Universidad de Granada
Departamento de Física Aplicada

Complejidad composicional en Secuencias de ADN

TESIS DOCTORAL
por
Pedro A. Bernaola Galván

Memoria presentada para aspirar
al grado de Doctor en Ciencias Físicas

Codirigida por los profesores:

Dr. D. Ramón Román Roldán
Catedrático de Universidad
Departamento de Física Aplicada
Universidad de Granada

Dr. D. José Luis Oliver Jiménez
Profesor Titular de Universidad
Departamento de Genética
Universidad de Granada

Granada septiembre de 1997

A Gloria

A mis padres

Agradecimientos

Deseo expresar mi más sincero agradecimiento a todos aquellos que, de una forma u otra han contribuido a la realización tanto de esta memoria como del trabajo de investigación que en ella se recoge.

En especial a mis directores, D. Ramón Román Roldán y D. José Luis Oliver Jiménez que han sabido en todo momento encaminar adecuadamente mis esfuerzos. Su ayuda y paciencia han sido inestimables.

También quiero agradecer el apoyo a los miembros del Grupo de Procesamiento de Imágenes, Aureliano, Chakir, Lina, Kiko, Mª José, Pedro, Pepe y Vicente. Muchas de sus sugerencias están reflejadas en esta memoria.

Asimismo, no puedo dejar de mencionar a mis compañeros del Departamento de Física Aplicada II de la Universidad de Málaga, en especial a Cristóbal Carnero que, con su fino humor y sabios consejos me ha animado en todo momento a la realización de esta memoria y a Pedro Carpena por las fructíferas discusiones científicas que ha mantenido conmigo.

Por último, quisiera agradecer a mi familia el ánimo que me han dado con su ilusión por ver finalizada esta memoria, y en especial a Gloria que ha sabido prescindir de mi durante muchas horas.

Fábula Gratulatoria

Estaba un conejo escribiendo en su portátil y, tan absorto se encontraba con su trabajo, que no se percató de que un lobo se le acercaba sigilosamente entre los matorrales. Cuando llegó hasta él, soltó un fuerte gruñido enseñando los dientes, a lo que el conejo se dio la vuelta y, con gesto indiferente, le contestó:

- Hola ¿qué te trae por aquí?

El lobo, bastante desconcertado, fijó la vista en la pantalla del ordenador y preguntó:

-¿Qué escribes?

-Mi tesis doctoral - contestó el conejo.

-Y, ¿de qué trata?

-De cómo los conejos comen lobos.

El lobo empezó a reir soltando grandes carcajadas revolcándose por el suelo. En estas, pasaba por allí un zorro que, al escuchar las carcajadas del lobo se acercó a ver qué ocurría.

-¿Qué te hace tanta gracia amigo lobo?

El lobo, entre carcajada y carcajada, secándose las lágrimas, contestó:

-Éste, que dice que está escribiendo una tesis sobre cómo los conejos se comen a los lobos.

-De cómo los conejos comen lobos y zorros -apostilló el conejo-. Creo que he decidido cambiar el título.

El zorro también empezó a reir, a lo que el conejo, algo molesto, respondió:

-Si no os lo creéis, acompañadme dentro de la cueva.

El conejo se encaminó con diligencia hacia la cueva seguido por el lobo y el zorro que iban dándose codazos e intentando, sin mucho éxito, aguantar la risa.

A los pocos minutos salió el conejo llevando, en una mano, el cráneo del lobo y en la otra el del zorro, seguido de un león y un tigre.

Moraleja: Sea cual sea el tema de tu tesis, elige bien a tus directores.

Índice

Capítulo 1

Introducción

1.1 Antecedentes y planteamiento

Desde 1944, año en que Avery demostró que el ADN era capaz de transformar otras células, y luego con el modelo de Watson y Crick, que justificaba 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 Termodinámica de procesos irreversibles y la Teoría de la información. En palabras de Werner Ebeling y Mijail Volkenstein [Eb 90] "...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 habilidad 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.

El párrafo anterior sugiere que el mantenimiento de la actividad vital (visto como un sistema de procesamiento de información), tiene, por una parte naturaleza física y por otra, afecta a otros fenómenos de largo alcance, como la evolución biológica o el origen de la vida. Como hemos dicho, estos problemas se abordan desde la doble perspectiva de la Termodinámica y la Teoría de la Información. Muchos autores han intentado unificar estos dos tratamientos para resolver el problema. [Bri 62] [Wi 87] [Br 88] [We 88] [Cav 92].

Aquí nos centraremos en un campo mucho más limitado. Las secuencias de nucleótidos serán examinadas desde un punto de vista externo, como un mensaje, sin tener en cuenta los detalles de los mecanismos físico-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. Una cuestión básica es obtener medidas significativas y fiables de parámetros tales como orden, regularidad, estructura, complejidad, etc., en una secuencia de ADN dada. Esto permitirá la comparación con otras secuencias, o con otros segmentos de la misma secuencia, pudiendo derivar por tanto, resultados de interés para estudios de evolución (filogenia molecular), identificación de segmentos codificadores (encontrar genes, exones,

señales de transcripción), etc. En general, el objetivo, es encontrar medidas capaces de indicar cómo de lejos está una secuencia natural de una aleatoria.

El descubrimiento reciente de la presencia de correlaciones de largo alcance y estructura fractal en las secuencias de ADN [Li 92c], [Vo 92], [Pe 92], en cierto sentido, ha dado un vuelco al planteamiento del problema. Como consecuencia de estos estudios se ha puesto de manifiesto la importancia de las heterogeneidades presentes en la composición de las secuencias primarias de nucleótidos; no parece que las medidas "clásicas" de la Teoría de la Información, que en la mayoría de los casos están pensadas para el análisis de secuencias estacionarias, sean relevantes para diferenciar entre las secuencias de distintos organismos o entre secuencias con distinta función biológica.

Justamente, este nuevo enfoque es el que nos llevó a proponer una medida de la complejidad de las secuencias de ADN basado en la heterogeneidad composicional; lo que denominaremos "Complejidad Composicional".

1.2 Descripción de los contenidos

El capítulo 2 contiene una revisión de las técnicas estadísticas y de Teoría de la Información más relevantes utilizadas hasta el momento para cuantificar la complejidad de las secuencias de ADN. Se ha hecho una exposición más o menos cronológica, comenzando por los primeros intentos basados en la Teoría de la Información que, dicho sea de paso, no tuvieron gran éxito. Parte de esta revisión está recogida en un artículo publicado por nuestro grupo [Ro 96]. A continuación se presentan los estudios más recientes basados, en su mayor parte, en técnicas derivadas de la teoría de la señal y que han puesto de manifiesto la presencia de estructura fractal en estas secuencias. Los métodos que aquí se desarrollan no son el soporte teórico de los capítulos posteriores sino que se introducen para fundamentar las referencias a los

resultados obtenidos por otros autores y así, poder compararlos con los que aquí obtendremos.

También se ha incluido una sección con algunos de los modelos propuestos para simular las propiedades estadísticas de las secuencias de ADN. No se ha pretendido hacer un desarrollo riguroso de todas las propuestas encontradas en la bibliografía, sino un resumen de los más relevantes en relación con los temas que se tratarán en capítulos posteriores y las cuestiones que actualmente se debaten sobre el tema. Para terminar el capítulo se de comenta brevemente el estado actual de la cuestión.

En el capítulo 3 se describen los primeros resultados obtenidos por nuestro grupo en el análisis estadístico de secuencias de ADN con histogramas de segundo orden [Be 94], [Ol 93] que, posteriormente serían generalizadas al análisis de otro tipo de secuencias [Ro 93b], [Ro 94]. El capítulo se dedica principalmente a la descripción de los métodos desarrollados ya que las conclusiones obtenidas sobre la complejidad de las secuencias de ADN no son tan relevantes como las que presentaremos en capítulos posteriores. Esencialmente, con estos métodos se comprueba la importancia de la heterogeneidad composicional presente en estas secuencias. Téngase en cuenta que cuando fueron presentados estos trabajos, gran número de autores sostenían que los métodos basados en Teoría de la Información no podían distinguir las secuencias de ADN de secuencias aleatorias. Además, los resultados aquí obtenidos ponen de manifiesto la necesidad de identificar zonas de composición homogénea dentro de estas secuencias, lo que se llevará a cabo en los capítulos sigu-ientes.

En el capítulo 4 se presenta el algoritmo de segmentación desarrollado por nuestro grupo para localizar zonas con composición homogénea en secuencias de ADN, aplicable, en principio, a cualquier tipo de secuencia simbólica. [Be 96]. Se

describen tanto el método, que es original, como la medida utilizada para la segmentación (divergencia de Jensen-Shannon), incluyendo las principales propiedades de esta medida, algunas de las cuales ya eran conocidas y otras suponen una contribución original.

Los resultados obtenidos y las conclusiones que se desprenden de la aplicación a las secuencias de ADN de este método de segmentación se recogen en el capítulo 5. Se presta especial atención a las cuestiones que en la actualidad centran el interés de los estudios composicionales del ADN, comparando los resultados con los obtenidos por otros métodos de análisis.

También se dedica un epígrafe al estudio comparativo de los 16 cromosomas de la levadura de la cerveza (*Saccharomyces cerevisiæ)*, primer genoma de un organismo eucariota que ha sido secuenciado por completo. Este estudio comparativo es, en la actualidad, es una de nuestras líneas de trabajo, en colaboración con otros grupos [Li 97c].

En el capítulo 6 se propone una medida de la complejidad composicional aplicable a secuencias simbólicas, basada en el algoritmo de segmentación [Ro 97]. Se compara con otras medidas y definiciones genéricas de complejidad existentes en la bibliografía, prestando especial interés a las últimas definiciones propuestas para cuantificar la complejidad de los sistemas biológicos. También se presentan algunos ejemplos con secuencias artificiales para discutir las propiedades de esta medida.

En el capítulo 7 se aplica la medida de la complejidad composicional a secuencias de ADN. Al igual que en el capítulo 5, se presta especial atención a la comparación de los resultados obtenidos con los presentados por otros autores con técnicas diferentes y se aplica la medida a secuencias generadas con los modelos descritos en el capítulo 2. Con esto último se comprueba hasta qué punto dan cuenta estos modelos de la complejidad composicional, tal y como la definimos aquí.

Finalmente, en el capítulo 8 se enumeran las principales resultados aportados en esta memoria y un comentario sobre algunos problemas abiertos y futuras líneas de trabajo.

Capítulo 2

Revisión Bibliográfica

2.1 Introducción

En esta capítulo se hará, en primer lugar, una revisión [Ro 93a], [Ro 96] de los antecedentes de la aplicación de la Teoría de la Información a las secuencias de ADN junto con algunas aportaciones recientes que pueden englobarse dentro de este campo (sección 2.2). En la sección 2.3 veremos los métodos de análisis, basados en la Teoría de la Señal y la Geometría Fractal, que recientemente se están aplicando al análisis de secuencias de ADN y que, en la actualidad, suponen uno de los campos de mayor interés en el análisis estadístico de estas secuencias. A continuación (sección 2.4) haremos una revisión de los modelos de secuencias de ADN más relevantes que han sido propuestos en la bibliografía y finalmente (sección 2.5) un breve resumen del estado actual de la cuestión.

2.2 Aplicación de la Teoría de la Información a Secuencias de ADN

La aplicación de la Teoría de la Información a las secuencias de ADN comienza en los años 70. Se pueden distinguir dos períodos, el primero alrededor de 1970-77, en el que aparecen las primeras publicaciones. Varios autores ([Ga 72], [Re 73], [Gui 77]) desarrollaron métodos para estimar parámetros tales como información, redundancia o divergencia en secuencias de ADN. El objetivo de todos estos estudios fue obtener una expresión cuantitativa de la complejidad de estas secuencias. En la sección 2.2.2 se describen los trabajos pioneros de Gatlin [Ga 72], junto con algunas modificaciones más recientes.

A pesar del hecho de que las secuencias de ADN deben contener toda la información relevante para los seres vivos, los intentos comentados no tuvieron, del todo, éxito en la obtención de una medida cuantitativa para tal información. En algunos de estos estudios, el ADN era virtualmente indistinguible de una secuencia aleatoria. El mayor exponente de esta visión pesimista fue el artículo de Hariri y colaboradores [Har 88]. Después de cierto tiempo en el que no se publica sobre el tema, se abre un segundo período (desde 1987 hasta la actualidad) en el que se retoma el interés, en parte, a causa de la avalancha de datos, disponiéndose cada vez de secuencias más y más largas. Las bases de datos contenían ya cadenas suficientemente largas como para soslayar el mayor inconveniente que se tenía para poder aplicar correctamente la Teoría de la Información a las secuencias de ADN.

2.2.1 El sistema de comunicación ADN-Proteína

Toda la información necesaria para controlar las reacciones biológicas en las células y tejidos, incluida la síntesis de proteínas, está contenida en el ADN. Somos conscientes de que esta afirmación supone sólo una primera aproximación, ya que los

ácidos nucleicos presentan una estructura química muy compleja* y el modelo de cadena (representación del ADN como una cadena de letras que representan el orden en el que los nucleótidos aparecen en la secuencia) no expresa tal estructura, siendo una representación muy aproximada de la molécula. La cuestión central de la aplicación de técnicas tales como la Teoría de la Información puede ser, por tanto, una medida de hasta qué punto es buena esta primera aproximación. La cadena de ADN se puede describir esquemáticamente como una cadena de nucleótidos, formando una doble hélice, en la que se encuentran cuatro nucleótidos distintos, *adenina*(A), *timina o uracilo*† (T ó U), *citosina*(C) *y guanina* (G), por lo que podemos considerarla como un mensaje que procede de un alfabeto de cuatro símbolos. Por otra parte, en este contexto, las proteínas, son también cadenas lineales (aquí sí que es extremadamente simplista esta primera aproximación que comentábamos antes) formadas por 20 constituyentes básicos (los aminoácidos). Cada secuencia de ADN, con función codificadora, determina una única cadena proteínica, aunque cada proteína puede ser codificada por varias secuencias de ADN, ligeramente diferentes. El *Código Genético* establece la correspondencia entre la secuencia de nucleótidos y la secuencia de aminoácidos. Puesto que hay 20 aminoácidos y sólo cuatro nucleótidos, se necesita una combinación de varios nucleótidos, concretamente tres, para codificar cada aminoácido; estos grupos de tres nucleótidos son lo que se conoce como *codones.*

La información genética siempre fluye de forma irreversible desde los ácidos nucleicos a las proteínas en todos los seres vivos (no se conocen mecanismos de "retro-traducción" de proteínas en ácidos nucleicos) y esencialmente, a través de los mismos mecanismos básicos. No se pretende describir aquí los detalles de estos

---

*Incluso recientemente se especula con la posibilidad de que las moléculas de ADN sean capaces de conducir electricidad, al contrario de otros biopolímeros como las proteínas, y la relación de esta propiedad con la posibilidad de reparación de daños a distancia [Wil 97],[Da 97].

†La timina y el uracilo juegan esencialmente el mismo papel a efectos de transmisión de información, la primera aparece en el ADN y la segunda en el ARN (ácido ribonucleico).

mecanismos, salvo hacer notar que la transferencia de la información biológica se puede se puede tratar como un canal de comunicación. La entrada es la secuencia de ADN y la salida es la cadena de aminoácidos en la proteína. La fuente de información y el canal de transmisión del sistema de comunicación que proponemos se describen a continuación.

La fuente de información

Es frecuente en el estudio de secuencias hablar de un dispositivo abstracto (que se suele denominar fuente) que genera secuencias de símbolos (mensajes) escogiéndolos de un alfabeto finito. Tal selección de los símbolos puede ocurrir de acuerdo con gran variedad de mecanismos. Son especialmente importantes las fuentes ergódicas, en las que, mecanismos aleatorios llevan a la producción de mensajes "típicos" (i.e. estadísticamente homogéneos) con probabilidad bastante alta (próxima a 1) y de secuencias "atípicas" con probabilidad despreciable. Es bien conocido que las distribuciones de probabilidad de elementos textuales (tales como letras) se conservan a lo largo de textos suficientemente grandes, y por tanto se pueden suponer que estos lenguajes se pueden modelar como fuentes ergódicas. Como veremos más adelante, justamente esta suposición de ergodicidad (que lleva implícita la estacionariedad) no está justificada en el ADN en general [Ka 93]; aunque para las zonas codificadoras podría no ser descabellada [Che 96]. Tal vez sea este una de las causas principales de los malos resultados obtenidos con estos primeros análisis.

Admitiendo la ergodicidad, para el ADN la fuente de información se puede definir como:

(1) El alfabeto: $\mathbf{B} = \{A, T, C, G\}$

(2) En general, los símbolos de **B** no se emiten con la misma frecuencia. La distribución de probabilidad sobre el alfabeto es una propiedad de la fuente:

$$\mathrm{P}(A) + \mathrm{P}(T) + \mathrm{P}(C) + \mathrm{P}(G) = 1 \tag{2.1}$$

Tabla I. Código Genético

GCU, GCC, GCA, GCG	Alanina
$CGU, CGC, CGA, CGG, AGA, AGG$	Arginina
AAU, AAC	Asparagina
GAU, GAC	Ácido aspártico
UGU, UGC	Cisteína
CAA, CAG	Ácido glutámico
GAA, GAG	Glutamina
GGU, GGG, GGA, GGC	Glicina
CAU, CAC	Histidina
AUU, AUC, AUA	Isoleucina
$UUA, UUG, CUU, CUC, CUA, CUG$	Leucina
AAA, AAG	Lisina
AUG	Metionina
UUU, UUC	Fenilalanina
CCU, CCC, CCA	Prolina
UCU, UCC, UCA, UCG	Serina
ACU, ACC, ACA, ACG	Treonina
UGG	Triptófano
UAU, UAC	Tirosina
GUA, GUC, GUG, CUU	Valina
UAA, UAG, UGA	STOP

(3) Las bases no son independientes en el mensaje genético. La fuente no se puede considerar como de Bernuilli, sino que debe ser modelada como una fuente de Markov, con una matriz estocástica:

$$[\mathrm{P}(B_i \mid B_j)], \quad \sum_i \mathrm{P}(B_i \mid B_j) = 1 \tag{2.2}$$

Se supone también que esta fuente de Markov es estacionaria y ergódica, y por tanto, la distribución de probabilidad puede ser a) Obtenida a partir de las probabilidades condicionales, b) determinada experimentalmente:

$$\mathrm{P}(B_i) = \sum_j \mathrm{P}(B_i \mid B_j)\mathrm{P}(B_j) \tag{2.3}$$

El canal de transmisión ADN-Proteína

Éste corresponde al código genético mostrado en la tabla I, fuertemente degenerado, y conocido desde 1961. Lo supondremos como estacionario y sin memoria; se puede representar por la correspondencia aleatoria entre el conjunto de codones $\mathbf{B}^3 = \{B_1, B_2, B_3\}$ y el conjunto de aminoácidos $\mathbf{A} = \{A_i\}$:

$$\mathbf{B}^3 \xrightarrow{\mathrm{P}(A_i/B_1B_2B_3)} \mathbf{A}$$

Para un canal sin ruido (mutaciones), las probabilidades de entrada/salida son:

$$\mathrm{P}(A_i/B_1B_2B_3) = \begin{cases} 1, \text{ si el par } (A_i/B_1B_2B_3) \text{ pertenece al código} \\ 0, \text{ si no} \end{cases}$$

2.2.2 Medidas de orden y complejidad

A continuación vamos a exponer la contribución original de Gatlin [Ga 72], junto con las modificaciones introducidas por Sibbald [Si 89]. Para mayor claridad, se ha modificado algo la presentación. El análisis de secuencias tiene en cuenta tanto la composición de bases como la ordenación de éstas, dando lugar a dímeros, trímeros, etc.,..., n-tuplas, siendo n tan grande como sea posible. Una secuencia de ADN se considera bajo los dos puntos de vista complementarios que se citan a continuación:

Orden o regularidad en la secuencia

Como un ejemplo teórico extremo, la secuencia más ordenada posible sería, por ejemplo, *AAAAAA*..., mientras que la más desordenada (¿compleja?), una que hubiese sido generada al azar (dicho sea de paso que la verificación experimental de esta propiedad no es ni mucho menos trivial [Cp 91]). Como medida de orden se adopta la divergencia entre una secuencia que es tan aleatoria como sea posible y la secuencia natural; para tal divergencia se toma la diferencia entre las correspondientes entropías de Shannon:

$$\mathrm{D}_n = \mathrm{H}_n^{(r)} - \mathrm{H}_n^{(n)} = -\sum_{B_n} \frac{f_i f_j}{f_{i\otimes j}} \log \frac{f_i f_j}{f_{i\otimes j}} - \left(-\sum_{B_n} f_k \log f_k \right) \tag{2.4}$$

donde f_s son las frecuencias relativas y los índices recorren los siguientes conjuntos:

k: Las B_n subsecuencias de longitud n;

i: Las subsecuencias de longitud $n-1$, obtenidas eliminando la primera base de B_n;

j: Las subsecuencias de longitud $n-1$, obtenidas eliminando la última base de B_n;

$i \otimes j$: Las subsecuencias de longitud $n-2$, obtenidas eliminando la primera y la última base de B_n.

Para $n = 1$, y una secuencia de longitud infinita, $\mathrm{H}_{esperada} = 2 \log 2 = 2$ bits, para cualquier otro nivel, $\mathrm{H}_{esperada}$ es la entropía de la distribución teórica de $n-$tuplas, condicionadas por las observadas para $n-1$, suponiendo independencia. Por tanto, $\mathrm{H}_{esperada}$ representa la máxima entropía teórica al nivel n compatible con lo observado en la secuencia para el nivel $n-1$.

La complejidad

Conceptualmente, es la complementaria de la medida anterior. Numéricamente, lo sería también si el análisis se llevase a cabo sólo para un nivel pero, en ese caso, sería superflua. Sin embargo, el estudio multinivel hace que ambas medidas sean útiles. La complejidad se define a partir de las divergencias, de forma recurrente:

$$\mathrm{C}_n = \mathrm{C}_{n-1} - \mathrm{D}_n \tag{2.5}$$

Nótese que la complejidad aumenta con las entropías de los correspondientes histogramas:

$$\begin{aligned} \mathrm{C}_n &= \mathrm{C}_{n-1} - \mathrm{D}_n = (\mathrm{C}_{n-1} - \mathrm{D}_{n-1}) - \mathrm{D}_n = ... = \\ &= \mathrm{C}_0 - \sum_{i=1}^{n} \mathrm{D}_i = \mathrm{C}_0 + \sum_{i=1}^{n} \mathrm{H}_i^{(n)} - \sum_{i=1}^{n} \mathrm{H}_i^{(r)} \end{aligned} \tag{2.6}$$

Las dos primeras divergencias tienen un significado simple en el contexto del mensaje genético: mientras D_1 mide la distancia de la equiprobabilidad, D_2 refleja la distancia de la independencia entre bases contiguas. La *redundancia* $\mathbf{R}$ del código (hasta ese nivel) viene dada por [Gui 77]:

$$2\mathbf{R} = \mathrm{D}_1 + \mathrm{D}_2 \tag{2.7}$$

La Figura 2.1 muestra en cascada las medidas multinivel de la divergencia y la complejidad. El significado es el siguiente: antes de observar la secuencia de ADN, tenemos una complejidad "disponible" de 2 bits/base. La primera medida (composición en bases) detecta una divergencia D_1 con respecto a $\mathrm{C}_{\mathrm{max}}$, que nos lleva a una complejidad C_1 (en este caso, igual a la entropía). Al segundo nivel, se detecta una divergencia D_2 con respecto a C_1, lo que nos lleva a una complejidad C_2,

y así sucesivamente. A cada nivel n, la complejidad C_n tiene un carácter residual, en el sentido de que es lo que queda después de sucesivas reducciones debidas al orden introducido en la secuencia desde $n = 1$. Sin embargo, al mismo tiempo, sigue estando disponible para posteriores reducciones al pasar a niveles superiores.

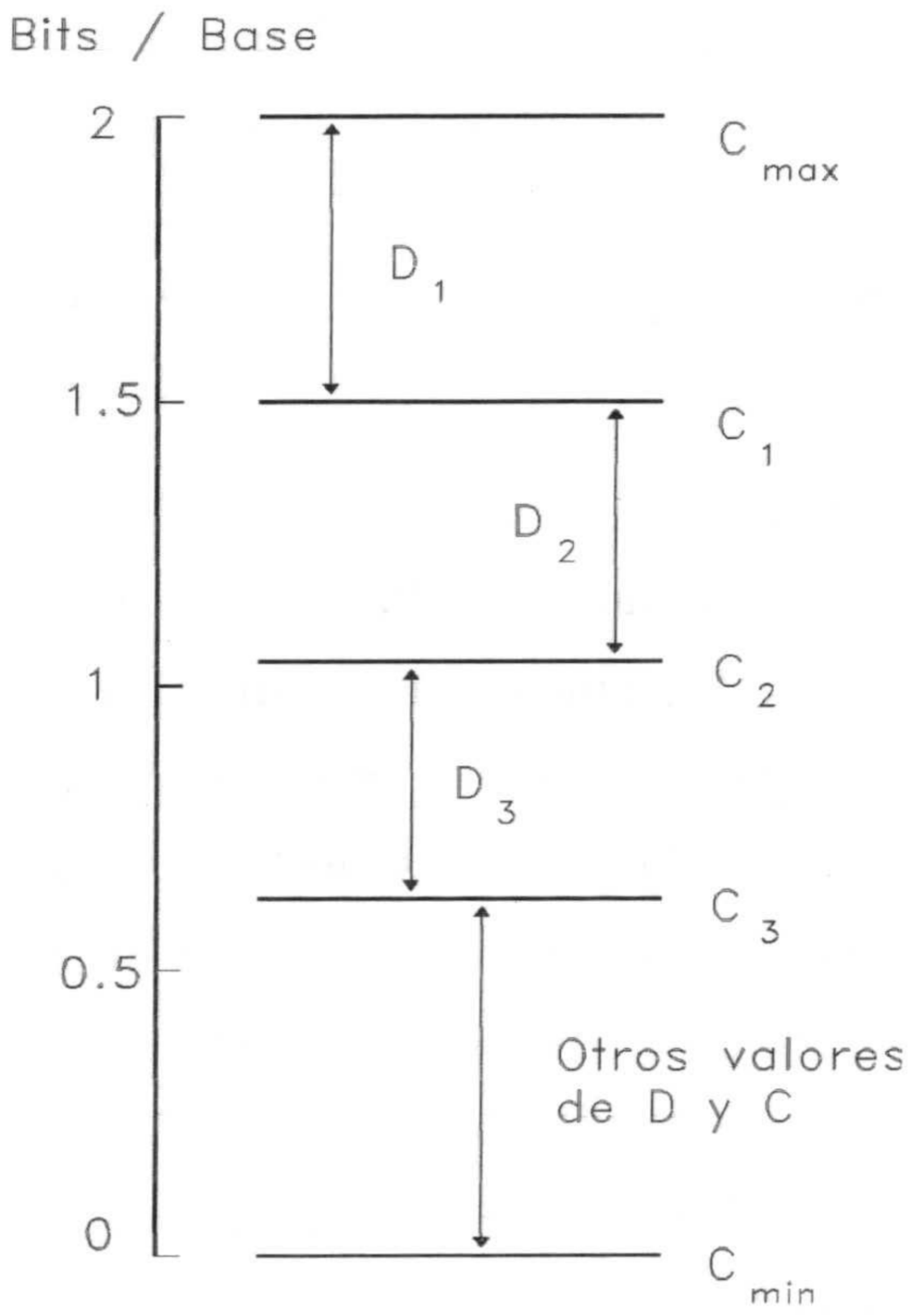

Figura 2.1: Escala de complejidad en el ADN

En condiciones de equiprobabilidad e independencia de los símbolos (lo que implica la máxima entropía por símbolo) tenemos que la fuente puede emitir la máxima diversidad de mensajes y por tanto tiene la máxima riqueza en información. Cualquier desviación de estas condiciones disminuye la diversidad pero aumenta la

fiabilidad del mensaje (ésta viene medida por la redundancia). Como resultado, el sistema es capaz de detectar errores. Las cadenas de ADN, como los mensajes en cualquier lenguaje natural y en cualquier sistema de comunicación artificial, adoptan soluciones de compromiso entre lo que puede entenderse como *cantidad* y *calidad* de información. En el caso que nos ocupa, destacan dos hechos experimentales: 1) en los vertebrados, la redundancia se debe básicamente a D_2, mientras que en los animales inferiores, aparece gracias a D_1[Ga 72] b) consistente con lo anterior, Reichert y Wong [Re 73] observaron que la linealidad de $\mathbf{R}$ con respecto a D_2 en un conjunto de proteínas de algunos organismos; los autores llamaron a este hallazgo la "verdadera flecha del tiempo".

Como se ha comentado anteriormente, los resultados globales de estos intentos no fueron execesivamente alentadores [Har 88], lo cual es bastante sorprendente, ya que las cadenas de ADN contienen toda la información acerca de los seres vivos. Es difícil creer que no seamos capaces de expresar en forma de relaciones matemáticas todo lo que sabe la Biología actual acerca de la información genética. Es probable que el motivo de este fracaso esté en la forma de abordar el problema. Hasta el momento todos los tratamientos que hemos visto utilizan una sola entropía (la de Shannon), una distribución de probabilidad (la de las frecuencias relativas de aparición de oligómeros en la cadena) y una única medida de divergencia respecto a la aleatoriedad (la redundancia).

Por otra parte, se usan cadenas de ADN excesivamente cortas como para que la estadística sea suficiente. En la Figura 2.2 hemos representado la divergencia condicional (dividida por $\log 4$ para expresarla en bits/base) para algunas secuencias de ADN de longitud similar a las que usa Hariri (ver ilustración 3 de [Har 88]), junto con ellas se ha incluido una secuencia aleatoria y un trozo de 800 pares de bases de la secuencia HUMHBB con una longitud total de 73326 bp.

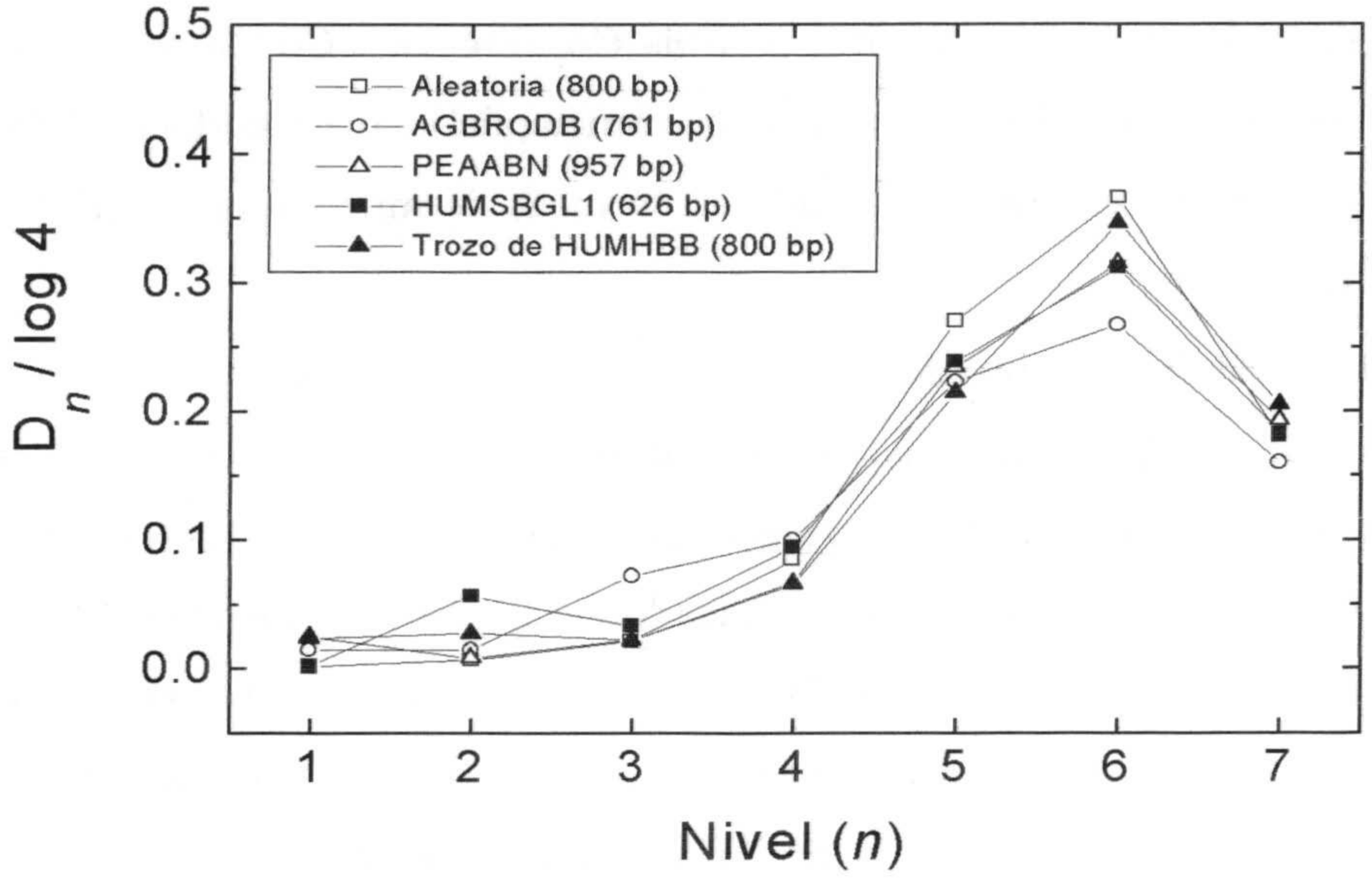

Figura 2.2: Perfiles de Divergencia condicional en función del nivel de agrupamiento para algunas secuencias cortas de ADN.

Vemos claramente que se cumplen los augurios pesimistas que habíamos citado antes; las secuencias (que son de diversos organismos) aparecen completamente indistinguibles y, es más, indistinguibles de la secuencia aleatoria.

