley *et al.* 1999 a, 1999 b, 2000 a, 2000 b).

4. Adicionalmente a las RDFs existen otros elementos repetidos y altamente redundantes en el genoma. De particular interés en fechas recientes han resultado las secuencias ALU (Dagan *et al.* 2006; Umynly *et al.* 2007). Las secuencias ALU son elementos cortos interdispersos de alrededor de 300 nucleótidos de longitud. En el genoma humano, por ejemplo, se han hallado más de un millón de ALUs. Se considera que estas secuencias pueden afectar la estructura de los genes, las secuencias de proteínas, los motivos de *splicing* y los patrones de expresión.

Las ALUs suelen tener estructura dimérica y se cree que son ancestralmente derivadas del gen que especifica al 7SL RNA, un componente citoplásmico abundante de la partícula de reconocimiento de señales que es mediadora de la traslocación de proteínas secretadas a través del retículo endoplásmico.

Se ha sugerido que partes de los elementos ALU pueden insertarse en ARNms por un mecanismo de "corte y empalme" (*splicing*) en un proceso llamado exonización (Dagan *et al.* 2004). De hecho, más del 5% del empalme alternativo de exones en el genoma humano está causado por elementos ALU. Por ejemplo, repeticiones ALU en la región promotora distal de la proteína de transferencia del colesteril ester en humanos (CETP) y se ha encontrado que actúa como represor de los elementos regulatorios de la actividad del promotor. Además, se ha relacionado la presencia de ALUs con apoptosis y muerte celular inducida por taxol. Los ALUs se encuentran altamente organizados en cúmulos de genes que se relacionan con los procesos de metabolismo, señalización y transporte, y son menos abundantes en genes que codifican para proteínas estructurales o componentes de la ruta de información (Dagan *et al.* 2004). Dado que las estructuras ALU son secuencias repetidas dispersas, contribuyen a aumentar la redundancia estadística global del genoma.

5. De enorme importancia son también otros elementos estructurales redundantes denominados islas CpG; éstos son segmentos que poseen un gran número de citosina y guaninas mutuamente adyacentes unidas mediante enlaces fosfodiester. Se encuentran tanto dentro como cerca de aproximadamente el 40% de los promotores de los genes de mamíferos (cerca del 70% en el caso humano). La longitud típica de una isla CpG es de entre 300 y 3000 bases.

Estas regiones están caracterizadas por un contenido del dinucléotido CG mayor al estadísticamente esperado (aprox. 6%), mientras que el resto del genoma tiene niveles por debajo de lo esperado (cercanos al 1%) debido a un fenómeno llamado supresión CG. A diferencia de los sitios CpG en las regiones codificantes de un gen, la mayoría de los sitios CpG en las islas de regiones promotoras no están metilados en presencia de genes expresados (Saxonov *et al.* 2006). Al tratarse de regiones altamente repetidas de un dinucleótido técnicamente estamos tratando de RDFs por lo que su tratamiento lingüístico-estadístico será

bajo ese marco; sin embargo, por razones funcionales muchas veces es conveniente considerar tales regiones como casos separados. En el marco del presente trabajo que aún no toma en cuenta la naturaleza funcional de tales repeticiones, las islas CpG son un caso especial de las RDFs consideradas.

6. Este trabajo pretende mostrar, entre otras cosas, por medio del análisis de la función de correlación entre *palabras genéticas*, que son los segmentos de ADNnc y no de ADNc los más resistentes a cambios a lo largo del tiempo (H. E. Stanley *et al.* 2001 b, N. V. Dokholyan 1999). Esto pudiera ser aplicado a investigaciones sobre tasas de mutación, ya que el investigador sabría con qué tipo de ADN trabajar desde el principio. Por otra parte, sería también posible determinar cuáles son los genes que codifican para ciertas adaptaciones. Las tasas de acumulación pueden cuantificar la tendencia de las secuencias del ADN a conglomerarse y pueden ser utilizadas en estudios posteriores de agregados de nucleótidos en las secuencias del código genético (H. E. Stanley *et al.* 2001 a). Por ejemplo, diferentes tasas de dímeros variados pueden sugerir diferentes tasas de mutación, específicas para cada dímero y organismo.

