Aplicación del análisis de rango reescalado R/S para la predicción de genes en el genoma vegetal

Rescaled range R/S analysis application for genes prediction in the plant genome

Martha Isabel Almanza Pinzón[1], Karina López López [2], Carlos Eduardo Téllez Villa[3]

[1]Bióloga, M.Sc., Candidata a Ph.D. en Ciencias Agropecuarias, Universidad Nacional de Colombia, sede Palmira. A.A. 237. Docente Universidad del Cauca.
[2] Ingeniera Bioquímica. Ph.D. en Biotecnología de Plantas. Docente Universidad Nacional de Colombia, sede Palmira.
[3] Ingeniero de Sistemas. Universidad del Cauca.
Autores para correspondencia: klopezl@palmira.unal.edu.co, mialmanzap@palmira.unal.edu.co, ctellez@unicauca.edu.co

Recibido: 22-10-2008 Aceptado: 2-11-2010

Resumen

La predicción de genes es en la actualidad uno de los principales desafíos de la genómica. La predicción permite realizar experimentos con alta probabilidad de encontrar genes de interés y comparar regiones de ADN de importancia agronómica entre genomas; además, ayuda a restringir los espacios de búsqueda en las bases de datos. Un procedimiento estadístico con base en el análisis R/S y el coeficiente de Hurst fue desarrollado para caracterizar y predecir genes y los componentes estructurales de estos (exones e intrones) en los genomas eucariotas completos de *Arabidopsis thaliana*, *Oriza sativa* y *Mus musculus*. Algoritmos en lenguaje de programación Python fueron desarrollados para extraer, filtrar y modelar más del 80% de las secuencias de genes registradas para estos genomas en la base de datos del GeneBank del NCBI. El análisis R/S permitió demostrar que existe un orden estructural en la distribución de los nucleótidos que constituyen las secuencias en las que predominan los fenómenos de memoria o dependencia de largo alcance. La estructura de memoria varía según el tipo de secuencias y el genoma de la especie. Las secuencias de los genes y exones de los genomas vegetales analizados presentaron comportamiento persistente mientras que las de los intrones tuvieron un comportamiento antipersistente, en comparación, al genoma animal en el cual los tres tipos de secuencias presentaron comportamiento persistente. De acuerdo con los parámetros provenientes del análisis R/S, el patrón de distribución de las secuencias del genoma se repitió de manera estadísticamente similar en cada uno de los cromosomas que pertenecen a una especie, constituyéndose en evidencias fundamentales de invarianza por cambio de escala; es decir, cada cromosoma por sí solo es una réplica estadística a menor escala del genoma completo. Los parámetros constituyeron criterios compactos para derivar predictores (clasificadores) de secuencias que alcanzaron promedios de sensibilidad y especificidad mayor del 81% y 70%, respectivamente. Este procedimiento podría ser probado en otros genomas y utilizado como criterio para incrementar la eficiencia de la selección en los programas de mejoramiento genético vegetal.

Palabras clave: Genómica comparativa, predicción de genes, análisis R/S, coeficiente de Hurst, *Arabidopsis thaliana, Oryza sativa*, *Mus musculus*.

Abstract

Currently gene's prediction problem is one of the main genomic challenges. Prediction allows performing experiments with high probability of interesting genes to be found and compare DNA regions of agronomic importance among genomes; besides, it helps to restrict the searching spaces into the data

bases. A statistical procedure based on the R/S analysis and the Hurst coefficient was developed in order to characterize and predict genes and their structural components (exones and introns) in the whole eukaryotic genomes of *Arabidopsis thaliana*, *Oriza sativa* and *Mus musculus*. Python programming language algorithms were developed with the purpose of extract, screen and modeling more than 80% of the registered gene sequences for these genomes in the NCBI Gene Bank data base. The R/S analysis allows to demonstrate that a structural order do exist in the distribution of the nucleotides which are constituting sequences with the memory or long range dependence phenomena predominance. The memory structure varies according to the sequences type and the species genome. The genes and exones sequences from the analyzed plant genomes showed a persistent behavior whereas those from the introns had an anti-persistent behavior, in comparison with animal genome in which the three type of sequences showed persistent behavior. According to R/S analysis out coming parameters the genome sequences distribution pattern was replicated in a statistically similar manner in each chromosome belonging to one species, constituting fundamental evidences of invariance by scale change; it means each chromosome by itself is a statistical replication to a minor scale of the whole genome. The parameters constituted compact criteria in order to derivate sequences predictors (classifiers) which reached sensibility and specificity averages higher than 81% and 70% respectively. This procedure could be tried in other genomes and be used as a criterion in order to increasing selection efficiency in plant genetic breeding programs.

Key words: Comparative genomics, gene's prediction, R/S analysis, Hurst coefficient, *Arabidopsis thaliana, Oryza sativa*, *Mus musculus*.

Introducción

Analizar genomas completos es de fundamental importancia para entender procesos biológicos como la herencia, la evolución y el mejoramiento genético. La predicción de genes consiste en identificar regiones codificantes (exones) en una secuencia de ADN anónima. La identificación no es de ningún modo una tarea trivial debido a la complejidad estructural y funcional de los genomas eucariotas como de las bases de datos, depositarias de la información biológica. La predicción permite realizar experimentos con alta probabilidad de encontrar genes de interés o alguno de sus componentes; comparar regiones de ADN entre genomas de importancia agronómica, patológica o taxonómica; además de ayudar a restringir los espacios de búsqueda en las bases de datos. El gran reto de la predicción es lograr establecer procedimientos matemáticos para descubrir patrones, propiedades o reglas generales que determinan la información, estructura y organización de los genomas o para romper el texto cifrado en el ADN.

Desde la década de los sesenta en el siglo XX se han realizado diversos análisis estadísticos en secuencias de ADN mediante métodos de la lingüística, la teoría de la información y métodos de escalamiento, como la complejidad y el análisis fractal. La primera compilación de estos análisis fue publicada por Hawkins en 1988, la cual indica que las secuencias de exones e intrones presentan comportamientos estadísticos diferentes y enfatiza la existencia de dos propiedades básicas de dependencia o correlación de los nucleótidos en las secuencias: correlaciones de corto y largo alcance. Según Li y Kaneko (1992), Peng *et al*. (1992), Karlin y Brendel (1993), Karlin y Cardon (1994), Buldyrev *et al*. (1995) estas propiedades son indicadoras del nivel de complejidad y estructura jerárquica del ADN. Sin embargo, existe desacuerdo en relacionar las propiedades de correlación con el tipo de secuencia. El debate continúa abierto y aspectos interesantes de éste se encuentran en Yu *et al* (2001). El argumento principal de la discusión se basa en la propuesta de la existencia de más de un nivel de información en el lenguaje biológico del ADN.

Hao (2000) argumenta que aunque los análisis estadísticos han sido herramientas útiles en el estudio del ADN, las evidencias señalan que los métodos no son lo suficientemente buenos para detectar diferencias entre y dentro de grupos de secuencias de ADN; y, afirma que existe la necesidad urgente de desarrollar nuevas aproximaciones conceptuales y experimentales a nivel biológico, matemático y computacional para el análisis e interpretación de patrones en los genomas.

El análisis del recorrido estandarizado, también llamado análisis de rango reescalado, o análisis R/S, es una prueba estadística utilizada para cuantificar la dinámica de una serie temporal y determinar la existencia de características fractales en un sistema (Hoop *et al.,* 1993). El valor de esta prueba es su sensibilidad para distinguir correlaciones o dependencias estadísticas de corto o largo alcance, en procesos aleatorios; correlaciones que se presentan como consecuencia de la tendencia que presentan las observaciones a desviarse del valor medio durante un tiempo más o menos prolongado. Así, el estadístico R/S mide el rango de las desviaciones de las sumas parciales de una serie temporal respecto de su media, reescalado por la desviación estándar de la serie.