Las causas de estos malos resultados, aparte de la suposición inicial de estacionariedad de la fuente, están muy posiblemente en los efectos causados por el pequeño tamaño de las secuencias: nótese el comportamiento de la secuencia aleatoria en la figura 2.2. Incluso con secuencias de mayor tamaño hay que introducir correcciones a causa de la dependencia con la longitud que tienen los estimadores de estadísticos basados en la entropía de Shannon [He 94b], y que estos estudios no tienen en cuenta.

2.2.3 Caracterización de secuencias

Otra aportación, relativamente reciente, es la de Cosmi et al. [Co 90], que intenta clasificar las secuencias de acuerdo con la frecuencia del uso de codones. Estos autores usan un método estadístico basado en técnicas de máxima entropía de estadísticos multivariados.

En ella se aborda el problema de clasificar secuencias de ADN en clases predefinidas y caracterizadas por las frecuencias relativas de aparición de codones en cada una de ellas. El método es como sigue. Las secuencias disponibles se dividen en los siguientes grupos: 1) secuencias codificadoras; 2) genes eucariotas; 3) genes procariotas; 4) genes *Escherichia coli*, 5) intrones. Para el conjunto completo de secuencias que componen cada clase S, se obtiene el vector de frecuencias relativas de codones, $\mathcal{G}^s = (g_1^s, g_2^s, ..., g_r^s)$, $r = 64$. Dada una secuencia experimental j a reconocer, se obtiene también su vector frecuencias relativas $\mathcal{F}^j = (f_1^j, f_2^j, ..., f_r^j)$. Como cuestión esencial al método, los autores determinan la distancia entre $\mathcal{F}^j$ y cada una de las clases $\mathcal{G}^s$ mediante el estadístico entropía relativa [Cov 91] que los autores llaman "*maximum entropy statistic*". Se estudia la distribución de este estadístico por simulación, teniendo en cuenta que es función de la longitud de la secuencia. A partir de esta distribución, se realiza el reconocimiento de la secuencia, es decir, la asignación a una clase particular con un 99% de confianza.

Los porcentajes de secuencias correctamente asignadas es del 26% para zonas codificadoras y 35% para intrones, por tanto el método puede ser utilizado para distinguir entre zonas codificadoras y no codificadoras.

2.2.4 El papel de las repeticiones

Más recientemente, Herzel et al. [He 94a] estudian el papel del ADN repetido en los valores de las entropías de las $n-$tuplas (*word entropy* lo denominan ellos) y

encuentran valores mucho menores que los encontrados hasta el momento y, por tanto, valores mucho mayores de redundancia. La justificación que dan ellos para esta discrepancia es el hecho de que las secuencias disponibles hasta entonces estaban muy sesgadas hacia zonas codificadoras, dada la importancia de éstas, y justamente es en las zonas no codificadoras donde se encuentra principalmente el ADN repetido.

Cabe destacar que, en este estudio, se aplican técnicas para descontar el efecto debido a la longitud finita de las secuencias [He 94b], que no habían sido utilizadas en los trabajos anteriores. Esto les permite estimar entropías para valores de n mayores que hasta el momento.

Para estudiar el efecto de las repeticiones en palabras de longitud mayor de la que permitían las secuencias disponibles, proponen un modelo de ADN repetido consistente, esencialmente, en secuencias aleatorias en las que se introducen al azar múltiples copias de una o varias subsecuencias. Obtienen que los valores máximos de redundancia se alcanzan para longitudes relativamente cortas por lo que llegan a la conclusión de que las repeticiones insertadas al azar en las secuencias no pueden dar cuenta de las correlaciones de largo alcance observadas en algunas secuencias de ADN (que comentaremos detenidamente en las secciones siguientes), aunque dejan abierta la posibilidad de que repeticiones no insertadas al azar puedan introducir correlaciones a distancias considerables.

2.3 Estadística fractal en secuencias de ADN

Hace unos años tres grupos de investigación descubrieron de forma independiente la existencia de correlaciones de largo alcance en las secuencias primarias de nucleótidos de algunas cadenas de ADN ([Li 92c], [Vo 92], [Pe 92]). Estadísticamente, estas correlaciones se pueden detectar midiendo la autocorrelación entre símbolos de la secuencia y comprobando si, al aumentar la distancia entre símbolos, ésta decae más

despacio que una función exponencial. Pero lo más sorprendente no es la presencia de estas correlaciones, sino la forma particular en que decaen con la distancia, en forma de ley de potencia, comportamiento que, por el momento, no es muy bien conocido y que, por lo general, va asociado a fenómenos tremendamente complejos (ver p.e. [Vo 88],[Bu 94] para una revisión). Entre otras cosas, la presencia de estas correlaciones implica la inexistencia de una longitud de correlación característica.

En este punto es importante destacar, que lo interesante en el estudio de las secuencias de ADN es la presencia de este tipo particular de correlaciones y no la mera presencia de correlaciones a larga distancia, lo que revela cierta complejidad en la estructura. Hay casos en los que existen correlaciones a larga distancia, pero son fácilmente explicables y no requieren de la presencia de una estructura sin longitud de escala característica. Por ejemplo, una secuencia con una función de autocorrelación exponencial puede presentar correlaciones a distancias considerables si la longitud característica de la función exponencial es grande, o una secuencia formada por la alternancia de bloques iguales puede presentar correlaciones que incluso crecen al aumentar la distancia, pero en ambos casos existe una explicación muy simple que justifica estas correlaciones.

Otra implicación importante de la existencia de este tipo de correlaciones es la relación que tienen con la estructura fractal, invarianza ante cambio de escala, de las secuencias que las presentan. Sobre esta cuestión volveremos con más detenimiento en las secciones siguientes.

2.3.1 Métodos de análisis

La mayoría de los métodos de análisis que se describen en esta sección tienen su origen en el análisis de series temporales numéricas, ampliamente utilizados en disciplinas que van desde la teoría de la señal a la predicción meteorológica por citar sólo dos ejemplos [Rei 65], [Rob 74]. A la hora de aplicar estos métodos a una serie

simbólica se plantea el problema de convertirla en una serie numérica, pero hay un inconveniente, resulta que cualquier asignación numérica a una serie simbólica con una alfabeto de más de dos símbolos introduce correlaciones, que no existen en la propia secuencia y que son debidas únicamente a la asignación. Se puede comprobar que no es posible una asignación numérica que no introduzca correlaciones, como mínimo hay que asignar a cada nucleótido un vector en un espacio de 4 dimensiones [He 95].

Para soslayar estas dificultades se ha optado por convertir la serie de nucleótidos en una serie binaria a la que se pueden aplicar técnicas de análisis de series numéricas, sin importar la asignación numérica concreta. Formalmente la asignación es una aplicación del conjunto $\mathbf{B}$ ={A,T,C,G} en los números reales:

$$\{\mathrm{A,T,C,G}\} \longrightarrow \Re \tag{2.8}$$

$$B_i \longrightarrow u(i)$$

Puesto que tenemos 4 nucleótidos (A,T,C,G) son posibles 7 reglas de asignación binaria, 3 de ellas agrupan los nucleótidos por parejas[‡]:

1. $u\,(\mathrm{A}) = u\,(\mathrm{T}) = 1$, $u\,(\mathrm{C}) = u\,(\mathrm{G}) = -1$. Conocida como *ligadura de hidrógeno (hidrogen bound)*. Agrupa los nucleótidos unidos por tres puentes de hidrógeno (C,G) y los unidos por dos puentes de hidrógeno (A,T), se conoce también como Regla S-W (de Strong=fuerte, Weak=débil).

2. $u\,(\mathrm{A}) = u\,(\mathrm{G}) = 1$, $u\,(\mathrm{T}) = u\,(\mathrm{C}) = -1$. Regla R-Y, asocia purinas (A,G) por una parte y pirimidinas (T,C) por otra.

3. $u\,(\mathrm{A}) = u\,(\mathrm{C}) = 1$, $u\,(\mathrm{T}) = u\,(\mathrm{G}) = -1$. Sin significado biológico.

[‡]En este tipo de asignaciones el valor numérico es completamente irrelevante. Aquí se ha optado por 1 y -1, pero también es frecuente encontrar en la bibliografía la asignación a 1 y 0.

Y otras cuatro que enfrentan a un nucleótido con el resto:

1. $u(\mathrm{A}) = 1$, $u(\mathrm{T}) = u(\mathrm{C}) = u(\mathrm{G}) = -1$.
2. $u(\mathrm{T}) = 1$, $u(\mathrm{A}) = u(\mathrm{C}) = u(\mathrm{G}) = -1$.
3. $u(\mathrm{C}) = 1$, $u(\mathrm{A}) = u(\mathrm{T}) = u(\mathrm{G}) = -1$.
4. $u(\mathrm{G}) = 1$, $u(\mathrm{A}) = u(\mathrm{T}) = u(\mathrm{C}) = -1$.

Algunos autores analizan directamente la información obtenida por el análisis de la secuencia con una regla concreta, mientras que otros optan por promediar los resultados de varias de estas asignaciones. Aunque los mejores resultados se obtienen analizando las secuencias con la regla R-Y, además algunos autores justifican en términos biológicos [Bu 93a] la existencia de correlaciones de largo alcance cuando se analizan determinadas secuencias con esta regla de asignación y su ausencia cuando se analizan con la regla S-W.

Junto con estas reglas, se podrían añadir otras (menos utilizadas en la bibliografía) como por ejemplo asignar a cada nucleótido la masa molecular de la base correspondiente:

$u(\mathrm{A}) = 134$, $u(\mathrm{T}) = 125$, $u(\mathrm{C}) = 110$ y $u(\mathrm{G}) = 150$.

En este tipo de asignaciones sí que importa el valor numérico asociado a cada nucleótido, pero aquí estos valores tienen significado biológico.

Un problema similar aparece intentando representar gráficamente la secuencia (refs [7-12] de [Vo 92]), de hecho para una secuencia de ADN (4 símbolos) es necesario como mínimo un espacio de 4 dimensiones para obtener un gráfico en el que las propiedades geométricas, entre ellas las fractales, no dependan de la técnica utilizada [Vo 92].

Con la asignación binaria desaparece este problema y se puede representar la secuencia en un espacio bidimensional (aunque por supuesto, no se representa

toda la información almacenada en la secuencia, sólo la debida a la alternancia de los nucleótidos agrupados en la asignación). Una representación muy utilizada es la conocida como *DNA-walk,* que tiene su origen en el conocido *random-walk,* utilizado para representar, por ejemplo, el movimiento browniano.

Una vez hecha la asignación de la secuencia a una secuencia binaria, ésta se representa gráficamente tomando como eje x un índice que se mueve a lo largo de la secuencia y representando en el eje y el valor acumulado de $u(B_i)$ con $B_i \in$ {A,T,C,G}:

$$w(x) = \sum_{i=1}^{x} u(B_i) \tag{2.9}$$

Por ejemplo, la secuencia:

{A,T,T,A,T,C,G,A,T,A,C,G,C,A,C,G,G,T,G,G}

que convertida en secuencia binaria con la regla S-W, quedaría:

$$\{1,1,1,1,1,-1,-1,1,1,1,-1,-1,-1,1,-1,-1,-1,1,-1,-1\}$$

tendría un DNA-walk como el que vemos en la figura 2.3.

Esta representación, dejando aparte las limitaciones impuestas por la necesidad de utilizar alfabeto binario, da una visión muy directa de la secuencia, por ejemplo, nótese como se aprecia sensiblemente la abundancia de A+T en la primera zona (gráfica ascendente) mientras que en la segunda, la abundancia de C+G provoca un descenso en la gráfica, además, en este caso el hecho de que el walk termine en 0 indica que hay una composición global de A+T igual a la de C+G. En general, zonas con pendiente positiva indican abundancia de 1's, zonas con pendiente negativa abundancia de -1's y zonas más o menos horizontales una composición alternada.

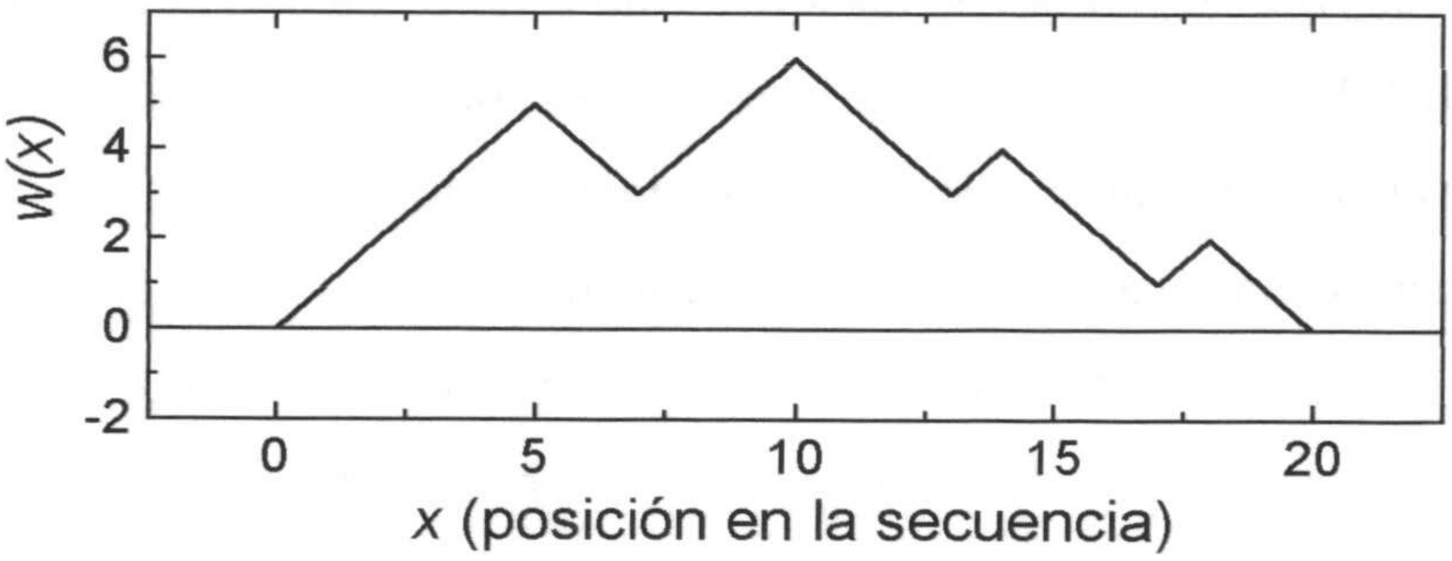

Figura 2.3: Ejemplo de DNA-walk

Medida directa de la autocorrelación

Una de las consecuencias más importantes de la presencia de correlaciones que decaen siguiendo leyes de potencia en los sistemas físicos es la inexistencia de tamaños característicos, esto es, no se pueden aislar zonas del sistema y estudiarlos por separado ya que hay interacciones entre todas las partes, desde distancias muy pequeñas hasta distancias comparables con el tamaño del conjunto. Una magnitud que caracteriza estas situaciones, muy utilizada en Mecánica estadística, es la autocorrelación. Es una medida que nos da idea de la "relación" que hay entre los valores de una magnitud en puntos del sistema situados a una distancia determinada. Para una secuencia numérica $\{u(i)\}$ se define como:

$$C(\ell) = \langle u(i)\, u(i+\ell) \rangle - \langle u(i) \rangle^2$$

donde $\langle \cdot \rangle$ indica promedio sobre todas las posiciones i en la secuencia.

Para una secuencia aleatoria $C(\ell) = 0$ para $\ell \neq 0$ y $C(0) = 1$, y para secuencias con longitudes características, por ejemplo una secuencia generada con un modelo de Markov, o cualquier otro que produzca dependencias entre los símbolos

de la secuencia pero hasta una distancia finita, aparece un comportamiento exponencial decreciente, dependiendo los parámetros de la exponencial de las longitudes características de correlación de la secuencia.

Por el contrario, una secuencia en la que la autocorrelación sea constante, independientemente de la distancia a la que se mida, tiene propiedades fractales. En concreto, se puede comprobar [Vo 88] que aparece la propiedad conocida como *autosemejanza estadística.* Esto es, si observamos tramos de la secuencia de longitud ℓ y otros de longitud $\lambda\ell$, en promedio las fluctuaciones aumentan un factor λ:

$$\langle \Delta w(\lambda\ell) \rangle = \lambda \langle \Delta w(\ell) \rangle \tag{2.10}$$

Y no sólo es cierto esto para el valor medio de las fluctuaciones, de hecho se comprueba que la distribución de probabilidad de las fluctuaciones también es invariante ante cambio de escala:

$$\mathrm{P}\{\Delta w(\lambda\ell)\} = \mathrm{P}\{\lambda \Delta w(\ell)\} \tag{2.11}$$

De forma intuitiva, esta propiedad supone que si ampliamos una parte del random-walk obtenemos algo que, en términos estadísticos es igual a lo que veíamos para la secuencia entera.

Si en lugar de tener una función de autocorrelación constante, ésta decae como una ley de potencia $C(\ell) \sim (1/\ell)^{\gamma}$, con $0 \leq \gamma \leq 1$ (que para $\ell \gg 1$, decae mucho más lentamente que una exponencial) ya no aparece autosemejanza estadística, pero en su lugar aparece un tipo de invarianza ante cambio de escala conocido como *autoafinidad estadística.* Ahora al pasar de observar un tramo de longitud ℓ a uno de longitud $\lambda\ell$, las fluctuaciones, en lugar de aumentar un factor λ lo hacen en un factor λ^{α} con $0 \leq \alpha \leq 1$, esto es [Wes 95]:

$$\mathrm{P}\{\Delta w(\lambda\ell)\} = \mathrm{P}\{\lambda^{\alpha} \Delta w(\ell)\}$$

Gráficamente en el random-walk esto se traduciría en que si aumentamos en un factor λ el eje x de una parte de la gráfica, observaremos lo mismo, en términos estadísticos si aumentamos, en un factor λ^{α} el eje y.

Este modelo matemático describe muy bien, por ejemplo, las fluctuaciones de la temperatura en un lugar concreto o las irregularidades de una cadena montañosa [Vo 88].

El exponente de escala de las fluctuaciones, α está relacionado con el exponente de la función de autocorrelación, γ; de hecho se puede comprobar [St 94] que para $1/2 \leq \alpha \leq 1$ se tiene que $C(\ell) \sim (1/\ell)^{\gamma}$ con:

$$\gamma = 2(1-\alpha) \tag{2.12}$$

Si en lugar de tener correlaciones la secuencia presenta anticorrelaciones, esto es, valores negativos de $C(\ell)$ debidos a la presencia de una alternancia mayor que la que cabría esperar para una secuencia aleatoria, y estas anticorrelaciones decrecen en forma de ley de potencia, se comprueba [Pe 94] que se obtienen valores de α en el rango $[0, 1/2]$ y ahora se cumple la relación:

$$\gamma = 2\alpha \tag{2.13}$$

Esta medida tiene la ventaja de ser directa, mide justamente lo que buscamos, la dependencia entre los nucleótidos a una distancia dada y no requiere que la secuencia analizada sea estacionaria. Por otra parte, el principal inconveniente está en que, a causa de la longitud finita de las secuencias, los valores obtenidos para $C(\ell)$ suelen fluctuar mucho y, en la mayoría de los casos, es difícil estimar el tipo de dependencia funcional para valores grandes de ℓ. Incluso llegan a obtenerse ajustes igualmente buenos para leyes exponenciales o de potencia [Pe 93a] [Sh 94].

Análisis de la fluctuación

En vista de la relación entre las correlaciones presentes en una secuencia y la forma en que se comportan las fluctuaciones de ésta, un método alternativo para analizar las correlaciones presentes en una secuencia numérica es realizar un estudio de las fluctuaciones del *random-walk* de la secuencia (eq. 2.9) [Pe 92]:

$$F^2(\ell) \equiv \left\langle [\Delta w(\ell) - \langle \Delta w(\ell) \rangle]^2 \right\rangle = \left\langle \Delta w(\ell)^2 \right\rangle - \langle \Delta w(\ell) \rangle^2 \tag{2.14}$$

donde la cantidad $\Delta w(\ell)$ se define como:

$$\Delta w(\ell) \equiv w(x+\ell) - w(x) = \sum_{i=x}^{x+\ell} u(B_i) \tag{2.15}$$

y $\langle \cdot \rangle$ denota el promedio sobre todas las posiciones x a lo largo de la secuencia. En la bibliografía es frecuente el uso de esta magnitud utilizando como regla de asignación la R-Y o la S-W, aunque algunos autores (p.e. [Vo 94]) la calculan promediando sobre las cuatro reglas de asignación que enfrentan un nucleótido con el resto (reglas 4,5,6,7).

En esencia, esta medida nos dice cómo dependen las fluctuaciones del tamaño de la porción de secuencia analizado, ya que no es otra cosa que el cálculo de la varianza de los incrementos de la cantidad $u(B_i)$ en todas las posibles "ventanas" de tamaño ℓ. Por lo general, valores grandes de $F(\ell)$ se corresponden a *random-walks* con perfiles abruptos.

Además, se puede comprobar que para *secuencias estacionarias*, esta cantidad esta relacionada con la autocorrelación por la expresión [Ka 93]:

$$F^2(\ell) = \ell C(0) + 2\sum_{i=1}^{\ell-1} (\ell - i) C(i) \tag{2.16}$$

Si la secuencia estudiada es aleatoria, tenemos que $C(\ell) = 0$ si $\ell \neq 0$ y $C(0) = 1$, por tanto tendríamos que $F(\ell) \sim \ell^{1/2}$. Si tenemos correlaciones de corto

alcance, como por ejemplo las producidas con un modelo de Markov de orden 1, tendremos un comportamiento similar al anterior pero con una pequeña corrección exponencial, que será despreciable para ℓ suficientemente grande y el comportamiento asintótico será de nuevo $F(\ell) \sim \ell^{1/2}$. Si, por el contrario, la correlación decae más lentamente que una exponencial $F^2(\ell)$ ya no tendrá un comportamiento lineal. En particular, una dependencia en forma de ley de potencia para $C(\ell)$ se traduce en una ley de potencia para $F(\ell)$ como se vio en la sección anterior (eqs. 2.12, 2.13).

Esta medida, al estar basada en una variable acumulada, presenta muchas menos fluctuaciones que la medida directa de la autocorrelación y permite hacer una estimación mucho más precisa de los exponentes α ó γ, pero presenta el problema de que la relación (2.16) sólo es válida para secuencias estacionarias, propiedad que no se le puede suponer de ninguna forma a las secuencias de ADN [Ka 93].

Los mismos autores que la propusieron introdujeron posteriormente una modificación que hace posible eliminar el efecto de la no estacionariedad de la secuencia y que denominaron *Detrended Fluctuation Analysis* (DFA) [Pe 94], [Bu 95] (en castellano sería algo así como "análisis de la fluctuación sin tendencia local"). Este método está basado en una de las técnicas habituales para el tratamiento de series temporales no estacionarias [Man 96] y ha sido también utilizado con buenos resultados en la estimación de espectros multifractales [Ca 97]. Lo describiremos brevemente ya que es el método más utilizado en la bibliografía y será el que tomaremos como referencia al referirnos a la presencia de estructura fractal en secuencias de ADN.

El procedimiento es como sigue:

1. Se calcula $w(x)$ para cada posición en la secuencia.

2. Se toma una ventana deslizante de ℓ símbolos que recorre cada posición i en la secuencia con $i \in \{1, 2, ..., N - \ell\}$. En cada ventana se calcula el ajuste por mínimos cuadrados de $w(x)$, $w_{i,\ell}(x) = xa + b$ con $x \in \{i, i+1, ..., i+\ell\}$.

3. Finalmente se calcula la fluctuación como la suma de los cuadrados de la diferencia entre el walk original y el ajuste por mínimos cuadrados en cada ventana, y se promedia para todas las ventanas:

$$F_d^2(\ell) \equiv \frac{1}{(N-\ell+1)(\ell-1)} \sum_{i=1}^{N-\ell} \sum_{x=i}^{i+\ell} [w(x) - w_{i,\ell}(x)]^2 \qquad (2.17)$$

Se comprueba que, para ℓ suficientemente grande, esta cantidad también sigue una ley de potencia si la función de autocorrelación lo hace y siguen siendo válidas las relaciones (2.12) y (2.13). En [Bu 95] se presenta también una correción que permite estimar los exponentes incluso para valores pequeños de ℓ.

Transformada de Fourier

A causa de varias razones, entre ellas una interpretación más sencilla de los resultados (ver p.e. [Per 93]), se usa bastante el espectro de potencia en lugar de la autocorrelación para caracterizar la estructura estadística de la secuencia.

Dada la secuencia numérica $u(x)$ se define su transformada de Fourier como[§]:

$$q(f) = \frac{1}{N} \sum_{x=1}^{N} u(x) \exp\left(2\pi i \frac{fx}{N}\right) \qquad (2.18)$$

donde la frecuencia es $f = x/N$ $(x = 1, 2, ..., N/2)$. El espectro de potencia se define como el cuadrado del módulo de la transformada de Fourier:

$$S(f) = \frac{1}{N^2} \left| \sum_{x=1}^{N} u(x) \exp\left(2\pi i \frac{fx}{N}\right) \right|^2 \qquad (2.19)$$

Por el teorema de convolución (también conocido como teorema de Wiener-Khinchin [Per 93]) se tiene que, para una secuencia estacionaria, el espectro de potencia es la transformada de Fourier de la correspondiente función de autocorrelación.

[§]Al igual que con el análisis de fluctuación, también algunos autores usan la suma de las transformadas de Fourier de las secuencias obtenidas con las reglas 4,5,6 y 7.

Por tanto el conocimiento de $S(f)$ también nos permite obtener los exponentes α y γ.

Para una secuencia aleatoria se deben producir fluctuaciones a todas las longitudes y, por tanto, su espectro debe contener por igual todas las frecuencias. El resultado es un espectro de potencia plano. Por este se denomina a veces *ruido blanco* a este tipo de secuencias, por analogía con la luz blanca que contiene todas las frecuencias del visible. Por otra parte, si la secuencia tiene correlaciones de corto alcance (tipo exponencial) entonces el espectro de potencia tiene forma de curva lorentziana. Otro ejemplo es el caso de una secuencia formada por la sucesiva alternancia de un par tramos aleatorios con diferente composición, que lleva a un espectro de potencia $\sim 1/f^2$ [Li 94]. En todos estos ejemplos no aparece estructura fractal, aunque en el último sí pueden existir correlaciones a largo alcance, pero trivialmente explicadas por la alternancia de los tramos. Cuando en la secuencia existen correlaciones de largo alcance no triviales, i.e. funciones de autocorrelación de la forma $C(\ell) \sim (1/\ell)^\gamma$ con $0 \leq \gamma \leq 1$ se comprueba que el espectro de potencia tiene un comportamiento del tipo $S(f) \sim 1/f^\beta$ con $0 \leq \beta \leq 2$. Además el exponente β se puede relacionar con α y γ por la expresión [St 94]:

$$\beta = 2\alpha - 1 = 1 - \gamma \tag{2.20}$$

Los casos extremos corresponderían a las situaciones que hemos comentado antes. El valor intermedio $\beta = 1(\gamma = 0)$ sería el más interesante, ya que se correspondería a una secuencia con autosemejanza. Para el resto de valores tendríamos autoafinidad. Es por este motivo que las secuencias que presentan un espectro de potencia de este tipo también reciben el nombre genérico de *ruido* $1/f$ [Wei 88].

Al igual que ocurre con la medida directa de la autocorrelación los resultados suelen estar afectados por muchas fluctuaciones por lo que los autores que la utilizan suelen presentar las gráficas de $S(f)$ frente a f suavizadas promediando sobre varios

valores consecutivos. Además, la limitación más importante del método está en que el teorema de convolución, que nos da la relación entre el espectro de potencia y la autocorrelación, sólo es aplicable a secuencias estacionarias.

Información mutua

Estas tres medidas que hemos visto están basadas en técnicas de teoría de la señal, que han sido pensadas para analizar secuencias numéricas. Como ya se ha visto al principio, para ser aplicables a secuencias simbólicas necesitan una asignación numérica, que para no introducir falsas correlaciones, tiene que ser necesariamente binaria. Para evitar estas limitaciones, algunos autores, entre ellos los de nuestro grupo, proponen el uso de medidas derivadas de la teoría de la información, especialmente pensadas para el análisis de secuencias simbólicas.

Entre estas medidas, la información mutua entre nucleótidos ha sido la más utilizada ya que está muy relacionada con la autocorrelación [Li 90], [He 95]. La información mutua entre nucleótidos separados por una distancia ℓ se define como:

$$I(\ell) = \sum_{i,j=(A,T,C,G)} f_{i,j}(\ell) \log\left(\frac{f_{i,j}(\ell)}{f_i \cdot f_j}\right) \tag{2.21}$$

donde f_i, es la frecuencia relativa de aparición del nucleótido B_i a lo largo de toda la secuencia y $f_{i,j}(\ell)$ es la frecuencia relativa de aparición a lo largo de toda la secuencia del par de nucleótidos B_iB_j separados por una distancia ℓ. La información mutua es la entropía relativa (o divergencia de Kullback) [Cov 91] entre la distribución de parejas de nucleótidos a distancia ℓ y el producto de las distribuciones individuales, y por tanto refleja cualquier diferencia entre ambas distribuciones, esto es, cualquier tipo de dependencia entre los nucleótidos situados a distancia ℓ ($I(\ell) = 0 \Leftrightarrow f_{i,j}(\ell) = f_i \cdot f_j \ \forall i,j \in \{A,T,C,G\}$), al contrario de la autocorrelación que sólo detecta dependencia o correlación lineal.

Para secuencias binarias, haciendo un desarrollo en serie de (2.21) se puede ver que $I(\ell) \sim C(\ell)^2$ [He 95], por lo que si la autocorrelación decae como una ley de potencia con exponente γ, la información mutua lo hará con un exponente 2γ. Para secuencias simbólicas con alfabetos de más de dos símbolos existen múltiples posibilidades para las funciones de autocorrelación y por tanto no existe la posibilidad de una comparación directa de éstas y la información mutua, aunque se puede comprobar [He 95] que esta última es una forma cuadrática de las funciones de autocorrelación con lo que se cumpliría también una relación similar a la que hemos visto para secuencias binarias.

Como principal ventaja de esta medida estaría la aplicabilidad directa a secuencias simbólicas y el hecho de que detecta todo tipo de correlaciones, no sólo las lineales. En cuanto a los inconvenientes, en primer lugar, al tratarse de una medida siempre positiva, las desviaciones de cero pueden ser debidas tanto a la existencia de correlaciones como de anticorrelaciones, que no serían distinguibles. De todas formas, para el análisis de secuencias de ADN, no es un inconveniente importante ya que, como veremos, prácticamente en todos los casos, son correlaciones lo que se detecta. Otro inconveniente, más relevante en este caso, es el hecho de que, implícitamente, se supone la secuencia como estacionaria desde el momento en que se habla de la distribución global de un nucleótido, distribución que no tiene sentido si la secuencia no es estacionaria.

Wavelet Analysis

Este método, que traducido al castellano sería algo así como "análisis de onditas", es una generalización de la transformada de Fourier [Dau 88], [Dau 92] que, en lugar de usar las funciones seno y coseno, utiliza un conjunto de funciones localizadas. Es un método ideal para estudiar secuencias heterogéneas (no estacionarias), al contrario de lo que ocurre con la transformada de Fourier [Ts 96]. Además el método, en cierto

sentido, es también una generalización del análisis de fluctuación que vimos en las secciones anteriores, y también permite hacer una estimación del exponente de escala α. Escogiendo adecuadamente el conjunto de funciones para el análisis, se pueden eliminar, no sólo las heterogeneidades lineales (como hace el DFA), sino cualquier tipo de heterogeneidad, obteniendo valores más fiables de α [Ar 95], [Ar 96], [Zha 95].

Este método también permite diferenciar entre una secuencia con un único exponente de escala y otra con un espectro multifractal, aunque por el momento los resultados obtenidos corroboran la idea de la presencia de un único exponente de escala en las secuencias analizadas. Por el contrario parecen haberse encontrado espectros multifractales que no corresponden a una ley de escala homogénea analizando algunas secuencias de proteínas [Str 95].

Para terminar con esta sección hay que comentar que todos los resultados que se han presentado aquí son válidos, en principio para secuencias infinitas por lo que habrá que tener en cuenta los posibles efectos debidos al tamaño finito de las secuencias [Pe 93a], [Ts 95], [He 94b], [Sh 94].

2.4 Modelización de secuencias de ADN

2.4.1 Cadenas de Markov

Uno de los primeros intentos de modelizar secuencias de ADN fue utilizando cadenas de Markov de orden 1 [Ga 66], [Ga 72], [El 74]. Ya entonces se vio que no eran un buen modelo. El primer motivo es que, en las zonas codificadoras, la posición dentro del codón es importante, por tanto las probabilidades de transición deberían depender de esta posición , [Bo 87a],[Bo 87b], [Tav 89]. Este problema es fácil de solucionar, se pueden usar modelos de Markov con probabilidades de transición que dependan de la posición dentro del codón. El segundo inconveniente es debido a la heterogeneidad: a escala global las probabilidades de transición dependen sobre qué

zona homogénea de la secuencia nos encontramos [Bo 87a],[Bo 87b]. Este problema no se puede resolver dentro del marco de los modelos de Markov.

Las correlaciones de largo alcance encontradas en las secuencias de ADN nos dan otra confirmación de que los modelos de Markov no son muy adecuados. La función de correlación de una cadena de Markov de orden 1 siempre decae de forma exponencial [Ta 94]. De hecho, se comprueba que la longitud típica de correlación viene dada en términos del mayor autovalor no trivial de la matriz de transición. En concreto, para secuencias binarias, la longitud típica de correlación es [Bu 93b]:

$$\ell_{\text{corr}} = -\frac{1}{\ln|\lambda|} \tag{2.22}$$

siendo λ el autovalor de la matriz de transición distinto de 1. Si $\lambda > 0$ la secuencia presenta correlaciones ($\alpha > 0.5$ en el DFA) y si $\lambda < 0$, presenta anticorrelaciones ($\alpha < 0.5$ en el DFA).

También se han intentado modelizar las secuencias con cadenas de Markov de orden superior [Raf 94], y modelos ocultos de Markov [Chu 89], [Chu 92], pudiendo estos últimos generar secuencias con grados de heterogeneidad relativamente altos, aunque no tanto como se observa en las secuencias naturales. Pero, en general, estos modelos presentan el inconveniente de tener demasiados parámetros libres.