Estructura lingüística del código genético

Cadenas de Markov, complejidad lingüística y ADN

Las secuencias de ADN han sido analizadas utilizando una gran variedad de modelos que pueden ser considerados básicamente en dos categorías. Los del primer tipo son análisis locales que toman en cuenta el hecho de que las secuencias del ADN son producidas en orden progresivo, de manera tal que los pares de bases vecinas afectan el enlazamiento del siguiente par de bases. Este tipo de análisis tales como los modelos Markovianos de n pasos pueden describir algunas de las correlaciones de corto alcance observadas en las secuencias de ADN. La segunda categoría es de naturaleza más global y se concentra en la presencia de patrones que pueden ser encontrados en la mayoría de las secuencias genómicas.

Los lenguajes naturales están caracterizados por estructuras determinadas por las reglas de la gramática. Las palabras enlazadas mediante el uso de estas reglas tienen significado, es decir expresan ideas, sen-

timientos y emociones, de manera que es posible para los receptores del mensaje entenderlo. Las reglas gramaticales dan, pues, coherencia y significado a los textos extensos; de esta manera los lenguajes tienen un orden de largo alcance. Los espectros de frecuencia de aparición de palabras muestran la presencia de periodos largos. Éstos son identificados por un comportamiento del tipo de ley de potencia en la región de baja frecuencia del espectro. Las palabras colocadas al azar tendrían un aspecto muy diferente sin orden de largo alcance. La secuencia de las letras A, C, G, T, en el ADN tiene un espectro de frecuencias del mismo tipo de ley de potencias. Es posible, por lo tanto, que estas secuencias presenten un orden de largo alcance que posea "reglas gramaticales subyacentes". Las opiniones a este respecto continúan divididas, algunos han tomado el punto de vista de que el ADN tiene una estructura similar a la de un lenguaje. En las regiones codificadas, los periodos largos tienen una menor incidencia que en las partes no codificadas. El análisis de frecuencias de "palabras" (también llamado análisis de Zipf) en las regiones diversas del ADN no-codificante (intrones, ADN separador, regiones reguladoras, etcétera) ha mostrado que el exponente de la ley de potencias tiene un valor más alto en los segmentos no codificantes y este valor es más cercano al de los lenguajes naturales que en el ADN codificante.

Metodologías aplicadas a la descripción estadística de genomas y segmentos genómicos

A continuación describiremos brevemente las metodologías utilizadas en el análisis estadístico; si se desea conocer los aspectos más técnicos asociados con tales cálculos véase (B. Cantú-Bolan y E. Hernández-Lemus 2006).

Análisis de Zipf

Con la finalidad de demostrar si el ADNnc se encuentra realmente dispuesto aleatoriamente o no se eligieron los siguientes fragmentos de genoma. Para los fines del presente trabajo fueron tomados como textos genéticos los siguientes segmentos genómicos:

1. *Mycoplasma pneumoniae* (en este caso, los datos obtenidos fueron del genoma completo)

2. *Drosophila melanogaster*
3. *Felis catus*
4. *Pinus thunbergii*
5. Generador aleatorio (utilizado a manera de control)

Considerando que al analizar los genomas de una bacteria, un invertebrado, un vertebrado y una planta es posible abarcar a rasgos generales de una amplia gama de formas de vida presentes en la naturaleza, adicionalmente se tomó este criterio debido a que las especies aquí consideradas han sido anotadas tanto en sus regiones codificantes como no-codificantes (S. Matsutani 2005, Dewey *et al.* 2006, Hubbard *et al.* 2006). La inclusión de genomas eucariontes y no eucariontes tuvo por objetivo desarrollar criterios básicos para encontrar diferencias entre regiones codificantes y no codificantes.

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

Se realizó el conteo total de palabras genéticas, ya obtenidas las combinaciones para dímeros, trímeros, tetrámeros, pentámeros y RDFs; se utilizó un procedimiento computacional para determinar sus frecuencias de aparición. Se les asignó rangos a las palabras, de forma tal que la más frecuente tenía rango 1, la siguiente más frecuente rango 2, etcétera. Posteriormente se analizó el comportamiento de la función que determina la relación entre frecuencia y rango para cada "palabra" en el genoma a fin de determinar si el comportamiento era del tipo ley de potencia, indicando así coherencia y orden de largo alcance o de tipo exponencial que representaría incoherencia, desorden y posible aleatoriedad.