El análisis R/S fue desarrollado por Harold Hurst (1951; 1956); sin embargo, no fue sino hasta 1968, cuando Mandelbrot y Van Ness (1968); Mandelbrot y Wallis, 1969a,b,c) lo introdujeron como herramienta de análisis estadístico para registros fractales. El fundamento matemático de la prueba se presenta en Mandelbrot (1982). Una consecuencia importante del análisis R/S es obtener el coeficiente de escalamiento, también denominado, de Hurst, o exponente R/S, que puede tomar cualquier valor entre 0 y 1 (Mandelbrot, 1982).

Xiao *et al* (1995) encontraron que las secuencias de nucleótidos en animales, plantas y humanos presentan propiedades fractales y demostraron que las secuencias de exones e intrones difieren en sus propiedades fractales. Yu y Chen (2000), Yu *et al* (2001) y Yu y Wang (2001) propusieron modelos de series de tiempo basados en las longitudes de las secuencias de ADN de genomas completos; calculando correlaciones, análisis R/S y coeficientes de Hurst encontrando que estas medidas podrían dar una señal del orden de los genes en el cromosoma.

Materiales y métodos

Las secuencias completas y anotaciones de los genes con sus respectivas secuencias de exones e intrones por cromosoma de los genomas: *A. thaliana*, *O. sativa* y *M. musculus*, fueron obtenidas directamente de los archivos en formato gbk disponibles vía ftp en la base de datos pública GeneBank perteneciente al Centro Nacional de Información Biotecnológica NCBI (ftp://ftp.ncbi.nih.gov/genomes/) en marzo del 2009.

Algoritmos en lenguaje de programación Python fueron desarrollados, tanto para extraer las secuencias de ADN del formato gbk a un formato Fasta para depurar, caracterizar, clasificar y evaluar, biológica y estadísticamente los datos, por conjunto de secuencias (genes, exones e intrones), por cromosoma y por genoma.

La información extraída de la base de datos se organizó en tres archivos Fasta por cromosoma (secuencias completas de genes, exones e intrones) para facilitar el manejo de la información de las secuencias por genoma y mayor rapidez de procesamiento. Dos restricciones fueron hechas: las secuencias de genes que contenían nucleótidos desconocidos (representados por la letra N) o longitudes menores de 20 nucleótidos en sus secuencias fueron excluidas del estudio. La primera restricción, se hizo porque existen criterios biológicos para asignarles valores a los cuatro nucleótidos que constituyen una secuencia, y así, transforma la secuencia de nucleótidos en una secuencia numérica (Stanley *et al.*, 1994; Yu y Chen, 2000) pero, no se encontró, un criterio biológico o estadístico confiable para asignarle valor a una posición N en una secuencia. En la segunda restricción, se tuvo en cuenta dos hallazgos: el de Deutsch y Long (1999) indicando que las señales de corte y empalme (*splicing*) de genes codifican adecuadamente cuando la longitud de los intrones es mayor de 20 nucleótidos; y el de Peters (1991), quien encontró mediante simulaciones de Montecarlo desviaciones sistemáticas en el valor esperado del Log(R/S) para n(observaciones)<20.

M. musculus fue el genoma de referencia para validar las dos estrategias propuestas debido: primero, a que la información biológica de una buena cantidad de sus secuencias ha sido obtenida experimentalmente (Mouse Genome Sequencing Consortium, 2002); y segundo, porque todos sus cromosomas autosómicos son morfológicamente similares (Craig y Bickmore, 1993). Además, Kolbe *et*

al., (2004) señalan que la adición de una tercera especie mejora notablemente los análisis comparativos globales y locales de secuencias genómicas.

El estudio consideró todos los cromosomas de los únicos genomas vegetales eucariotas secuenciados disponibles en la base de datos (*A. thaliana*: 2n=10; *O. sativa*: 2n=24) y únicamente los 19 cromosomas autosómicos del genoma del ratón (*M. musculus*: 2n=42). Los cromosomas sexuales (X y Y) fueron excluidos porque presentan características estructurales, genéticas, anatómicas y fisiológicas particulares que podrían ocasionar sesgos estadísticos (Gu *et al*., 2000).

El procedimiento del análisis R/S aplicado a cada secuencia completa de ADN (gen, exón e intrón) para la estimación del respectivo coeficiente de Hurst (H), se desarrolló en cuatro etapas: la primera, la sucesión de nucleótidos individuales de cada secuencia se traduce a una sucesión numérica; la segunda, procedimiento matemático del análisis R/S; tercera, estimación del coeficiente de Hurst (H); y cuarta, prueba de significancia del coeficiente de Hurst (H). A continuación se presentan los criterios biológicos, propiedades matemáticas y notación del análisis R/S aplicado a secuencias de ADN:

Las cuatro letras (A, C, G y T) que representan las cuatro clases de nucleótidos presentes en una secuencia de ADN fueron sustituidas por cuatro valores: -2, -1, 1 y 2, obteniéndose una secuencia numérica unidimensional. Esta representación numérica de los nucleótidos fue utilizada posteriormente por Yu y Chen (2000). El criterio para asignar estos valores consiste en discriminar las purinas (A y G) de las pirimidinas (C y T) porque proporciona resultados robustos debidos probablemente a la complementariedad fisicoquímica de las purinas y pirimidinas (Stanley *et al*., 1994).

El procedimiento matemático del análisis R/S consta de los siguientes pasos:

Promedio: $< x >_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sumas parciales: $X(i,n) = \sum_{u=1}^{i} [x_u - < x >_n]$

Rango: $R(n) = \max_{1 \le i \le n} X(i,n) - \min_{1 \le i \le n} X(i,n)$

Desviación estándar:

$$S(n) = [\frac{1}{n}\sum_{i=1}^{n} (x_i - < x >_n)^2]^{\frac{1}{2}}$$

Generalización: $\frac{R(n)}{S(n)} \approx cn^H$ donde $0 \le H \le 1$

En la generalización: c, es una constante; n, es el número de nucleótidos que constituyen la secuencia; y H: es el coeficiente de escalamiento o de Hurst (H); R/S es el estadístico que depende del tamaño de la serie y que se define como el rango de variación de la serie expresada en términos de su desviación estándar. El coeficiente H corresponde a la pendiente obtenida por el modelo clásico de regresión lineal ajustado por un estimador de mínimos cuadrados ordinarios ($Log(R_{(n)}/S_{(n)})$ *vs* Log(n)) con base en la transformación logarítmica de la ecuación de generalización, de la siguiente manera: $Log(R_{(n)}/S_{(n)})=log\ c+HLog(n)$.

Este artículo propone un procedimiento estadístico basado en el análisis de rango reescalado R/S y el coeficiente de escalamiento por intervalos. La idea subyacente del estudio es que toda la información necesaria para una proteína está en la propia estructura y contenido de nucleótidos de las secuencias que componen los genes de un genoma.

El coeficiente de Hurst (H) es empleado usualmente como medida de complejidad de un sistema (Mandelbrot, 1982). H mide la aleatoriedad aparente de los nucleótidos en secuencias altamente ordenadas como los exones e intrones a nivel de cromosomas y genomas. Un exponente de Hurst bajo (0<H<0.5) indica presencia de una complejidad alta y "comportamiento antipersistente"; la fuerza de la antipersistencia aumenta a medida que H tiende a cero. Un exponente de Hurst alto (0.5<H<1) indica complejidad baja y "comportamiento persistente"; cuanto mayor sea el valor de H, más fuerte es la tendencia de persistencia. H=0.5, corresponde a series de datos completamente aleatorias, en donde la correlación de un periodo actual tiene una correlación nula con un periodo anterior o

futuro. H=1: indica un comportamiento determinístico. En general, series de datos con un H≠0.5 sugieren que aunque las observaciones estén suficientemente distantes unas de otras en el espacio no son estadísticamente independientes, característica importante de los sistemas biológicos.