2.4.2 Mutación-duplicación

Este modelo estaría dentro de lo que, en general, se conoce como sistemas de re-escritura (*rewriting systems*). Estas técnicas para generar secuencias con estructura jerárquica compleja reciben varios nombres en la bibliografía aunque, salvo ligeras modificaciones, todos responden a procesos similares: lenguajes libres de contexto [Cho 56], cuando se elimina la distinción entre símbolos terminales y no terminales, Sistemas-L [Ld 68], [Roz 80], sistemas de desarrollo [Weg 90], secuencias substitucionales [Chg 90].

Todos estos sistemas generan secuencias de forma iterada sustituyendo cada símbolo de la secuencia (o cada subsecuencia, dependiendo del modelo) por otro símbolo o conjunto de símbolos.

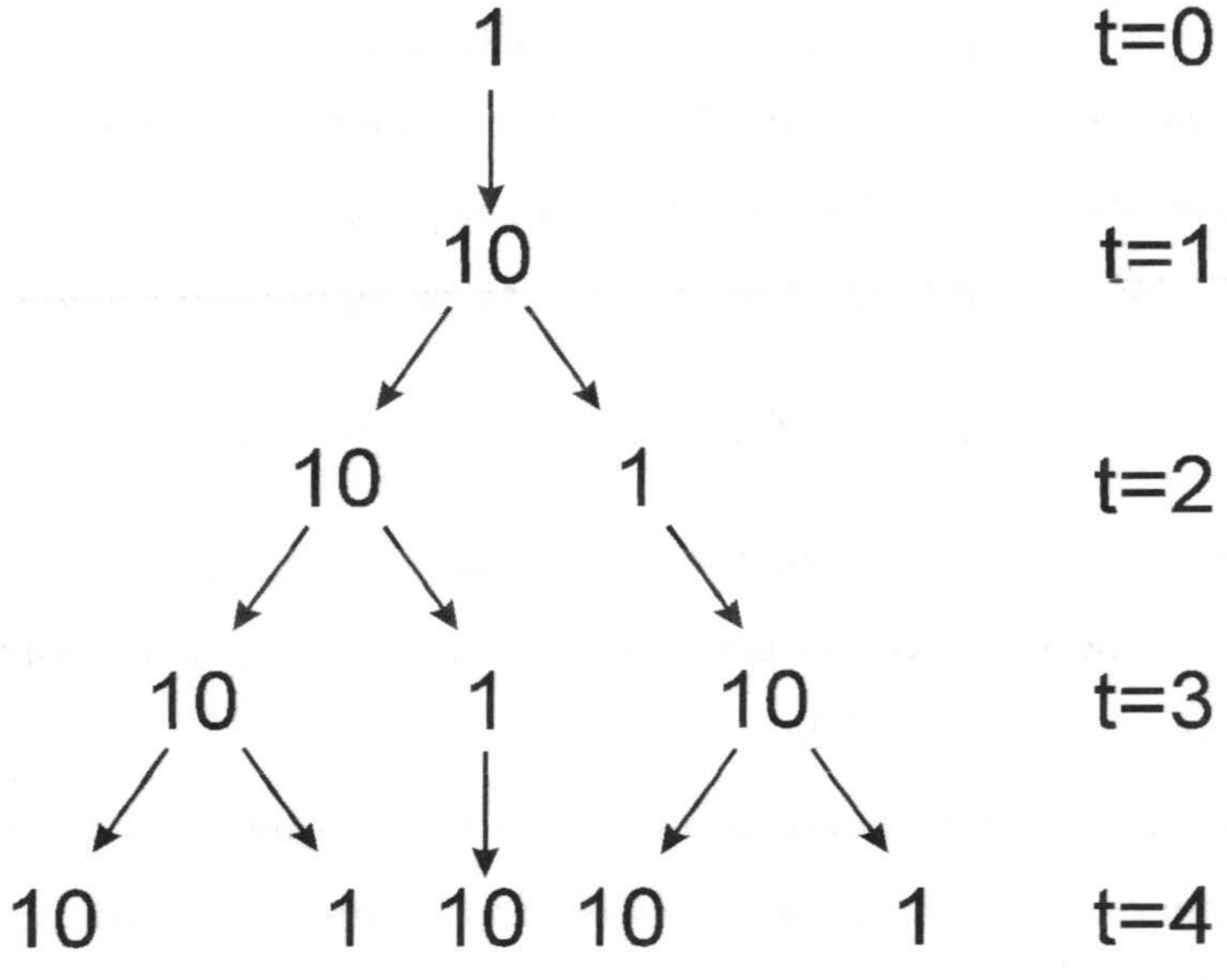

Figura 2.4: Primeras iteraciones de la regla de reescritura que da lugar a la secuencia de Fibonacci.

Por ejemplo, la conocida secuencia de Fibonacci se genera aplicando de forma iterada la siguiente regla de reescritura:

$$\begin{aligned} 0_t &\rightarrow (1)_{t+1} \\ 1_t &\rightarrow (10)_{t+1} \end{aligned} \tag{2.23}$$

En la figura 2.4 vemos las primeras iteraciones del proceso, tomando como secuencia inicial $\{1\}$.

Wentian Li [Li 89], [Li 91a], propuso uno de estos modelos, en los que se per-

miten varias posibilidades de sustitución cada una de ellas con cierta probabilidad:

$$0_t \rightarrow \begin{cases} (00)_{t+1} & \text{con probabilidad } 1-p \\ (1)_{t+1} & \text{con probabilidad } p \end{cases}$$

$$1_t \rightarrow \begin{cases} (11)_{t+1} & \text{con probabilidad } 1-p \\ (0)_{t+1} & \text{con probabilidad } p \end{cases} \tag{2.24}$$

Con este modelo intentaba simular las mutaciones y duplicaciones que tienen lugar en las cadenas de ADN. Además es fácilmente generalizable para simular mutaciones o duplicaciones concretas, asignando cierta probabilidad a cada una de ellas:

$$T_t \rightarrow A_{t+1}, \quad (ACC)_t \rightarrow (ACCACC)_{t+1}, ..., \text{etc.} \tag{2.25}$$

Inicialmente lo denominó modelo de expansión-modificación, aunque más tarde, introdujo la denominación (quizá más apropiada dentro de un contexto biológico) de mutación-duplicación [Li 92b].

Un hecho curioso en relación con este modelo es que, al contrario de los que veremos a continuación, cuando fue propuesto no pretendía justificar las correlaciones de largo alcance en las secuencias de ADN, que por aquel momento aún no habían sido encontradas (nótese que las referencias que hemos citado antes son de los años 89 y 91). Sucedió justo al revés, este modelo que pretendía simular las mutaciones y duplicaciones, producía secuencias con correlaciones de largo alcance, y esto hizo pensar a los autores en la posibilidad de que estas correlaciones estuviesen también en las secuencias reales [Li 92a].

Este modelo reproduce las propiedades de la autocorrelación de las secuencias reales y puede explicar ciertas características de ADN altamente repetido, pero no tiene en cuenta muchos otros procesos importantes en la evolución del ADN, como pueden ser las inserciones de retrovirus y las delecciones que son probablemente una de las fuentes más importantes de la rápida evolución de las secuencias de ADN. Los modelos que vamos a ver a continuación sí tienen en cuenta estos factores.

2.4.3 Lévy-walk generalizado

El modelo clásico del Lévy-walk describe una amplia variedad de fenómenos que presentan correlaciones de largo alcance [Shl 86] (refs. 9-15 [Bu 93b]). En la figura 2.5.a) se define esquemáticamente: consideremos una partícula que se mueve durante ℓ_1 pasos en una dirección fija escogida al azar, a continuación ℓ_2 pasos en otra dirección, etc. Las longitudes de los desplazamientos se escogen al azar de una distribución de probabilidad hiperbólica, esto es, una ley de potencia:

$$\mathrm{P}\left\{\ell_i = x\right\} = \frac{(\mu - 1)\,\ell_0^{\mu-1}}{x^{\mu}} \tag{2.26}$$

donde ℓ_0 es la longitud mínima de los recorridos y μ el exponente de la ley de potencia.

Se puede comprobar [Shl 86] que si el exponente cumple $2 \leq \mu \leq 3$, la serie numérica que nos da la posición de la partícula en cada paso presenta correlaciones de largo alcance y el desplazamiento medio de la partícula sigue una distribución de probabilidad invariante ante cambio de escala (autoafinidad estadística):

$$\mathrm{P}\left\{\lambda^{\alpha}\Delta w\left(\ell\right)\right\} = \mathrm{P}\left\{\Delta w\left(\lambda\ell\right)\right\} \tag{2.27}$$

con $0.5 \leq \alpha \leq 1$.

Buldyrev et al. [Bu 93b] proponen una generalización de este modelo para justificar las correlaciones de largo alcance observadas en algunas secuencias de ADN. En lugar de tomar una partícula que hace recorridos con direcciones elegidas al azar, se toman subsecuencias binarias de longitudes ℓ_i, también obtenidas de una distribución hiperbólica, donde los símbolos se escogen al azar con una probabilidad p_i de escoger un 1 y una probabilidad $1 - p_i$ de escoger un -1. Nótese que la representación de estas secuencias en un random-walk estaría formada de tramos que, salvo fluctuaciones estadísticas, tendrían una pendiente $2\left(p_i - \frac{1}{2}\right)$ (figura 2.5.b).

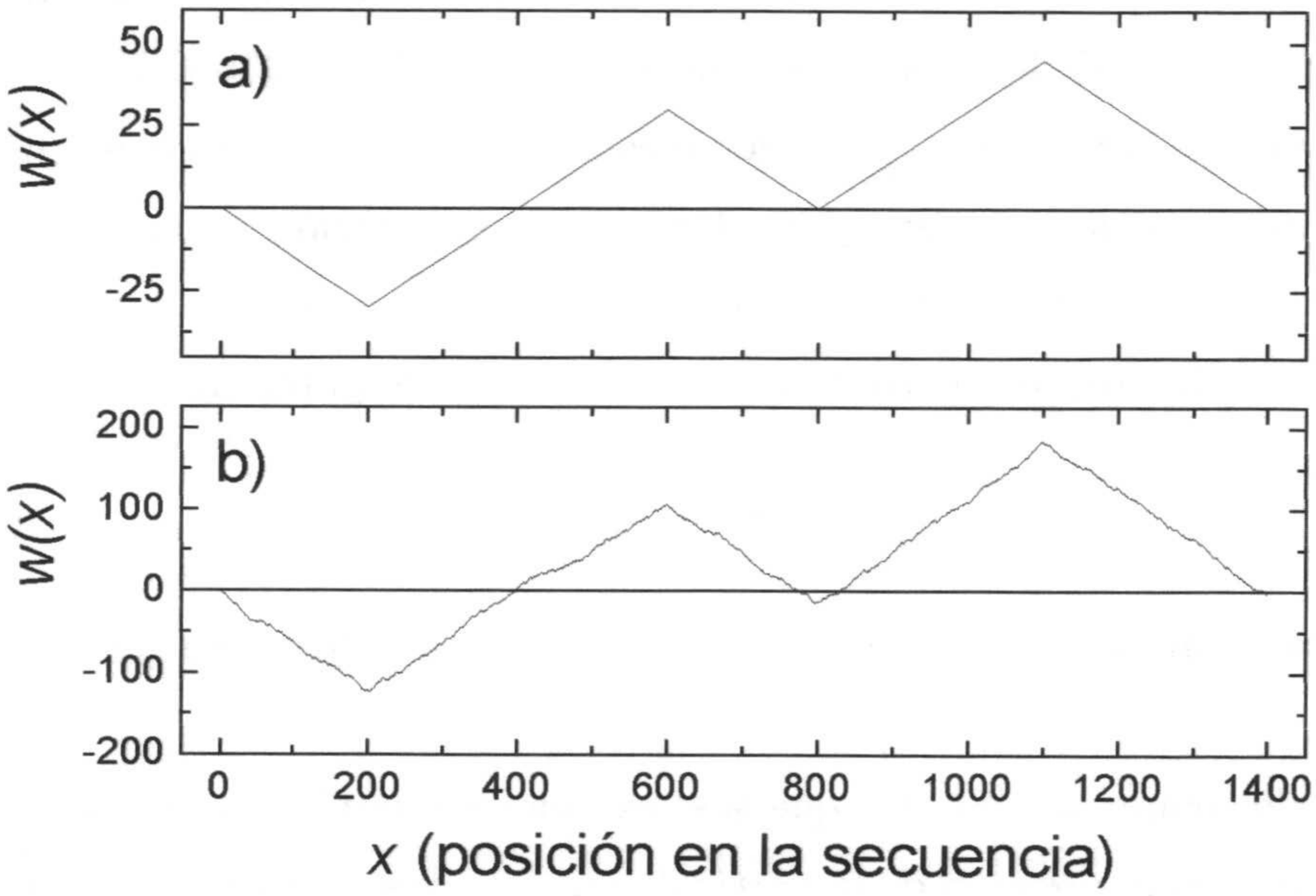

Figura 2.5: Ejemplo de Lévy-walk (a) y Lévy-walk generalizado (b).

El caso más simple, que es el que proponen los autores, consiste en tomar las probabilidades p_i de forma que los tramos tengan pendientes alternadas, i.e.

$$\begin{aligned} p_{2i} &= \frac{1+\epsilon}{2} \\ p_{2i-1} &= \frac{1-\epsilon}{2} \end{aligned} \quad \text{con } i = 1, 2, ... \text{ y } \ 0 \leq \epsilon \leq 1 \tag{2.28}$$

Se comprueba [Bu 93b] que, al igual que con el modelo original, se obtienen secuencias con correlaciones de largo alcance para $2 \leq \mu \leq 3$ y que, independientemente de la elección de los parámetros ℓ_0 y ϵ, para valores suficientemente grandes de ℓ, estas secuencias presentan un exponente α en el análisis DFA, que viene dado por:

$$\alpha = 2 - \frac{1}{2}\mu \tag{2.29}$$

que da valores de α en $[0.5, 1]$.

Con este modelo, que considera la secuencia de ADN formada por la concatenación de tramos no correlacionados con diferente composición, los autores intentan justificar la presencia de correlaciones de largo alcance como debidas a la inserción de retrotransposones. Éstos parecen jugar un papel importante en la evolución de las secuencias y se pueden considerar, en primera aproximación como secuencias no correlacionadas con diferentes composiciones de purinas-pirimidinas y que se insertan al azar. El uso de una ley de potencia (2.26) para escoger las longitudes de los tramos se justifica porque, para insertar un retrotransposón en una cadena de ADN, éste debe formar un bucle, y la probabilidad de encontrar un bucle de una determinada longitud en una cadena polimérica en disolución viene dada por una distribución del tipo de (2.26) [Des 80]; además parecen haberse encontrado indicios de que la distribución de tamaños de inserciones y delecciones es aproximadamente una ley de potencia [Gu 95].

El hecho de que las diferencias de composición más relevantes en estos elementos sean las debidas a la diferente proporción de purinas-pirimidinas es lo que justifica, para los autores, el uso de un modelo basado en secuencias binarias. Por último, decir que el modelo, da buenos resultados con la medida directa de la autocorrelación, el DFA y los espectros de potencia, utilizando valores de los parámetros similares a los observados en las cadenas reales.

2.4.4 Inserción-delección

Los mismos autores del modelo anterior proponen este otro [Bu 93a], algo más sofisticado, que tiene en cuenta la delección de ciertas subsecuencias de ADN, y el intercambio de material genético entre las dos hebras de la doble hélice, así como la inserción de retrotransposones que, según ellos son responsables del aumento de las correlaciones de largo alcance que se observa a medida que aumenta el grado de

complejidad del organismo cuyas secuencias de ADN se consideran.

El modelo parte de una secuencia binaria de longitud N, generada de forma aleatoria pero con composición sesgada (que se correspondería, según los autores, a un tramo de ADN codificador con exceso de purinas sobre pirimidinas o viceversa). Los autores toman secuencias con un porcentaje del orden del 60% de purinas, como suele ocurrir con las zonas codificadoras. A continuación, en sucesivos pasos, la secuencia va mutando de acuerdo con el siguiente procedimiento:

a) Se escoge un punto al azar en la secuencia y, a partir de él, se corta una subsecuencia de longitud ℓ, escogida al azar siguiendo una distribución hiperbólica (2.26)¶ con $\mu \approx 2$ y $\ell_0 = 20$. A continuación se escoge otro punto al azar en la secuencia y se inserta esta subsecuencia de longitud ℓ.

b) Al realizar esta inserción, con una probabilidad de 0.5, en lugar de la subsecuencia extraída, se inserta la complementaria (esto es, se cambian purinas por pirimidinas).

c) Para simular las inserciones de retrovirus, con cierta probabilidad $p_i \approx 0.2$, la subsecuencia que se debía insertar es sustituida por una secuencia aleatoria de la misma longitud y con el sesgo que tenía la secuencia de partida ($\sim 60\%$ de purinas).

Con este modelo se parte de una secuencia aleatoria ($\alpha = 0.5$) que, tras sucesivas iteraciones presenta exponentes cada vez más altos en el análisis con DFA, hasta alcanzar un estado estacionario. La aparición de correlaciones de largo alcance resulta del balance de dos tendencias opuestas: el cortar un segmento y colocarlo en

¶La elección de la distribución hiperbólica se justifica de la misma forma que en el modelo anterior.

otro sitio tiende a barajar la secuencia, mientras que la inserción de subsecuencias aleatorias sesgadas tiende a organizarlo.

Aunque no hacen un análisis teórico sobre la influencia de los parámetros sobre el resultado final, sí comprueban que no todos los posibles valores llevan a secuencias con correlaciones de largo alcance. En concreto, la exigencia de que la secuencia de partida sea sesgada es decisiva.

2.4.5 Pseudo-cromosomas

Como veremos más adelante, hay autores que buscan el origen de las correlaciones de largo alcance en efectos, que podríamos denominar de corto alcance como las periodicidades introducidas por el uso no homogéneo de codones dentro de las zonas codificadoras [Che 95], [He 97]. Para apoyar esta idea Herzel et al. [He 97] proponen lo que ellos llaman el modelo de pseudo-cromosomas.

En esencia, este modelo supone la secuencia formada por la alternancia de zonas codificadoras y no codificadoras, cuyos tamaños se extraen al azar de las distribuciones de longitudes observadas en secuencias reales. Las zonas codificadoras son una sucesión de tripletes de nucleótidos (codones) escogidos al azar, pero de forma que las probabilidades de aparición de cada nucleótido en cada una de las tres posiciones del codón sea la observada en secuencias reales; ya que, como es sabido, dentro de un misma fase cada nucleótido no aparece con la misma frecuencia en las tres posiciones. Por ejemplo, en la tabla II tenemos la frecuencia de aparición de cada nucleótido en cada una de las tres posiciones en la zona codificadora de la cadena pesada de la β-miosina humana (HUMBMYH7). Justamente son éstos los que se usan en algunos de los resultados de [He 97]. Además para simplificar el modelo y no introducir los efectos debidos a la diferencia de composición de zonas codificadoras y no codificadoras, la composición global de las no codificadoras es la misma que la de las codificadoras.

Tabla II.

	Posición 1	Posición 2	Posición 3
A	0.296	0.437	0.079
T	0.105	0.256	0.107
C	0.248	0.184	0.343
G	0.351	0.123	0.471

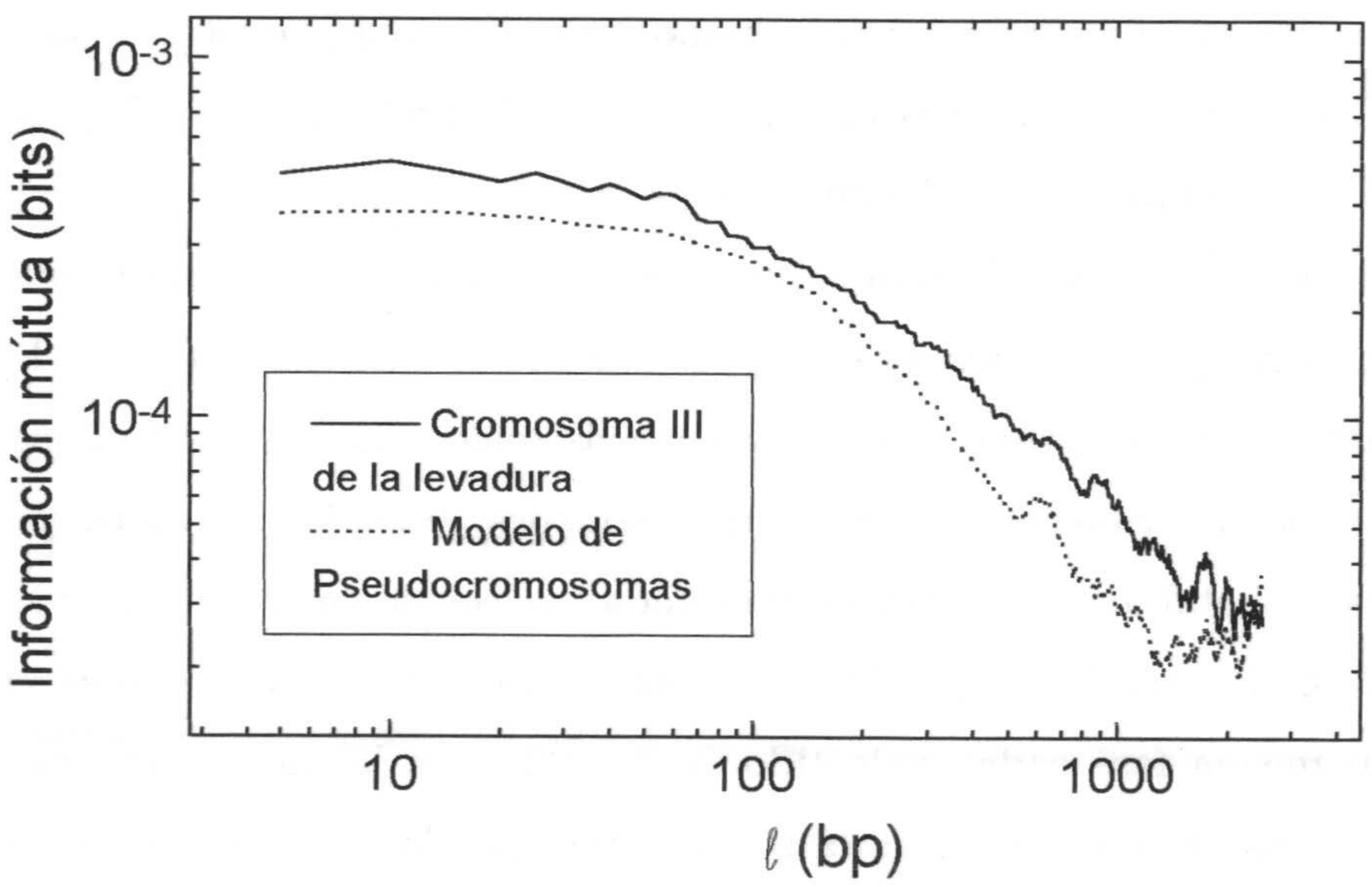

Figura 2.6: Información mútua en función de la longitud para el cromosoma III de la levadura y una secuencia generada con el modelo de pseudocromosomas.

En la figura 2.21 se han representado los valores de la información mutua en

función de la longitud para el cromosoma III de la levadura de la cerveza (*Saccharomyces cerevisiæ)* y una secuencia de la misma longitud generada con este modelo. Las probabilidades de cada posición se han tomado de la zona codificadora más larga encontrada la secuencia real (6504 bp, desde la posición 178.251 hasta 184.754) y los tamaños de zonas codificadoras y no codificadoras son los mismos que en la secuencia real. A causa de las correlaciones introducidas para longitudes múltiplo de 3, que no aparecen para el resto de longitudes, este tipo de gráficas presentan muchas fluctuaciones. Para poder observar la tendencia global los datos del gráfico se han suavizado usando 20 valores consecutivos, técnica que también se usa en [He 97]. Como se puede ver, el acuerdo entre la secuencia real y el modelo es bastante razonable, teniendo en cuenta las simplificaciones de éste.

Sin embargo, el modelo no reproduce los resultados obtenidos con el DFA (figura 2.7). Según los autores, esto es debido a que el DFA es una magnitud acumulada y no tiene en cuenta el efecto de las correlaciones que aparecen sólo a distancias múltiplo de 3.

2.5 El estado de la cuestión

2.5.1 ¿Existe realmente estructura fractal en el ADN?

Desde el año 92 en que se publicaron los primeros trabajos referentes a la existencia de correlaciones de largo alcance en secuencias de ADN ha sido grande el interés y la controversia que ha producido esta cuestión dentro de la comunidad científica (véase por ejemplo para una revisión [Li 94], [Li 97d], [Bu 94]).

La primera cuestión que se planteó fue si las correlaciones encontradas en las secuencias de ADN eran realmente fruto de la existencia de una estructura fractal, y que por tanto, podían tener relación con aspectos fundamentales del genoma, por el momento desconocidos [Li 92a], [Mad 92], [Mun 92], [Vo 92], [Ya 92], o, por

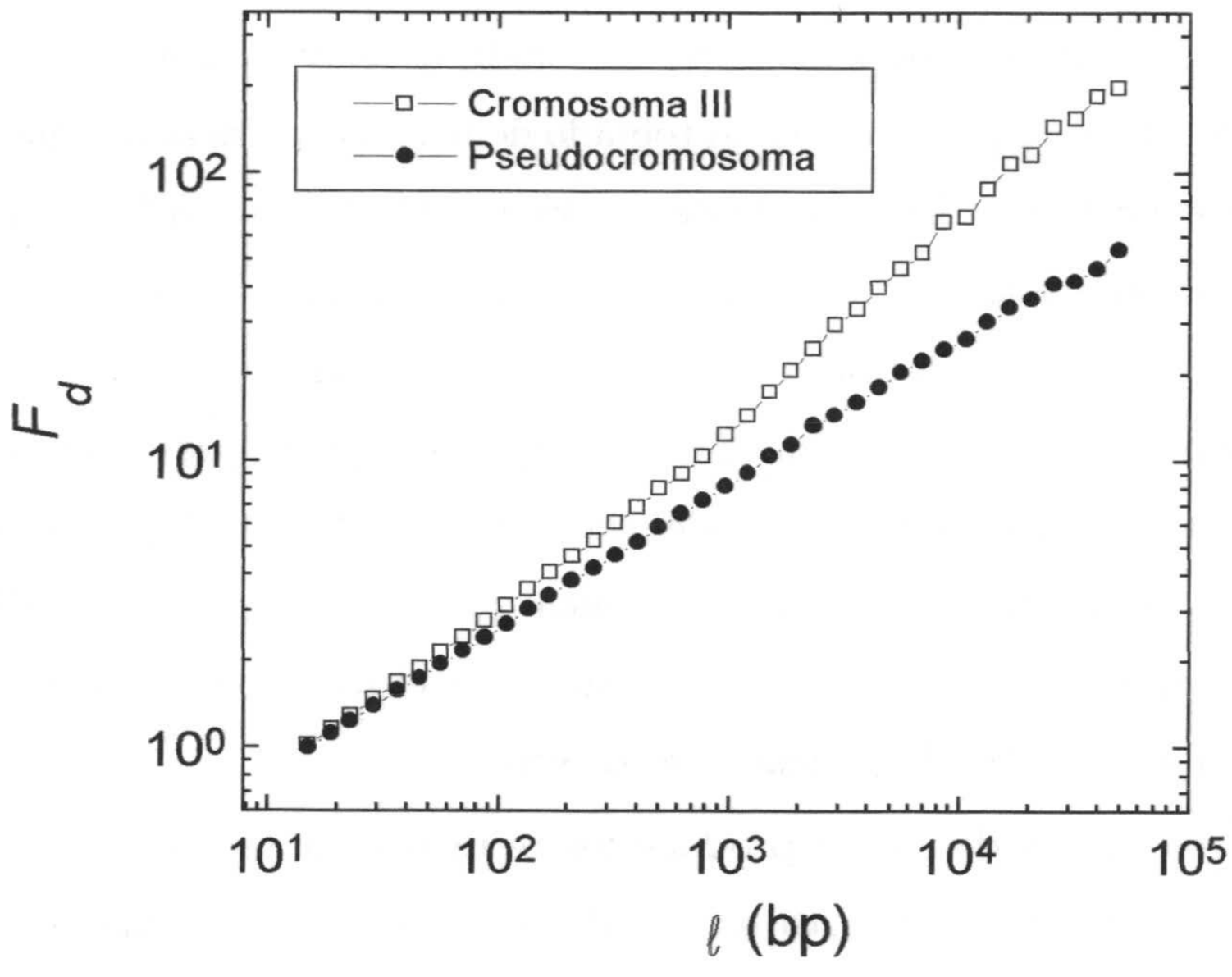

Figura 2.7: Análisis con DFA para el cromosoma III de la levadura y una secuencia generada con el modelo de pseudocromosomas.

el contrario, podían ser trivialmente explicadas por la presencia de heterogeneidades composicionales (zonas con diferente composición) [Ka 93], [Cha 94], [La 93], [Bor 93].

Karlin y Brendel [Ka 93], ponen serias objeciones a los resultados obtenidos, centrando la crítica en los métodos utilizados (previos al DFA) que, en principio, sólo son aplicables a secuencias estacionarias y las secuencias de ADN son heterogéneas a todas las escalas. Para ellos, las correlaciones encontradas son sólo un artefacto de cálculo causado por la no estacionariedad de las secuencias analizadas. Esta opinión es compartida por Larharmmar y Chatzidimitriou-Dreisman [Cha 94],

[La 93], aunque no tan bien fundamentada. Estos autores demuestran, mediante simulaciones por ordenador, que efectivamente las correlaciones observadas en algunas secuencias, pueden ser trivialmente explicadas por la existencia de zonas con diferente composición.

Por último Borŝtnic et al. [Bor 93] plantean la posibilidad de que los análisis realizados con espectros de potencia, que según sus autores, [Li 92c], [Vo 92] muestran la presencia de una componente del tipo $1/f$, son simplemente espectros lorentzianos, también causados trivialmente por la alternancia de zonas con diferente composición.

Esta controversia, por el momento, parece haber terminado con la introducción del DFA [Pe 94], que elimina el efecto de la no estacionariedad. Con este método las secuencias que, según los autores antes citados, tienen correlaciones trivialmente causadas por la heterogeneidad aparecen con exponente de escala $\alpha = 0.5$ (ausencia de correlaciones de largo alcance), pero aún eliminando el efecto de la no estacionariedad sigue habiendo gran número de secuencias que siguen presentado estas correlaciones. Es importante destacar que ninguna de estas secuencias, que siguen manteniendo las correlaciones una vez aplicado el DFA, fueron analizadas por los autores que querían justificarlas trivialmente por la sucesión de tramos de distinta composición.

De todas formas, no se ha descartado que el origen de las correlaciones de largo alcance esté en la alternancia de zonas con distinta composición, pero no cualquier alternancia. Buldyrev et al. [Bu 93b] propusieron el modelo del Lévy-walk generalizado (sec. 2.4.3) y comprobaron que un conjunto de dominios de diferente composición lleva a correlaciones de largo alcance tipo fractal si la distribución de tamaños de los dominios es una ley de potencia, además justifican la existencia de esta ley de potencia a causa de las inserciones de retrovirus, cuyos tamaños

característicos parecen seguir este tipo de distribución [Gu 95].

En esta misma línea algunos autores [Nee 92], [Du 95], [Ce 95], [Az 95] buscan el origen de estas correlaciones en la alternancia de zonas codificadoras y no codificadoras. En la mayoría de los casos los estudios se centran en los cromosomas de la levadura que fueron los primeros cromosomas completamente secuenciados y que contienen una proporción equilibrada de zona codificadoras y no codificadoras (70 y 30% respectivamente). Además, en este genoma la distribución de tamaños de las zonas codificadoras [Duj 96] se extiende bastante, pudiendo incluso ajustarse a una ley de potencia, aunque el mejor ajuste se obtiene para una ley exponencial. En cualquier caso, se trataría de una distribución exponencial con una longitud característica bastante grande (del orden de 1000 bp) que, dado el tamaño finito de las secuencias, podría producir resultados similares a los obtenidos con una distribución de dominios siguiendo una ley de potencia. De todas formas esta justificación no sería válida para los genomas de vertebrados superiores en los que la proporción de zona codificadora en muchos casos no alcanza el 5%.

Este último punto de vista ha sido recientemente rebatido por Viswanathan et al. [Vi 97], quienes encuentran que el efecto de la alternancia de zonas codificadoras y no codificadoras no es suficiente para explicar las correlaciones, ni siquiera en los cromosomas de la levadura. De hecho, encuentran que existe una dependencia tipo ley de potencia en la correlación que es "interrumpida" por la presencia de las longitudes características de las zonas codificadoras.

Las últimas aportaciones van algo más allá y sostienen que no es solamente alternancia de zonas con diferente composición lo que se encuentra en las secuencias de ADN que presentan propiedades fractales (lo que Wentian Li denomina heterogeneidad simple o dominios después de dominios [Li 97d]), sino que existe una estructura jerárquica de dominios dentro de dominios (heterogeneidad compleja), esto

es, las secuencias se pueden dividir en zonas de diferente composición, que pueden ser relativamente homogéneas, pero que, observadas con mayor detalle, también se pueden subdividir en zonas más pequeñas [Be 96]. Independientemente, otros autores [Ts 95], dudan que los resultados obtenidos con DFA indiquen la existencia de un único exponente de escala ya que en muchos casos el mejor ajuste de los datos en doble escala logarítmica no es una recta, lo que podría indicar la presencia de una estructura multifractal y que podría estar de acuerdo con la idea de una estructura jerárquica de dominios dentro de dominios pero, por el momento no hay evidencias claras [Ts 96].