Determinación de complejidad por dimensión fractal de Hausdorff

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

tiempo; con los demás nucleótidos se emplea el mismo algoritmo. Se obtuvo la primera derivada según la serie de derivadas fractales propuesta por Mandelbrot (B. Derrida 2000) y se continuó de la misma manera hasta llegar a la derivada número 10. Posteriormente, se construyó con tales derivadas la integral de correlación propuesta por Grassberger y Procaccia (P. Grassberger y I. Procaccia 1983).

Series de tiempo renormalizadas originales y conjuntos fractales asociados

Dado que el genoma es redundante y hasta el más pequeño muy largo, se procedió a renormalizarlo; esto significa que se sustituyó con un 1 cada vez que aparecía en la cadena genómica el nucleótido deseado y al resto de ellos con ceros; después se tomaron tríos de números que fueron a su vez sustituidos con ceros o unos dependiendo de si su suma era mayor o menor a 1. Se repitió el procedimiento hasta lograr invarianza para cada ejemplo experimental. Posteriormente, se tomó la serie de tiempo renormalizada de cada nucleótido y se construyó una gráfica lineal que maneja concentraciones de la base deseada *versus* posición dentro de la cadena genómica. Se repitió el procedimiento para los cuatro organismos vivos más el Generador Aleatorio. Se tomaron los valores de la Serie de Tiempo Renormalizada Original

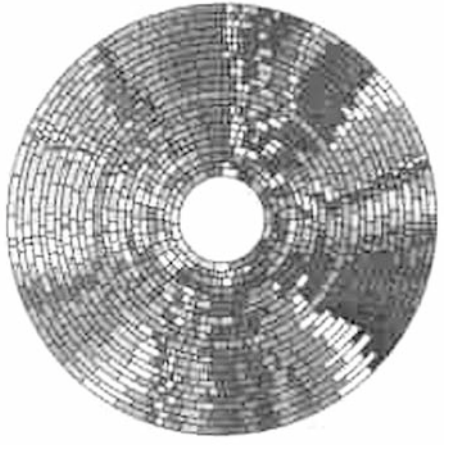


Figura 1. Conjunto fractal tipo anillo asociado con el nucleótido A en *Mycoplasma pneumoniae* (isocoras A).

Figura 2. Conjunto fractal tipo anillo asociado para *Mycoplasma pneumoniae* Serie A versión radial.

más los del conjunto de sus diez derivadas y se construyeron gráficas del tipo Anillos (figura 1) y Radial (figura 2).

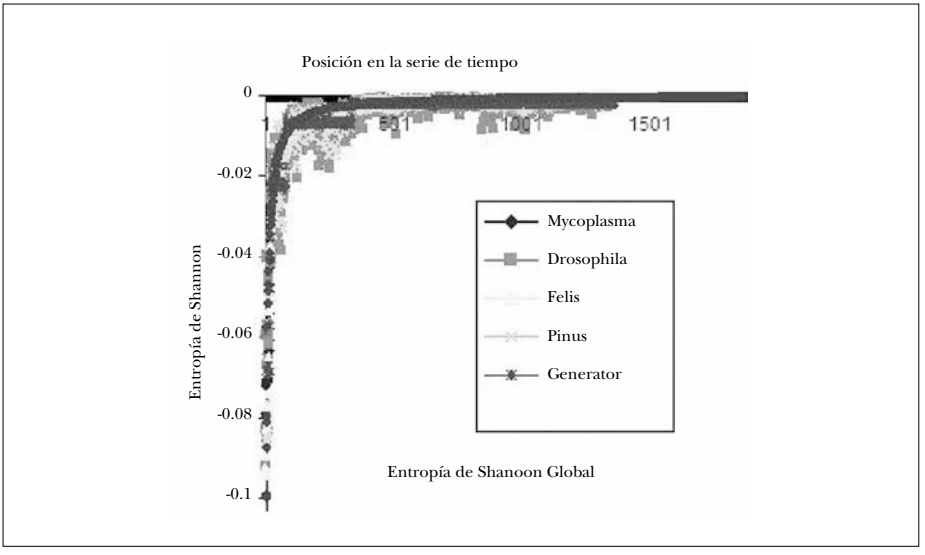
Proporciones de bases complementarias

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

Resultados

Entropía informacional de Shannon

En las gráficas de las funciones obtenidas para la entropía informacional de Shannon es posible notar (gráfica 1) el significativo contraste existente entre el Generador aleatorio y el resto de los elementos experimentales; ya que en estos últimos se observa, como podía esperarse,



Gráfica 1. Entropía informacional de Shannon.