El análisis de rango reescalado R/S se aplicó a cada una de las secuencias completas de genes, exones e intrones de los tres genomas de estudio, obteniéndose para cada secuencia su máximo, mínimo, rango y coeficiente de Hurst (H) y se adicionó la información de longitud de la secuencia. Estas medidas constituyen los parámetros o descriptores propios que definen cada secuencia de ADN y a partir de los cuales se clasifica o predice la secuencia. La decisión de incluir longitud como parámetro de definición de la secuencia de ADN, se basó en la afirmación de Mandelbrot y Wallis (1969b) respecto a que el análisis R/S de una serie de datos está asintóticamente relacionado con su longitud, por lo que formalmente plantearon la generalización, como: $R/S=cn^H$.

El coeficiente de Hurst se dividió en diez intervalos semiabiertos de igual amplitud, de tal manera que una secuencia pertenece a un intervalo, según el coeficiente obtenido. El Cuadro 1 presenta los resultados del procedimiento del análisis R/S por intervalos del coeficiente de Hurst, para las secuencias de exones e intrones del cromosoma 3 de *M. musculus*.

La calidad del modelo de predicción basado en el intervalo del coeficiente de Hurst, los parámetros del análisis R/S y la longitud de la secuencia, se calculó con base en su sensibilidad (probabilidad de clasificar correctamente un exón) y especificidad (probabilidad de clasificar correctamente un intrón) (Burset y Guigó, 1996):

Sensibilidad: = VP/(VP+FN) = FVP (fracción de verdaderos positivos)

Especificidad: = VN/(VN+FP) = FVN (fracción de verdaderos negativos)

En donde, VP: verdaderos positivos; FN: falsos negativos; VN: verdaderos negativos; FP: falsos positivos.

Resultados y discusión

Entre 0<H<1, cada grupo de secuencias de los genomas presenta una historia de complejidad diferente. Los coeficientes de Hurst (H) de las secuencias de genes, exones e intrones de los tres genomas fueron significativamente diferentes de H=0.5 con niveles altos de ajuste (R^2) de los modelos de regresión, denotando que el análisis R/S permite discriminar secuencias de ADN (Cuadro 2).

El coeficiente de Hurst (H) cuantifica y describe las relaciones entre los nucleótidos

Cuadro 1. Promedios para los parámetros del análisis R/S (Máximo, mínimo y rango) y longitud (pb) por intervalos del coeficiente de Hurst (H) de las secuencias de exones y de intrones del cromosoma 3 del genoma de *M. musculus*. 0≤H≤1.

Hurst	Frecuencias		Máximo		Mínimo		Rango		Longitud (pb)	
(H)	Exones	Intrones	Exones	Intrones	Exones	Intrones	Exones	Intrones	Exones	Intrones
[**0.1-0.2]**	12	7	9,12	66,01	-15,6	-28,48	24,72	94,49	518,42	2868,86
[**0.2-0.3]**	234	130	10,68	32,61	-8,4	-37,32	19,08	69,93	222,97	1899,43
[**0.3-0.4]**	1422	869	12,81	52,75	-9,93	-45,49	22,75	98,24	236,19	2857,05
[**0.4-0.5]**	2923	2136	13,88	63,98	-13,37	-60,90	27,25	124,89	263,15	3113,76
[**0.5-0.6]**	2736	2510	16,14	98,73	-16,75	-89,48	32,88	188,21	292,55	4563,18
[**0.6-0.7]**	1251	1545	18,30	141,28	-20,52	-136,34	38,81	277,62	305,2	6056,96
[**0.7-0.8]**	299	574	20,53	280,31	-26,38	-233,26	46,92	513,57	334,83	10873,24
[**0.8-0.9]**	42	162	20,82	550,84	-35,55	-490,97	56,36	1041,8	391,67	18102,3
[0.**9-1.0]**	8	31	10,12	635,35	-55,44	-781,00	65,56	1416,4	295,75	19809,58

Cuadro 2. Coeficientes de Hurst (H) y parámetros R/S, por conjunto de secuencias de ADN y por genoma.

		H	R^2	Parámetros R/S				
Genoma	**Frec. Abs.**			**b**	**Máximo**	**Mínimo**	**Rango**	**Long. (nt)**
At								
Genes	32066	0.53	0.89	0.48	77.22	-64.90	142.12	2119.17
Exones	146118	0.51	0.89	0.32	16.92	-21.65	38.57	338.60
Intrones	114052	0.48	0.88	0.43	13.02	-15.01	28.02	165.84
Os								
Genes	28134	0.52	0.88	0.52	83.33	-83.70	167.02	3230.71
Exones	135625	0.51	0.89	0.31	15.74	-19.05	34.79	328.64
Intrones	107491	0.49	0.88	0.42	22.62	-22.95	45.58	444.38
Mm								
Genes	24549	0.59	0.89	0.30	558.82	-494.09	1052.90	33400.71
Exones	189728	0.53	0.89	0.24	17.78	-17.82	35.60	299.25
Intrones	165179	0.55	0.89	0.32	118.16	-114.49	232.65	4800.25

Mm: *Mus Musculus*; *Os*: *Oryza sativa*; *At*: *Arabidopsis thaliana*; Frec. Abs.: Frecuencia Absoluta; R^2: Coeficiente de determinación; Long. (nt): Longitud en nucleótidos; b: intercepto de la regresión bilogarítmica. $H \neq 0.5$ ($p<0.05$).

de las secuencias genómicas. Los tres conjuntos de secuencias del genoma animal presentaron comportamiento persistente (H>0.5), mientras que en los genomas vegetales, los genes y exones presentaron comportamiento persistente y los intrones comportamiento antipersistente (H<0.5). Estos comportamientos indican que existe un orden estructural específico en la composición de nucleótidos que constituyen cada conjunto de secuencias en cada genoma. Los valores de Hurst fueron consistentes con los reportados por Yu y Chen (2000) al obtener H>0.5 para muestras de secuencias de exones e intrones de los genomas de *M. musculus* y del humano.

Las mismas características estadísticas generales del coeficiente de Hurst (H) observadas para los conjuntos de secuencias por genoma fueron encontradas por cromosoma ($H \neq 0.5$, $p<0.05$; $R^2 \geq 87\%$), lo que indica que también existe una estructura en las secuencias de los cromosomas de cada genoma. Sin embargo, el hallazgo más significativo fue que el valor del coeficiente de Hurst de cada conjunto de secuencias por genoma no era significativamente diferente del observado para cada cromosoma del genoma respectivo (datos no mostrados).

Estos primeros resultados revelan tres aspectos significativos: primero, según el valor de Hurst obtenido para cada conjunto de secuencias, las de genes y las de exones son las que presentan mayor tendencia a la persistencia en los tres genomas, este resultado es consistente con la función que realizan estas secuencias en el genoma. Segundo, pareciera que el efecto de persistencia se incrementa conforme las especies se tornan más complejas (nótese que los valores de Hurst de los tres conjuntos de secuencias son más altos en el genoma animal, Cuadro 2). Tercero, valores de Hurst estadísticamente similares entre cromosomas y el genoma respectivo son evidencias de invarianza de escala.

Significado matemático de los coeficientes de Hurst (H)

El comportamiento estadístico persistente (0.5<H<1) o memoria de largo alcance indica propagación de la información en el espacio genético que se manifiesta a través de la dependencia o correlación positiva entre los nucleótidos que contienen las secuencias y, entre las secuencias que contienen los cromosomas y los genomas. La información que llevan los nucleótidos sigue una caminata aleatoria sesgada, la fuerza del sesgo depende de qué tan lejos de 0.5 se encuentre el valor de H.

