Revista Española de Física **13** (2), 1999

Ramón Román Roldán

Complejidad de secuencias simbólicas. Aplicación a secuencias de ADN

Ramón Román Roldán, Pedro Bernaola Galván y José L. Oliver

ABSTRACT: This work examines a new method, based on Information Theory, to segment symbolic, complex, non-stationary sequences into homogeneous compositional domains. The method is mainly applied to DNA sequences, a complex natural system which is known to exhibit long-range fractal correlations. A new complexity measure, the *sequence compositional complexity*, is described, whose plot as a function of significance level provides a profile of sequence structure at different scales. Two outstanding features of DNA are uncovered: 1) non-coding sequences are more complex than the coding ones; and 2) a general agreement is found between the sequence and the biological complexities of the organism. The defined complexity has a broad range of potential applications; i.e., we use it here for the analysis of binary computer files.

EL ADN COMO SISTEMA FÍSICO

El ritmo vertiginoso con que se están secuenciando cadenas de ADN en los últimos años, junto con su disponibilidad para uso general en grandes bancos de datos, ha propiciado que el análisis de estas secuencias se haya convertido en el foco de atención de un pequeño, pero significativo, grupo de investigadores que publican sus resultados en las revistas científicas de máxima difusión. Y ello se debe también, naturalmente, a los resultados tan sorprendentes como alentadores, obtenidos con diversas técnicas estadísticas de análisis de secuencias que revelan la presencia en ellas de estructura compleja y correlaciones de largo alcance no triviales.

Desde que Avery (1944) demostró que el ADN es capaz de transformar otras células, y poco después Watson y Crick establecieron sus características como material genético (capaz de autoduplicarse y de contener la información para sintetizar las proteínas), el ADN pasó a tomar un papel central en el estudio de la Biología como soporte de la información *estable* esencial para la vida del individuo y para su transmisión hereditaria. Toda la información necesaria para el funcionamiento de la célula, así como su propia construcción se encuentra recogida en la cadena de ADN y por esto, no es de extrañar que haya suscitado el interés de muchas disciplinas, entre ellas la Física, que aborda su estudio desde dos puntos de vista: la Termodiná-mica de procesos irreversibles y la Teoría de la información, enfoques cuya unificación ha sido intentada por diversos autores. En palabras de Werner Ebeling y Mijail Volkenstein (1990) "...living beings are natural ordered and information-processing macroscopic systems originating from processes of self-organization and natural evolution... all proceses in living systems originate from physical processes. Living beeing are open thermodynamic systems wich permanently exchange matter, energy, entropy and information with their surrounding...". Muchos otros autores están de acuerdo con que los seres vivos están caracterizados principalmente por su capacidad para procesar información, y que, por tanto, pueden ser analizados desde esta perspectiva. El soporte físico de esta información es la doble hélice de ADN, que juega un papel fundamental tanto en la codificación como en la transmisión a la siguiente generación de toda la información necesaria para las funciones del ser vivo. Todo ello indica que el mantenimiento de la actividad vital (visto como un sistema de procesamiento de información) es de naturaleza física, además de afectar a otros fenómenos de largo alcance, como la evolución biológica o el origen de la vida.

Aquí nos situamos en un campo mucho más limitado. Las secuencias de nucleótidos se examinan desde un punto de vista externo, como un mensaje, sin tener en cuenta los detalles de los mecanismos fisico-químicos que intervienen en el procesamiento de la información. La síntesis de proteínas se modela como un sistema de procesamiento de la información, fuente más canal. Un objetivo básico es obtener medidas significativas y fiables de parámetros tales como orden, regularidad, estructura, complejidad, etc, de una secuencia de ADN dada. Ello permitiría realizar comparaciones cuantitativas con otras secuencias, o con otros segmentos de la misma secuencia, pudiendo deducir por tanto, resultados de interés para estudios de evolución (filogenia molecular), identificación de segmentos codificadores (encontrar genes, exones, señales de transcripción), etc. En general, la cuestión es encontrar medidas capaces de detectar estructuras estadísticamente significativas de la desviación de una secuencia dada con respecto a la esperada bajo la hipótesis de aleatoriedad y, en la medida de lo posible, asignar a estas estructuras significado biológico.