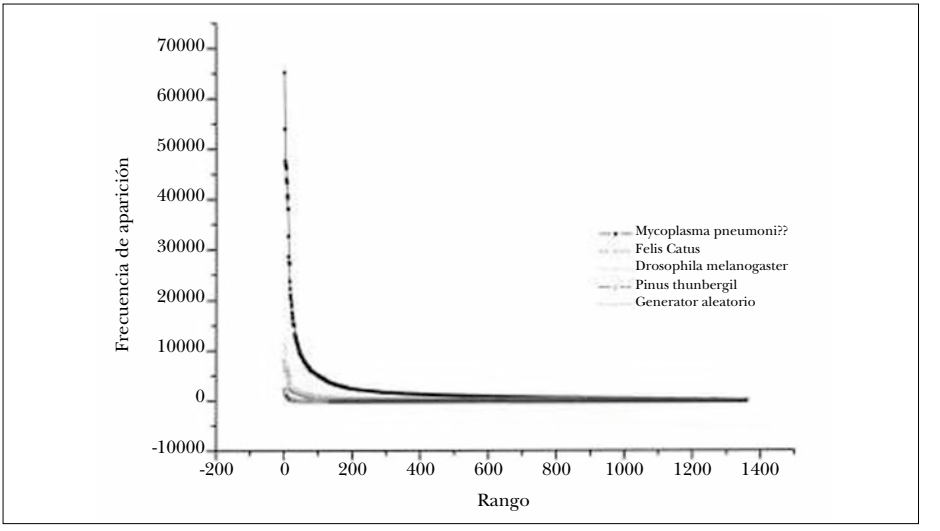
que comienzan con una entropía baja, o dicho de otra manera: un alto contenido de información que van perdiendo poco a poco a lo largo de la cadena, en el texto la información se pierde del transmisor a la transmisión, luego al receptor, y así sucesivamente. Es perceptible en ellos una curva continua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

Distribución de probabilidades (análisis lingüístico) de Zipf

Este método de estudio denota la coherencia y estructuración de un escrito, es decir, cuando un conjunto de reglas -genéticas, gramaticales o semánticas- hacen que las palabras o los cúmulos de nucleótidos de

un texto, lingüístico o biológico respectivamente, tengan un significado lógico por sí mismos y en interacción con otros elementos textuales. También muestra visualmente que los datos se corresponden con una Ley de Potencia (gráfica 2) que, como se explicó anteriormente, es el tipo de función que implica una correlación de largo alcance entre los nucleótidos o palabras y, por lo tanto, una memoria de largo plazo en la cadena (ya sea de bases o de letras). Se ampliará más esta noción al realizar el análisis de Markov. Los resultados señalan la escasa coherencia del Generador, se trata de un texto con letras intercaladas al azar sin estructura, significado ni lógica alguna, como puede suponerse en un texto producido artificialmente para un propósito como el de este trabajo. Por otro lado, son evidentes en la gráfica global los mismos cuatro escalones con aumento de tamaño periódico correspondientes a cúmulos de dos, tres, cuatro y cinco nucleótidos que ya se habían observado en su gráfica de Entropía Informacional de Shannon, evidenciando de nuevo su independencia estadística e inexistente interrelación. Aunado a que en la mayoría de los casos el coeficiente de correlación R^2 es más bajo, lo que implica el muy bajo carácter tipo Ley de Potencia (su distribución de frecuencias, para absolutamente todos los cálculos (dímeros, trímeros, tetrámeros, pentámeros, global y RDFs) el Generador Aleatorio muestra los exponentes de Zipf menos negativos; dicho comportamiento es más evidente en la gráficas global (gráfica 2) y en los valores obtenidos para RDFs (para éste y los si-guientes valores numéricos véase el cuadro 1), donde en ningún modo se acerca a los valores -1 en global (GA= -0.84. Hay que recordar que, dado que este método de lingüística estadística obedece a una ley de potencia, un cambio pequeño en el exponente significa un cambio importante en el carácter coherente del texto) y -2 en RDFs (GA= -0.16), que los organismos vivos sobrepasaron.

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



Gráfica 2. Gráfico del Zipf global.

rencia existente en el ADNnc que, como sabemos, comprende aproximadamente el 95% del genoma.