La intensidad del comportamiento persistente se incrementa cuando H se aproxima al valor de uno, es decir, mayor es el grado de correlación entre nucleótidos lejanos y las diferencias entre nucleótidos cercanos tienden al valor de cero; y es este efecto de memoria de largo alcance el que causa la aparición de tendencias y ciclos. La intensidad de la memoria se disipa a medida que H se acerca al valor de 0.5, es decir, mayor es el grado de correlación entre nucleótidos cercanos,

caracterizando un proceso de corto alcance. Un valor de H>0.5 indica la presencia de un sistema no lineal.

En contraste, la intensidad del comportamiento antipersistente (0<H<0.5) se incrementa cuando H se aproxima a cero, mayor es el ruido entre nucleótidos y mayor apariencia de aleatoriedad "errática" tienen los nucleótidos en las secuencias.

Entonces, una secuencia de ADN con un valor de H≠0.5 es el resultado de la interconexión entre los nucleótidos. Cada nucleótido acarrea memoria de los nucleótidos que le precedieron, existe un efecto de sesgo, tendencia o de memoria. Sin embargo, esta no es la memoria de corto alcance, comúnmente llamada Markoviana. Esta memoria es distinta, es de largo alcance. Nucleótidos contiguos tienen mayor impacto que los más distantes, pero estos últimos siguen teniendo un efecto residual que influye en la secuencia.

Conocida la interpretación matemática de los valores de Hurst desde la perspectiva del ADN se puede tomar el coeficiente de Hurst como una medida de la complejidad de la estructura de las secuencias de ADN en cada genoma. El valor de H indica el valor máximo que alcanzan las secuencias en algún punto entre el orden de lo completamente regular y el desorden de lo completamente aleatorio, siendo ambos extremos sistemas de complejidad nula (o extremadamente baja). En esta lógica, un coeficiente de Hurst alto indica que la estructura de la secuencia presenta una complejidad baja y viceversa, un coeficiente bajo revela que la estructura de la secuencia presenta una complejidad alta.

Dos conclusiones significativas surgen del análisis expuesto: la primera, el coeficiente de Hurst (H) es una medida de la información que los nucleótidos conservan a través del espacio genético; y, la segunda, los valores de H obtenidos indican que cada conjunto de secuencias de un genoma tiene su propia estructura organizada de memoria y, que además, esta estructura se repite estadísticamente en los cromosomas del genoma correspondiente.

Computacionalmente, la persistencia (correlación positiva entre nucleótidos) permite modelar fenómenos que tienden a agruparse primero a un lado de la media y luego al otro lado, mientras que la antipersistencia (correlación negativa entre nucleótidos) permite modelar fenómenos que fluctúan fuertemente alrededor de la media (Mandelbrot y Hudson, 2004).

Significado biológico de los coeficientes de Hurst (H)

Los comportamientos de persistencia o antipersistencia, indican que existen tendencias o ciclos de longitud variable en la composición de nucleótidos de las secuencias, las cuales se mantienen a lo largo de la secuencia y que realmente hay una estructura en la secuencia. La persistencia indica tendencias en la secuencia que se refuerzan continuamente, esto es, sí la tendencia de la secuencia ha sido positiva en el último trayecto observado, es probable que siga siendo positiva y no negativa en el siguiente trayecto. La persistencia indica estructuras estables con una alta probabilidad de cumplir funciones específicas, pero también revela que son insensibles a muchas alteraciones menores.

En contraste, el comportamiento antipersistente muestra tendencias que se revierten continuamente (los valores que toma la secuencia tienden a compensarse uno al otro) y se manifiestan como correlaciones negativas entre los nucleótidos que componen las secuencias. La antipersistencia indica estructuras inestables y sin función específica (Balbín y Andrade, 2004); sin embargo, Mattick *et al*., (2001), señalan que la antipersistencia en secuencias de ADN podría estar señalando plasticidad genética para acomodarse a la unión de ligados (proteínas, ADN, iones metálicos, etc.) que intervienen en la regulación génica.

Biológicamente, el predominio de valores altos de Hurst en las secuencias de los genomas superiores, como el del genoma animal, indica que la probabilidad de mutaciones es más baja que en las especies poco adaptadas con valores bajos H, donde la probabilidad de que las secuencias cambien continuamente es alta.

Las tendencias persistentes en secuencias de ADN pueden tener dos posibles orígenes: primero, son el resultado de las res-

tricciones físico-químicas que se establecen entre los nucleótidos de una secuencia y que dan lugar a un determinado orden con un rango de posibles variaciones en la secuencia. El orden estaría determinado, no sólo por el posible aumento o cambio de nucleótidos en la secuencia, sino también por la fijación o conservación de estos en posiciones determinadas de la secuencia (Carothers *et al*., 2004). Además, Zhang y Zhang (1991) afirman que el potencial codificante de una secuencia de ADN es debido a la estructura rígida y regular de los nucleótidos en el codón generada por las características físico-químicas de estos.

Las afirmaciones anteriores apuntan en el mismo sentido de las sugeridas por Denton et al. (2003) para las secuencias proteicas, en donde leyes físico-químicas y no únicamente la tríada: variación, herencia y presión de selección (natural o artificial) son los factores que más aportan al complejo orden biológico. Schultes *et al*. (1999) señalan que aunque la selección modifique las secuencias de ADN para adaptarse a funciones específicas, la mayor parte del orden estructural de un genoma es intrínseco. Bernardi (1995), señala que la composición de nucleótidos de las secuencias está sometida a reglas precisas que son decisivas en la organización, fisiología y evolución de un genoma.

El otro posible origen, está relacionado con el incremento de tamaño del genoma debido a procesos de poliploidía y transferencia horizontal que generan gran cantidad y variedad de estructuras de ADN repetitivas y de genes duplicados (Li y Kaneko (1992). Ordov *et al*. (2006) argumentan que la persistencia en secuencias de ADN puede estar relacionada con sesgos estadísticamente significativos originados por la presencia de estructuras repetitivas en la composición de nucleótidos de las secuencias, tales como: repeticiones dispersas, en tándem o altamente repetitivas, o una combinación de todas o algunas de estas estructuras. Según estos investigadores, los sesgos ocasionan que las variaciones de los valores de H en una secuencia persistente sean monótonas, correlacionen positivamente y sigan una caminata aleatoria sesgada.

La transferencia horizontal también duplica estructuras repetitivas, genes y produce redundancia génica funcional. Entendida ésta como la capacidad que tienen algunos genes de realizar la misma función, es una actividad que facilita el cambio evolutivo, puesto que se reduce la probabilidad de efectos letales por mutaciones en uno de los genes redundantes. La pérdida de funcionalidad de uno de los genes puede ser cubierta, al menos parcialmente, por otro que evite la aparición de alteraciones fenotípicas.

Las estructuras de ADN repetitivas y genes duplicados producidas por estos dos procesos biológicos (poliploidía y transferencia génica) son consideradas cruciales para explicar la complejidad y adquisición de nuevas funciones en las especies superiores; además de ayudar a explicar la propiedad de autosimilaridad estadística en genomas (Doolittle, 1981).

Mandelbrot y Hudson (2004) argumentan que los comportamientos de persistencia o antipersistencia en los sistemas naturales tienen un importante significado funcional, ya que están estrechamente relacionados con procesos de retroalimentación positiva o negativa, respectivamente. La retroalimentación positiva está orientada hacia la producción y cambio acumulativo de un estado dado. En secuencias de ADN se refiere al hecho de que los nucleótidos que componen una secuencia se influencian uno al otro, a los contiguos y aun a los más lejanos y, de esta manera, las secuencias logran mantener largos segmentos de la secuencia conservados. En contraste, la retroalimentación negativa consiste en alcanzar y/o mantener la funcionalidad o continuidad de un estado. El coeficiente de Hurst indica la medida de dicha retroalimentación. El genoma, entendido como un sistema natural complejo, tendría ambos procesos de retroalimentación para desarrollar sus funciones.