ANÁLISIS COMPOSICIONAL DE LAS SECUENCIAS DE ADN

Con pocas excepciones (Oliver et al. 1993), las aplicaciones de la Teoría de la Información al análisis de las secuencias de nucleótidos conducen a la conclusión de que éstas son indistinguibles de secuencias aleatorias (ver la revisión de Hariri et al. 1990). El descubrimiento reciente de la presencia de correlaciones de largo alcance y estructura fractal en las secuencias de ADN mediante análisis de fluctuaciones y técnicas espectrales (véase por ejemplo la revisión de Li y Kaneko (1994)), ha dado un vuelco al planteamiento del problema. Este tipo de correlaciones sugieren la presencia de una estructura autosemejante, con una composición que varía considerablemente de unas zonas a otras de la secuencia (heterogeneidad espacial), y ponen claramente de manifiesto el motivo por el cual los trabajos pioneros no dieron los resul-

tados esperados: las medidas clásicas de la Teoría de la Información, que en la mayoría de los casos están pensadas para el análisis de secuencias estacionarias, son irrelevantes para diferenciar las secuencias ADN, no estacionarias, de distintos organismos o con distinta función biológica. Es justamente este enfoque el que nos ha llevado a proponer una medida de la complejidad de las secuencias de ADN basada en la heterogeneidad composicional.

En el método que proponemos, la estructura particular de las secuencias ADN se pone de manifiesto segmentándolas mediante un algoritmo conceptualmente simple y computacionalmente eficiente, basado en la divergencia entrópica de Jensen-Shannon (Bernaola-Galván et al. 1996). Este algoritmo permite descomponer la secuencia en subsecuencias homogéneas, conocidas como *dominios composicionales*. Así se ha comprobado que no sólo existe una gran heterogeneidad espacial en la composición de las secuencias de ADN sino que también hay una compleja organización de las zonas con diferente composición que ya había sido intuida por algunos autores (Li y Kaneko 1994). En la segunda parte de este trabajo proponemos una medida de la complejidad composicional de una secuencia simbólica, basada también en la Teoría de la Información, que se aplica al resultado de la segmentación de la secuencia en dominios composicionales (Román-Roldán et al. 1998).

Aunque la motivación de este trabajo ha sido el análisis de secuencias ADN, los métodos aquí expuestos son de aplicación general a secuencias simbólicas no estacionarias. Como ejemplo de esta generalidad, se dan los resultados para algunas secuencias de ficheros de ordenador.

MÉTODO ENTRÓPICO DE SEGMENTACIÓN DE SECUENCIAS

El término *dominio composicional* es usado a veces en el análisis de secuencias simbólicas sin una definición precisa. Para nosotros, un dominio composicional (DC) es una subsecuencia que se caracteriza por tener una composición diferente de las subsecuencias adyacentes, con un cierto nivel dado de fiabilidad estadístico. Esta fiabilidad indica la frecuencia con la que dicha diferencia de composición (ponderada con los tamaños) no se habría obtenido en una realización aleatoria de la secuencia, repetida un gran número de veces.