Es importante subrayar que los exponentes de Zipf para las RDFs fueron significativamente más grandes en todos los organismos, salvo en el Generador Aleatorio; esto resalta una poderosa estructura coherente. Se sabe por autores como Stanley (*et al.* 1999 a, 1999 b, 2000 a, 2000 b) y Dokholyan 1999) que la presencia de RDFs es dramáticamente superior en el ADNnc en comparación con el ADNc. En este caso fue posible comprobar que esa superestructura gramatical efectivamente presenta los valores máximos de coherencia (alrededor de -2), así como su presencia a gran escala en el ADNnc de los cuatro organismos vivos analizados, no así en ADNc (para trímeros el exponente de Zipf más alto se obtuvo en el caso de Felis catus, Fc= -0.5) ni en Generador Aleatorio.

Procesos de Markov

En el ADNc las unidades de tamaño tres son muchas y existen numerosas posiciones posibles que pueden ocupar estos codones a lo largo de la cadena genómica (en genomas diferentes), de tal suerte que puede

Cuadro 1
Valores obtenidos para RDFs

	Mycoplasma pneumoniae	*Drosophila melanogaster*	*Felis Catus*	*Pinus thunbergii*	Generador aleatorio
Zipf dímeros	R^2= 0.8061 y= 72142X^(-0.3156)	R^2=0.8323 y= 14426X^(-0.4171)	R^2=03865 y=12391X^(-0.4725)	R^2= 0.8234 y=13202X^(-0.4156)	R^2= 0.4456 y=6789.6X^(-0.0849)
Zipf trímeros	R^2= 0.8799 y= 39307X^(-0.4408)	R^2= 0.8729 y=7049.2X^(-0.4609)	R^2=0.4428 y=6813.3X^(-0.5861)	R^2= 07632 y= 5794.5X^(-0.4662)	R^2=0.4507 y= 1715.3X^(-0.0533)
Zipf tetrámeros	R^2= 0.8609 y= 25600X^(-0.5325)	R^2= 0.9000 y=4184.1X^(-0.5189)	R^2= 0.5044 y= 5183.7X^(-0.7024)	R2=0.7684 y=4139.9X-0.5285	R^2= 0.6296 y= 467.18x^(-0.0589)
Zipf pentámeros	R^2=0.8244 y= 20812X^(-0.6212)	R^2= 0.8953 y=2908.9X^(-0.5775)	º	R^2=0.7791 y=2028.9X^(-0.6044)	R^2= -0.7624 y= 157.15X^(-0.0995)
Zipf global	R^2=0.9519 y=724665X^(-1.0893)	R^2= 0.9708 y= 99613X^(-1.0468)	R^2= 0.793 y= 88740X^(-1.1572)	R2=0.9482 y=178273X-1.2068	R^2= 0.9054 y= 25561X^(-0.8418)
Zipf RDFs	R^2=0.7703 y=221172X^(-2.7899)	R^2=0.8537 y=24396X^(-2.2736)	R^2=0.8682 y=16545X^(-1.9198)	R^2=0.8602 y=17555X^(-2.1617)	R^2=0.716 y=9575.6X^(-2.0819)
Shannon dímeros	-2.73855944	-2.71470778	-2.69845225	-2.71557437	-2.7682107
Shannon trímeros	-4.06800757	-4.058805744	-4.0256709	-4.0601716	-4.15659888
Shannon tetrámeros	-5.3902669	-5.38882369	-5.3416738	-5.3980608	-5.54280927
Shannon pentámeros	-6.71078407	-6.71456862	-6.64790101	-6.72826455	-6.92533465
Shannon global	-6.12614279	-6.143244837	-5.87160893	-5.9747581	-6.272359234
Shannon RDFs	-2.7085	-2.8363	-3.2475	-2.8719	-2.9937
Contenido A/T	59.25%	63.55%	61.82%	61.50%	50.23%
Contenido G/C	40.75%	36.45%	38.18%	38.50%	49.77%