Adicionalmente, algunos grupos de investigación señalan que las condiciones de persistencia le confieren al genoma importantes ventajas relacionadas con procesos de adaptación a condiciones ambientales; también, que la información biológica proveniente del análisis de las correlaciones de largo alcance puede estar relacionada con

los mecanismos de las funciones saludables o patológicas de un organismo (Goldberger et al., 2000).

Finalmente, dos posibles explicaciones biológicas tendría la invarianza de escala que se expresa entre cromosomas y el genoma respectivo, la primera: disminuir la cantidad de información genética que se necesita para modelar una especie al reutilizar sus propios principios estructurales o biológicos fundamentales. Las regularidades estadísticas de persistencia o antipersistencia en una secuencia representan ahorro en la capacidad de almacenar información y llevan implícito cierto equilibrio de las fuerzas resultantes de la interacción genotipo-ambiente en una especie. Cada cromosoma contiene en sí mismo la estructura de toda la información genética del genoma al cual pertenece.

La segunda explicación es la homogeneización del genoma que da lugar a mayor diversidad entre que dentro de especies. Brown *et al*. (1972) encontraron que dentro de una especie de anfibios, las diferencias entre las secuencias repetitivas presentes en los intrones de una familia de genes de ARNr fueron mínimas; mientras que entre especies, estas secuencias en intrones presentaron diferentes mutaciones y evolucionan libremente. Existe una evolución concertada que resulta en que estas secuencias repetitivas (independientemente del número de copias, función o distribución) conserven su identidad dentro de una especie, frente a la variación entre especies Dover (1982).

Representación gráfica del Coeficiente de Hurst (H)

La Figura 1 es la representación gráfica del análisis R/S (denominada *pox plot*) para obtener el valor del coeficiente de Hurst (H) para una secuencia de ADN o un genoma. En la figura también se puede visualizar, para una mejor comprensión intuitiva, los comportamientos de persistencia y antipersistencia en cualquier secuencia de ADN. El eje de las abscisas representa el logaritmo del valor numérico asignado a cada nucleótido (*n*) de una secuencia, y el eje de las ordenadas el logaritmo del valor R/S alcanzado por la secuencia, a medida que la caminata avanza nucleótido por nucleótido. El coeficiente de Hurst corresponde a la pendiente del ajuste lineal de los puntos en el diagrama. Las secuencias persistentes presentan menor cantidad de escalones largos que se caracterizan por pequeñas pero constantes variaciones de los valores máximo y mínimo de los nucleótidos a lo largo de la secuencia. El desplazamiento vertical más que horizontal de las secuencias se refleja en una pendiente menos inclinada y por lo tanto, en un valor del coeficiente de Hurst más alto. Nótense diferentes pendientes en diversos lugares de las curvas, lo que implica la existencia de más de un coeficiente por secuencia.

Selección de parámetros

El coeficiente de Hurst es un parámetro necesario pero no suficiente para clasificar secuencias de ADN. Dos secuencias pueden

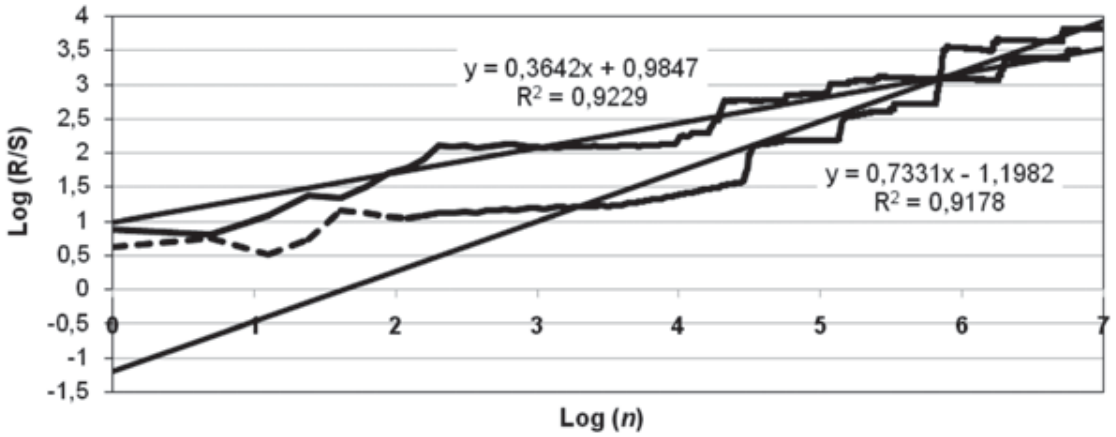

Figura 1. Diagrama de *Pox* del análisis R/S y estimación del coeficiente de Hurst (H) para dos secuencias de ADN.

tener el mismo coeficiente de escalamiento o de Hurst pero diferentes patrones de autosimilaridad. Por lo tanto, se analizaron estadísticamente las medias obtenidas durante el procedimiento matemático del análisis R/S para la estimación del Coeficiente de Hurst.

Estos parámetros R/S describen regularidades biológicas, denotan la complejidad de un conjunto de secuencias en términos de las tendencias generales o locales de las fluctuaciones de los nucleótidos que las constituyen y no de la posición de estos en las secuencias o de señales en las secuencias, cuantifican el comportamiento espacial de las secuencias y presentan características estadísticas robustas para derivar algoritmos de clasificación de secuencias entre y dentro de genomas, conceptualmente simples y computacionalmente eficientes.

Significado biológico de los parámetros R/S

Los parámetros Máximo (Máx.) y Mínimo (Mín.) representan medidas de la máxima y mínima variabilidad de nucleótidos alcanzada por un conjunto de secuencias en un genoma. El parámetro rango representa la diferencia entre la máxima y la mínima variabilidad estructural que puede soportar un conjunto de secuencias en un cromosoma o en un genoma. Nótese que el conjunto de genes y el de intrones del genoma de *M. musculus* presentan los rangos más altos de variabilidad 1.052,9 y 232,7 nucleótidos, respectivamente) y aparentemente estos rangos se reducen a medida que las especies descienden en la escala evolutiva, como así lo indican los valores bajos del parámetro rango obtenidos por los genomas vegetales para estas mismas secuencias (Cuadro 2). En contraste, el rango de variabilidad de los exones permanece constante a través de los genomas (el valor del parámetro rango para el conjunto de exones del genoma de *A. thaliana* fue de 38.6 mientras que para los genomas de *O. sativa* y *M. musculus* fue de 34.8 y 35.6, respectivamente.

Nótese que los valores de los parámetros Máximo y Mínimo son de igual longitud dentro de los conjuntos de secuencias de un mismo genoma (por ejemplo, en el conjunto de intrones del genoma de *O. sativa*, el Máximo óptimo es 22.6 y el Mínimo óptimo es -22.9 nucleótidos (Cuadro 2). Tres interpretaciones surgen al respecto: la primera, las variaciones de nucleótidos en las secuencias se autorregulan a través de algún patrón básico intrínseco de retroalimentación o proceso recursivo que incorpora autocorrección e indica cuánta variación soportan las secuencias de un determinado genoma. La segunda, las variaciones de determinadas posiciones de los nucleótidos en las secuencias están estrechamente correlacionadas o parecen influir en la aceptación de cambios en otras posiciones que podrían estar lejos en la secuencia, pero que cabe esperar estén próximas en la estructura tridimensional o de empaquetamiento del ADN (Gobel *et al*, 1994); y, la tercera, un aumento en la información genética de una secuencia siempre debe compensarse con una pérdida equivalente de información.