2.5.2 Secuencias codificadoras y no codificadoras

Otra de las cuestiones que más llamó la atención fue que, según dos de los tres trabajos pioneros [Li 92c], [Pe 92], las correlaciones de largo alcance se encontraban preferentemente en secuencias de ADN que contenían zona no codificadora (intrones o regiones intergénicas). Esta afirmación produjo cierta controversia [Bu 93c], [Vo 93], [Vo 94] ya que el tercero de estos autores, Richard Voss, sostenía que las correlaciones de largo alcance aparecían en todo tipo de secuencias de ADN, encontrando sólo diferencia entre secuencias pertenecientes a distintos organismos. Según él, la estructura fractal es una propiedad común a todas las secuencias de ADN, independientemente de su función, y supone un mecanismo inmunidad ante el error a todas las escalas. Hay que decir que los trabajos de Voss son previos al DFA, con lo que en sus análisis las secuencias con estructura trivial de dominios pueden presentar correlaciones de largo alcance (que analizadas con el DFA desaparecerían). Otra cuestión, que puede ser la responsable de las discrepancias, es que este autor analiza las secuencias promediando sobre las reglas de asignación 4,5,6,7 (sec. 2.3.1), con lo que tiene en cuenta las dependencias entre los cuatro nucleótidos y las correlaciones que mejor se observan, y que más diferencia muestran entre zona codificadora y no

codificadora, son las correspondientes a la regla R-Y.

Una de las últimas aportaciones con respecto a esta cuestión [Bu 95] presenta un análisis de todas las secuencias disponibles en el *Gen-Bank* hasta ese momento y obtiene resultados bastante convincentes que apoyan la idea de las correlaciones de largo alcance aparecen preferentemente en zonas no codificadoras. Además, otro dato en favor de esta hipótesis son los resultados, relativamente buenos, que han obtenido otros autores del mismo grupo utilizando el análisis de estas correlaciones como método para identificar zonas codificadoras [Os 94].

Por el momento no se ha encontrado una justificación comúnmente aceptada para esta observación, quizá la más simple sea que las zonas no codificadoras tienen una mayor tolerancia evolutiva a las mutaciones, duplicaciones e inserciones, por lo que parece razonable que puedan tener una estructura más redundante [Li 97d]. De todas formas, como se verá en el capítulo 7, y de acuerdo con algunos resultados previos [Vi 97], [Be 96], la existencia de ADN repetido no es suficiente para justificar la complejidad encontrada en las secuencias no codificadoras. Otra explicación a esta mayor complejidad, que ha sido recientemente propuesta, es la posibilidad de la existencia de algún tipo de lenguaje oculto en las zonas no codificadoras [Ma 94], [Cz 95], [Cz 96]. Sin embargo, este punto ha sido fuertemente cuestionado [Bon 96], [Cha 96], [Is 96], [Li 96], [Vo 96].

Por otra parte las zonas codificadoras, además de tener mayores restricciones evolutivas, deben tener estructuras más homogéneas composicionalmente (más cercanas a la aleatoriedad) con objeto de producir proteínas estables [Che 96], de hecho, para que una cadena de aminoácidos pueda alcanzar una configuración de mínima energía estable es suficiente que esté formada por una secuencia no correlacionada de aminoácidos [Sha 90].

2.5.3 Evolución molecular

Por último vamos a ver otra cuestión interesante planteada en la bibliografía: la posibilidad de que la complejidad fractal de las secuencias de ADN incremente con la evolución. Buldyrev et al. [Bu 93a] observan que las secuencias que contienen la zona codificadora de una misma proteína en distintos organismos, a medida que el organismo es más complejo, presentan una mayor proporción de zona no codificadora y además los perfiles del DNA-walk de estas secuencias se van haciendo más abruptos. Para confirmar estas observaciones, calculan el exponente de escala con el DFA para las secuencias que contienen la zona codificadora de algunas proteínas de diversos organismos, encontrando que el exponente de escala aumenta con la complejidad organísmica. También observan que el incremento de la complejidad fractal no sólo depende de la mayor proporción de zona no codificadora, ya que secuencias con una misma proporción tienen exponentes distintos dependiendo del organismo al que pertenecen, sino que parece haber un incremento de la complejidad intrínseca de estas zonas con la evolución.

Para justificar esto proponen el modelo de inserción-delección (sec. 2.4.4) que explicaría el incremento del exponente de escala con la evolución, independientemente de la proporción de zona no codificadora. Otra cuestión importante que ponen de manifiesto estos autores es la importancia del análisis con una u otra regla de asignación. Todos estos resultados se obtienen con la regla R-Y, mientras que con la regla S-W no. Para ellos, esto justifica el modelo de inserción delección, ya que la copia de la secuencia complementaria modificaría sólo las proporciones de purinas-pirimidinas no de AT-CG (a causa de la complementariedad de las bases A-T y C-G).

Capítulo 3

Análisis de Secuencias con Perfiles Entrópicos

3.1 Introducción

En este capítulo se describen las primeras aportaciones de nuestro grupo al análisis de secuencias de ADN [Be 94], [Ol 93] que, posteriormente serían generalizadas al análisis de otro tipo de secuencias [Ro 93b], [Ro 94]. En esencia, el método consiste en el análisis de lo que en Estadística se suele llamar histogramas de segundo orden. El histograma de primer orden o simplemente, histograma, de una fuente tiene en cuenta la frecuencia de aparición de cada símbolo o conjunto de símbolos emitidos, mientras que el histograma de segundo orden, cuenta el número símbolos que han sido emitidos un número dado de veces. En términos más intuitivos, si el histograma de primer orden nos dice la frecuencia con la que se emite cada símbolo, el de segundo orden, además nos da cuenta de cómo fluctúan estos valores alrededor del valor medio cuando consideramos un gran número de secuencias emitidas por la misma fuente.

Para cuantificar las diferencias o similitudes entre los histogramas de diversas secuencias o para comparar con modelos teóricos utilizaremos la entropía de Shannon aplicada sobre el histograma de segundo orden, convenientemente normalizado.

El desarrollo del método que aquí se presenta parte de la aplicación de métodos de procesamiento de imágenes digitales ciertas imágenes fractales que se obtienen a partir de las secuencias. Aunque la introducción de estas imágenes no es estrictamente necesaria para el desarrollo del método, facilitan considerablemente la interpretación de algunos resultados y hacen más intuitivas determinadas deducciones de distribuciones de probabilidad. Además este fue el enfoque original de los trabajos de nuestro grupo que hemos citado antes.

El capítulo se organiza como sigue: en primer lugar se describen algunos elementos de la Teoría del Caos que se utilizan para introducir las imágenes fractales de las que antes hemos hablado y se describen algunas de sus propiedades (secciones 3.2 y 3.3). A continuación se introduce el histograma de segundo orden que, en este tratamiento, denominaremos *histograma del CSR* y se calculan los histogramas esperados para secuencias aleatorias equiprobables y no equiprobables (sección 3.4). En la sección 3.5 se introduce el uso de la entropía de Shannon del histograma y se calculan y sus cotas, así como una normalización que elimina la dependencia de esta medida con la longitud de la secuencia analizada. La sección 3.6 se dedica al estudio multi-longitud de la secuencia y finalmente se muestran algunos ejemplo de la aplicación del método a secuencias de ADN (sección 3.7).

3.2 Sistemas de Funciones Iteradas (IFS)

Los aspectos de la Teoría del Caos utilizados en este capítulo se describen en [Bar 88] (véase también [Pei 92], para una introducción más intuitiva, aunque no menos rigurosa), y en esencia, es lo que se conoce como *Sistema de Funciones Iteradas*

(IFS), que son un conjunto de aplicaciones $\{w_i\}$ con $i = 1, 2, .., n$ sobre un espacio métrico (**X**,d), que cumplen la propiedad:

$$\begin{aligned} \forall i \;\; \exists c_i \text{ con } 0 \;\; &\leq \;\; c_i < 1 \;\; / \\ \forall x, y \;\; &\in \;\; \mathbf{X} \qquad \mathrm{d}(w_i(x), w_i(y)) \leq c_i \; \mathrm{d}(x, y) \end{aligned} \tag{3.1}$$

donde los c_i se conocen como *factores de contracción.* A consecuencia de esta propiedad reciben el nombre de *contracciones* o *aplicaciones contractivas.* Se puede probar de forma inmediata que estas aplicaciones son continuas, de hecho son mucho más que continuas. Por simplicidad, se suele restringir la atención a los IFSs en los que el espacio métrico es $\Re^M$, siendo entonces las funciones, aplicaciones afines contractivas, y además, en muchos casos se hace $M = 2$, lo que permite una representación visual simple. En estas condiciones los IFS se pueden representar como:

$$w_i \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_{11}^i & a_{12}^i \\ a_{21}^i & a_{22}^i \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} b_1^i \\ b_2^i \end{bmatrix} \tag{3.2}$$

Cuando se aplican estas funciones forma iterada sobre un punto, el conjunto de puntos resultante, dependiendo del tipo de funciones y a veces del orden en que se aplican, puede formar imágenes tipo fractal. Por ejemplo, el conocido Triángulo de Sierpinsky se puede obtener escogiendo los parámetros de (3.2) como:

i	a_{11}^i	a_{12}^i	a_{21}^i	a_{22}^i	b_1^i	b_2^i
1	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	0
2	$\frac{1}{2}$	0	0	$\frac{1}{2}$	1	0
3	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$

(nótese que los $\mathbf{b}^i$ son los vértices de un triángulo equilátero de lado 1) y eligiendo al azar la función que se aplica en cada caso, p.e.:

$$p_0 = (\frac{1}{2}, \frac{1}{2})$$

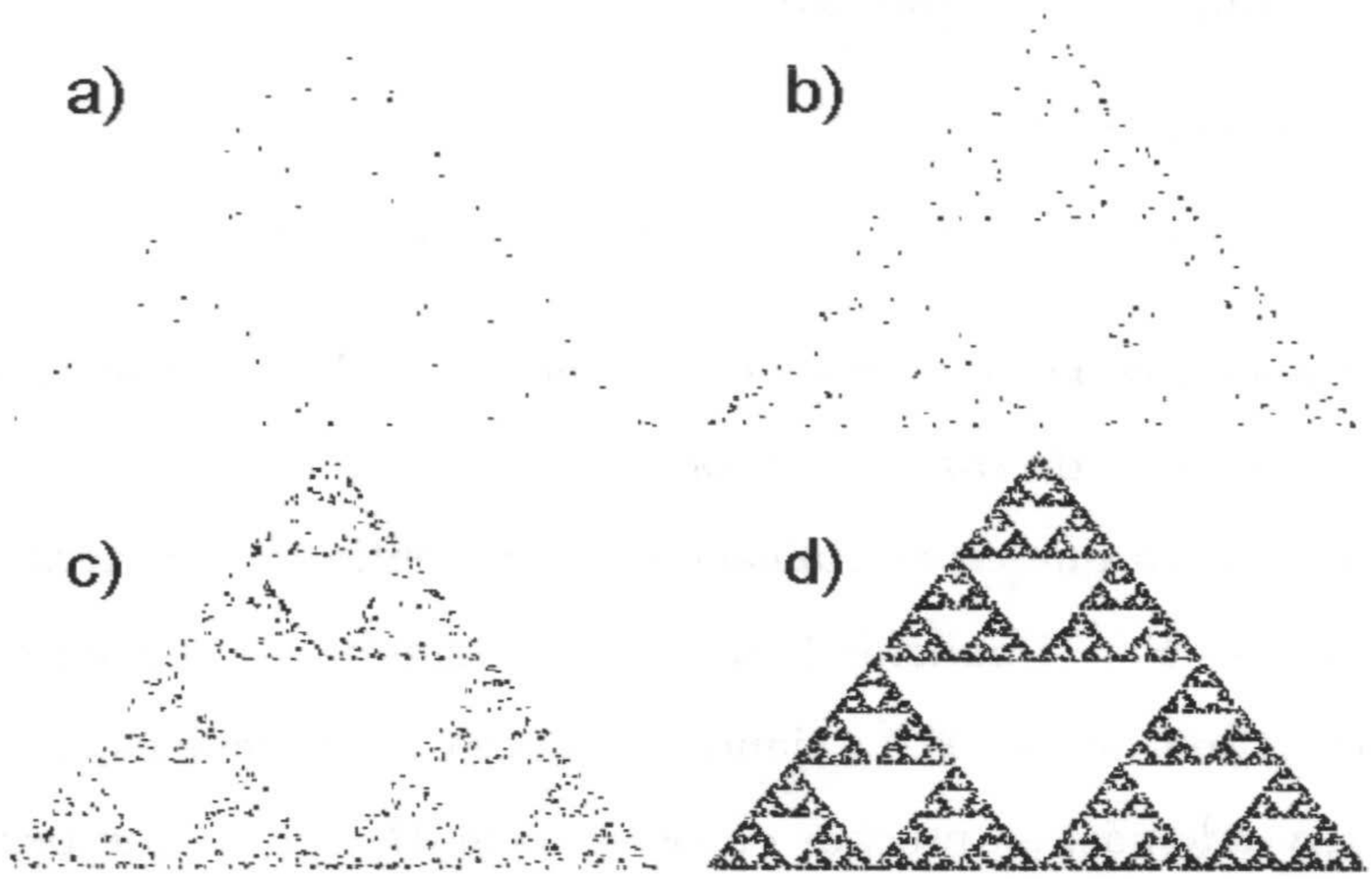

Figura 3.1: Generación del triángulo de Sierpinsky con un IFS: a) 150, b) 500, c) 2000 y d) 15000 iteraciones.

$$\begin{aligned} p_1 &= w_2(p_0) = (\frac{7}{4}, \frac{1}{4}) \\ p_2 &= w_1(p_1) = (\frac{7}{8}, \frac{1}{8}) \\ &\ldots \end{aligned}$$

Es fácil comprobar que el efecto de estas contracciones es equivalente a escoger uno de los tres vértices del triángulo y tomar como nuevo punto el punto medio entre éste vértice y el punto anterior. Además se puede comprobar que la imagen que resulta tras un número suficientemente grande de iteraciones es independiente de la elección del punto inicial. La figura 3.1 representa los puntos obtenidos con este método después de 150, 500, 2000 y 15000 iteraciones.

3.3 Representación Caótica de Secuencias (CSR)

Puesto que, en principio, parece que la imagen resultante de la aplicación iterada del IFS depende de la forma en que se escogen las funciones, surgió la idea de aplicar los IFS como técnica de análisis de secuencias [Je 90]: En lugar de escoger las funciones al azar, se toma una secuencia de símbolos pertenecientes a un alfabeto con n elementos (número de funciones del IFS) de forma que sea la secuencia la que *controle* el IFS, i.e. cada símbolo de la secuencia determina la función que se ha de aplicar.

Con el IFS anterior, se obtiene el triángulo de Sierpinsky independientemente de la forma de escoger las funciones, siempre que el número de apariciones de una función no sea excesivamente pequeño, en cuyo caso, seguirá apareciendo el triángulo pero con la zona cercana al vértice correspondientes algo más "clara". Por tanto, no parece que sea muy útil para extraer información sobre la secuencia que lo controla, sin embargo, esto no ocurre siempre. En concreto los IFSs que se definen a continuación, son especialmente interesantes para este objetivo:

1. $\left[a^i_{jk}\right] = \frac{1}{2}I$ con $i = 1, ..., n$ y $j, k = 1, ..., M$. (como en el ejemplo anterior).
2. El número de funciones (n) será el número de vértices de un hipercubo M-dimensional (2^M).
3. Los vectores $\mathbf{b}^i$ serán las posiciones de los vértices del hipercubo. Se puede comprobar que en estas condiciones, la figura resultante llena el hipercubo si las funciones se escogen de forma aleatoria, asignando la misma probabilidad a todas ellas, mientras que en otro caso aparecen figuras tipo fractal. En este caso también es fácil comprobar que la aplicación de una de las contracciones a un punto, equivale a calcular el punto medio entre éste y el vértice asociado a la contracción.

En lo sucesivo, nos referiremos a este tipo de IFSs, controlados por una secuencia como **Representación Caótica de la Secuencia** o, simplemente, **CSR de la secuencia** (las siglas provienen del Ingles: Chaos Sequence Representation).

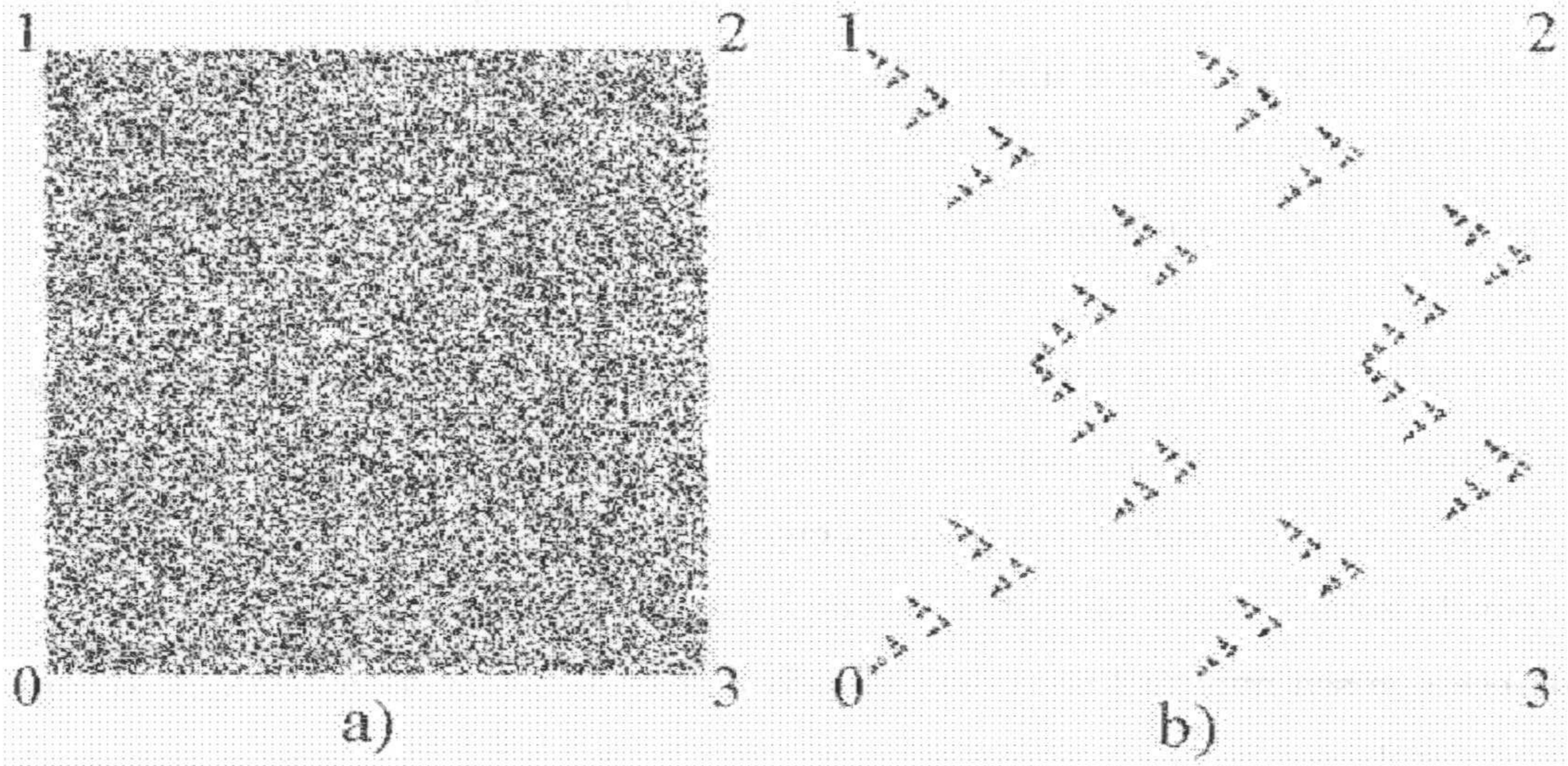

Figura 3.2: CSR para dos secuencias pseudoaleatorias de enteros módulo 4. a) Generador A-Carry. b) Generador Logis-1.

Veamos ahora algunos ejemplos que ponen de manifiesto la utilidad de este IFS para el análisis de secuencias. En primer lugar se han generado dos secuencias de 30.000 enteros módulo 4 (equiprobables) primero con un generador pseudoaleatorio considerado como muy bueno en la bibliografía, el A-Carry [Ja 90] y después con otro generador basado en una transformación de la variable logística (logis-1) [Col 92], que presenta bastante correlación entre sucesivas llamadas, pero que, sin embargo no presenta ningún problema a la hora de generar reales equidistribuidos en [0,1). En la figura 3.2 aparecen las imágenes generadas en ambos casos por el IFS (con $M = 2$, ya que el alfabeto en este caso es 4). El A-Carry llena el cuadrado de forma

homogénea sin ningún patrón visible, mientras que el Logis-1 presenta unos patrones muy claros. Estas desviaciones son debidas a la correlación entre sucesivas "tiradas" del Logis-1, esto se comprueba viendo como se hacen menos notorios estos patrones si en lugar de generar la secuencia con todos los valores que produce Logis-1, se toma, p.e. sólo uno de cada 10. De hecho este "espaciamiento" entre las salidas del generador es recomendado por el autor que lo propone [Col 92] para mejorar la calidad del algoritmo.

En la figura 3.3 vemos otro ejemplo, ahora se han generado dos secuencias pseudoaleatorias (ambas con el A-Carry), y ahora en la primera se hace que los símbolos no sean equiprobables sino que se asignan las probabilidades: $p_0 = p_1 = p_2 = 0.32$ y $p_3 = 0.04$; y en la segunda los cuatro símbolos son equiprobables pero se han modificado las probabilidades condicionadas de forma que cuando aparece un 1, en la tirada siguiente el 3 es mucho menos probable.

3.3.1 Propiedades del CSR.

En los ejemplos anteriores vemos que, al utilizar secuencias en las que los símbolos no son equiprobables (fig. 3.3 a) o no son independientes (3.2 b, 3.3 b), aparecen curiosas figuras en el CSR. En este apartado vamos a justificar el aspecto de estas figuras y la interpretación del CSR en términos de la estadística de la secuencia. Para la discusión que sigue, nos limitaremos a la situación con $M = 2$, por la sencillez de representación, aunque los resultados son fácilmente generalizables.

- Todos y cada uno de los símbolos de la secuencia corresponden a un punto del CSR. El recíproco no es cierto para secuencias finitas, por ejemplo, dado un punto inicial, p_0 un número finito de iteraciones nunca puede hacerlo coincidir con un vértice (a no ser que p_0 lo fuese), la única forma de "llegar" un vértice sería mediante una serie infinita de repeticiones del símbolo correspondiente a

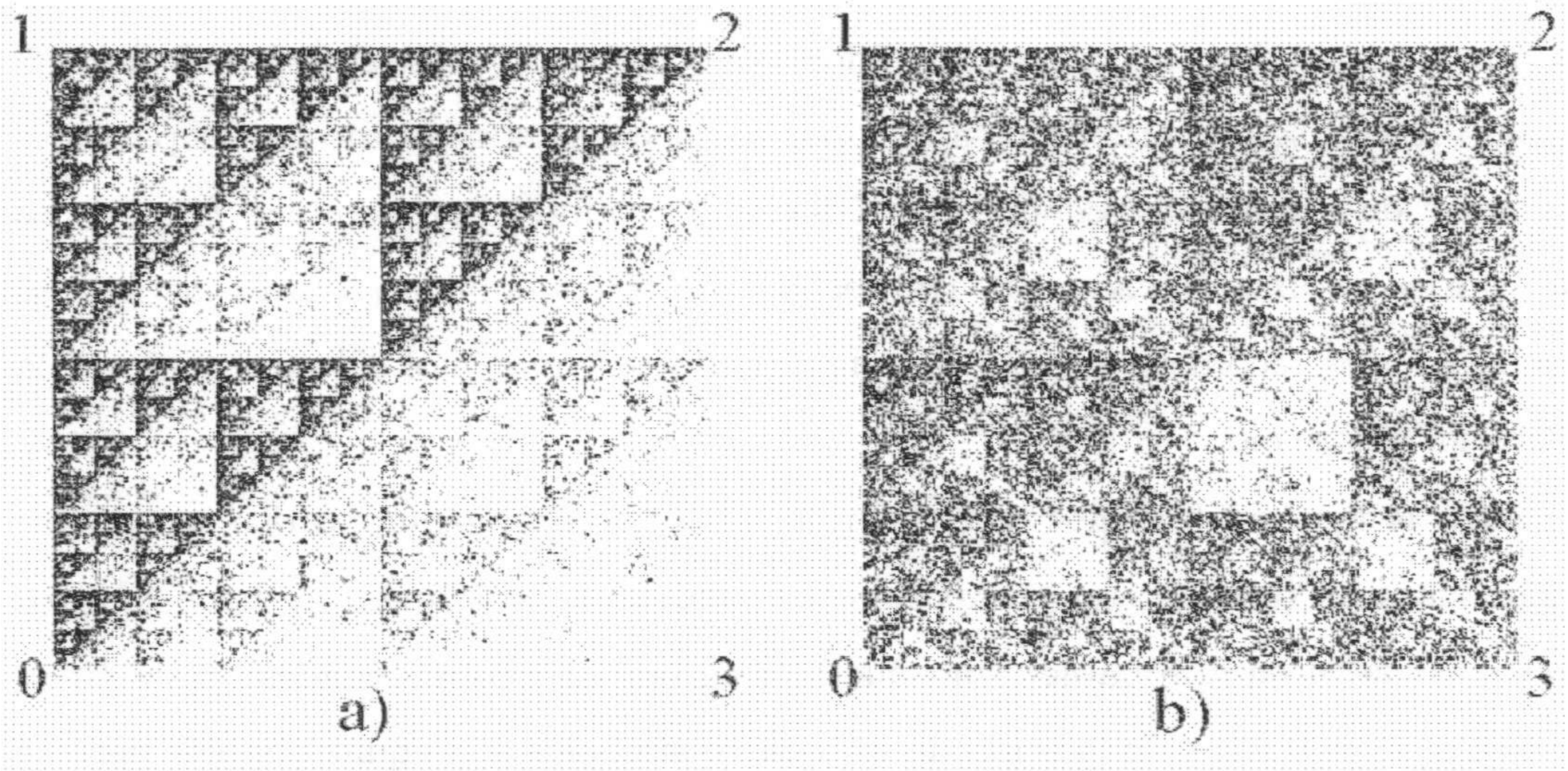

Figura 3.3: CSR para dos secuencias pseudoaleatorias de enteros módulo 4, en ambos casos se utiliza el generador A-Carry. a) Las cuatro contracciones no son equiprobables. b) Se modifican las probabillidades condicionadas.

ese vértice.

- En el proceso de trazado del CSR, cada punto se puede situar conociendo el punto anterior y el correspondiente símbolo que sigue en la secuencia.
- Las coordenadas de un punto del CSR permiten recuperar la secuencia completa que nos llevó a él, ya que las aplicaciones son inyectivas y siempre se pueden invertir. Por tanto, el último punto del CSR determina la secuencia completa. Esta afirmación puede resultar extraña, ya que implica que se puede almacenar toda la secuencia en sólo dos números reales, pero no hay que olvidar que dos números reales, conocidos con toda precisión, contienen infinito número de cifras decimales.

- Dividiendo el CSR en 4^k subcuadrados de lado $1/2^k$ (suponiendo un cuadrado inicial de lado 1), el número de puntos contenidos en cada uno de estos subcuadrados, es el número de veces que ha aparecido en la secuencia cada una de las 4^k subsecuencias posibles de longitud k (teniendo en cuenta solapamientos). De hecho, en relación con la propiedad anterior, la localización de la posición de un punto del CSR dentro de uno de los subcuadrados (i.e. conocer su posición con una precisión de $1/2^k$), permite reconstruir los últimos k símbolos de la secuencia.

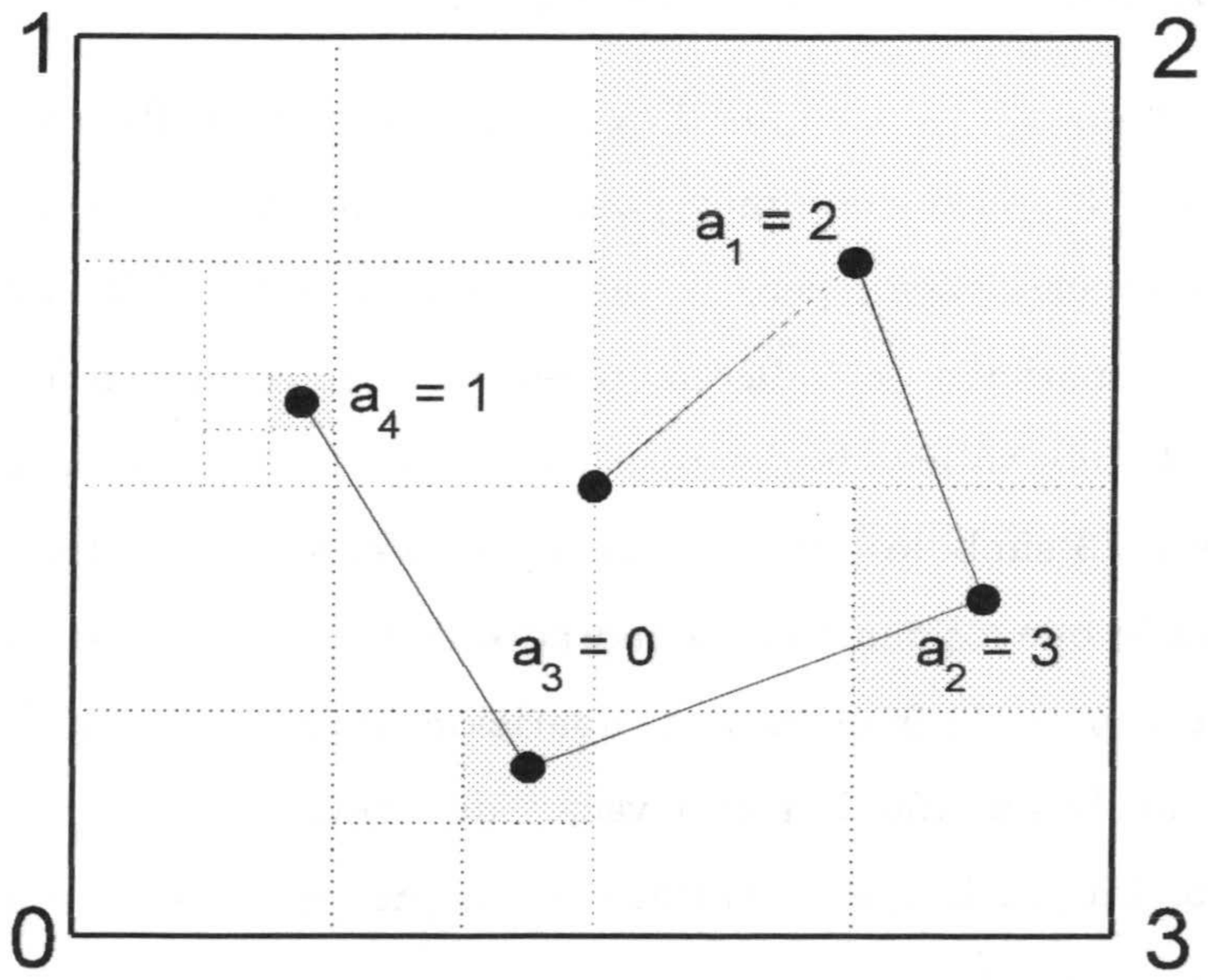

Figura 3.4: Primeras iteraciones de un CSR controlado por la secuencia {2,3,0,1,...}.

Con el esquema de la figura 3.4 comprenderá mejor esta afirmación. Supongamos una secuencia de enteros módulo 4 que comienza {2,3,0,1,...}, tomamos el

centro como punto inicial (no está localizado en ningún subcuadrado), a continuación se aplica la contracción 2, que nos da el punto medio entre el centro y el vértice 2. Ahora el punto está localizado en un subcuadrado de lado 1/2. La siguiente contracción (3) nos lleva a un punto situado en el subcuadrado inferior derecho, pero ahora está localizado dentro de este subcuadrado en el superior derecho (de lado 1/4), si proseguimos con la secuencia, los puntos que van apareciendo se encuentran localizados en subcuadrados cada vez menores. Además es fácil comprobar que, independientemente del punto de partida, esto es, de los símbolos anteriores, si aparece un 2 en la secuencia, el siguiente punto caerá dentro del cuadrado grande sombreado, si aparecen 23, caerá en el más pequeño, etc.

Teniendo en cuenta esto, se justifica el aspecto de los CSRs de la figura 3.3. En el primer caso, el 3 aparece con muy poca frecuencia, por tanto el subcuadrado inferior derecho está casi vacío, pero la pequeña aparición del 3 influye también en la frecuencia de aparición de todas las parejas que lo contienen lo que también están casi vacíos algunos subcuadrados más pequeños, lo mismo se puede decir para las subsecuencias de 3 símbolos, etc. En el segundo caso, los cuatro símbolos son equiprobables, por lo que los cuatro subcuadrados de tamaño 1/2 tienen un número de puntos parecidos pero, la frecuencia de aparición de la pareja {1,3} es pequeña, de ahí el cuadrado de tamaño 1/4 casi vacío; de nuevo el efecto se transmite a subsecuencias más largas, lo que se traduce en la presencia de subcuadrados más pequeños con pocos puntos.

Puesto que, según lo visto, el número de puntos dentro de cada subcuadrado de lado $1/2^k$ se corresponde con el número de apariciones de cada una de las 4^k subsecuencias de longitud k, cada uno de estos subcuadrados se pueden identificar con una de las subsecuencias. La forma natural de hacerlo es identificando cada uno de los cuatro símbolos del alfabeto $\{A_0, A_1, A_2, A_3\}$ con los enteros $\{0, 1, 2, 3\}$. De esta

forma, dada una subsecuencia de longitud k, $A_{i_1}, A_{i_2}, ..., A_{i_k}$ y su correspondiente secuencia de enteros $s_{i_1}, s_{i_2}, ..., s_{i_k}$, le asignamos el entero:

$$4^0 s_{i_1} + 4^1 s_{i_2} + ... + 4^{k-1} s_{i_k} \tag{3.3}$$

que evidentemente está comprendido entre 0 y el número de subcuadrados -1 ($4^k - 1$), y es único para cada una de las posibles subsecuencias de longitud k. Este "etiquetado" de las casillas, no es demasiado relevante para el desarrollo teórico que viene a continuación, aunque sí es bastante conveniente a la hora de implementar los cálculos.