pensarse en ellas como moléculas distribuidas de manera aparentemente aleatoria e inestable porque, finalmente, cada unidad es estadísticamente independiente y sólo se afecta a sí misma, de ahí la memoria markoviana que presenta el ADNc (como evidencia el carácter exponencial de su decaimiento en la correlación (H. E. Stanley *et al.* 1999), mientras que en el ADNnc hay regiones de hasta cientos de bases de longitud que están relacionadas estadísticamente, debido a una memoria no markoviana o de largo alcance (como evidencia su comportamiento, que obedece a una Ley de Potencia). Así, grandes unidades constitutivas no pueden estar dispuestas de manera azarosa, sino que deben tomar el papel correspondiente dado por esta memoria de largo alcance. En el ADNnc, como hay menos maneras de ordenar los nucleótidos, el carácter aleatorio de este ordenamiento disminuye, y podemos hablar de dependencia estadística y de estabilidad molecular. A causa de su tamaño pequeño puede cambiar fácilmente de posición en el genoma (molécula inestable) sin afectar al resto de las bases (independiente estadísticamente). El desplazamiento de los codones puede ocurrir de tantas maneras en la cadena que tiene un comportamiento aparentemente aleatorio (caso del ADNc).

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

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

Dimensión fractal de Hausdorff

Los valores d (dimensión de Hausdorff) son indicadores de la magnitud de complejidad en un objeto matemático como puede ser un cuasifractal (estrictamente hablando, un fractal posee un número infinito de puntos); entre más grande sea d, la complejidad de éste es alta y viceversa.

Por otro lado, cuando d no pertenece al conjunto de números enteros, sino al de las fracciones, se dice que el cuasifractal no es trivial. De este modo, si una Ley de Potencia tiene asociada aparte una dimensión fractal, la información que alberga es aún mayor, es decir, coherente, de largo alcance y compleja. En este caso lo que se busca conocer es la cantidad máxima de información que existe en un genoma relativa a un nucleótido y qué tan compleja es. Se obtuvieron exclusivamente cifras fraccionarias, lo cual implica la cuasifractalidad (y consecuentemente un aumento en la cantidad de información) del conjunto de datos de los cinco sujetos experimentales. Al observar los resultados para la dimensión fractal en el cuadro 1 es patente que en todos los genomas hay dos valores altos y dos bajos, coincidiendo los primeros para la proporción de adeninas y timinas (m= 0.7-0.8 para organismos vivos) y los segundos para la de citosinas y guaninas (m= 0.07-0.13 para organismos vivos). La periodicidad en todos es la misma porque en todos tenemos cuatro bases. Sin embargo, en Generador Aleatorio se aprecian diferencias en cuanto al promedio de las sumas de A-Ts y C-Gs (0.36 y 0.33 respectivamente), esto es porque su proporción de concentraciones nucleotídicas difiere significativamente del resto de los casos (alrededor de 25% para cualquiera de los cuatro; es significativo que todas las bases del Generador tuvieron la misma d, lo cual indica que, a pesar de existir complejidad, ésta es redundante en los cuatro nucleótidos). También la R^2, que es proporcional a la probabilidad, es prácticamente la misma para cada una de las bases, es decir, están equiprobablemente distribuidas en el genoma de manera homogénea. Lo anterior significa que, aunque el Generador posee información con cierto nivel de complejidad (en el caso de la proporción de G-Cs, supera la de todos los demás), no está bien estructurada pues no se observa ni la complementariedad en las bases ni una distribución de valores d que indiquen algún tipo de dinámica en el espacio de fases de la serie de tiempo. Su concentración de nucleótidos es la misma y, por lo tanto, la probabilidad de encontrar alguna de ellos en el genoma, igual. Esto sería comparable a recorrer un largo y tortuoso camino para llegar al punto de partida, un tipo de complejidad redundante y azarosa.

En el resto de los organismos, por el contrario, se aprecia una complejidad dirigida y persistente (como en el caso de un caminante

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

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

Proporciones de bases complementarias

Chargaff descubrió que la concentración de adeninas es, aproximadamente, la misma que la de timinas y que lo mismo ocurre con las citosinas y las guaninas (Gribbin 1986, Resendis O. y L. S. García-Colín 2001). Una consecuencia directa de esta observación es que las tasas (G+A)/(C+T) y (G+T)/(A+C) tienen un valor cercano a uno. Se procedieron a hacer estas mismas operaciones para los organismos vivos y el Generador Aleatorio; los resultados para todos fueron los obtenidos por el investigador austriaco. No obstante, al sumar la cantidad de adeninas y timinas y posteriormente la de citosinas y guaninas (bases complementarias en ambos casos) se observa una tendencia positiva en los porcentajes de A-Ts (59 a 63%) y otra conse-cuentemente negativa en C-Gs (37-41%). A causa de esta desigualdad compartida por los organismos vivos, al dividir el número de CGs entre el de ATs se obtiene un valor casi constante (0.61-0.68). Al dividir de manera inversa es posible encontrar también valores muy semejantes. Se hace evidente que la actividad química entre los nucleótidos de la cadena genómica no es uniforme, lo cual es absolutamente indispensable para dar lugar a interacciones moleculares (puentes de Hidrógeno, fuerzas de Van der Waals, enlace químico, etcétera).