Los tres parámetros (Máximo, Mínimo y Rango) caracterizan la autoorganización o complejidad de las secuencias de un genoma, mientras que el parámetro intercepto b representa la constante característica de cada conjunto de secuencias en un genoma. La inclusión de la longitud de la secuencia como parámetro se interpreta como que el sistema posee algo o existe algo que es capaz de actuar a larga distancia. En general, valores altos de estos parámetros en secuencias de genes e intrones en el genoma animal contrastaron con valores bajos de estas secuencias en los genomas vegetales.

El hecho de que los parámetros R/S tengan una interpretación biológica puede considerárseles como variables biológicas dado que estos varían de manera específica entre conjuntos de secuencias y entre genomas. Por lo tanto, a partir de estos, podrían construirse criterios de mejoramiento vegetal para la selección de genes, parentales o poblaciones.

Finalmente, la evidencia presentada de que secuencias no codificantes como los intrones puedan ser descritas matemáticamente y analizadas biológicamente mediante los parámetros R/S indica que el procedimiento del análisis R/S potencialmente puede ser aplicado a las otras secuencias de ADN que constituyen la estructura de un gen, tales como regiones promotoras, reguladoras y terminadoras, lo que permite completar

generalizaciones respecto a la dinámica y funcionamiento de los genes en los genomas.

Caracterización de las secuencias

Una vez seleccionados los parámetros R/S, examinado su comportamiento global estadístico y explicado su significado biológico por conjunto de secuencias, por cromosoma y por genoma, se realizaron análisis locales de los parámetros por frecuencias de ocurrencia de las secuencias en los intervalos de clase de Hurst. Este análisis se hizo para perfeccionar la caracterización de los conjuntos de secuencias y optimizar las posibilidades de discriminación de los parámetros R/S.

El análisis mostró que los promedios de los parámetros que definen las secuencias contenidas en un determinado intervalo de clase de Hurst fueron estadísticamente diferentes de los promedios de los parámetros de los intervalos previos o contiguos dentro de un mismo genoma; por ejemplo, el promedio del parámetro rango de las 552 secuencias de genes del genoma de *A. thaliana* con valores del coeficiente de Hurst en el intervalo [0.2-0.3] es de 96.6±D.S difiere estadísticamente del promedio del parámetro rango del siguiente intervalo [0.3-0.4] que es de 105±D.S. con 3.558 secuencias de genes (Cuadro 3). A la

Cuadro 3. Distribución de las secuencias de ADN según su coeficiente de Hurst en intervalos de clase con los respectivos parámetros R/S, genoma de *A. thaliana*.

a) Genes

H	Frecuencia			Parámetros R/S				
Int. Clase	Absoluta	H	R^2	b	Máximo	Mínimo	Rango	Long. (nt)
[0.1-0.2]	12	0.18	0.57	2.41	25.92	-55.33	81.25	2.145.92
[0.2-0.3]	552	0.27	0.70	1.97	36.43	-60.15	96.58	2.351.61
[0.3-0.4]	3558	0.36	0.82	1.40	46.30	-58.68	104.98	2.140.15
[0.4-0.5]	8820	0.45	0.88	0.89	62.46	-60.45	122.91	2.147.25
[0.5-0.6]	10567	0.55	0.91	0.38	79.06	-64.21	143.27	2.107.84
[0.6-0.7]	6228	0.64	0.92	-0.12	99.29	-69.66	168.96	2.077.76
[0.7-0.8]	1993	0.74	0.93	-0.63	120.67	-79.90	200.56	2.101.42
[0.8-0.9]	308	0.83	0.92	-1.13	134.39	-97.21	231.59	1.991.61
[0.9-1.0]	28	0.93	0.92	-1.72	152.24	-141.41	293.65	2.189.27

a) Exones

H	Frecuencia			Parámetros R/S				
Int. Clase	Absoluta	H	R^2	b	Máximo	Mínimo	Rango	Long. (nt)
[0.1-0.2]	143	0.18	0.60	1.24	10.08	-6.04	16.12	222.48
[0.2-0.3]	3863	0.27	0.75	1.11	12.27	-10.08	22.35	282.39
[0.3-0.4]	21595	0.36	0.84	0.84	13.84	-13.42	27.27	300.52
[0.4-0.5]	44463	0.45	0.89	0.53	15.62	-17.36	32.98	323.16
[0.5-0.6]	44653	0.55	0.91	0.20	17.52	-22.67	40.19	343.45
[0.6-0.7]	23444	0.64	0.92	-0.13	19.72	-30.20	49.92	372.40
[0.7-0.8]	6810	0.74	0.92	-0.48	22.60	-41.06	63.66	424.91
[0.8-0.9]	1044	0.83	0.92	-0.85	26.66	-51.80	78.46	450.18
[0.9-1.0]	103	0.93	0.92	-1.21	25.86	-68.98	94.84	555.47

a) Intrones

H	Frecuencia			Parámetros R/S				
Int. Clase	Absoluta	H	R^2	b	Máximo	Mínimo	Rango	Long. (nt)
[0.1-0.2]	188	0.18	0.58	1.25	9.18	-4.85	14.03	142.11
[0.2-0.3]	4189	0.27	0.76	1.05	9.41	-8.25	17.66	144.67
[0.3-0.4]	21118	0.36	0.84	0.81	10.34	-11.47	21.81	156.91
[0.4-0.5]	39055	0.45	0.88	0.54	12.15	-14.01	26.16	164.96
[0.5-0.6]	32831	0.55	0.90	0.25	14.43	-16.61	31.04	171.57
[0.6-0.7]	13519	0.64	0.91	-0.05	16.24	-19.71	35.95	173.06
[0.7-0.8]	2850	0.73	0.91	-0.36	18.20	-23.19	41.39	175.90
[0.8-0.9]	287	0.83	0.90	-0.67	18.79	-29.26	48.05	178.58
[0.9-1.0]	15	0.94	0.87	-0.99	14.85	-27.23	42.09	115.14

Int. Clase: Intervalo de clase de Hurst (H); Frec. Absol.: Frecuencia absoluta; R^2: coeficiente de determinación; b: intercepto de la regresión log-log; Long. (n): longitud en nucleótidos.

misma conclusión se llega con los otros parámetros del conjunto de secuencias de genes, con las secuencias de exones e intrones del genoma de *A. thaliana* (Cuadro 3) y con los parámetros de los conjuntos de secuencias de los genomas de *O. sativa* y *M. musculus* (datos no mostrados).

La Figura 2 ilustra con contundencia en el plano cartesiano, la trayectoria específica y el espacio claramente diferenciado tomado por los parámetros según el intervalo de clase de Hurst y el conjunto de secuencias del genoma de *M. musculus*. De estos pocos parámetros se derivaron los algoritmos de clasificación de secuencias entre y dentro de genomas que alcanzaron altos promedios de sensibilidad y especificidad.

Cabe anotar que éste es el primer estudio que no se limita a asignarle un valor puntual de escalamiento o coeficiente de Hurst a las secuencias genómicas, muy por el contrario, explora con detalle el potencial del procedimiento matemático del análisis R/S e interpreta su significado biológico para obtener la metodología que se ha expuesto de los parámetros R/S acotados por los intervalos de clase de Hurst. En esencia, el análisis R/S en secuencias genómicas se hizo de manera similar a como lo planteó originalmente Hurst, quien no solamente lo desarrolló para demostrar que hay dependencia entre los eventos de un sistema sino que lo utilizó para encontrar el diseño óptimo de una represa para un sistema tan complejo como el del rio Nilo (Mandelbrot, 1997).