3.4 Histograma del CSR

El CSR tal y como se ha descrito en el capítulo anterior, nos ofrece la posibilidad de un tratamiento similar al de las imágenes digitales. Para analizar imágenes por ordenador, en primer lugar éstas se digitalizan, esto es: la imagen se divide en pequeñas regiones llamadas píxeles (del inglés picture elements), siendo la división más común en rectángulos iguales mediante un enrejillado. Si la imagen es en blanco y negro a cada pixel le asociamos un número entero que nos da información sobre el brillo u oscuridad de este pixel (en la imagen original el brillo no estaría cuantizado, sino que sería un continuo), este número se suele conocer como nivel de gris del pixel. Con esta representación, a efectos de procesamiento digital, una imagen no es más que un array de enteros, de forma que cada elemento del array representa un pixel y su valor el nivel de gris correspondiente. En procesamiento digital de imágenes se utiliza el término nivel de resolución para indicar de alguna forma el tamaño de los píxeles. En una imagen dada, la máxima resolución con la que podemos observarla es aquella en la que cada pixel tiene solamente dos posibles niveles de gris, 0 ó 1 (que correspondería a negro y blanco respectivamente); tomando n de estos píxeles con niveles 0 ó 1 tendríamos un pixel del nivel de resolución inmediatamente inferior y

así iríamos agrupando hasta que la imagen quedase formada por un solo pixel. Como convenio se suele decir que en este caso nos encontramos a un nivel de resolución 0.

El CSR lo podemos dividir en píxeles de forma natural: a resolución 0 tenemos un único píxel, que corresponde a la imagen completa, a resolución 1, cuatro, que serían las cuatro subimágenes de la 1a iteración, y así sucesivamente; en un nivel de resolución k, tendríamos 4^k píxeles que se corresponderían con los 4^k subcuadrados de la k-ésima iteración. Tenemos una identificación entre nivel de resolución (en términos de imágenes) y nivel de construcción del CSR, además esta división en píxeles de la imagen del CSR no es arbitraria, aquí cada píxel a un nivel de resolución k está relacionado con una de las 4^k subsecuencias de longitud k. Como nivel de gris de cada píxel tomaremos el número de puntos del CSR que caen dentro de la casilla correspondiente, por tanto el nivel de gris de un píxel a resolución k será el nº de veces que se repite una de las subsecuencias de longitud k.

Aprovechando esta similitud del CSR con los tratamiento usuales de imágenes, utilizaremos algunos elementos desarrollados recientemente en este campo, más concretamente un método de análisis de información multi-resolución de Román et al. (1991) [Ro 91]. Este método es particularmente interesante ya que el estudio de una imagen a diferentes resoluciones se traduce en el CSR en un estudio de la composición de la secuencia en subsecuencias de distinta longitud (*multi-length analisys*) que parece adecuado a la hora de buscar correlaciones y estructuras ordenadas en la secuencia.

De todas formas, hay una diferencia entre el tratamiento de imágenes y del CSR. En imágenes los valores de resolución de interés son los cercanos al valor máximo (píxeles binarios) que corresponderían a una imagen con mayor detalle. Pero el CSR binario correspondería a cada casilla con uno o ningún punto, el valor de esta resolución máxima debería ser tal que los dos puntos más próximos del CSR

ocupasen dos casillas diferentes (y vecinas), y tendría que ser definido para cada secuencia en particular. Además si la secuencia tiene una longitud medianamente grande, el problema se complica computacionalmente por el número de casillas del CSR que aparecen. Por tanto aquí las resoluciones de interés van a ser las que nos llevan a píxeles próximos al CSR completo, ya que hay que tener en cuenta que el análisis de subsecuencias de longitud mayor de 8 ó 10 ya supone un problema bastante complicado.

Denominaremos Histograma del CSR al equivalente a lo que, en procesamiento de imágenes, se conoce como Histograma de Niveles de Gris [Go 87]. Para una resolución dada el histograma de niveles de gris es una función que muestra para cada nivel de gris el número de píxeles en la imagen que tienen ese nivel de gris, la abcisa será el nivel de gris y la ordenada la frecuencia de aparición (nº píxeles). El histograma puede estar normalizado dividiendo las ocurrencias por el nº total de píxeles, para que sea independiente del tamaño de la imagen.

A partir de la analogía del CSR con una imagen digital, es evidente la forma de calcular su histograma, que aquí notaremos como $\mathcal{Z}^{(k)}$. El superíndice k se ha colocado entre paréntesis para recordar que no indica la dimensión del vector, que será siempre $N+1$, sino el nivel de resolución:

$$\mathcal{Z}^{(k)} = \{z_0, z_1, ..., z_N\} \text{ con } z_j = \frac{n_i}{4^k}$$

siendo n_j el número de k-casillas con un número j de puntos. O sea que z_j mide la frecuencia relativa de aparición de casillas con j puntos; y se puede interpretar como la probabilidad de encontrar una casilla con j puntos al escoger una al azar.

En algunos aspectos el tratamiento será esencialmente distinto al del histograma de niveles de gris. Para una imagen, los niveles de gris permitidos suelen estar fijados de antemano. Durante el proceso de digitalización se divide el continuo de niveles de gris en una serie de valores discretos, quedando por tanto fijado

el rango del histograma. En el CSR el rango del histograma irá desde 0 (o desde el menor número de puntos que ocupen una casilla) hasta el máximo número de puntos en una casilla, dependiendo por tanto de la secuencia y siendo imposible de determinar a priori. En principio el rango del histograma puede ir desde 0 a N (N, número de símbolos de la secuencia), aunque evidentemente, este caso extremo será difícil de encontrar, ya que supondría tener todos los puntos en una casilla y el resto vacías. De todas formas, en la práctica, para secuencias bastante alejadas de la aleatoriedad, se encuentran histogramas con rangos bastante más amplios que los usuales en imágenes, sobre todo para resoluciones grandes.

Las conclusiones que se pueden extraer sobre el CSR a partir del histograma son similares a las que se extraen de un histograma de grises para una imagen, sólo que en este caso habrá que traducir su significado al correspondiente comportamiento de la secuencia que lo genera. En primer lugar, veamos cuál es el histograma esperado para secuencias aleatorias.

3.4.1 Histograma esperado para secuencias aleatorias.

Consideremos el CSR generado por una secuencia aleatoria de enteros módulo 4, $s_1, s_2, ..., s_N$, independientes e idénticamente distribuidos (i.i.d.) y supongamos que los 4 enteros son igualmente probables, la secuencia de puntos del CSR determina una nueva variable aleatoria $X = (X_1, X_2, ..., X_N)$ tomando valores en $\{0, 1, 2, ..., 4^k - 1 = a - 1\}$, donde $X_i = x_i$ representa el suceso consistente en que el i-ésimo punto del CSR ha caído dentro de la casilla x_i del CSR, con la ordenación de casillas descrita anteriormente, aunque en este caso será irrelevante ya que estamos tratando con una secuencia aleatoria y todas las casillas son indistinguibles.

Aunque las variables X_i no son independientes, sí podemos asegurar que son equiprobables, o lo que es lo mismo, que todas las casillas tienen igual probabilidad de recibir puntos, ya que cada casilla del CSR se corresponde con una subsecuencia

de la secuencia de enteros y, ya que estamos suponiendo ésta como i.i.d. y los cuatro símbolos son igualmente probables, todas estas subsecuencias serán también igualmente probables.

Por tanto, para una secuencia aleatoria el CSR sería uniforme, esto significa que para todas las resoluciones $k = 0$ a ∞, deberíamos tener estrictamente la misma densidad de puntos en todas las casillas, lo que implicaría un histograma degenerado, esto es, con un sólo valor distinto de cero; pero esto sólo es cierto para una secuencia de longitud infinita.

El histograma normalizado del CSR es un conjunto de números $\mathcal{Z}^{(k)} = (z_0, z_1, ..., z_N)$, con $z_j = n_j/a$ y, de nuevo, podemos considerar una secuencia de variables aleatorias $\mathcal{Z}^{(k)} = (Z_0, Z_1, ..., Z_N)$ para la que cada histograma es una realización particular. Estas variables aleatorias no son equiprobables ni son independientes, tenemos dos restricciones: $\sum z_j = 1$ (Número fijo de casillas en el CSR) y $\sum i \cdot z_j = N/a$ (número fijo de puntos en la secuencia).

Vamos a calcular el valor esperado de las z_j, $\mathrm{E}[z_j]$, bajo la hipótesis de que la secuencia es una serie de variables aleatorias i.i.d. y que todos los símbolos son igualmente probables:

Recordemos que $Z_j = z_j = n_j/a$ supone que hay n_j casillas con j puntos de un total de $a = 4^k$ casillas y un total de N puntos; y consideremos los sucesos:

$$A_j = \left\{ \begin{array}{c} \textbf{Sacamos una casilla al azar} \\ \textbf{y contiene } j \textbf{ puntos} \end{array} \right\} \tag{3.4}$$

$$A_{j,m} = \left\{ \begin{array}{c} \textbf{Entre las } a \textbf{ casillas} \\ m \textbf{ contienen } j \textbf{ puntos} \end{array} \right\} \tag{3.5}$$

Es fácil comprobar que el suceso A_j se puede escribir como unión de a sucesos:

$$A_j = \bigcup_{m=0}^{a} \left(\left\{ \begin{array}{c} \textbf{La casilla sacada} \\ \textbf{contiene } j \textbf{ puntos} \end{array} \right\} \cap A_{j.m} \right) \tag{3.6}$$

y a partir de esta descomposición del suceso A_j, su probabilidad vendrá dada por:

$$\mathrm{P}(A_j) = \sum_{\substack{m \text{ con} \\ \mathrm{P}(A_{j,m})\neq 0}} \mathrm{P}\left(\begin{matrix}\textbf{La casilla sacada} \\ \textbf{contiene } j \textbf{ puntos}\end{matrix} \middle| A_{j.m} \right) \cdot \mathrm{P}(A_{j,m}) \Rightarrow \tag{3.7}$$

$$\begin{aligned} \mathrm{P}(A_j) &= \sum_{\substack{m \text{ con} \\ \mathrm{P}(A_{j,m})\neq 0}} \frac{m}{a} \cdot \mathrm{P}(A_{j.m}) = \sum_{m=0} \frac{m}{a} \cdot \mathrm{P}(A_{j.m}) = \\ &= \sum_{m=0}^{a} \frac{m}{a} \cdot \mathrm{P}\left(Z_j = \tfrac{m}{a}\right) = \mathrm{E}[Z_j] \end{aligned} \tag{3.8}$$

Ahora bien, como los puntos caen al azar dentro de las casillas, el suceso A_j tiene la misma probabilidad que el suceso {la casilla i tiene j puntos}. Esta probabilidad viene dada por la distribución binomial; en efecto, puesto que todas las casillas son equiprobables, la probabilidad de que un punto caiga en cualquiera de ellas será $1/a$, y por tanto la probabilidad de que caigan j puntos en la casilla i, será para cualquier casilla $\binom{N}{j} \cdot \left(\frac{1}{a}\right)^j \left(1-\frac{1}{a}\right)^{N-j}$ y de aquí:

$$\mathrm{E}[Z_j] = \mathrm{P}(A_j) = \binom{N}{j} \cdot \left(\frac{1}{a}\right)^j \left(1-\frac{1}{a}\right)^{N-j} \tag{3.9}$$

Teniendo en cuenta que $\mathrm{E}[Z_j]$ es la proporción esperada de casillas del CSR con j puntos, supongamos que tenemos un histograma $\mathcal{Z}^{(k)*}$, construido de acuerdo con la condición $n_j = a \cdot \mathrm{E}[Z_j]$, $\forall j$; esto es posible ya que:

$$\sum_{j=0}^{N} n_j = a \cdot \sum_{j=0}^{N} \mathrm{E}[Z_j] = a \tag{3.10}$$

y siempre que supongamos $a = 4^k$ lo suficientemente grande como para hacer los n_i enteros. Un histograma tal será por supuesto el histograma esperado:

$$\mathrm{E}[Z^*]_j = z_j^* = \mathrm{E}[Z_j] = \binom{N}{j} \cdot \left(\frac{1}{a}\right)^j \left(1-\frac{1}{a}\right)^{N-j} \tag{3.11}$$

Este resultado nos dice que si calculamos el histograma, normalizado, del CSR para un buen número de secuencias aleatorias, el promedio de estos histogramas debe coincidir con la distribución binomial.

De hecho, según lo visto arriba, si el número de casillas es suficientemente grande en comparación con el número de puntos de la secuencia (media no excesivamente grande), un único histograma debe asemejarse bastante a la distribución binomial, aunque no sería correcto hacer estadística con un solo histograma ya que, en este contexto un histograma es una única medida. Sacar conclusiones de un histograma aislado sería como estimar el valor medio de una variable sobre una población con una sola medida, pero en algunos casos esto no es disparatado, por ejemplo al medir la energía de un sistema macroscópico una sola medida es suficiente ya que la distribución de energías se asemeja mucho a una delta de Dirac; en el CSR, en estas condiciones, ocurre algo parecido.

Más adelante daremos un criterio razonable para decidir cuándo es suficientemente grande el número de casillas.

3.4.2 Secuencias con símbolos no equiprobables

Supongamos ahora que el CSR está generado por una secuencia aleatoria en la que los símbolos no son equiprobables. Esto se traduce en que la probabilidad de que un punto caiga en una casilla determinada no será $1/a$ para todas ellas sino que cada una tendrá una probabilidad p_i con $i = 0, 1, ..., 4^k - 1$. Nótese que, en la demostración que sigue, no se exige que secuencia sea una serie de variables aleatorias independientes, aunque sí idénticamente distribuidas. Es fácil comprobar que, en estas condiciones sigue siendo cierta la expresión (3.8), esto es, el valor esperado del j-ésimo elemento del histograma, $\mathrm{E}[Z_j]$, sigue siendo la probabilidad de que al sacar una casilla al azar, ésta contenga j puntos (eq. 3.4). Lo que ya no es cierto es que esa probabilidad venga dada por la distribución binomial. Veamos qué forma tiene ahora, y para ello

haremos uso de los sucesos:

$$A_{i,j} = \left\{ \begin{array}{c} \textbf{La casilla extraída al azar} \\ \textbf{es la } i\textbf{-esima y contiene } j \textbf{ puntos} \end{array} \right\} \tag{3.12}$$

que son mutuamente excluyentes, ya que aunque pueda haber varias casillas con j puntos, el suceso consiste en extraer una de ellas en concreto. Se tiene, por tanto, que el suceso que nos interesa (3.4) se puede expresar como unión disjunta de los $A_{i,j}$:

$$A_j = \bigcup_{i=0}^{4^k-1} A_{i,j} \text{ con } A_{i,j} \cap A_{i',j} = \emptyset$$

$$\forall i \neq i';\ i, i' \in \{0, 1, ..., 4^k - 1\} \tag{3.13}$$

y por tanto podemos escribir su probabilidad como:

$$\mathrm{P}(A_j) = \sum_{i=0}^{4^k-1} \mathrm{P}(A_{i,j}) \tag{3.14}$$

A su vez los sucesos $A_{i,j}$, se pueden descomponer como la intersección de dos sucesos independientes:

$$A_{i,j} = \left\{ \begin{array}{c} \textbf{La casilla escogida} \\ \textbf{es la } i\textbf{-ésima} \end{array} \right\} \cap \left\{ \begin{array}{c} \textbf{La casilla } i\textbf{-ésima} \\ \textbf{contiene } j \textbf{ puntos} \end{array} \right\} \tag{3.15}$$

con lo que:

$$\mathrm{P}(A_{i,j}) = \frac{1}{4^k} \mathcal{B}_j(N, p_i) \tag{3.16}$$

donde $\mathcal{B}_j(N, p_i)$, es el elemento j-ésimo de la distribución binomial de parámetros N y p_i, que como ya se vio, es la que nos da la probabilidad de que una casilla en concreto tenga un número de puntos j, cuando la probabilidad del suceso elemental (que el punto caiga dentro de la casilla) es p_i y $1 - p_i$ la del suceso contrario (que el punto caiga fuera). Finalmente, a partir de (3.14), se obtiene:

$$\mathrm{E}[Z_j] = \mathrm{P}(A_j) = \frac{1}{4^k} \sum_{i=0}^{4^k-1} \mathcal{B}_j(N, p_i) \tag{3.17}$$

En particular, si $p_i = \frac{1}{a}\ \forall i = 1, 2, ..., 4^k$, se recupera el resultado obtenido en (3.9) para secuencias con símbolos equiprobables.

3.5 Medidas Entrópicas

Para comparar los histogramas del CSR de diferentes secuencias utilizaremos la entropía de Shannon, aplicada sobre el histograma del CSR normalizado:

$$\mathrm{H}\left(\mathcal{Z}^{(k)}\right) = -\sum_{j=0}^{N} z_j \log z_j \tag{3.18}$$

$\mathrm{H}\left(\mathcal{Z}^{(k)}\right)$ toma su valor mínimo (0) para un histograma puntual, que correspondería a un CSR uniforme. Este valor será alcanzable si el número de elementos de la secuencia es divisible entre el número de casillas del CSR.

Para obtener el valor máximo tenemos que maximizar la expresión (3.18) sujeta a las restricciones:

$$\sum_{j=0}^{N} z_j = 1 \text{ y } \sum_{j=0}^{N} j \cdot z_j = \frac{N}{4^k} \equiv \mu_k \tag{3.19}$$

La primera es la condición de normalización del histograma y la segunda indica que el número de elementos de la secuencia es fijo, o si se prefiere, condición de media de ocupación de las casillas (μ_k) constante. Aplicando multiplicadores de Lagrange [Gui 77] se obtiene que el histograma que maximiza (3.18) es:

$$z_j = e^{-\alpha - \beta j} \ con\ j = 0, 1, ..., N$$

que es la conocida *Distribución de Gibbs* que, en Mecánica estadística corresponde a la colectividad canónica, donde la temperatura constante hace el papel de la restricción de media fija que imponemos aquí. Ahora hay que determinar el valor de los multiplicadores de Lagrange. Si llamamos:

$$\Phi(\beta) = \sum_{j=0}^{N} e^{-j\beta} \tag{3.20}$$

y, teniendo en cuenta que $\sum_{j=0}^{N} e^{-\alpha-j\beta} = 1 \Rightarrow e^{\alpha} = \sum_{j=0}^{N} e^{-j\beta} \Rightarrow \alpha = \ln \Phi(\beta)$, se puede escribir el histograma que maximiza la entropía como:

$$z_j = \frac{1}{\Phi(\beta)} e^{-j\beta} \tag{3.21}$$

y sustituyendo en la condición de media constante:

$$\mu_k = \frac{1}{\Phi(\beta)} \sum_{j=0}^{N} j \cdot e^{-j\beta} \Rightarrow \mu_k = -\frac{1}{\Phi(\beta)} \frac{\mathrm{d}}{\mathrm{d}\beta} \sum_{j=0}^{N} j \cdot e^{-j\beta} \Rightarrow$$

$$\frac{\mathrm{d} \ln \Phi(\beta)}{\mathrm{d}\beta} = \frac{\Phi'(\beta)}{\Phi(\beta)} = -\mu_k \tag{3.22}$$

Se puede comprobar [Gui 77] que en las condiciones en las que nos encontramos, esta ecuación tiene solución única, aunque difícilmente se puede encontrar de forma analítica para N medianamente grande. Teniendo en cuenta que (3.20) es la suma de una progresión geométrica podemos expresar (3.22) como:

$$\mu_k = \frac{e^{-\beta}\left[1 - (N+1)\, e^{-\beta N} + N e^{-\beta(N+1)}\right]}{\left(1 - e^{-\beta(N+1)}\right) \cdot \left(1 - e^{-\beta}\right)} \tag{3.23}$$

y, para simplificar la notación, definimos $x \equiv e^{-\beta}$, con lo que la expresión anterior queda:

$$\mu_k = N + \frac{1}{1-x} - \frac{N+1}{1-x^{N+1}}$$

Si suponemos que N es grande comparado con μ_k, y teniendo en cuenta que $0 \leq x \leq 1$, tenemos:

$$\mu_k \approx \frac{1}{1-x} - 1 \Rightarrow x \approx \frac{\mu_k}{1+\mu_k} \tag{3.24}$$

En la mayoría de los casos de interés esta aproximación será suficiente. Cuando no lo sea, se podrá resolver (3.23) numéricamente sin ningún problema.

El histograma que maximiza la entropía se puede escribir de forma más compacta usando x:

$$z_j = (1-x) \cdot x^j \tag{3.25}$$

Finalmente, sustituyendo en la expresión de la entropía de Shannon, obtenemos, para la entropía máxima:

$$\begin{aligned} \mathrm{H_{max}}\left(N,k\right) &= \log\frac{1}{1-x}+\frac{\left[x-\left(N+1\right)x^{N+1}+Nx^{N+2}\right]}{1-x}\log\frac{1}{x}= \\ &= \log\frac{1}{1-x}-\mu_k\left(1-x^{N+1}\right)\log\frac{1}{x}\approx-\log\left(1-x\right)x^{\mu_k}\approx \\ &\approx \left(\mu_k+1\right)\log\left(\mu_k+1\right)-\mu_k\log\mu_k \end{aligned} \tag{3.26}$$

Cuando hablamos del histograma no hay que perder de vista que si lo tratamos como una distribución de probabilidad, no nos estamos refiriendo a un histograma concreto obtenido a partir de una secuencia, sino de un histograma promedio, o dicho de otro modo, el histograma visto como una distribución de probabilidad es una propiedad de la fuente y no de la secuencia. Esta cota máxima que hemos obtenido se refiere al histograma promedio; si estamos tratando con el histograma de una secuencia, por supuesto que el valor de la entropía no puede ser superior al valor de (3.26) ya que se cumplen las condiciones de normalización y de media constante, pero vamos a ver que existe también otra cota superior que limita el valor de la entropía cuando no es un promedio de histogramas (en cualquier caso la entropía debe ser menor que ambas cotas y la cota efectiva será la menor de las dos).

Al calcular el histograma del CSR de una secuencia, como el número de casillas con un número de ocupación dado debe ser entero ($z_j 4^k = n_j$) puede que los valores del histograma de máxima entropía (3.25) sean inalcanzables. A lo sumo puede haber 4^k valores del histograma distintos de cero (no puede haber más valores de ocupación distintos que el número total de casillas), si llamamos I al conjunto de índices para los que $z_j \neq 0$, el valor máximo de la entropía se obtendrá para $z_j = \frac{1}{4^k}$ $\forall j \in I$, por tanto tenemos que:

$$\mathrm{H_{max}}\left(N,k\right) = -\sum_{j\in I}\frac{1}{4^k}\log\frac{1}{4^k} = 2k \tag{3.27}$$

Pero esta cota no siempre es alcanzable, de hecho para k suficientemente grande puede ser mayor que (3.26). Para que sea alcanzable es necesario que se cumpla la condición de media constante:

$$\begin{aligned} \sum_{j\in I} j \cdot z_j &= \sum_{j=1}^{4^k} \frac{j}{4^k} = \mu_k \Rightarrow \sum_{j=1}^{4^k} j = 4^k \mu_k \Rightarrow \\ \frac{4^k\left(4^k+1\right)}{2} &= 4^k \mu_k \Rightarrow 4^k = 2\mu_k - 1 \end{aligned} \tag{3.28}$$

de donde se tiene que, para $k \geq \frac{1}{2}\log\left(2\mu_k - 1\right)$ la cota de $2k$ para el histograma no es alcanzable. Aproximadamente para este valor de k la entropía máxima se hace menor que $2k$. Por tanto, parece razonable imponer como condición para que tenga sentido utilizar la entropía de un único histograma:

$$\mu_k \leq \frac{4^k+1}{2} \tag{3.29}$$

Si en lugar de tratarse del histograma de una única secuencia, se ha promediado sobre M secuencias, se comprueba fácilmente que esta condición se convierte en:

$$\mu_k \leq \frac{4^k M+1}{2} \tag{3.30}$$

3.5.1 Entropía del histograma binomial. Aproximación normal

Hemos visto que la entropía del histograma está acotada entre 0, que corresponde a la secuencia más ordenada posible (o menos compleja, si se quiere) y un valor máximo que depende de la longitud de la secuencia y de la resolución empleada, este segundo valor se puede interpretar como el correspondiente a la secuencia menos ordenada (o más compleja), en el sentido de que es la más alejada de la uniformidad. Entre estos dos extremos estaría la entropía del histograma binomial que corresponde a una fuente que emite secuencias aleatorias. A la hora de decidir si una fuente emite

secuencias más o menos ordenadas, la entropía del histograma binomial puede ser una buena referencia: Entropías inferiores a la binomial corresponden a secuencias *excesivamente uniformes* para ser aleatorias, es decir, los valores de ocupación de las casillas fluctúan alrededor del valor medio menos de lo que cabría esperar para una secuencia aleatoria. Por el contrario, entropías superiores a la binomial corresponden a secuencias en las que las fluctuaciones de los valores de ocupación de las casillas son superiores a lo admisible para suponer una distribución uniforme del vector de frecuencias.

Que tengamos noticia, no existe una expresión analítica explícita para calcular la entropía de Shannon de la distribución binomial, por lo que su manejo se hace bastante engorroso si N es medianamente grande. En muchos casos, si la media de ocupación es suficientemente grande, la entropía de la distribución binomial se puede aproximar con bastante precisión por la entropía de la distribución normal con la misma media y varianza, siendo mejor la aproximación si además de ser grande la media, también lo es el número de puntos.

Para una distribución normal con media μ y varianza σ, la entropía de Shannon (en bits) viene dada por [Gui 77]:

$$\mathrm{H}\left(\mathcal{N}\left(\mu,\sigma\right)\right) = \log\left(\sigma\sqrt{2\pi e}\right) \tag{3.31}$$

para la distribución binomial correspondiente se tiene que:

$$\sigma = \sqrt{\mu\left(1-P\right)} = \sqrt{NP\left(1-P\right)} \tag{3.32}$$

donde $P = 1/4^k$. Por tanto, si suponemos válida la aproximación tenemos que:

$$\mathrm{H}_{\mathbf{bin}}\left(N,k\right) = \mathrm{H}\left(\mathcal{B}\left(N,P\right)\right) \approx \frac{1}{2}\log\left[2\pi eP\left(1-P\right)\right] + \frac{1}{2}\log N \tag{3.33}$$

En la figura 3.5 hemos representado la entropía máxima y la entropía binomial para $k = 1$, que es uno de los casos más desfavorables para estas aproximaciones,

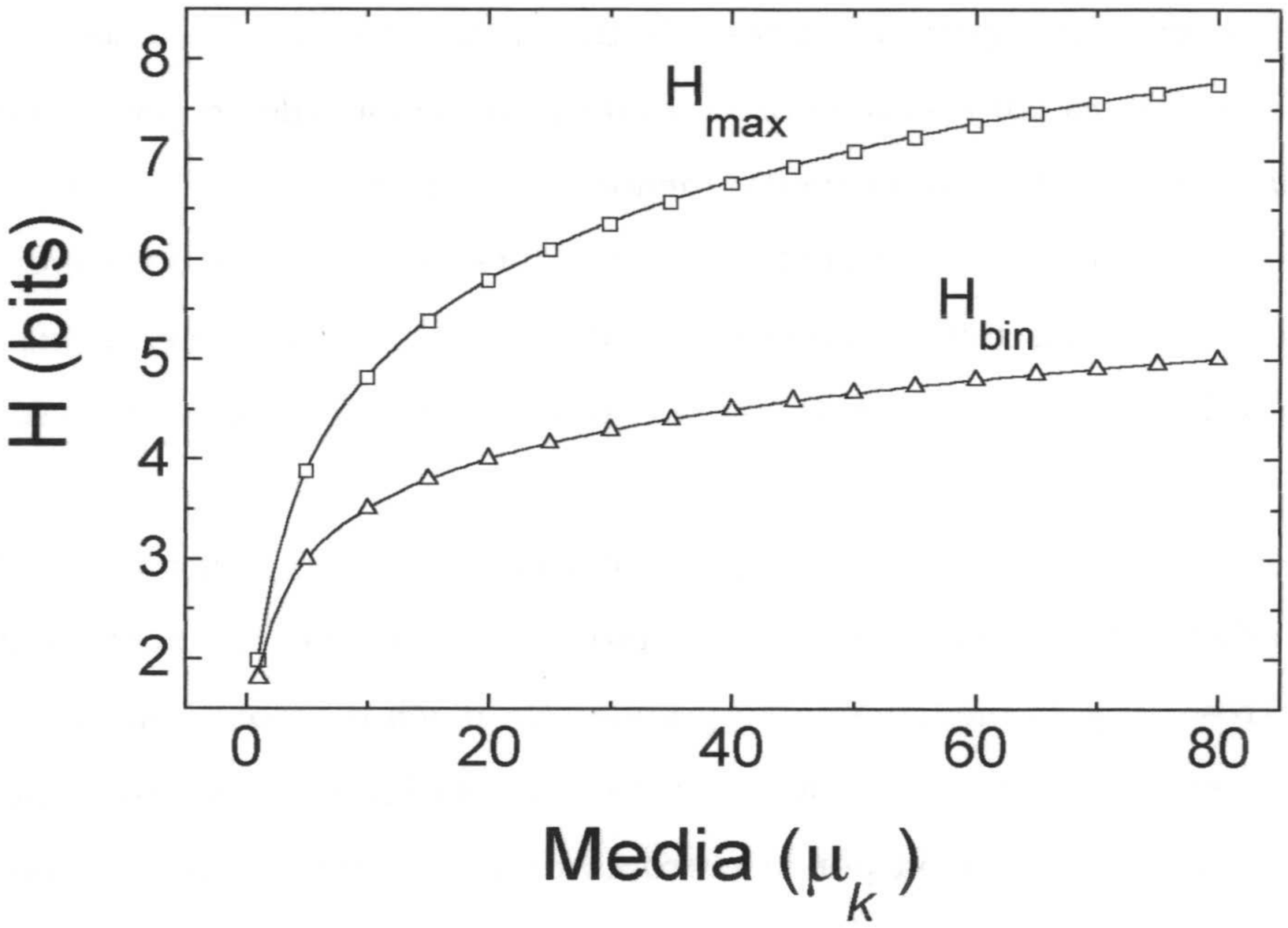

Figura 3.5: Entropía del histograma binomial y entropía máxima para $k = 1$ en función de la media de ocupación. Los trazos continuos corresponden a las aproximaciones, las marcas a los valores exactos.

en función de la media de ocupación. La línea continua representa en ambos casos la aproximación y las marcas los valores exactos. Como se puede ver, el error es bastante pequeño.

3.5.2 Entropía normalizada

Hemos visto que la entropía del histograma del CSR depende de la longitud de la secuencia incluso para el caso del histograma binomial y el de entropía máxima. Por esto, los valores absolutos de entropía no serán muy significativos cuando se comparen secuencias de diferente tamaño. Para evitar este problema definimos la

Entropía Normalizada para una resolución dada k, como:

$$\mathrm{h}\left(\mathcal{Z}^{(k)}\right) = \frac{\mathrm{H}\left(\mathcal{Z}^{(k)}\right) - \mathrm{H}_{\mathbf{bin}}}{\mathrm{H}_{\mathbf{max}} - \mathrm{H}_{\mathbf{bin}}} \tag{3.34}$$

Ésta no es una normalización en sentido estricto, no obtenemos un valor entre cero y uno, ya que para los valores de la entropía del histograma menores que el correspondiente valor del histograma binomial obtenemos valores negativos de la entropía normalizada, que tampoco tienen que estar comprendidos entre 0 y -1, pero en la mayoría de los casos de interés, los valores se encuentran entre la entropía máxima y la binomial.

Lo más interesante de esta normalización es que, en todos los casos, el 0 corresponde al histograma binomial que, como vimos, se correspondía con la aleatoriedad y el 1 a la entropía máxima.

3.6 Perfiles Entrópicos

Puesto que, en el caso de las secuencias de ADN no tenemos acceso a la fuente emisora, no podemos promediar sobre los histogramas de varias secuencias. En su lugar podremos calcular el histograma promediando sobre varios tramos. De esta forma, además de la información puramente composicional sobre la secuencia podremos obtener información sobre la heterogeneidad espacial, o dicho de otro modo, cómo se aparta la secuencia de la estacionariedad.

Otro aspecto interesante es la posibilidad de realizar un análisis multiresolución. Representado los valores de la entropía del histograma para distintos niveles resolución, lo que denominaremos **Perfiles Entrópicos,** podremos ver cómo varia la organización o estructura de la secuencia al ser descompuesta en símbolos, parejas, tríos, etc.