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

Conclusiones

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

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

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

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

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

5. Por otro lado, los conteos de nucleótidos indican la presencia de reglas específicas en su distribución similares a las propuestas por

Chargaff (Gribbin 1986, Resendis O. y L. S. García-Colín 2001) en el contexto de cadenas complementarias. Esto resalta un principio subyacente en el texto genómico que indica que la proporción de nucleótidos posee un significado relacionado con su función. Todo esto lleva a concluir que el ADN llamado no codificante contiene dentro de sí una gran cantidad de información dispuesta de manera coherente, estructurada y compleja, cuyo carácter lingüístico (comunicativo) hace pensar en alguna clase de función biológica. Queda desde luego mucha investigación por hacer para dilucidar los complejos mecanismos bioquímicos y genéticos que se encuentran presentes en la información que nuestras pruebas estadísticas han revelado.

REFERENCIAS

CANTÚ-BOLÁN, B. Y E.HERNÁNDEZ-LEMUS
2006 Statistical properties and linguistic coherence in noncoding DNA sequences, *Revista Mexicana de Física E,* 51 (2): 118-125 (2005).

DAGAN, T. *ET AL.*
2004, LUGene: a database of ALU elements incorporated within protein-coding genes, *Nucleic acids research,* vol.32, Database issue D489-D492.

DERRIDA, B. *ET AL.*
2000 Distribution of Repetitions of Ancestors in Genealogical trees, *Physica A,* 281: 1-16.

DEWEY, C. N., P. M. HUGGINS, K. WOODS, B. STURMFELS Y L. PACHTER
2006 Parametric alignement of drosophila genomes, PLOS *Computational Biology*, vol.2, no. 6, e73: 606-614.

DOKHOLYAN, N. V
1999 *Applications of statistical mechanics to biological macromolecules,* Boston University, Graduate School of Arts and Sciences, doctoral these (Physics).

GRASSBERGER, P. Y I. PROCACCIA
1983 Characterization of Strange attractors, *Physical review,* Letters, 50: 346-349.

GRIBBIN, J.
1986 *En busca de la doble hélice,* Biblioteca Científica Salvat, 2a. ed., Barcelona, 287 p.

GRIFFITS, A.
2001 *Modern genetic analysis,* 4a. ed., W. H. Freeman, Nueva York, 675 p.

HATTORI, M. A. *ET AL.*
2000 The DNA Sequence of Human Chromosome 21, *Nature,* 405 (6784): 311.

HUBBARD, T. J. P. *ET AL.*
2007 *Nucleic acids research,* vol. 35, doi:10.1093/nar/gkl996.

LI, W.
2001 *Zipf's Law in importance of genes for cancer classification using microarray data,* arxiv:physics/0104028 v1 6/Apr/2001.

MATSUTANI, S.
2005 Links between repeated sequences, *Journal of biomedicine and biotechnology,* vol. 2006, ID 13569: 1-3.

RESENDIS, O. Y L. S. GARCÍA-COLÍN
Application of the theory of stochastic processes to the configuration of biological systems, *Physica A* 290 (2001): 203-210.

SAXONOV, S., P. BERG Y D. L. BRUTLAG
2006 *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters,* PNAS 103 (5), PMID 16432200.

STANLEY, H.E. *ET AL.*
1999a Scaling features of noncoding DNA, *Physica A* 273: 1-18.
2000a Exotic statistical physics: applications to biology, edicine, and Economics, *Physica A* 285, 1-17.
2000b Scale invariance and universality: organizing principles in complex systems, *Physica A* 281: 60-68.

STANLEY, H.R.R. *ET AL.*
1999b Clustering of identical oligomers in coding and noncoding DNA sequences, *Journal of Biomolecular Structure & Dynamics,* 17 (1): 79-87.

UMYNLY, B. *ET AL.*
2007 Most human ALU and Murine B1 repeats are unique, *J. Cell. biochem,* Apr. 3, ID, 17407136.

YANAI, I. *ET AL.*
2000 Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification, *Physical review letters,* 85 (12): 2641.