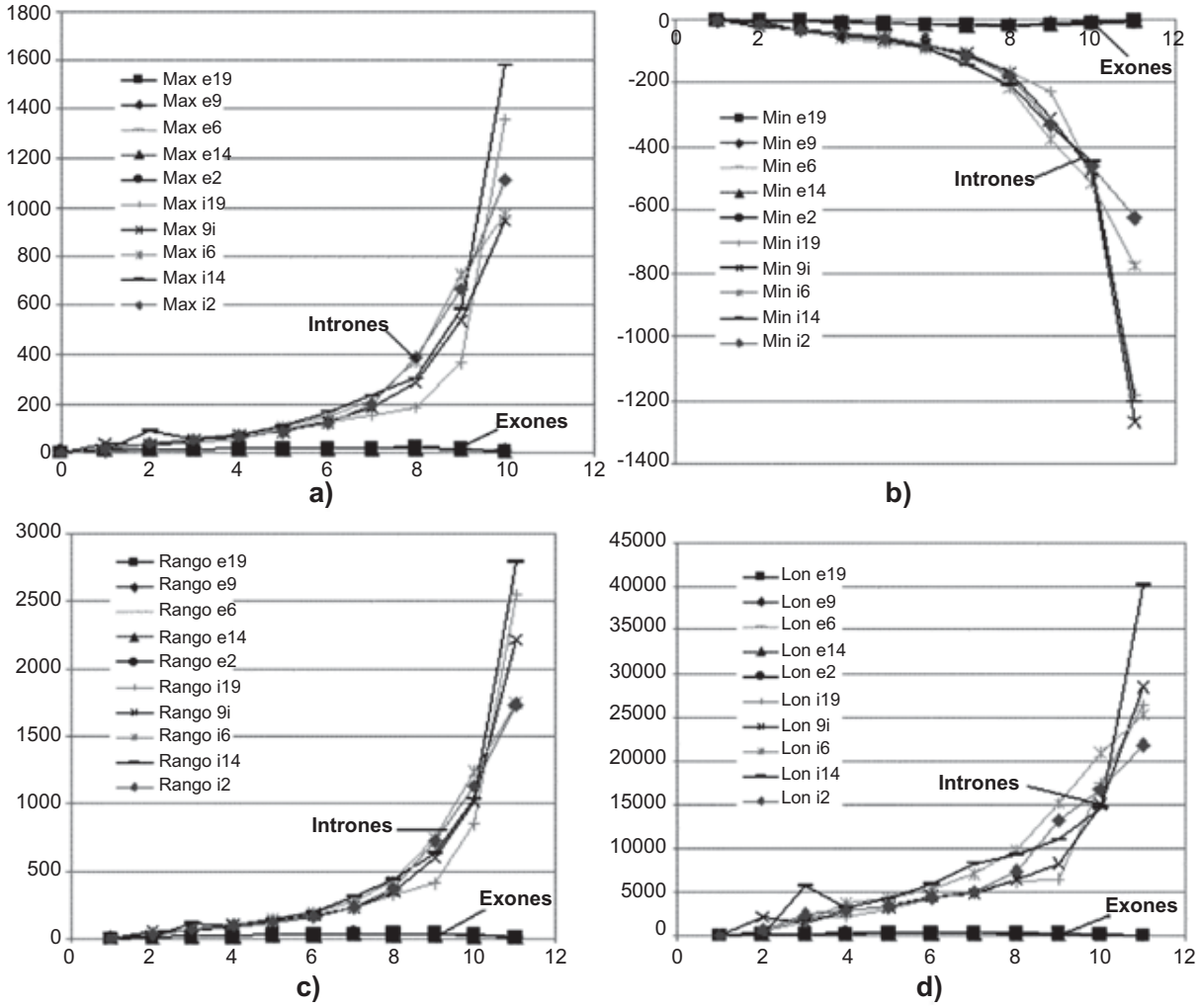

Figura 2. Comportamiento promedio de los parámetros del análisis R/S (eje Y): a) máximos, b) mínimos, c) rango, y d) longitud por intervalos del coeficiente de Hurst (Eje X) para cromosomas del genoma de *Mus musculus*.

Predicción de secuencias

Los análisis de predicción permiten verificar si los parámetros seleccionados corresponden a características estadísticas generales y robustas de las secuencias, los cromosomas y los genomas. A partir de los Cuadros 1 y 3 que indican valores promedio de los parámetros R/S±D.S. acotados por intervalos de clase del coeficiente de Hurst, por cromosoma y por genoma, respectivamente, según los conjuntos de secuencias, se elaboraron los algoritmos para la clasificación de secuencias entre y dentro de genomas.

En cada Cuadro se explota la posibilidad de predecir cómo una determinada secuencia puede encajar en uno de los intervalos del coeficiente de Hurst definido por los parámetros R/S, sin necesidad de que la secuencia tenga que tener ninguna homología o señal biológica con las secuencias incluidas en este intervalo, basándose exclusivamente en el comportamiento estadístico de los parámetros de dicha secuencia. Cada secuencia a clasificar debe contener la información de los parámetros: valor de H, máximo, mínimo, rango, intercepto b y longitud de la secuencia.

El algoritmo para la clasificación de genes, exones e intrones en genomas comprende dos grandes etapas: primera, estimar el valor del coeficiente de Hurst y de los parámetros derivados del análisis R/S (máximo, mínimo, rango, longitud de la secuencia y el intercepto de la ecuación) para cada secuencia de estudio; y segunda, construir los cuadros de clasificación mencionados por conjuntos de secuencias y por genoma.

El resultado general de sensibilidad de los predictores para clasificar secuencias de genes, exones e intrones dentro del mismo genoma fue del 81%, mientras que para clasificar secuencias de otros genomas fue mayor del 70% porque varía significativamente según el predictor y el genoma.

Resultados publicados en la literatura científica reportan porcentajes de sensibilidad similar o inferior a los de esta investigación pero, definitivamente, las medidas de especificidad son más bajas o no aparecen en las publicaciones. En conclusión, las medidas de desempeño de los predictores desarrollados en nuestro estudio son altas y robustas, como puede deducirse de los análisis de los artículos de Majoros et al. (2003) y Rojic *et al*. (2008). Majoros y colaboradores (2003) compararon el rendimiento de diferentes algoritmos para la predicción de genes y encontraron que el porcentaje de genes exactamente detectados no superaba el 50% y las medidas de desempeño para detectar exones tenían menos del 75%, tanto en especificidad como en sensibilidad. Los resultados de estos investigadores fueron confirmados por Rojic et al (2008) que demuestran, al analizar los mejores programas actuales de predicción de genes en mamíferos, que aunque se ha incrementado la exactitud en la predicción de la composición de nucleótidos en las secuencias (la especificidad y sensibilidad alcanzan 95% y 93%, respectivamente), realmente, las medidas de desempeño de estos programas para predecir secuencias exones no llega al 70%.

Finalmente, cuando se estableció un predictor único de exones a partir de todas las secuencias de exones de los tres genomas (datos no mostrados), el desempeño del predictor alcanzó una sensibilidad del 91% y de especificidad del 75% y 91% para intrones y genes, respectivamente.

Conclusiones

A pesar de las grandes diferencias en tamaño y en el número de cromosomas y secuencias por genoma, existen regularidades estadísticas básicas en la distribución de las secuencias genómicas que se manifiestan a través de los cromosomas de un mismo genoma y entre genomas, lo que corrobora que la complejidad estructural de un genoma no está relacionada con estas características.

El análisis R/S permitió demostrar que ambas distribuciones, la de los nucleótidos que constituyen las secuencias de los genes, exones e intrones y la de estas secuencias en los genomas están caracterizadas por el fenómeno de memoria o dependencia de largo alcance. La estructura de memoria varía según el tipo de secuencia y el genoma.