La segmentación de secuencias en DC, en particular las de ADN, requiere una medida de diferencia entre las composiciones de las subsecuencias bajo ensayo. Nuestro método de segmentación de secuencias de ADN adopta la *divergencia de Jensen-Shannon* (JS) como medida de la diferencia entre las frecuencias relativas (consideradas como distribuciones de probabilidad) de los símbolos en los segmentos S_1 y S_2 de la secuencia $S = S_1 + S_2$, pesados con sus longitudes relativas:

$$JS_2(S_1, S_2) = \\ = H[S] - \left(\frac{l_1}{L} H[S_1] + \frac{l_2}{L} H[S_2] \right), \quad (1)$$

donde $H[\cdot] = -\Sigma\, p \log p$ es la entropía de Shannon del histograma correspondiente. Aunque JS_2 no es una verdadera medida de distancia probabilística, por no cumplir la propiedad triangular, ha resultado idónea para segmentar secuencias simbólicas no estacionarias gracias a sus propiedades interesantes (no negativa, simétrica, acotada, pesada, ramificable y extensible a un número cualquiera de distribuciones).

Para una secuencia dada existen muchas segmentaciones en DC, es decir, muchas particiones que cumplen con la propiedad de que todos los pares de subsecuencias adyacentes tienen una divergencia que cumple el requisito impuesto de fiabilidad estadística. Entre todas las que cumplen esta condición, definimos como segmentación óptima la que maximiza la función

$$JS_n = H[S] - \sum_{i=1}^{n} \frac{l_i}{L} H[S_i] = \\ = \sum_{i=1}^{n} \frac{l_i}{L} (H[S] - H[S_i]) \quad (2)$$

que es una generalización de JS_2, aplicada a los n dominios resultantes de la segmentación.

El problema de encontrar la segmentación óptima es de máxima complejidad, por lo que seguimos un método heurístico de *splitting* (Bernaola-Galván et al. 1996) que, en esencia, opera como sigue: Inicialmente se recorre la secuencia entera, se localiza la posición de máxima divergencia JS_2 y, si ésta satisface la condición de fiabilidad, la secuencia se corta en dos segmentos que son DC. El procedimiento se reitera con los DC sucesivamente obtenidos hasta que no es posible un nuevo corte con la fiabilidad especificada. Antes de decidir cada nuevo corte, se comprueba que éste y también los dos cortes contiguos una vez modificados por éste, satisfacen la condición de fiabilidad. El procedimiento es global en el sentido de que se examina la secuencia (o subsecuencia) completa, no una porción delimitada por ventana, para decidir cada corte. Ello se justifica por el objetivo de la segmentación: se busca estructura (correlaciones) de largo alcance que, de estar presentes, implican la inexistencia de una longitud o escala característica.

Al aplicar el método a secuencias reales de ADN se ha encontrado una mayor heterogeneidad (mayor número de dominios) en las secuencias que, según otros autores, presentan correlaciones de largo alcance de tipo fractal (Bernaola-Galván 1997). Además se han obtenido distribuciones de tamaño de los dominios claramente diferentes en unos casos y otros. De hecho, en las secuencias con propiedades fractales se obtienen distribuciones que siguen una ley logarítmica normal, que podría corresponder a una ley de potencia para el caso de una secuencia finita y discreta.

Otro resultado muy ilustrativo, y que pone claramente de manifiesto la diferencia en la organización de los dominios en las secuencias con y sin propiedades fractales es lo que se obtiene al segmentar las secuencias de forma "recursiva", esto es: se segmenta una secuencia a un nivel de significación, y a continuación se vuelve a segmentar el mayor de los dominios obtenidos, ahora a una significación inferior. En las figuras 1 y 2 se muestran los resultados de este proceso con las dos secuencias utilizadas con mayor frecuencia en la bibliografía como ejemplos de presencia de estructura fractal (HUMTCRAD, secuencia humana) y ausencia de ella (ECO110K, bacteria). A simple vista se aprecian claras diferencias entre ambas: aunque las dos secuencias son heterogéneas, en la bacteriana pronto aparecen dominios homogéneos y relativamente grandes, que no siguen subdividiéndose. Sin embargo, en la humana siguen apareciendo un número considerable de nuevos dominios dentro de los dominios obtenidos a niveles superiores, cubriendo todo el rango de posibles longitudes; es más, la forma en que se subdividen estos dominios es similar a como lo hace la secuencia completa. De hecho la figura 1 recuerda el aspecto de los conjuntos fractales en una dimensión con autosemejanza estadística, en los que al ampliar una zona se obtiene una imagen muy parecida al conjunto global

(aunque no exactamente, sí en términos estadísticos). Una cuestión interesante que hay que notar es que, en este caso, la ampliación se lleva a cabo mediante algo que podríamos denominar "zoom estadístico", esto es, al reducir el nivel de significación, el algoritmo detecta detalles menos relevantes que, por lo general, suelen ser de menor tamaño.