Como vimos en la sección anterior, los requerimientos estadísticos para poder

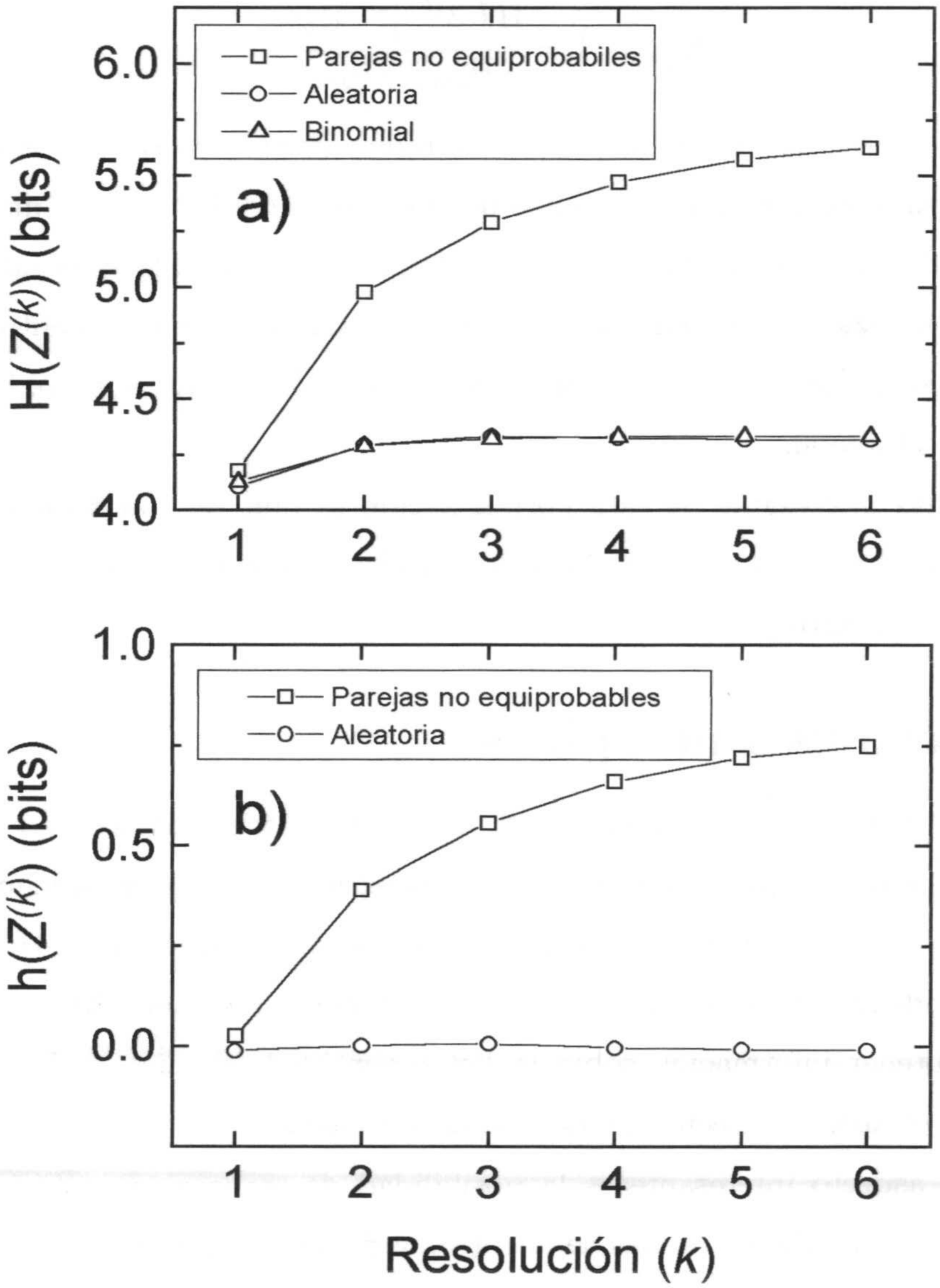

Figura 3.6: Perfiles entrópicos de una secuencia aleatoria ($N = 98304$) y de una secuencia de la misma longitud generada de forma que los símbolos son equiprobables pero las parejas de símbolos no. Como comparación se han incluido las entropías binomiales correspondientes.

comparar los histogramas con los modelos teóricos no sólo dependen de la longitud de la secuencia sino, también del nivel de resolución (eqs. 3.29, 3.30). Por ello, para tener una misma fiabilidad a todas las resoluciones proponemos el **CSR desdoblado** que consiste en lo siguiente:

Dada una secuencia de longitud N, escogemos la resolución máxima a la que vamos a analizarla, $K_{\max}$. Esta elección se hará de forma que la media de ocupación $\mu = \frac{N}{4^{K_{\max}}}$ tenga un valor adecuado, es conveniente que sea entero por lo que, si es necesario, se puede recortar ligeramente la longitud de la secuencia. Para cada una de las resoluciones inferiores $k = 1, 2, ...K_{\max} - 1$, realizamos $4^{K_{\max}-k}$ CSRs con una longitud de subsecuencia $N_k = \frac{N}{4^{K_{\max}-k}} = \mu 4^k$, calculamos los correspondientes histogramas y los promediamos. De esta forma, según el criterio de (3.30), es fácil comprobar que las comparaciones con histogramas teóricos son igualmente fiables para todas las resoluciones.

Como ejemplo, en la figura 3.6 vemos los perfiles entrópicos de una secuencia aleatoria, otra secuencia en la que los símbolos son equiprobables pero con probabilidades condicionadas que hacen que las parejas no lo sean (la misma utilizada para el CSR de la figura 3.3) y, como referencia, los valores de la entropía de los correspondientes histogramas binomiales. La secuencia aleatoria tiene, para todas las resoluciones, una entropía muy similar a la binomial (y en caso de la entropía normalizada, muy próxima a cero). Por otra parte, la secuencia con parejas no equiprobables presenta diferencias notables a todas las resoluciones, salvo a resolución 1 ya que, al tratarse de símbolos equiprobables no debe haber diferencia con la secuencia aleatoria. Podría pensarse que esta secuencia debería presentar diferencias notables sólo para resolución 2, ya que para resoluciones superiores no se introduce ningún otro condicionamiento, pero hay que tener en cuenta que si se modifican las probabilidades de aparición de las parejas se modifican también las de los tríos, cuartetos,

etc.

3.6.1 Heterogeneidad espacial

Cuando un histograma tiene un valor de entropía normalizada distinto de cero, esto quiere decir que es distinto del binomial. Pero, sobre todo a resoluciones pequeñas, cuando un histograma difiere del binomial puede ser por dos motivos:

a) Las probabilidades de aparición de las subsecuencias de longitud k se apartan de la equiprobabilidad.

b) Aún cuando las subsecuencias sean equiprobables, existe cierta heterogeneidad espacial en la composición de la secuencia, esto es, la composición no es la misma en todas las zonas de la secuencia. Esto hace que las fluctuaciones resultantes al promediar varios CSRs no sean las correspondientes al histograma binomial.

Ambos efectos se pueden considerar como desviaciones del comportamiento aleatorio, pero quizá los segundos sean más relevantes como representativos de una cierta estructura en la secuencia: Una secuencia que presenta patrones similares a cierta distancia, o que presenta faltas de homogeneidad que no están distribuidas al azar se parece más a lo que, intuitivamente, entendemos por algo complejo que otra en la que las subsecuencias de una determinada longitud aparecen con frecuencias diferentes, pero que se encuentran distribuidas al azar a lo largo de la secuencia.

En la figura 3.7 se muestran los perfiles de entropía normalizada para una secuencia natural de ADN (HUMHBB con 73326 bp) y dos secuencias generadas por ordenador, la primera tiene las mismas frecuencias de aparición de parejas de símbolos que la secuencia natural (modelo de Markov de orden 1), mientras que la segunda es una secuencia i.i.d. en la que se han escogido las frecuencias de aparición de los símbolos (0.31, 0.12, 0.22 y 0.35) de forma que la entropía a resolución 1 coincida con la de la secuencia natural. En primer lugar vemos que la primera secuencia artificial, a pesar de tener las mismas frecuencias globales de aparición de

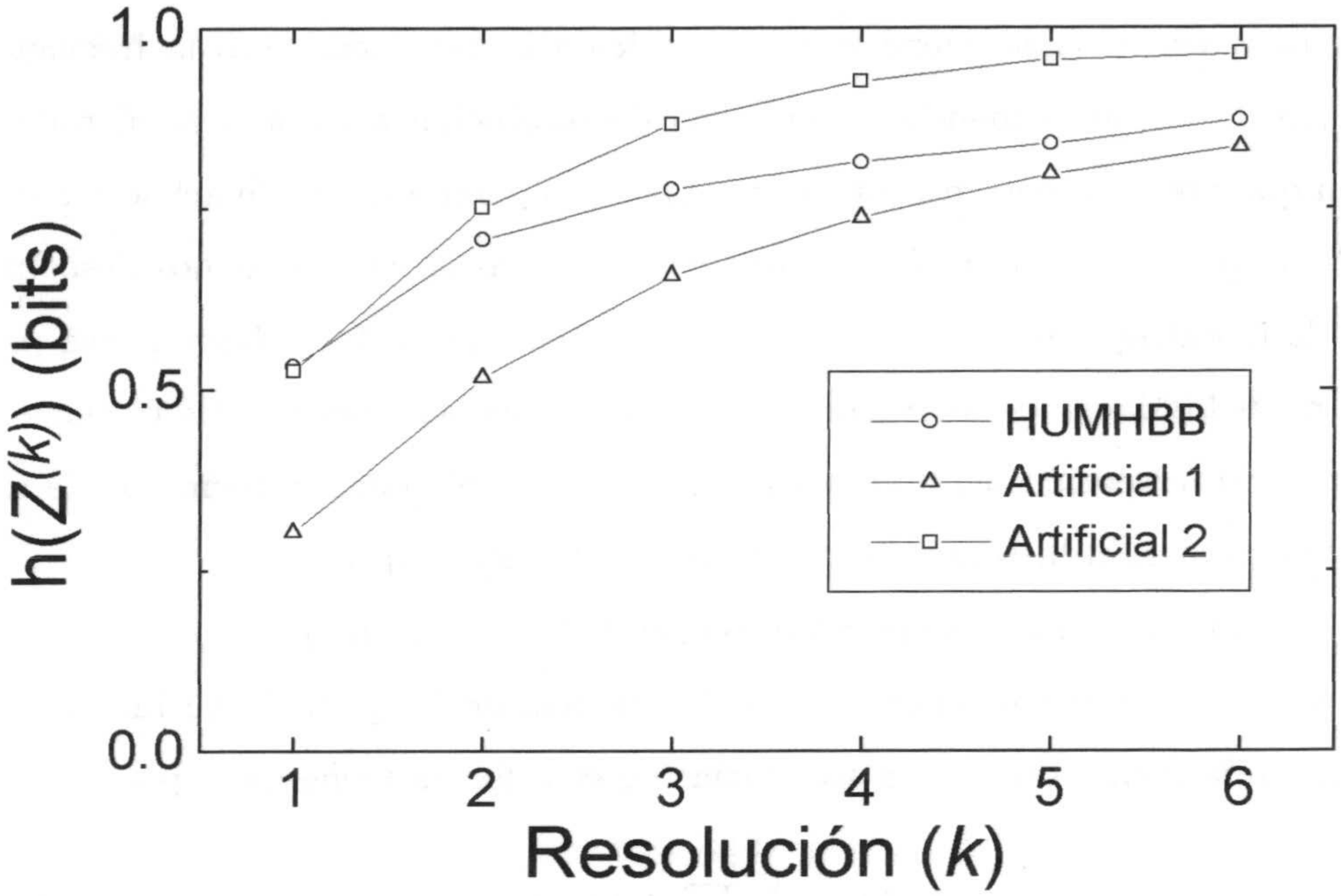

Figura 3.7: Perfiles de entropía normalizada para una secuencia de ADN (HUMHBB) y dos secuencias simuladas (ver texto).

símbolos y parejas que la natural, muestra diferencias claras a resoluciones 1 y 2. Esto es debido a que la secuencia natural no tiene la misma composición en todas las zonas (heterogeneidad espacial), mientras que la artificial sí. Pero, por otra parte, la segunda secuencia simulada alcanza el mismo valor de entropía para resolución 1 que la natural, a pesar de tener una composición homogénea; es más, para resoluciones superiores tiene una entropía mayor que la natural.

En vista de este ejemplo, parece interesante separar de alguna forma la contribución debida simplemente a la no uniformidad del vector de frecuencias de la debida a la existencia de heterogeneidad espacial. Para ello aquí proponemos lo

siguiente:

Si una secuencia no tiene una composición homogénea, pero la frecuencia de aparición de las subsecuencias a un nivel de resolución k es la misma, toda la desviación que presente con respecto a la entropía binomial se puede achacar a esta falta de homogeneidad. Si las frecuencias no son todas iguales podemos descontar del valor de la entropía el que cabría esperar si la secuencia fuese homogénea pero con unas probabilidades de aparición de las subsecuencias iguales a las frecuencias observadas. En la sección 3.4.2 vimos cómo se podía calcular, en términos de una suma de binomiales, el histograma esperado en este caso (eq. 3.17).

Consideremos una secuencia de longitud N, y sean $f_0, f_2, ..., f_{4^k-1}$ las frecuencias relativas de aparición de cada subsecuencia de longitud k. El histograma teórico para una fuente que emita secuencias de esta forma viene dado por:

$$\mathcal{Z}^{T(k)} = \frac{1}{4^k} \sum_{i=0}^{4^k-1} \mathcal{B}(N, f_i) \tag{3.35}$$

De aquí, una estimación de la entropía debida a la existencia heterogeneidad espacial, a una resolución dada, puede ser la diferencia entre la entropía del histograma y la entropía del histograma teórico calculado en la expresión anterior, que denominaremos **entropía espacial**:

$$\mathrm{H}_{\mathbf{espacial}} = \mathrm{H}\left(\mathcal{Z}^{(k)}\right) - \mathrm{H}\left(\mathcal{Z}^{T(k)}\right) \tag{3.36}$$

Esta cantidad tiene las desventajas de las diferencias de entropías, que no son medidas muy aceptadas, por lo general, en la literatura de Teoría de la Información [Gui 77], pero nos permite seguir usando la normalización que hemos introducido antes.

Para ilustrar cómo esta medida puede poner de manifiesto la presencia o no de heterogeneidad espacial, se ha generado, en primer lugar, una secuencia de 75.000

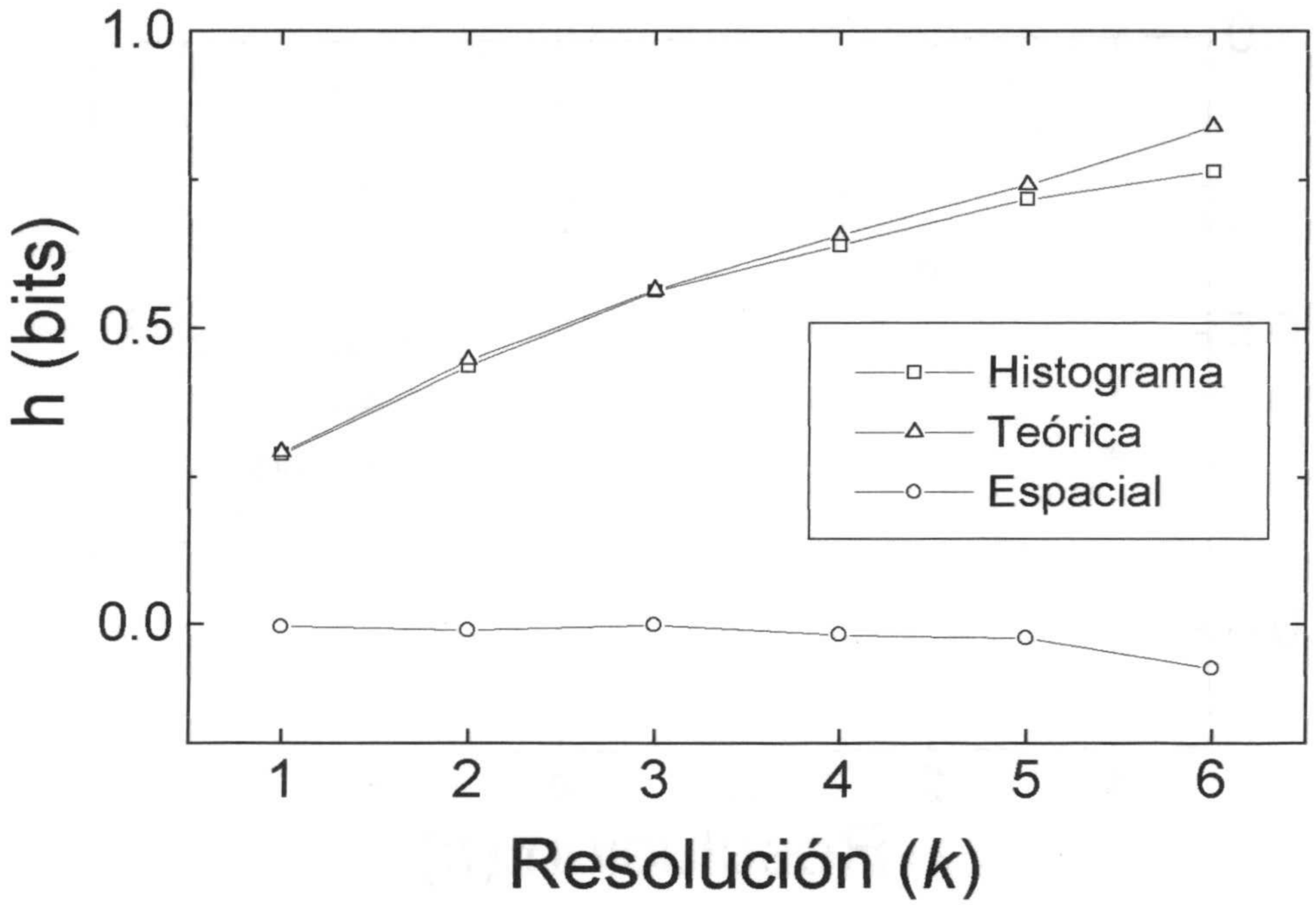

Figura 3.8: Entropía normalizada del histograma, teórica y espacial para una secuencia con símbolos no equiprobables.

enteros módulo 4, de forma que las probabilidades de cada símbolo son 0.29, 0.28, 0.27 y 0.19 respectivamente, y no se ha introducido ningún otro condicionamiento. En la figura 3.8 vemos los perfiles de la entropía del histograma, la entropía del histograma teórico y la entropía espacial (todas ellas normalizadas). Como se puede ver, aunque los valores de la entropía del histograma son considerablemente altos, la entropía teórica toma valores muy próximos a ella, lo que indica que toda la desviación con respecto al histograma binomial es debida a las distintas frecuencias de aparición de las subsecuencias, no a la existencia de heterogeneidad espacial.

Por otra parte, se ha generado otra secuencia de la misma longitud y asig-

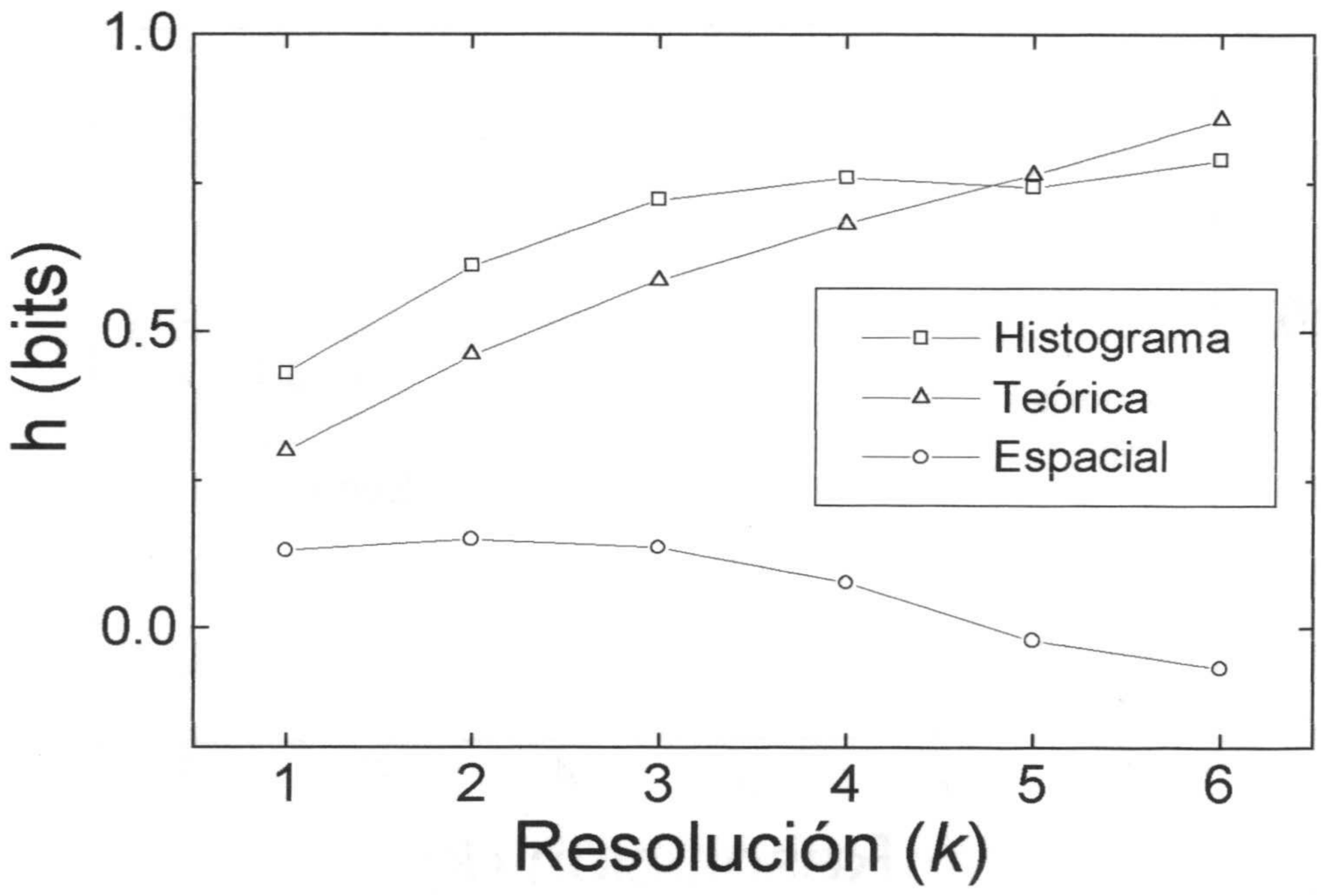

Figura 3.9: Entropía normalizada del histograma, teórica y espacial para una secuencia con símbolos no equiprobables y con cierta estructura espacial (ver texto).

nando las mismas probabilidades globales a los símbolos, pero las probabilidades en cada tramo de 5000 símbolos son distintas. Los perfiles de esta secuencia aparecen en la figura 3.9. En este caso la entropía teórica coincide con la del ejemplo anterior, pero ahora la entropía del histograma difiere notablemente de ésta. Los perfiles ponen de manifiesto, en este caso, la heterogeneidad que hemos introducido al generar la secuencia.

Por último, veamos algunos inconvenientes del tratamiento que hemos desarrollado aquí. En las gráficas anteriores se puede ver cómo para las resoluciones

superiores la entropía del histograma teórico se hace mayor que la entropía del histograma sin que haya ningún motivo lógico para ello. Este efecto es un artefacto del cálculo, hay que tener en cuenta que, para estas resoluciones, hay pocos CSRs sobre los que promediar (de hecho, para la resolución máxima, hay sólo un CSR) y, por tanto, las frecuencias relativas que se utilizan como probabilidades de las binomiales son el resultado de un promedio sobre pocos valores (o sobre uno solo) y si aparece algún valor disparatado, cosa que puede ocurrir a causa de las fluctuaciones, no hay posibilidad de que sea compensado con otros valores. Además, existe un problema de fondo, para estas resoluciones pierde un poco el sentido la comparación con un histograma esperado ya que promediamos sobre pocos histogramas. Se podría pensar que este mismo inconveniente debe aparecer cuando se compara el histograma, por ejemplo a la resolución máxima con el histograma binomial que, en definitiva, también es un histograma esperado, pero aquí se suple de alguna forma la falta de estadística: el histograma binomial asigna una misma probabilidad a todas las casillas y, a la resolución máxima, aunque tengamos sólo un CSR, tenemos muchas frecuencias relativas que estamos comparando con una probabilidad, con lo que los valores disparatados quedan amortiguados; sin embargo cuando comparamos el histograma real con el teórico, estamos asimilando cada frecuencia relativa a una probabilidad.

Otro inconveniente es la complejidad del cálculo, por ejemplo, para resolución 6 hay que calcular 4096 distribuciones binomiales, que además de ser complejas de evaluar, son muy dadas a introducir errores de redondeo en los cálculos.

Estas complicaciones hacen algo molestos estos tipos de perfiles, pero de todas formas hay que tener en cuenta que, para las resoluciones superiores, la heterogeneidad difícilmente puede aparecer ya que la longitud de la secuencia analizada se aproxima a la longitud total de la secuencia.

3.6.2 Entropía incremental

Hemos visto como la estructura que aparece para cada resolución se puede desglosar en un parte debida a la no uniformidad global de la distribución de las $k-$tuplas en la secuencia y otra parte debida a la no uniformidad en la distribución espacial de éstas a lo largo de la secuencia. Otro tipo de desglose de la entropía del histograma, que puede ser también interesante, consiste en separar, la entropía debida a cada resolución de la *arrastrada* de resoluciones anteriores. Es evidente que si para una resolución $k-1$ tenemos un valor de entropía distinto de cero, ya sea por falta de uniformidad global o por heterogeneidad espacial, para la resolución siguiente k, tendremos normalmente un valor también distinto de cero sin necesidad de que aparezcan nuevas restricciones a esta resolución, ya que las $k-$tuplas se construyen a partir de las $(k-1)-$tuplas.

Para descontar este efecto podemos hacer lo siguiente: calculamos el histograma esperado a resolución k suponiendo que sólo existen las restricciones impuestas a resolución $k-1$, y comparamos su entropía con la del histograma de la secuencia.

Consideremos una secuencia con unas frecuencias relativas para las $(k-1)-$tuplas $f_0, f_1, ..., f_{4^k-1}$. Admitiendo las restricciones impuestas por el conocimiento de estas frecuencias, la secuencia más aleatoria posible a la resolución siguiente sería aquella que no impone nuevas restricciones en la aparición de símbolos, aparte del condicionamiento con los $k-1$ símbolos anteriores. En estas condiciones, una secuencia de longitud k, $s_1, s_2, ..., s_k$, tendrá una probabilidad:

$$\begin{aligned} \mathrm{P}(s_1, s_2, ..., s_k) &= \mathrm{P}(s_1, s_2, ..., s_{k-1}) \cdot \mathrm{P}(s_k / s_1, s_2, ..., s_{k-1}) = \\ &= \frac{f(s_1, s_2, ..., s_{k-1}) \cdot f(s_2, s_3, ..., s_k)}{f(s_2, s_3, ..., s_{k-1})} \end{aligned} \tag{3.37}$$

donde se han identificado las probabilidades de aparición de las subsecuencias de

longitud $k-1$ y $k-2$ con las frecuencias relativas reales. Por sencillez nos referiremos a la probabilidad de (3.37) como P_i^c, donde $i = 0, 2, ..., 4^k - 1$, es la etiqueta de la correspondiente $k-$tupla (sección 3.3 al final). La expresión anterior es válida para $k \geq 3$, para $k = 2$ es fácil comprobar que la expresión es la misma, eliminando el denominador que carece de sentido y para $k = 1$ es obvio que $\mathrm{P}_i^c = 0.25\ \forall i \in \{0, 1, 2, 3\}$.

Conocidas estas probabilidades, el histograma esperado para resolución k, que denominaremos *histograma condicionado*, vendrá dado por:

$$\mathcal{Z}^{c(k)} = \frac{1}{4^k} \sum_{i=0}^{4^k-1} \mathcal{B}\left(N, \mathrm{P}_i^c\right) \tag{3.38}$$

Al igual que se hizo en la sección anterior con la entropía debida a heterogeneidades espaciales, aquí podemos calcular la diferencia de la entropía del histograma real con el histograma condicionado, que denominaremos **Entropía Incremental**. Esta medida nos dará idea de la estructura de la secuencia a cada resolución, evitando la influencia de las resoluciones inferiores:

$$\mathrm{H}_{\mathrm{incr}} = \mathrm{H}\left(\mathcal{Z}^{(k)}\right) - \mathrm{H}\left(\mathcal{Z}^{c(k)}\right) \tag{3.39}$$

Para ilustrar la utilidad de esta medida se han simulado dos secuencias, ambas de 75.000 símbolos; en la primera se han introducido distintas probabilidades para los símbolos, pero ninguna otra restricción. En la segunda las restricciones se han introducido sobre los símbolos y parejas de símbolos.

En la figura 3.10 se han representado los perfiles de entropía y de entropía incremental de estas secuencias. Para la primera secuencia vemos que la entropía es distinta de cero para todas las resoluciones, sin embargo la entropía incremental sólo difiere significativamente para resolución 1, que es donde se ha introducido condicionamiento. Para la segunda secuencia tenemos que la entropía incremental

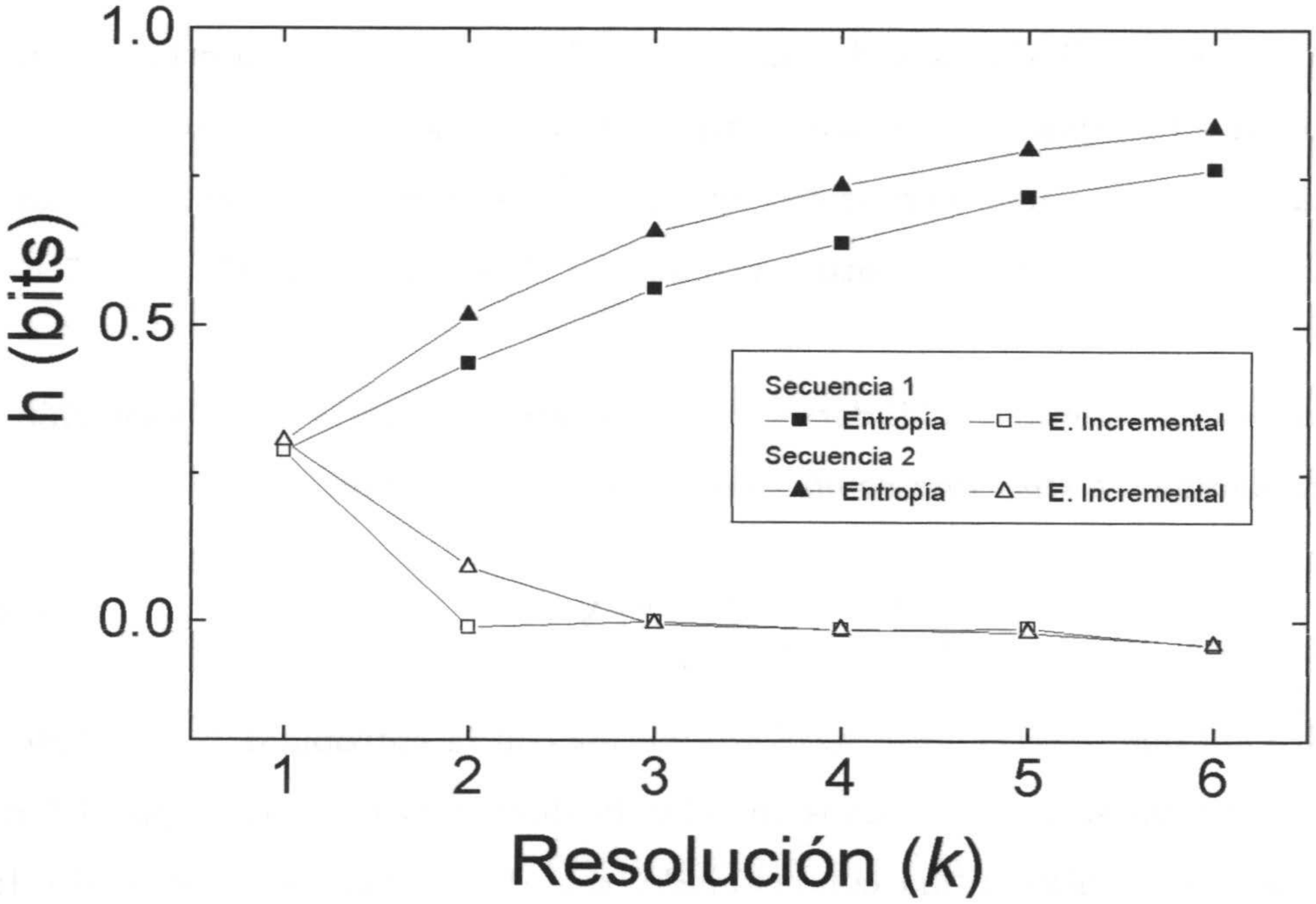

Figura 3.10: Perfiles de entropía y entropía incremental para una secuencia con distintas probabilidades para los símbolos (secuencia 1) y para una secuencia con distintas probabiliades para los símbolos y condicionamiento para las parejas de símbolos (secuencia 2).

difiere notablemente de cero para resoluciones 1 y 2, como era de esperar. Nótese que para resoluciones superiores la entropía es diferente para las dos secuencias, pero la entropía incremental es prácticamente cero para las dos, indicando la ausencia de condicionamientos para estas resoluciones en ambas secuencias.

Este tipo de perfiles reflejan la entropía propia de cada resolución descontando la "heredada" de resoluciones anteriores, pero la entropía que reflejan puede corresponder tanto a faltas de uniformidad como a la existencia de heterogeneidad composicional. En los ejemplos anteriores no existe lugar a dudas ya que se trata

de secuencias generadas por ordenador en las que no se ha introducido ningún tipo de estructura espacial y toda la entropía que aparece en cada resolución se debe a la falta de uniformidad introducida a esa resolución.

3.7 Aplicación a Secuencias de ADN

Como hemos visto, aunque el CSR no tiene ningún contenido teórico importante para el estudio que hemos realizado, se trata de una forma novedosa de representar el vector de frecuencias de una secuencia. Han sido muchos los intentos de representar gráficamente la información contenida en las cadenas de ADN, la mayoría de ellos basados en variantes del *random-walk,* pero estos métodos suelen introducir correlaciones cuando la dimensión utilizada es menor que cuatro [Vo 92]. Puesto que estas secuencias se pueden considerar, en esencia, mensajes emitidos por una fuente con un alfabeto de cuatro símbolos, se pueden utilizar de forma natural para generar CSRs bidimensionales, asociando a cada una de las cuatro bases una de las esquinas del atractor.

En 1990 Jeffrey [Je 90] propuso por primera vez el uso de cadenas de ADN para controlar el IFS que genera un CSR bidimensional*, con lo que se consigue una representación gráfica de las mismas. Jeffrey se limita a presentar gráficos de CSRs con distintas cadenas de ADN y ARN, en las que se aprecian diferencias notables entre secuencias de distinta naturaleza y estudia algunas de las propiedades generales del CSR (tratadas al principio este capítulo). Jeffrey pone de manifiesto las diferencias que se aprecian a simple vista entre los CSRs de distintas cadenas de ADN y, por supuesto, las diferencias con respecto a secuencias aleatorias. Por ejemplo, la figura 3.11 muestra los CSRs de una secuencia humana (HUMHBB) y una del virus bacteriófago T7 (PT7CG). En la humana se observan claros patrones,

*En la referencia original de Jeffrey, así como en [Ol 93] se utiliza la terminología *Chaos Game Representation* (CGR).

tipo fractal, debidos principalmente a la distribución de frecuencias de dinucleótidos, mientras que en la bacteriana estos patrones no son tan evidentes. Por otra parte, en 1992 Kathlen et al. [Kat 92] también utilizan estos diagramas, aunque a nuestro modo de ver usan secuencias demasiado cortas para el análisis que pretenden.