Las estructuras de memoria de las secuencias de los genes e intrones fueron específicas para el genoma de cada especie, mientras que la estructura de las secuencias de los exones fue estadísticamente similar en

los tres genomas. La estructura de memoria de los tres tipos de secuencias del genoma animal presentaron comportamiento persistente, mientras que las de los genes y de los exones de los genomas vegetales tuvieron comportamientos persistentes y la de los intrones, comportamiento antipersistente. Estos resultados fueron consistentes con la complejidad de las especies y con la función que realizan estas secuencias en el genoma. La persistencia está asociada a estructuras estables con alta probabilidad de cumplir funciones específicas, mientras que la antipersistencia se relaciona con estructuras inestables que buscan funcionalidad.

Los parámetros R/S permiten definir biológica y matemáticamente conjuntos de secuencias entre y dentro de genomas, lo que indica que el procedimiento del análisis R/S potencialmente podría utilizarse como una herramienta fenética de inferencia filogenética molecular, que complemente los métodos clásicos de comparación de secuencias o especies; o podría implementarse en los algoritmos de uso común de las herramientas bioinformáticas para mejorar la capacidad de predicción; o podría utilizarse para influir sustancialmente en las estadísticas de las puntuaciones del alineamiento de secuencias, o complementar los análisis de comparación de secuencias que sean difíciles de alinear.

Referencias

Balbín, A.; Andrade E. 2004. Protein Folding and Evolution are driven by the Maxwell Demon activity of Proteins. *Acta Bio theor*. 52 (3): 173-200.

Bernardi, G. 1995.The human genome: organization and evolutionary history. *Annu Rev. Genetics*. 29: 445-476.

Brown, D.D.; Wensink, P.C.; Jordan, E.A. 1972. A comparison of the ribosomal DNA´s of *Xenopus laevi* and *Xenopus mulleri*, the evolution of tandem genes. *J. Mol. Biol*. 63: 57-73.

Buldyrev, S.V., Golberger A.L., Havlin S., Mantegna R.N., Matsa M.E., Peng C.K., Simons M., y Stanley H.E. 1995. *Phys. Rev. E.* 51, 5084-5091.

Burset, M.,; Guigo. R. 1996. Evaluation of Gene Structure Prediction Programs. Genomics 34**,** 353–367.

Carothers, J.M.; Oestreich, S.C.; David, J.H.; Szostak, J.W. 2004. Informational complexity and functional activity of RNA structures. *J. Am. Chem*. Soc. 126: 5130-5137.

Craig, J.M.; Bickmore, W.A. 1993. Chromosome bands: flavors to savor. *BioEssays*. 15: 349-354.

Denton, M.J.; Dearden, P.K.; Sowerby, S.J. 2003. Physical law not natural selection as the major determinant of biological complexity in the subcellular realm: new support for the pre-Darwinian conception of evolution by natural law. *BioSystems* 71: 297-303

Deutsch, M.; Long, M. 1999. Intron–exon structures of eukaryotic model organisms. Nucleic Acids Research. Vol. 27(15): 3219-3228.

Doolittle, R.F. 1981. Similar amino acid sequences: chance or common ancestry? *Science*. 214: 149-159.

Gobel, U. et al. 1994. Correlated mutations and residue contacts in proteins. *Proteins: Struct. Funct. Genet*. 18: 309-317.

Goldberger, A et al. 2000. PhysioBank, physioToolkit, and physioNet: components of a new research resource for complex physiologic signals. *Circulation*. 101(23): e215-220.

Gu, Z.; Wang, H.; Nekulenko, A.; Li, W.L. 2000. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 Mb of genomic sequence. *Gene*. 259: 81-88.

Hao, B.L. 2000. fractals from genomes –exact solutions of a biology– inspired problem. *Physica* A 282, 225-246.

Hawkins, J.D. 1988. A survey on intron and exon lengths. *Nucleic Acids Re*s. 16, 9893-9906.

Hoop, B., Kazemi, H., Leibovitch, L. 1993. Rescaled range analysis of resting respiration. Chaos, 3(1), 27-29.

Hurst, H.E. 1956. Methods of using long-term storage in reservoirs. Part 1.Proc. Inst. Civ. Eng. Part I. 519 p.

Hurst, H.E. 1951. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civil Engineers*. 116: 770-808.

Karlin, S. and V. Brendel. 1993. Patchiness and correlations in DNA sequences. *Science*. 259: 677-680.

Karlin, S. and LR. Cardon. Computational DNA sequence analysis. *Annu Rev Microbiol* 1994. 48: 619-54.

Kolbe *et al*. 2004. Genetic variation increases during biological invasion by Cuban lizard. *Nature*. 431. 177-181.

Li, W., K. Kaneko. 1992. Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence. *Europhysics. Lett*. 17(7), 655-660.

Majoros, W.H.; Pertea, M.; Antonescu, C; Salzberg, S.; Glimmer, M. 2003. Exonomy and Unveil: three *ab initio* eukariotic genefinders. *Nucleic Acids Res*. 31: 3601-3604.

Mandelbrot, B. y Hudson, R.L. 2004. The (miss) behavoir of markets. A fractal view of risk, ruin and reward. Tusquets editors, S.A. 322p.

Mandelbrot, B., Van Ness, J.W. 1968. Fractional Brownian motions, fractional noises and applications, SIAM Review 10, 422-437.

Mandelbrot, B., Wallis J.R. 1969a. Some long-run properties of geophysical records. Water Resources Res. 5: 321-340.

Mandelbrot, B., Wallis J.R. 1969b. Robustness of the rescaled range R/S in the measurement of non-cyclic long-run statistical dependence. Water Resources Research, 5: 967-988.

Mandelbrot, B., Wallis J.R. 1969c. Computer experiments with fractional Gaussian noises. Water Resources Research, 5: 228-267.

Mandelbrot, B. 1982. Cambios de escala y leyes potenciales sin geometría. En: The Fractal Geometry of Nature. San Francisco: V.H. Freeman. P. 477-487.

Mattick, S.; John, Y.; Gagen, M. J. 2001. The evolution of a controlled multitasked gene Network: The role of introns and other Noncoding RNA in the development of complex organism. *Molecular Biology Evolution* 18(9): 1611-1630.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420: 520-562.

Ordov, YL; Boekhorst, R.; Abnizova II. 2006. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J Bioinform Comput Biol.* 4(2):523-36.

Peng, C.K., Buldyrev S., Golberg A.L., Havlin S., Sciortino, F., Simons, M., y Stanley H.E. 1992. Long-Range Correlations in Nucleotide Sequences. *Nature* 356, 168.

Peters, E. 1991. Chaos and order in the Fractal Market Analysis (Applying chaos theory to investment an economic). New York: John Wiley and Sons.

Rojic, S.; Mackworth, A.K.; Qullette, B.F. 2008. Evaluation of gene finding programs on mammalian sequences. Genome Res. 11: 817-832.

Schultes, E.; Hraber, P.; Labean, T. 1999. Estimating the contributions of selection and self-Organitation in ARN Secondary Structure. *J. Mol. Evol.* 49: 76-83.

Stanley, H.E.; Buldyrev, S.V.; Goldberger, A.L.; Goldberger, Z.D.; Havlin, S.; Mantegna, R.N.; Ossadnik, S.M.; Peng, C.K.; Simons, M. 1994. Statistical mechanics in biology: how ubiquitous are long-range correlations? *Physic*. A205: 214-253.

Xiao, Yi., Chen, R. Jian S. Jun X. 1995. Fractal dimension of Exon and Intron Sequences. *J. Theor. Biol.* 175, 23-26.

Yu, Z.G., Chen G.Y. 2000. Rescaled range and transition matrix analysis of DNA sequences. *Comm. Theor. Phys*. 33(4), 673-678.

Yu, Z.G., Anh V.V., Lau K.S. 2001. Multifractal characterisation of length sequences of coding and noncoding segments in a complete genome. *Physica* A. 301 (1-4), 351-361

Yu, Z.G., Wang B. 2001. Chaos, Solitons Fractals 12, 519.

Zhang, C.T.; Zhang, R. 1991. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, 19: 6313-6317.