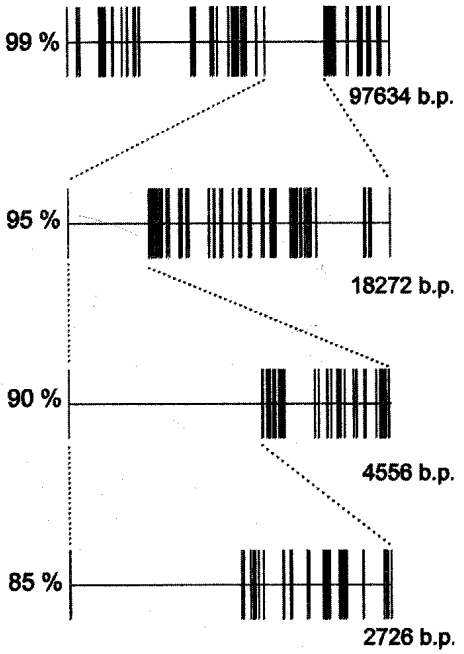

Figura 1. Segmentación recursiva de HUMTCRAD para los niveles de significación del 99, 95, 85 y 80%. Las líneas verticales indican la posición de los cortes en la secuencia. En todos los casos el dominio de mayor tamaño obtenido a un nivel de significación se vuelve a segmentar al nivel siguiente. (b.p. = pares de bases).

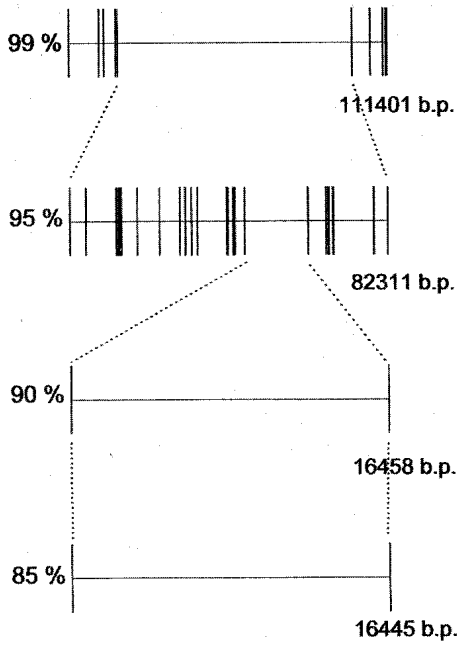

Figura 2. Segmentación recursiva de ECO110K para los niveles de significación del 99, 95, 85 y 80%. Las líneas verticales indican la posición de los cortes en la secuencia. En todos los casos el dominio de mayor tamaño obtenido a un nivel de significación se vuelve a segmentar al nivel siguiente. (b.p. = pares de bases).