En 1993 nuestro grupo [Ol 93] propuso el estudio del histograma del CSR generado con una secuencia de ADN. En este primer trabajo no se utiliza el CSR desdoblado ni ningún tipo de normalización de la entropía del histograma por lo que los perfiles que se obtienen no son muy fiables, aunque sí ponen claramente de manifiesto las diferencias entre cadenas de ADN de organismos superiores y microorganismos así como las similitudes entre regiones codificadoras de proteínas similares en diversas especies de mamíferos.

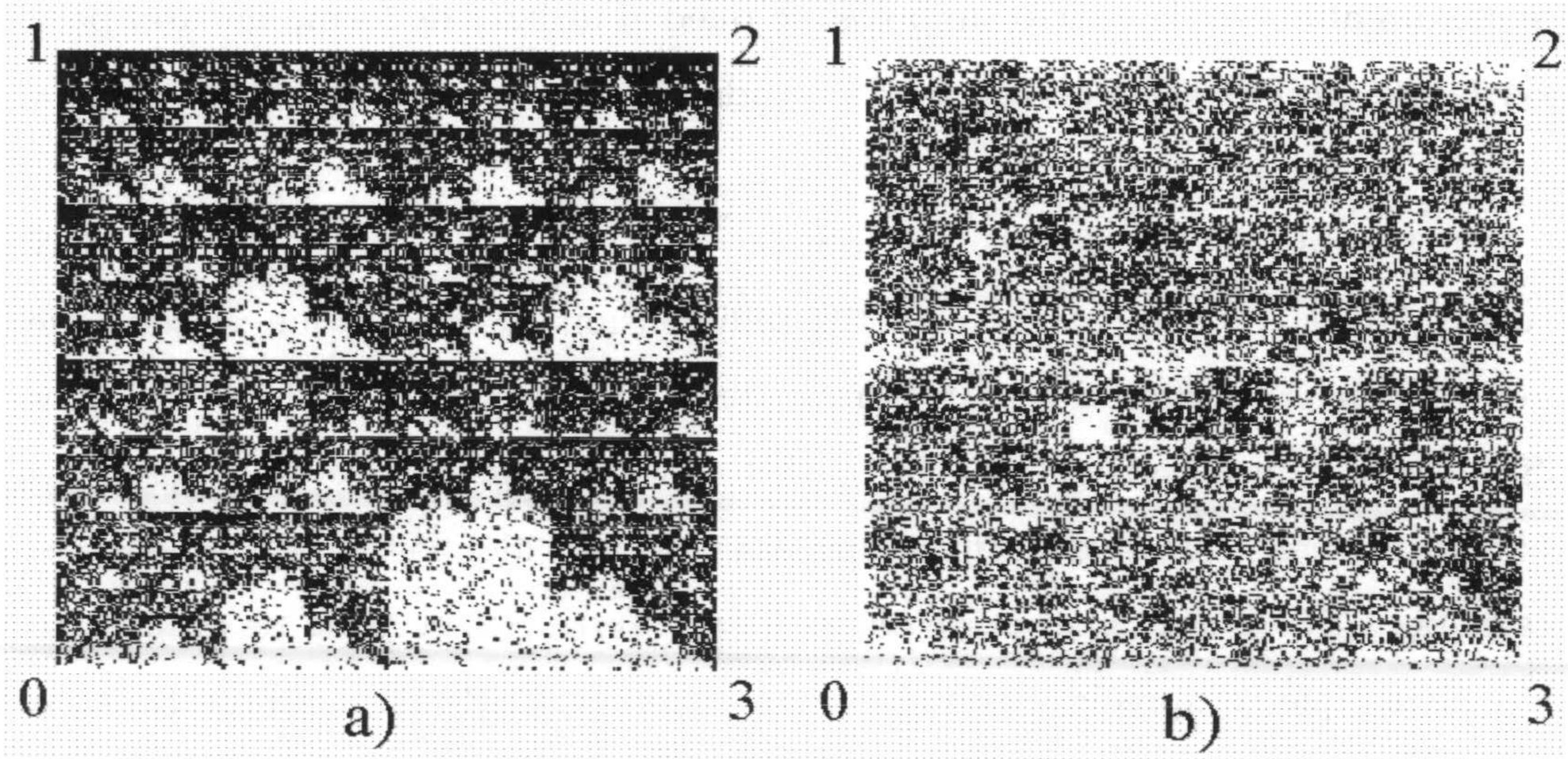

Figura 3.11: a) CSR de una secuencia humana (HUMHBB). b) CSR de una secuencia del virus Bacteriófago T7 (PT7CG).

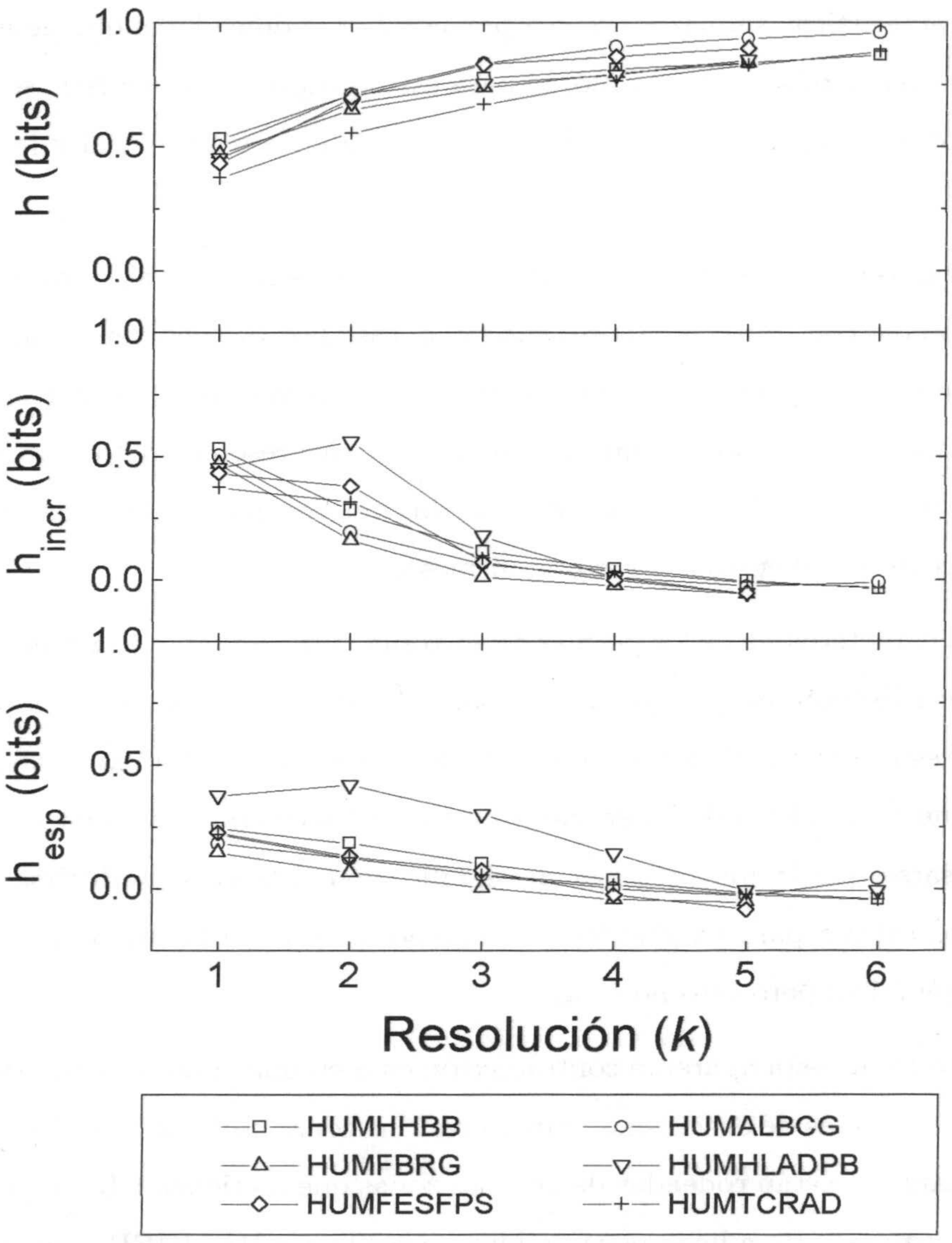

Figura 3.12: Pefiles entrópicos para algunas secuencias humanas.

Como ilustración a los métodos desarrollados en este capítulo, vamos a ver los perfiles entrópicos de algunas secuencias de ADN. Se han agrupado algunas secuencias con características similares (ya sea por la labor codificadoras que desempeñan o por pertenecer al mismo organismo) y, para estos grupos de secuencias, se han calculado los perfiles con la entropía del histograma, la entropía espacial y la entropía incremental.

En primer lugar, en la figura 3.12 tenemos los perfiles de algunas secuencias de ADN humano con diversas labores codificadoras. Los perfiles tienen un comportamiento similar para todas ellas, caben destacar los valores altos de la entropía del histograma para resoluciones pequeñas, de hecho, para resolución 1 se empieza con valores muy cercanos a 0.5 (compárense con los valores que se observan en figuras siguientes, correspondientes a otros organismos).

Una característica de los perfiles de entropía incremental que, a primera vista, pueda parecer extraña es que los valores máximos se obtengan en la mayoría de los casos para resolución 1. Teniendo en cuenta la estructura del código genético (pág. 15), en la que los tripletes de bases son los encargados de codificar cada aminoácido podría pensarse que la mayor carga de significación debiera introducirse para resolución 3, o tal vez para resolución 2 si tenemos en cuenta la fuerte degeneración del código genético; pero esto no es así.

La causa de esta aparente contradicción está en una característica bien conocida de las secuencias de organismos superiores: las zonas codificadoras (*exones*) son poco abundantes y están rodeadas de grandes zonas que no tienen labor codificadora conocida (*intrones*). Para hacernos una idea, la secuencia HUMHBB, que codifica la beta-globina humana, con una longitud total de 73326 pares de bases, contiene zonas codificadoras con una longitud total de unos 6000 pares de bases. Evidentemente, la existencia de estas grandes zonas no codificadoras hace que nuestra suposición an-

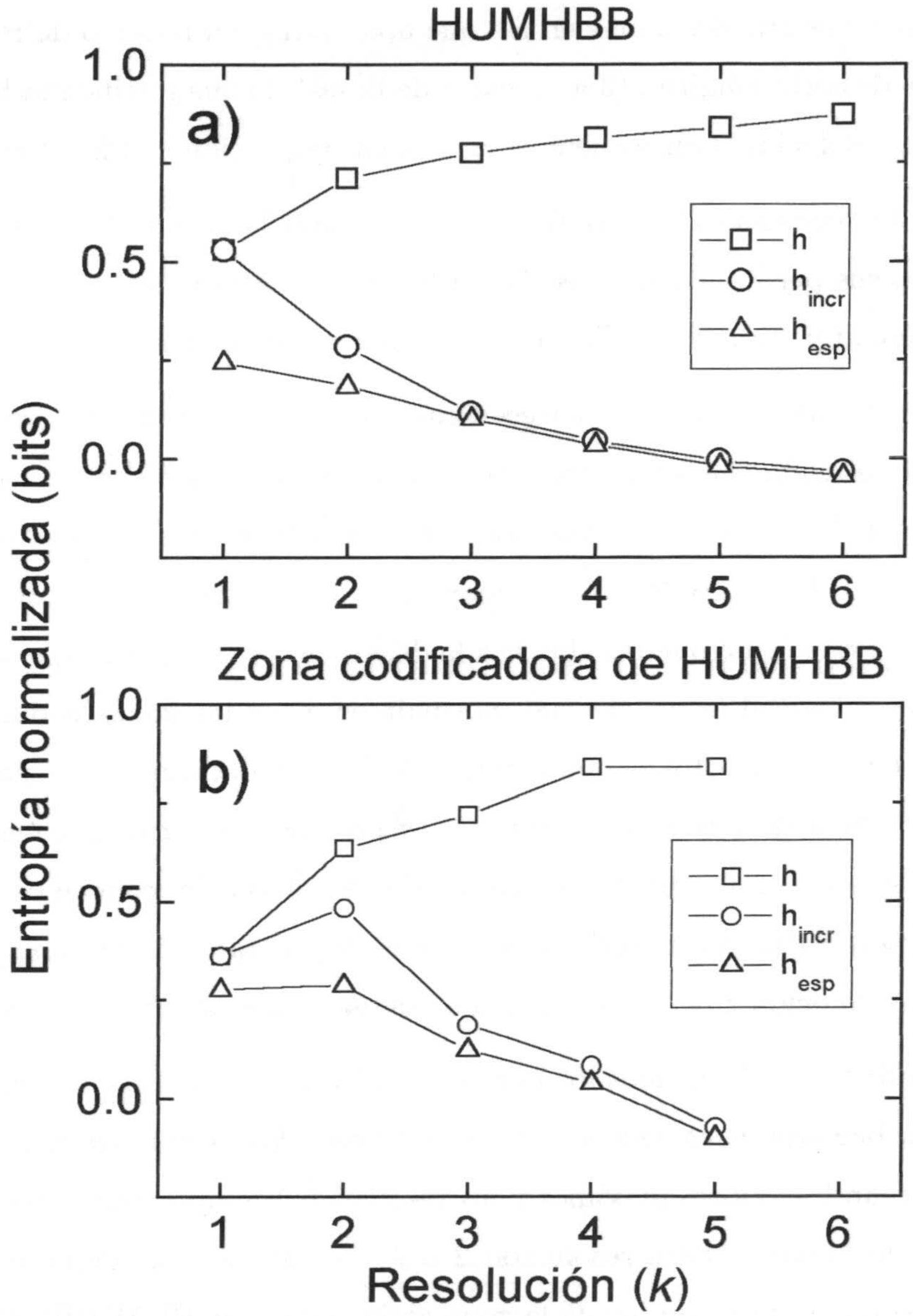

Figura 3.13: Perfiles entrópicos para de HUMHBB: a) Secuencia completa. b) Zona codificadora.

terior no estuviese bien fundamentada. Además, en estas zonas no codificadoras es frecuente encontrar grandes zonas en las que aparecen gran número de repeticiones de secuencias de corta longitud (3 a 10 pares de bases), lo que justificaría las bajadas de entropía a resolución 3 en secuencias con poca proporción codificadora.

Para la secuencia HUMHBB se han separado las zonas codificadoras y se han calculado sus perfiles entrópicos. En la figura 3.13.a) aparecen los perfiles de la secuencia original y en la 3.13.b) los de estas zonas codificadoras.

Entre las diferencias que aparecen cabe destacar, en primer lugar, el menor valor que toman todas las entropías para las resoluciones inferiores y sobre todo, cómo en los perfiles de zona codificadora se aprecia claramente un máximo en h_{incr} para resolución 2. El hecho de que este máximo se encuentre a resolución 2 en lugar encontrarse a resolución 3 puede deberse a la degeneración del código que se comentó anteriormente o bien al hecho de que, normalmente, en las zonas codificadoras el dinucleótido CG es poco frecuente a causa de la facilidad que tiene para mutar, además la ausencia de este dinucleótido no se presenta de forma homogénea, sino que quedan reductos en los que no ha disminuido su presencia, conocidos como *islas CG*. Esta característica puede ser la responsable de que también para h_{esp} aparezca un máximo a resolución 2, ya que esta medida detectaría esta falta de homogeneidad.

En la figura 3.14 tenemos los perfiles de algunas secuencias pertenecientes la genoma de la bacteria *Escherichia coli*. En este caso los perfiles de entropía incremental no toman los valores máximos para resolución 1 aunque tampoco se observa claramente que lo hagan para resolución 2 o 3. Lo que sí es evidente es la mayor similitud de estos perfiles con los de la zona codificadora de HUMHBB (fig. 3.13.b) que con los de la secuencia completa (fig. 3.13.a). Esto se justificaría con el hecho, también conocido, de que en los organismos inferiores la información está mucho más "empaquetada", siendo mayor la abundancia relativa de zonas codificadoras.

Por otra parte, los perfiles de entropía espacial, son inferiores a los de las secuencias humanas, indicando una mayor heterogeneidad espacial en estas últimas. Sobre esta cuestión se volverá de forma más extensa en capítulos posteriores.

También caben destacar los valores pequeños que se obtienen para la entropía del histograma a resoluciones 1 y 2.

En la figura 3.15 están los perfiles de los genomas completos de los cloroplastos de tres plantas CHNTXX, *Nicotiana tabacum* (tabaco) con una longitud de 155844 bp, CHOSXX, *Oriza sativa* (arroz) con 134525 bp y CHMPXX, *Marchantia polymorpha* (planta talofita) con 121024 bp.

De nuevo se encuentran valores relativamente pequeños de h_{inc} para resoluciones 2 y 3, la causa más probable es que, al tratarse del genoma completo del cloroplasto, se encuentran zonas no codificadoras que hacen disminuir la importancia de las restricciones impuestas a estas resoluciones por el código genético.

También es interesante destacar cómo los perfiles de las dos plantas superiores (tabaco y arroz) son bastante parecidos entre sí para todas las resoluciones, mientras que los perfiles de *Marchantia polymorpha* presenta claras diferencias con las otras dos plantas.

En la figura 3.16 se han representado los perfiles entrópicos para los genomas mitocondriales completos de algunos vertebrados y el de un microorganismo (*Paramecium aurelia*).

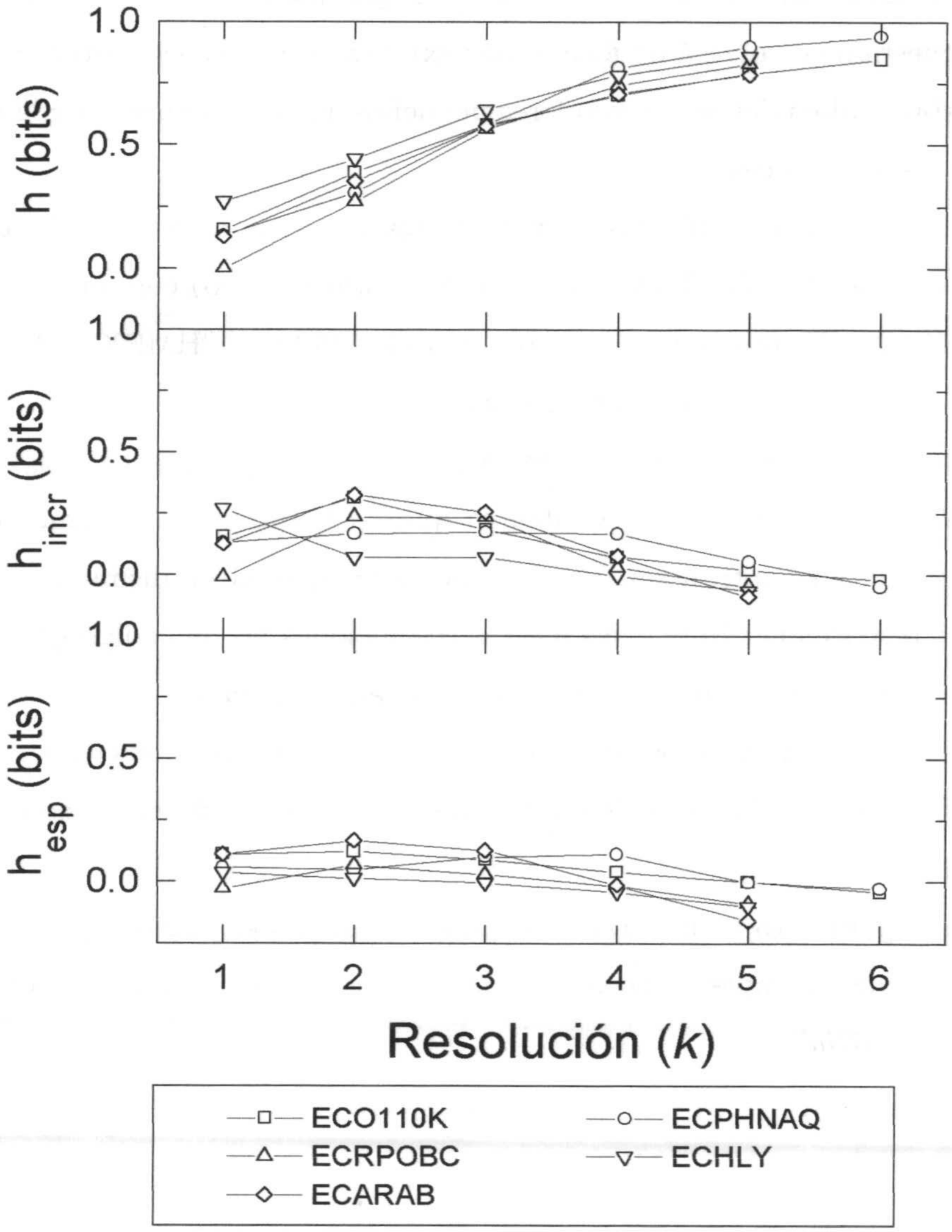

Figura 3.14: Perfiles entrópicos para algunas secuencias de *Escherichia coli*.

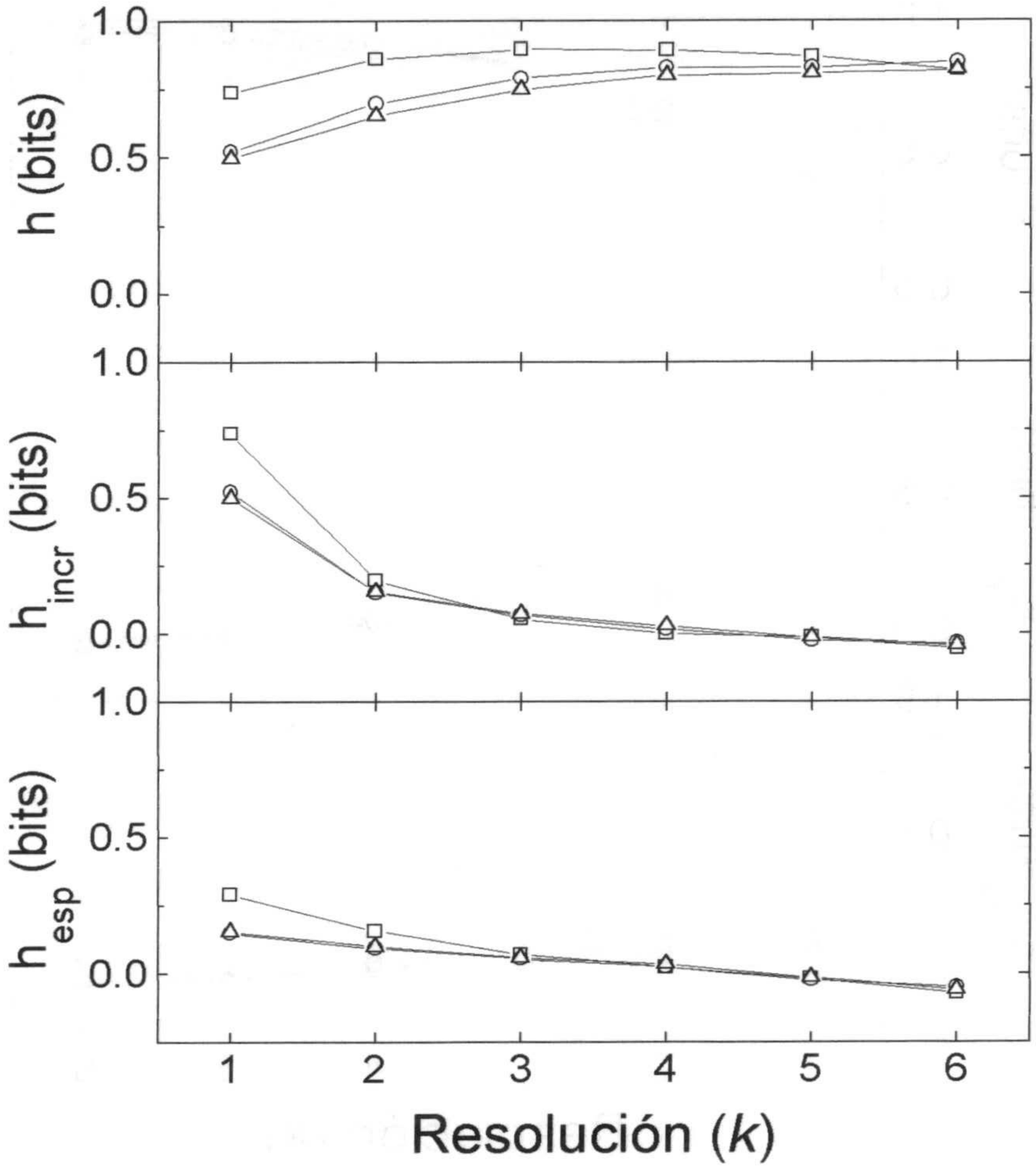

Figura 3.15: Perfiles entrópicos para las secuencias que codifican los cloroplastos.

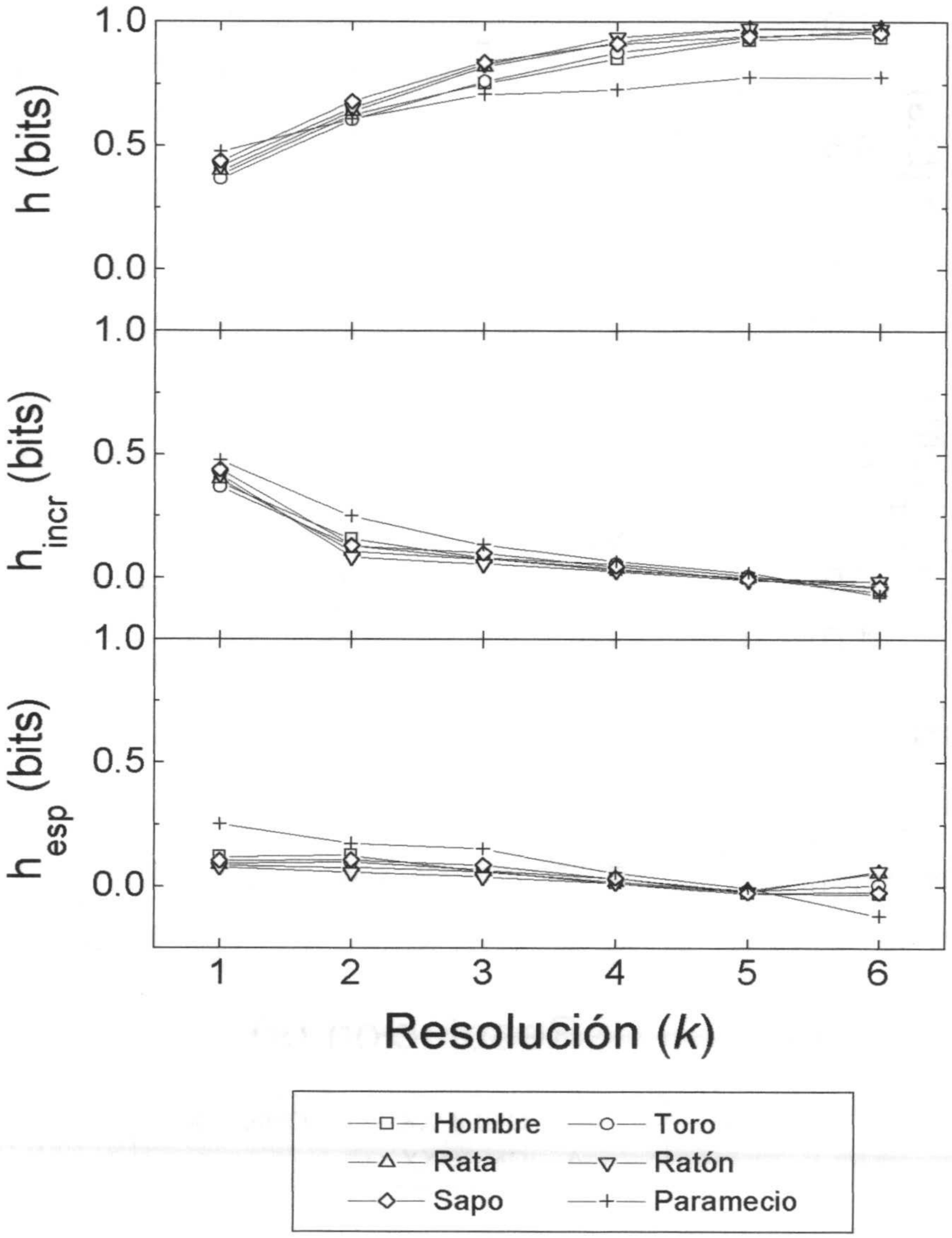

Figura 3.16: Perfiles entrópicos de las secuencias de ADN mitocondrial de algunos organismos.

Capítulo 4

Segmentación composicional de secuencias simbólicas

La determinación de dominios composicionales es una cuestión de gran importancia en disciplinas tan diversas como la segmentación de imágenes texturadas, fenómenos críticos, problemas de difusión, análisis de secuencias, reconocimiento de voz, y un largo etcétera. En concreto, como se vio en el capítulo 2, la organización de las zonas o dominios con diferente composición en las secuencias de ADN, es considerada por muchos autores como responsable de las correlaciones de largo alcance observadas en algunas de ellas; es por tanto de gran importancia, por una parte, una definición rigurosa de lo que se entiende por "zonas de diferente composición" y por otra, el desarrollo de técnicas que permitan determinar estos dominios. Esta cuestión, que resulta simple en algunas secuencias procariotas (por ejemplo [La 93], [Pe 92]), para secuencias más complejas, a veces fractales, se complica bastante a causa de la inexistencia de una longitud o escala característica, apareciendo irregularidades

que van desde algunos pares de bases hasta longitudes similares a las de la secuencia completa [Pe 94], es más, recientemente se habla de la existencia de dominios dentro de dominios ([Be 96], [Li 97a]), cuestión sobre la que se volverá más adelante.

En todos los casos el problema se puede plantear en términos similares, tenemos un conjunto de símbolos pertenecientes a cierto alfabeto, dispuesto según algún tipo de ordenación, y el objetivo es dividir el conjunto en zonas donde la composición sea lo más homogénea posible y diferente de las zonas que lo rodean; o lo que es lo mismo, encontrar los bordes que delimitan estas zonas. Por tanto, lo que se necesita es un método capaz de detectar variaciones locales de la composición. El proceso de segmentación, en cualquier caso, consta de tres elementos fundamentales: (1) *Estrategia general de segmentación:* forma en que se buscan los posibles bordes entre dominios. (2) *Distancia composicional.* Una medida que determine la diferencia de composición entre dominios, generalmente será una medida de distancia entre distribuciones de probabilidad. (3) *Criterio de significación.* Un criterio que decida hasta qué punto la diferencia de composición entre dos dominios es suficiente como para considerarlos distintos o no.

En este capítulo se presenta un algoritmo de segmentación basado en la divergencia de Jensen-Shannon, medida utilizada previamente por nuestro grupo para la segmentación de imágenes texturadas ([Ba 95a], [Ba 95b]), que coincide, salvo algunas mejoras recientemente introducidas, con el presentado, también por nuestro grupo, en [Be 96].

En primer lugar definiremos lo que vamos a entender por dominio composicional (sección 4.1), y a continuación desarrollaremos los tres elementos antes citados, que están presentes en cualquier técnica de segmentación: La medida de distancia composicional, que aquí será la divergencia de Jensen-Shannon (sección 4.2), el criterio de significación que, en este caso consiste en encontrar la distribución de

probabilidad que sigue la divergencia de Jensen-Shannon (sección 4.3) y el método de segmentación o, dicho de otro modo, la forma en que se buscan los posibles dominios composicionales (sección 4.4).

4.1 Dominio composicional

El concepto de dominio composicional es usado a veces en la descripción de secuencias de ADN, aunque no ha sido bien definido por el momento. Para nosotros un *dominio composicional es una subsecuencia que se caracteriza por tener una composición diferente de las subsecuencias adyacentes, con un cierto nivel de fiabilidad estadística dado.* Ésta, si bien, no es la única definición que podría darse, coincide con la idea intuitiva de "zonas con diferente composición", y parece bastante razonable y adecuada para los objetivos que aquí nos planteamos.

La fiabilidad estadística a la que se hace referencia en la definición, garantiza, hasta ese nivel dado, que los dominios en cuestión no se habrían obtenido en una generación aleatoria de la secuencia. Hay que tener en cuenta que esta posibilidad no puede ser excluida de manera rigurosa: cualquier par de dominios contiguos, por diferentes y regulares que sean, tienen alguna probabilidad de ser generados por azar.

Nótese, por último, que en la definición no se hace referencia a la composición interna del dominio, esto es, no se excluye la posibilidad de que un dominio pueda estar compuesto por varios dominios, incluso al mismo nivel de significación, con lo que la descomposición de una secuencia dada en dominios dependerá del método escogido para segmentarla. Como se verá más adelante, esta exigencia de homogeneidad interna, que también parece deseable para un dominio y que no se impone en la definición se verificará, de forma aproximada, en los dominios resultantes de nuestro algoritmo de segmentación.

4.2 Divergencia de Jensen-Shannon

4.2.1 Definición y primeras propiedades

Para cuantificar la diferencia entre la composición de segmentos adyacentes se puede usar cualquier medida de distancia entre distribuciones (varianza, divergencia de Kullback, etc. [Bov 84]) aplicada sobre los vectores de frecuencias de los segmentos, considerando éstos como distribuciones de probabilidad. Aquí adoptaremos la Divergencia de Jensen-Shannon (JS_2), medida introducida recientemente por J. Lin [Lin 91] y que también ha sido utilizada por nuestro grupo en los trabajos de detección de bordes en imágenes texturadas que antes se han comentado.