COMPLEJIDAD COMPOSICIONAL DE UNA SECUENCIA

La aplicación creciente de métodos científicos a sistemas y estructuras complejas ha planteado el problema de la determinación cuantitativa de la complejidad. Para ello se han propuesto diversas medidas dependientes del carácter del sistema (ser vivo, algoritmo, nicho ecológico, autómata, ...), de los factores que determinan su complejidad (estructura, función, composición,...) y de la finalidad perseguida. Por ejemplo, C. Benneth (1990) destaca la "... necesidad de formular definiciones adecuadas de complejidad: definiciones que, por un lado recojan la noción intuitiva de complejidad, y por otro sean suficientemente objetivas y matemáticas para probar teoremas acerca de las mismas". A pesar de las numerosas propuestas, el caso es que no existe acuerdo acerca de una definición general de complejidad y, por tanto, de una medida de la misma. Nuestro interés se centra en disponer de una medida de complejidad composicional aplicable a una secuencia simbólica y que sea indicativa de intervención, de relación, de evolución, etc. No vale una simple medida de aleatoriedad, tal como la complejidad algorítmica (Chaitin 1987), que tendría su máximo valor para una secuencia enteramente aleatoria. Tampoco es útil la "complejidad efectiva" o la "información total", propuestas por (Gel-Mann y Lloyd 1996), que derivan de la complejidad algorítmica, y que tienen en cuenta separadamente la descripción compacta de las regularidades identificadas y la aleatoriedad residual de una entidad. Para una revisión de este tema, ver el artículo de Benneth (1990) y las referencias allí contenidas.

La misma medida de divergencia probabilística utilizada por partida doble (como criterio de decisión para escindir un segmento en dos dominios, y también como criterio de óptimo de la segmentación), resulta ser una buena medida de complejidad composicional de la secuencia (Román-Roldán et al. 1998). En efecto, los criterios de complejidad antes expuestos se satisfacen bien gracias a las propiedades interesantes de la divergencia, tales como:

- Cuantifica la secuencia en función del número y sesgo de sus dominios, de acuerdo con la Eq. (2).
- Es independiente de la longitud total de la secuencia, por lo que sólo depende de la configuración interna. Por ejemplo, una secuencia concatenada con una copia de ella misma tiene su misma divergencia global.
- Es aplicable a secuencias no estacionarias, propiedad que no tienen algunas medidas derivadas de la teoría de la señal, y que son indebidamente utilizadas en el análisis de secuencias de ADN.
- La propiedad de ramificación de la divergencia (Bernaola-Galván 1997) permite desarrollarla en una suma ponderada, cuyos sumandos corresponden a subsecuencias más cortas o a alfabetos restringidos (el alfabeto binario es especialmente interesante, ya que ha sido el único abordable con los métodos anteriores).

En el método heurístico descrito antes, la divergencia total de la segmentación es inferior a la máxima teórica de la segmentación óptima, aunque los resultados experimentales muestran que las desviaciones son pequeñas. Llamaremos *complejidad composicional de la secuencia* (CCS) a la divergencia total de la segmentación resultante del procedimiento heurístico.

PERFILES DE COMPLEJIDAD COMPOSICIONAL

Puesto que el resultado de un proceso de segmentación es dependiente de la fiabilidad estadística impuesta (s), la complejidad composicional de la secuencia será también función de ésta; en lo sucesivo indicaremos esta dependencia con la notación: CCS(s). Esta función se puede representar gráficamente segmentando la secuencia para un conjunto seleccionado de valores de s, obteniéndose así el *perfil de complejidad* de la secuencia. Se demuestra (Román-Roldán et al. 1998) que la complejidad aumenta monótonamente cuando la fiabilidad disminuye, es decir, los perfiles de complejidad son decrecientes. A escala fina en el eje s, los perfiles aparecen escalonados irregularmente, con un escalón por cada corte dado a la secuencia cuya altura mide la complejidad añadida por dicho corte. A escala gruesa, y para una secuencia larga con muchos dominios, el perfil se presenta como una línea de aspecto más bien continuo, pero con tramos. La pendiente de un tramo refleja el ritmo de aumento de la divergencia (y de producción de cortes) con la disminución de s.

El ejemplo que sigue ilustra este comportamiento y ayuda a interpretar el perfil experimental de una secuencia. La figura 3, representa el llamado *camino aleatorio* (del inglés *random walk*)de una secuencia binaria sintética, construida

concatenando alternativamente tramos de longitud 10 con sesgos composicionales complementarios (es decir, si las frecuencias relativas de un tramo son $(a, 1-a)$, las de los tramos adyacentes son $(1-a, a)$. Además, estas frecuencias relativas están *moduladas* por otro par de frecuencias complementarias por tramos alternativos más amplios, lo que explica la forma de la gráfica del camino aleatorio ($W(x)$) de un diente de sierra fino sobreimpreso en otro más grueso. El perfil de complejidad de esta secuencia (Figura 4) contiene un pequeño escalón para $s = 92\%$ que corresponde a la segmentación en grandes tramos, sigue una amplia meseta significando que no hay cortes hasta el valor $s = 77$ %, en el que aparece un gran escalón correspondiente a la fractura simultánea de la secuencia en todos los tramos de 10 símbolos.

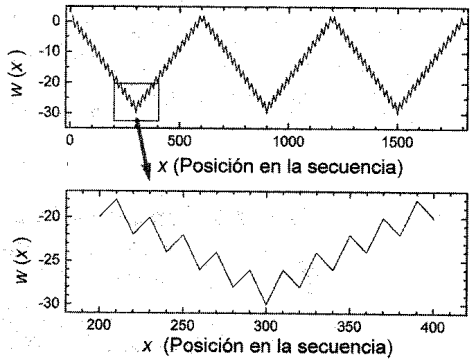

Figura 3. Camino aleatorio ($W(x)$) de una secuencia con tramos de longitud 10 organizados en dominios de mayor tamaño.

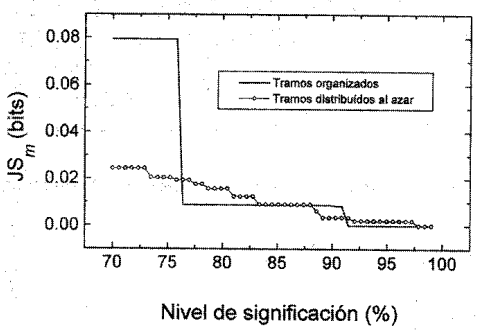

Figura 4. Perfil de complejidad de una secuencia con tramos de longitud 10 organizados en dominios de mayor tamaño y de la secuencia en la que los mismos tramos se han dispuesto al azar.

En general, la pendiente del perfil muestra, cambiada de signo, la velocidad de aumento de la divergencia con la disminución de la fiabilidad, lo que da idea del ritmo de producción de cortes.

Es interesante considerar el perfil correspondiente a una secuencia aleatoria muy larga. Los experimentos realizados (Bernaola-Galván 1997) muestran que la divergencia comienza a tener valores apreciables para $s < 80$, por lo que el intervalo de interés de Δs está restringido al (80,100).

PERFILES DE COMPLEJIDAD DE SECUENCIAS DE ADN

Hasta cierto punto, la conveniencia de la definición de una medida de complejidad viene justificada a posteriori por su utilidad. En esta sección vamos a ver cómo nuestros perfiles de complejidad permiten diferenciar entre las secuencias de ADN de diversos organismos, asignando niveles superiores de complejidad a secuencias que pertenecen a organismos, a priori, más complejos.

En la figura 5 vemos los perfiles de complejidad de secuencias de ADN de tres organismos bastante representativos: El hombre, la levadura de la cerveza (organismo eucariota) y la bacteria *Escherichia coli*. El perfil correspondiente a las secuencias humanas se ha obtenido promediando los perfiles de todas las secuencias disponibles con longitud superior a 100 kbases (alrededor de 20 Mbases), para la levadura de la cerveza se ha representado el promedio de los 16 cromosomas (un total de 12 Mbases), y para la bacteria se ha representado el perfil de su genoma completo (4.5 Mbases). Para las secuencias humanas y de la levadura, la zona sombreada representa la desviación estándar.

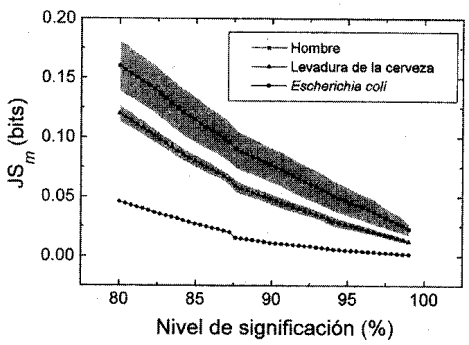

Figura 5. Perfiles de complejidad de secuencias humanas, de la levadura de la cerveza y de *Escherichia coli*.

En la figura 6 vemos otro ejemplo. En este caso se han tomado un conjunto de secuencias homólogas, esto es, secuencias que contienen la zona codificadora de un mismo gen para diversos organismos. De nuevo aparece un acuerdo bastante razonable entre complejidad composicional y complejidad del organismo. Tal vez cabe destacar como excepción el caso de la *Drosophila* (mosca del vinagre) que aparece con un valor de complejidad inferior al que cabría esperar.

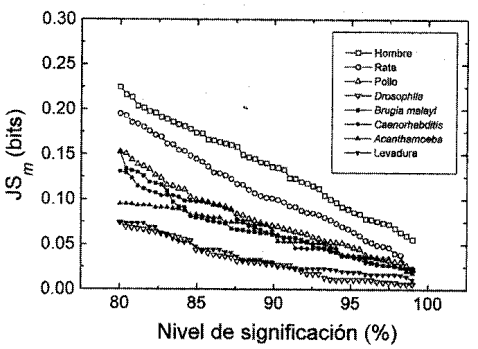

Figura 6. Pefiles de complejidad para las secuencias que contienen el gen de la cadena pesada de la miosina de diversos organismos.

PERFILES DE COMPLEJIDAD DE FICHEROS DE ORDENADOR

La medida de complejidad es aplicable a cualquier tipo de secuencias simbólicas. Para demostrar esta generalidad mostramos en la figura 7, los perfiles de complejidad correspondientes a ficheros de ordenador con distintos formatos, todos ellos convertidos previamente a secuencias binarias: programas ejecutables, archivos de sonido, archivos de Word (formato documento) y archivos de texto (ASCII).

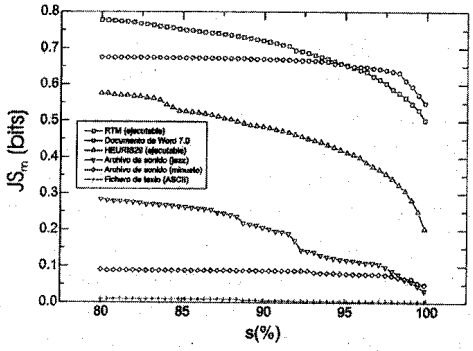

Figure 7. Perfiles de complejidad de ficheros de ordenador.

En primer lugar caben destacar los grandes valores que se obtienen para significaciones muy altas en el caso de los programas ejecutables y el fichero de Word que podrían deberse al frecuente uso que se hace en estos ficheros de largas cadenas de símbolos repetidos: se obtendrían dominios con una gran significación a causa de su tamaño. En el otro extremo tenemos el fichero en formato ASCII (que corresponde al mismo texto en formato Word). En este caso los valores de complejidad son muy bajos ya que, aunque no se usan todas las posibles combinaciones de octetos, los que sí se utilizan no suelen ir agrupados en largas series de repeticiones como ocurre en el caso anterior. Entre estos dos extremos tenemos los dos ficheros de música, con perfiles similares cuantitativamente a los de secuencias de ADN con propiedades fractales. Por otra parte, estos resultados concuerdan con la apreciación de algunos autores (Voss y Clarke 1975) en el

sentido de que las señales musicales presentan propiedades fractales, en concreto lo que se conoce como espectro de potencia de tipo $1/f$.

CONCLUSIONES

Se ha presentado un método nuevo de análisis de secuencias simbólicas, no estacionarias, basado en la segmentación de la secuencia en dominios composicionales. Se utiliza una medida entrópica de divergencia probabilística, la divergencia de Jensen-Shannon, que puede ser ajustada para operar con un nivel preestablecido de fiabilidad estadística. Se propone una definición de la complejidad composicional de una secuencia y se obtiene su perfil de complejidad, como una función de la fiabilidad. De las definiciones de complejidad propuestas hasta el momento, ésta es la única que presenta de forma explícita la complejidad como una función del nivel de detalle con el que se observa la secuencia.

El método se aplica principalmente al análisis de secuencias de ADN, encontrándose una autocorrelación de largo alcance tipo fractal con una distribución de longitudes de los dominios composicionales en ley de potencias. Se muestran algunos perfiles de complejidad con resultados de interés, tales como: 1) La presencia de una mayor heterogeneidad composicional en las secuencias de ADN no codificadoras que en las codificadoras y la relación de esta heterogeneidad con la presencia de estructura fractal en las primeras. 2) La concordancia entre los niveles de complejidad composicional y complejidad biológica. Este resultado indica la relevancia de la gran heterogeneidad composicional observada en las secuencias de organismos superiores y abre el camino a futuros estudios de evolución molecular y de genómica comparada.

AGRADECIMIENTOS

Quisiéramos agradecer a los profesores Dr. J. Aguiar y Dr. P. Carpena la revisión y lectura crítica del manuscrito.

BIBLIOGRAFÍA

[1] Bernaola-Galván, P., Román-Roldán, R. y Oliver J.L. (1996): "Compositional segmentation and long-range fractal correlations in DNA sequences." *Physical Rewiew E*, vol **53**(5), pp. 5181-5189.

[2] Bernaola-Galván, P. (1997): "Complejidad composicional en secuencias de ADN." Tesis Doctoral, Universidad de Granada.

[3] Benneth, C.H. (1990): "How to define complexity in Physics and why." pp. 137-148, en *Complexity, Entropy and the Physics of Information* Ed. W.H. Zurek, Addison-Wesley Press.

[4] Chaitin, G.J. (1987): *Algorithmic Information Theory*, Cambridge University Press.

[5] Ebeling, W. y Volkenstein, M.V. (1990): "Entropy and the evolution of biological information". *Physica A*, vol. **163**, pp. 398-402.

[6] Gell-Mann, M. and Lloyd, S. (1996): "Information Measures, Effective Complexity and Total Information. *Complexity* (1996), vol. **2**. pp. 44-52.

[7] Hariri, A., Weber, B. and Olmsted I. J. (1990): "On the validity of Shannon-information calculations for molecular biológical sequences". *Journal of Theoretical Biology*, vol. **147**, pp. 235-254.

[8] Li, W., Marr, T.G. and Kaneko, K. (1994): "Understanding Long-range Correlations in DNA Sequences". *Physica D*, vol. **75**, pp. 392-416.

[9] Oliver, J.L., Bernaola-Galván, P., Guerrero-García, J. and Román-Roldán, R. (1993): "Entropic Profiles of DNA Sequences Through Chaos-game-derived Images". *Journal of Theoretical Biology*, vol. **160**, pp. 457-470.

[10] Román Roldán, R, Bernaola Galván, P., Oliver J.L. (1998): "DNA Sequence compositional complexity through an entropic segmentation method". *Physical Review Letters* vol. **80**(6), pp. 1344-1347.

[11] Voss, R.F. and Clarke, J. (1975): "1/f noise in music and speech". *Nature*, vol. **258**, pp. 317-318.

Ramón Román Roldán
está en el Departamento de Física Aplicada de la Universidad de Granada,
Pedro Bernaola Galván
está en el Departamento de Física Aplicada II de la Universidad de Málaga y
José L. Oliver
está en el Departamento de Genética e Instituto de Biotecnología de la Universidad de Granada