Consideremos dos secuencias $S^{(1)} = \{a_1, a_2, ..., a_{n^{(1)}}\}$ y $S^{(2)} = \{b_1, b_2, ..., b_{n^{(2)}}\}$ formadas por símbolos pertenecientes a cierto alfabeto $\mathcal{A} = \{A_1, A_2, ...A_k\}$. Los respectivos vectores de frecuencias relativas los notaremos como:

$$\mathcal{F}^{(1)} = \{f_1^{(1)}, f_2^{(1)}, ..., f_k^{(1)}\} \text{ y } \mathcal{F}^{(2)} = \{f_1^{(2)}, f_2^{(2)}, ..., f_k^{(2)}\}$$

Se define la Divergencia de Jensen-Shannon entre $S^{(1)}$ y $S^{(2)}$, o más exactamente entre $\mathcal{F}^{(1)}$y $\mathcal{F}^{(2)}$, como:

$$\begin{aligned} &\mathrm{JS}_2\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \\ &= \mathrm{H}\left(\pi^{(1)}\mathcal{F}^{(1)} + \pi^{(2)}\mathcal{F}^{(2)}\right) - \left[\pi^{(1)}\mathrm{H}\left(\mathcal{F}^{(1)}\right) + \pi^{(2)}\mathrm{H}\left(\mathcal{F}^{(2)}\right)\right] \end{aligned} \tag{4.1}$$

donde

$$\mathrm{H}\left(\mathcal{F}\right) = -\sum_{i=i}^{k} f_i \log f_i \tag{4.2}$$

es la entropía de Shannon de la distribución* $\mathcal{F}$, y $\pi^{(1)}, \pi^{(2)} \geq 0, \pi^{(1)} + \pi^{(2)} = 1$ son los pesos asignados a $\mathcal{F}^{(1)}$ y $\mathcal{F}^{(2)}$ respectivamente.

*Mientras no se diga lo contrario, usaremos log para referirnos al logaritmo en base 2.

A partir de la desigualdad de Jensen [Cov 91] es fácil comprobar que $\mathrm{JS}_2 \geq 0$, dándose la igualdad, si y sólo si $\mathcal{F}^{(1)} = \mathcal{F}^{(2)}$.

Otras propiedades relevantes de esta medida son:

1. Es simétrica con respecto a sus argumentos (propiedad que, a pesar de su denominación, no cumplen muchas de las distancias entre distribuciones)

2. $\mathcal{F}^{(1)}$y $\mathcal{F}^{(2)}$ no tienen que ser absolutamente continuas (véase p.e. [Bov 84] para una definición rigurosa de esta propiedad).

3. Es generalizable como medida de distancia o dispersión entre un número arbitrario de distribuciones. Esta propiedad se utilizará en la sección 4.4.1 para definir la divergencia total de la segmentación.

4. Tiene una cota superior alcanzable, que viene dada por: $\min\{\log k$, log(n^0 de distribuciones)$\}$[RP 95].

5. Se pueden asignar pesos diferentes a las distribuciones.

4.2.2 Ventajas como medida de distancia entre segmentos

Además de estas propiedades que son convenientes, en general, para cualquier medida de distancia entre distribuciones, JS_2 tiene tres ventajas adicionales para la aplicación particular que nos ocupa.

En primer lugar, la posibilidad de asignar diferentes pesos a las distribuciones hace posible tener en cuenta la influencia de las longitudes de las subsecuencias comparadas. Para ello tomamos:

$$\pi^{(1)} = \frac{n^{(1)}}{N} \quad \text{y} \quad \pi^{(2)} = \frac{n^{(2)}}{N} \tag{4.3}$$

donde $N = n^{(1)} + n^{(2)}$. Por supuesto, ésta no es la única elección de pesos que, a priori, puede eliminar la influencia de las distintas longitudes, pero se verá más adelante que facilita considerablemente la estimación de la distribución de probabilidad de JS_2. Además, con esta elección, se tiene que:

$$\pi^{(1)}\mathcal{F}^{(1)} + \pi^{(2)}\mathcal{F}^{(2)} = \frac{n^{(1)}}{N}\mathcal{F}^{(1)} + \frac{n^{(2)}}{N}\mathcal{F}^{(2)} = \mathcal{F} \tag{4.4}$$

es el vector de frecuencias de la secuencia formada por la concatenación de los dos segmentos, con lo que (4.1) se puede reescribir como:

$$\mathrm{JS}_2\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \mathrm{H}\left(\mathcal{F}\right) - \left[\frac{n^{(1)}}{N}\mathrm{H}\left(\mathcal{F}^{(1)}\right) + \frac{n^{(2)}}{N}\mathrm{H}\left(\mathcal{F}^{(2)}\right)\right] \tag{4.5}$$

con lo que JS_2 queda como la diferencia entre la entropía de la secuencia completa (unión de los dos segmentos) menos la suma ponderada de las entropías de los segmentos, que se puede interpretar como el aumento de información sobre la secuencia que introduce la división de ésta en dos segmentos. Como curiosidad, citar que esta expresión coincide con la del aumento de entropía termodinámica en una mezcla de gases ideales.

En segundo lugar, otra ventaja de la medida es que, a causa del valor bajo de k que utilizaremos (tamaño del alfabeto), junto con el hecho de que tenemos sólo dos distribuciones, se evita la aparición de valores muy altos de JS_2, próximos a la cota superior, que provocaría una disminución del poder de discriminación de la medida.

Y por último, es posible encontrar una aproximación asintótica para la distribución de probabilidad que sigue JS_2 (vista como un estadístico), con una expresión muy simple en términos de la distribución χ^2 (ver sec. 4.3).

Antes de continuar vamos a ver un par de ejemplos que ponen de manifiesto la conveniencia de la elección de pesos que hemos hecho. En primer lugar, se

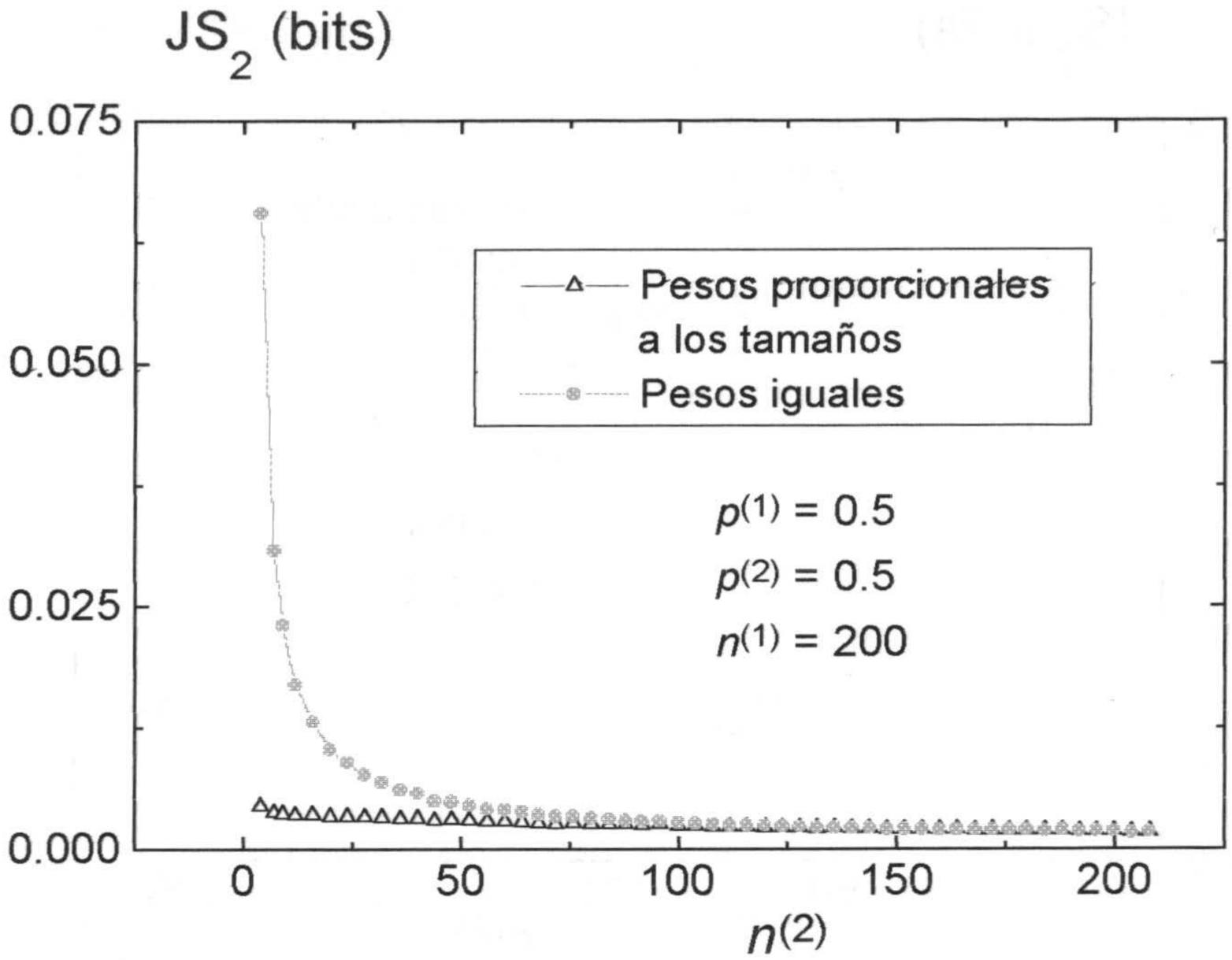

Figura 4.1: Comparación de los valores de JS_2 con los pesos proporcionales a los tamaños y con pesos iguales, para segmentos generados con igual distribución de probabilidad.

han generado secuencias binarias ($k = 2$) pseudoaleatorias equiprobables (generador *RAN*3 [Pr 94]) de longitud fija ($n^{(1)} = 200$) y se ha calculado la divergencia JS_2 de estas secuencias con otras también pseudoaleatorias y equiprobables de distintas longitudes ($n^{(2)}$), promediando para cada valor sobre 5000 experimentos. Los resultados se muestran en la figura 4.1. En principio cabría esperar que JS_2 fuese nula para cualquier valor de $n^{(2)}$ (incluso para cualquier elección de los pesos), ya que ambas secuencias han sido obtenidas a partir de la misma distribución de probabilidad,

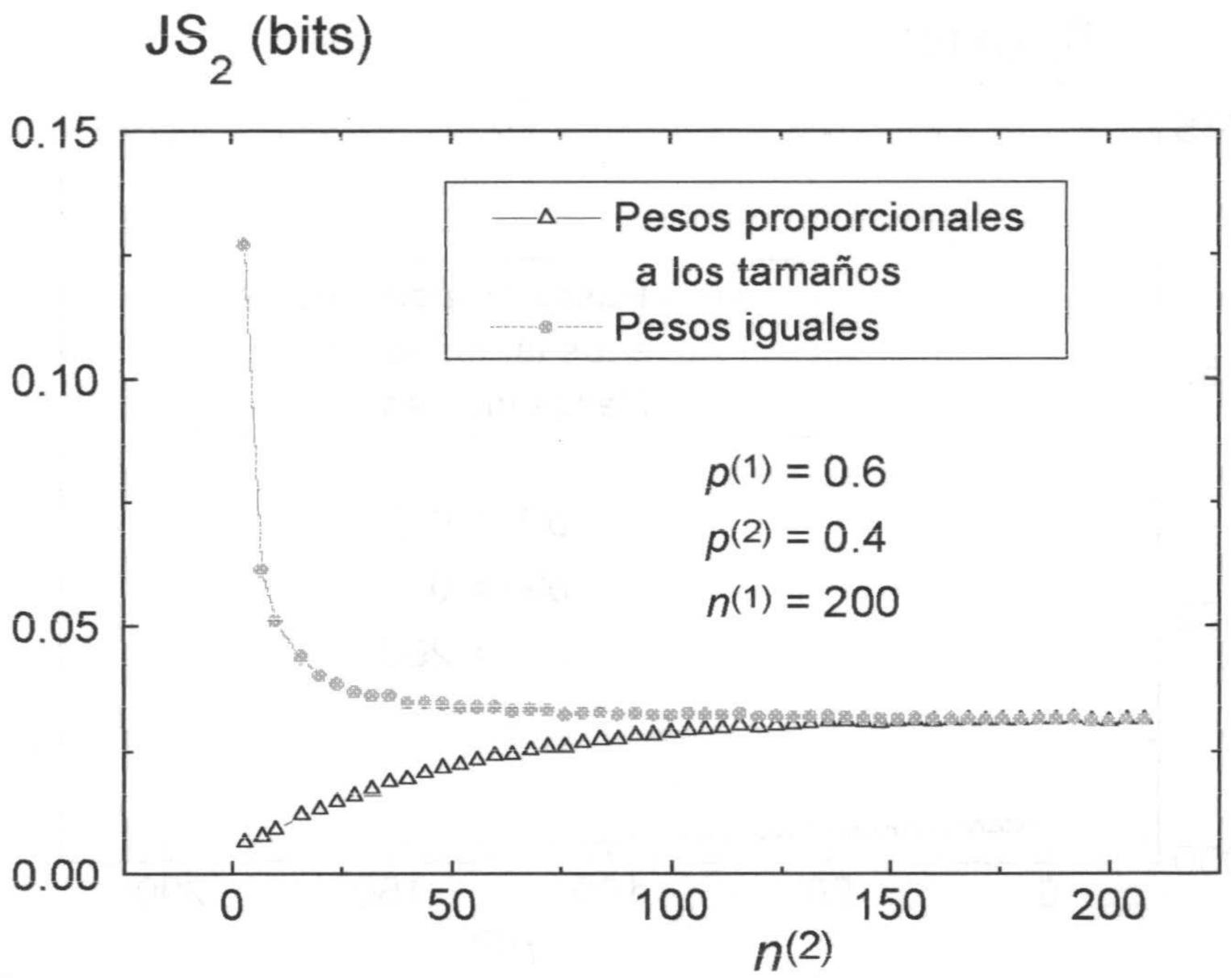

Figura 4.2: Comparación de los valores de JS_2 con los pesos proporcionales a los tamaños y con pesos iguales, para segmentos generados con distribuciónes de probabilidad diferentes.

pero esto no ocurre así a causa del tamaño finito de las muestras: en todos los casos la divergencia toma valores positivos distintos de cero (sobre esta cuestión se volverá con más detalle en la sección 4.3) . Sin embargo cuando se toman los pesos iguales ($\pi^{(1)} = \pi^{(2)} = \frac{1}{2}$), la diferencia de tamaño entre las secuencias provoca la aparición de valores "espúreos" de divergencia que son considerablemente grandes, para valores pequeños de $n^{(2)}$, es decir, cuando es más notoria la diferencia de tamaño; por lo que esta elección de pesos hace que JS_2 "detecte" como una diferencia entre

las secuencias su distinto tamaño, cosa que no ocurre, o al menos no de forma tan evidente, con la elección de pesos proporcional a los tamaños.

Por otra parte, consideremos ahora secuencias binarias generadas también de forma pseudoaleatoria pero ahora con diferentes distribuciones de probabilidad. En la figura 4.2 se representan los valores de JS_2 entre segmentos generados de esta forma: el primero tiene un tamaño fijo $n^{(1)} = 200$ y se ha generado con una probabilidad de aparición de 1's del 60%, mientras que el segundo tiene un tamaño que varía desde $n^{(2)} = 4$ hasta 200 y se ha generado con una probabilidad de aparición de 1's del 40 % (también se ha promediado sobre 5000 experimentos). De nuevo la elección de pesos iguales lleva a la aparición de valores altos de divergencia para $n^{(2)} \ll n^{(1)}$, sin embargo, con los pesos proporcionales a los tamaños la divergencia entre los segmentos disminuye al hacerlo $n^{(2)}$. Este comportamiento también es razonable para la medida que necesitamos: aunque la diferencia de composición sea la misma para todas las parejas de segmentos, cuando uno de ellos es muy pequeño, ésta diferencia podría ser debida a simples fluctuaciones estadísticas, y por tanto es más probable que, a pesar de la diferencia, ambos segmentos hubiesen sido generados con la misma distribución de probabilidad (o al menos con distribuciones más parecidas), y por tanto parece lógico que se asigne a esta situación una menor diferencia.

Esta dependencia de JS_2 con el tamaño relativo de los segmentos (para vectores de frecuencias dados) se ilustra en la figura 4.3 para secuencias binarias: La curva gruesa representa la entropía de Shannon, y los puntos A y B las correspondientes entropías de los dos segmentos; cualquier combinación lineal convexa de estas entropías estará sobre la línea que une los puntos A y B, mientras que el valor de la entropía de la correspondiente combinación lineal de vectores de frecuencias estará sobre la curva, por tanto, en cada caso los valores de JS_2 vendrán dados por las dis-

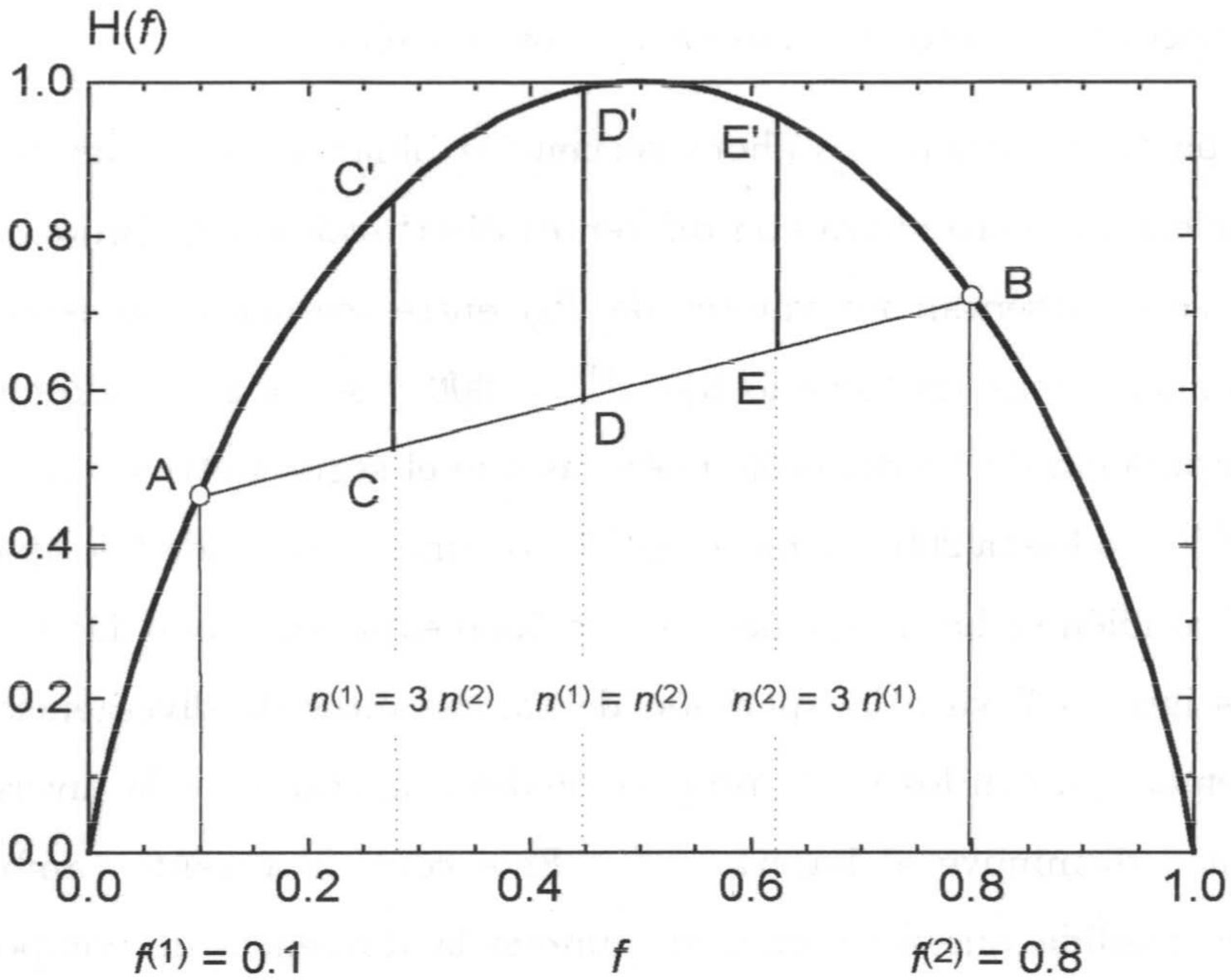

Figura 4.3: Representación gráfica de la dependencia de JS_2 con el tamaño relativo de los dos segmentos para la elección de pesos proporcionales a los tamaños.

tancias CC', DD' y EE' respectivamente. Sobre esta gráfica se puede ver con mayor claridad lo que comentábamos antes: la combinación lineal que corresponde al mismo tamaño para los dos segmentos está en el centro de AB, y alcanza el máximo valor de divergencia, mientras que los puntos desplazados a derecha e izquierda corresponden a situaciones en las que el primer segmento es menor o mayor que el segundo respectivamente, y como se puede observar en estas situaciones la divergencia es menor.

4.2.3 Agrupamiento

Otra propiedad interesante de JS_2, con esta elección de pesos, es la relación que existe entre la divergencia entre dos segmentos, considerando un alfabeto dado y la que se obtiene cuando los elementos del alfabeto se agrupan en clases.

Consideremos una secuencia generada con un alfabeto $\mathcal{A}$, con un vector de frecuencias respecto de dicho alfabeto que notaremos como $\mathcal{F}$, y supongamos que agrupamos los elementos del alfabeto en m clases disjuntas. Para facilitar la notación, reordenamos los elementos del vector de frecuencias:

$$\begin{aligned} \mathcal{F} \quad = \quad & \{f_{1,1}, f_{1,2}, ..., f_{1,k_1}, \\ & f_{2,1}, f_{2,2}, ..., f_{2,k_2}, \\ & ..... \\ & f_{m,1}, f_{m,2}, ..., f_{m,k_m}\} \end{aligned} \tag{4.6}$$

El nuevo vector de frecuencias, con respecto al alfabeto *reducido* se podrá escribir como: $\mathcal{G} = \{g_1, g_2, ..., g_m\}$ donde $g_i = \sum_{j=1}^{k_i} f_{i,j}$ con $i = 1, 2, ..., m$.

Se puede ver [Cov 91] que la entropía de Shannon de este vector de frecuencias está relacionada con la del original por:

$$\mathrm{H}(\mathcal{F}) = \mathrm{H}(\mathcal{G}) + \sum_{i=1}^{m} g_i \mathrm{H}\left(\frac{f_{i,1}}{g_i}, \frac{f_{i,2}}{g_i}, ..., \frac{f_{i,k_i}}{g_i}\right) \tag{4.7}$$

que en el marco de la Teoría de la Información tiene una interpretación inmediata: la información necesaria para conocer qué símbolo ha emitido la fuente se puede desglosar en dos términos, uno que tiene en cuenta la información sobre a cuál de las m clases pertenece, $\mathrm{H}(\mathcal{G})$, más otro término que tiene en cuenta la identificación del símbolo dentro de la clase. Nótese que este último está pesado en proporción a la frecuencia relativa de aparición de los elementos de la clase (g_i).

Modificando la notación de (4.7), podemos obtener una interpretación ligeramente diferente, que va a ser más adecuada en lo que sigue. Si llamamos N_i al

número de símbolos de la clase i que hay en la secuencia y definimos:

$$\Phi_i = \{\varphi_{i,1}, \varphi_{i,2}, ..., \varphi_{i,k_i}\} = \{\frac{f_{i,1}}{g_i}, \frac{f_{i,2}}{g_i}, ..., \frac{f_{i,k_i}}{g_i}\} \text{ con } i = 1, 2, ..., m \qquad (4.8)$$

tenemos:

$$\mathrm{H}(\mathcal{F}) = \mathrm{H}(\mathcal{G}) + \sum_{i=1}^{m} \frac{N_i}{N} \mathrm{H}(\Phi_i) \qquad (4.9)$$

y ahora el segundo sumando se puede interpretar como la suma de entropías de las secuencias obtenidas extrayendo de la secuencia original sólo los elementos pertenecientes a cada clase, ponderadas según su tamaño relativo con respecto a la secuencia total.

A partir de (4.9) es fácil comprobar que la divergencia de Jensen-Shannon con la elección de pesos de (4.3) tiene una propiedad similar:

$$\mathrm{JS}_2\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \mathrm{H}(\mathcal{F}) - \left[\frac{n^{(1)}}{N}\mathrm{H}\left(\mathcal{F}^{(1)}\right) + \frac{n^{(2)}}{N}\mathrm{H}\left(\mathcal{F}^{(2)}\right)\right] =$$

$$= \mathrm{H}(\mathcal{G}) + \sum_{i=1}^{m} \frac{N_i}{N}\mathrm{H}(\Phi_i) -$$

$$-\frac{n^{(1)}}{N}\left[\mathrm{H}\left(\mathcal{G}^{(1)}\right) + \sum_{i=1}^{m} \frac{n_i^{(1)}}{n^{(1)}}\mathrm{H}\left(\Phi_i^{(1)}\right)\right] - \frac{n^{(2)}}{N}\left[\mathrm{H}\left(\mathcal{G}^{(2)}\right) + \sum_{i=1}^{m} \frac{n_i^{(2)}}{n^{(2)}}\mathrm{H}\left(\Phi_i^{(2)}\right)\right] =$$

$$= \mathrm{JS}_2\left(\mathcal{G}^{(1)}, \mathcal{G}^{(2)}\right) + \sum_{i=1}^{m} \frac{N_i}{N}\left[\mathrm{H}(\Phi_i) - \frac{n_i^{(1)}}{N_i}\mathrm{H}\left(\Phi_i^{(1)}\right) - \frac{n_i^{(2)}}{N_i}\mathrm{H}\left(\Phi_i^{(2)}\right)\right] \Longrightarrow$$

$$\mathrm{JS}_2\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = \mathrm{JS}_2\left(\mathcal{G}^{(1)}, \mathcal{G}^{(2)}\right) + \sum_{i=1}^{m} \frac{N_i}{N}\mathrm{JS}_2\left(\Phi_i^{(1)}, \Phi_i^{(2)}\right) \qquad (4.10)$$

Esto es: la divergencia global entre las dos secuencias se debe a la distinta composición en lo que a los elementos del alfabeto reducido se refiere, más un término, pesado, que da cuenta de la diferente composición con respecto a los símbolos agrupados.

Para ver la utilidad de esta propiedad, consideremos las dos secuencias de ADN:

$$S^{(1)} = \{A, G, G, T, C, A, C, G, A, G, C, A\}$$
$$S^{(2)} = \{T, C, C, A, C, T, G, G, C, C, C, A\}$$

El alfabeto "natural" es $\mathcal{A} = \{A, T, C, G\}$ y consideremos la agrupación de nucleótidos en purinas R $= \{A, G\}$ y pirimidinas Y$= \{T, C\}$. Los vectores de frecuencias son:

	f_A	f_T	f_C	f_G
$S^{(1)}$	$\frac{1}{3}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{3}$
$S^{(2)}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{1}{6}$
S	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{4}$

	g_R	g_Y
$S^{(1)}$	$\frac{2}{3}$	$\frac{1}{3}$
$S^{(2)}$	$\frac{1}{3}$	$\frac{2}{3}$
S	$\frac{1}{2}$	$\frac{1}{2}$

	$\varphi_{R,A}$	$\varphi_{R,G}$	$\varphi_{Y,T}$	$\varphi_{Y,C}$
$S^{(1)}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
$S^{(2)}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$
S	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{3}{4}$

y las divergencias:

$$\mathrm{JS}_2\left(\mathcal{F}^{(1)}, \mathcal{F}^{(2)}\right) = 0.0817 \quad \mathrm{JS}_2\left(\mathcal{G}^{(1)}, \mathcal{G}^{(2)}\right) = 0.0817$$

$$\mathrm{JS}_2\left(\Phi_R^{(1)}, \Phi_R^{(2)}\right) = 0 \qquad \mathrm{JS}_2\left(\Phi_Y^{(1)}, \Phi_Y^{(2)}\right) = 0$$

En este caso, el desglose de divergencias nos dice que la diferencia entre las dos secuencias es debida únicamente a la diferente composición en purinas/pirimidinas, no a la diferente frecuencia de aparición de cada purina o cada pririmidina, ya que en este caso tenemos $\Phi_R^{(1)} = \Phi_R^{(2)}$ y $\Phi_Y^{(1)} = \Phi_Y^{(2)}$.

En general, análisis de este tipo pueden ayudar a determinar a qué símbolos se le pueden achacar las variaciones de composición más significativas. En el análisis de secuencias de ADN, éste es un problema interesante, que se ha abordado comparando el comportamiento de distintas funciones de autocorrelación (correlaciones entre purinas/pirimidinas, correlaciones de un nucleótido frente al resto, etc. [He 95], [Te 96]), pero, hasta el momento no se han obtenido expresiones con las propiedades de aditividad de (4.10).

4.3 Criterio de fiabilidad estadística

Según la definición de JS_2, cualquier valor distinto de cero implica $\mathcal{F}^{(1)} \neq \mathcal{F}^{(2)}$, esto, en secuencias con longitud infinita implicaría que ambas han sido generadas con distribuciones de probabilidad distintas, con lo que serían distintos dominios composicionales. Pero cuando tenemos muestras de tamaño finito, nunca se puede asegurar que dos vectores de frecuencias distintos no hayan sido generados con la misma distribución de probabilidad, sólo se puede hablar de la probabilidad de que esto no haya ocurrido. Esta probabilidad justamente es lo que nos dará el *nivel de fiabilidad* (s), al que se hace referencia en la definición de dominio composicional (sec. 4.1)

Planteando el problema en términos estadísticos, el nivel de fiabilidad será la probabilidad de que al generar una secuencia aleatoria de símbolos independientes e idénticamente distribuidos (i.i.d. en lo sucesivo), y dividirla en dos subsecuencias por la posición dada, el valor de divergencia sea menor o igual que el obtenido; entonces

podremos decir que la división en dominios tiene una fiabilidad de s, puesto que s de cada 100 veces que se genere aleatoriamente una secuencia con los mismos $\mathcal{F}$, $n^{(1)}$ y $n^{(2)}$, se obtendrán valores menores de divergencia.

4.3.1 Sesgo del estimador de JS_2

No hay que olvidar que cuando calculamos JS_2 a partir de dos secuencias dadas, realmente estamos calculando un *estimador* de la divergencia y, como suele ocurrir con los estadísticos derivados directamente de la entropía de Shannon [He 94b], se trata de un estimador sesgado. Téngase en cuenta que no estamos calculando la entropía de la fuente (a la que no tenemos acceso) sino la entropía de una secuencia emitida por ella, esto es, una realización. Por ello vamos a hacer una estimación de este sesgo, que en determinadas circunstancias puede ser útil para comparar divergencias cuando éstas se han obtenido sobre muestras de diferente tamaño ya que, como veremos, éste depende principalmente de la longitud de la secuencia.

Si $\mathcal{F}$ es el vector de frecuencias de una realización de una secuencia de variables aleatorias i.i.d. que siguen una distribución de probabilidad $\mathcal{P}$, es sabido que la entropía de la muestra subestima de forma sistemática la entropía de la fuente, de hecho se puede probar [Mas 87] que, en promedio, las entropías muestrales cumplen:

$$\mathrm{E}\left[\mathrm{H}\left(\mathcal{F}\right)\right] \leq \mathrm{H}\left(\mathcal{P}\right) \tag{4.11}$$

En estas condiciones podemos considerar la fuente como de Bernuilli y se comprueba que, en primera aproximación (para secuencias suficientemente grandes) se cumple [He 94b]:

$$\mathrm{E}\left[\mathrm{H}\left(\mathcal{F}\right)\right] = \mathrm{H}\left(\mathcal{P}\right) - \frac{k-1}{2N\ln 2} \tag{4.12}$$

donde N es la longitud de la secuencia y k el número de elementos del alfabeto.

Trasladando este resultado al problema de la segmentación, aquí tenemos una secuencia S dividida en dos subsecuencias y se tiene que:

$$\begin{aligned} \mathcal{F} &= \frac{n^{(1)}}{N}\mathcal{F}^{(1)} + \frac{n^{(2)}}{N}\mathcal{F}^{(2)}, \\ \text{o bien } \mathcal{Q} &= \mathcal{Q}^{(1)} + \mathcal{Q}^{(2)}, \end{aligned} \tag{4.13}$$

donde $\mathcal{Q} = N\mathcal{F}$ es el vector de frecuencias absolutas. Tenemos, por tanto, una restricción (vector de frecuencias total dado), y si suponemos S como una fuente de Bernuilli, $S^{(1)}$ y $S^{(2)}$ no se pueden considerar como tales; la probabilidad de encontrar un vector de frecuencias determinado en una de las subsecuencias vendrá dado por la distribución *Hipergeométrica Multivariada* ($k-1$ variables). Para convencerse de esto, basta con tener en cuenta que obtener un vector de frecuencias dado, fijado el vector de frecuencias global, es un claro ejemplo de extracción de $n^{(1)}$ símbolos de un total de N, *sin reemplazamiento.*

Sin embargo, para $N \to \infty$ sí será válida la suposición y podremos aplicar (4.12) a las entropías de la secuencia completa y las subsecuencias:

$$\begin{aligned} \mathrm{E}\left[\mathrm{H}\left(\mathcal{F}^{(1)}\right)\right] &= \mathrm{H}(\mathcal{P}) - \frac{k-1}{2n^{(1)}\ln 2} \\ \mathrm{E}\left[\mathrm{H}\left(\mathcal{F}^{(2)}\right)\right] &= \mathrm{H}(\mathcal{P}) - \frac{k-1}{2n^{(2)}\ln 2} \\ \mathrm{E}\left[\mathrm{H}(\mathcal{F})\right] &= \mathrm{H}(\mathcal{P}) - \frac{k-1}{2N\ln 2} \end{aligned} \tag{4.14}$$

nótese que, al considerar S como la realización de una serie de variables aleatorias i.i.d. la entropía de la fuente es la misma en los tres casos. Por tanto el valor esperado de la divergencia que se obtiene al segmentar S, vendrá dado por:

$$\mathrm{E}\left[\mathrm{JS}_2\right] = \mathrm{E}\left[\mathrm{H}(\mathcal{F})\right] - \left(\frac{n^{(1)}}{N}\mathrm{E}\left[\mathrm{H}\left(\mathcal{F}^{(1)}\right)\right] + \frac{n^{(2)}}{N}\mathrm{E}\left[\mathrm{H}\left(\mathcal{F}^{(2)}\right)\right]\right) =$$

$$= \mathrm{H}\left[\mathcal{P}\right] - \frac{k-1}{2N\ln 2} - \left(\frac{n^{(1)}}{N}\mathrm{H}\left(\mathcal{P}\right) - \frac{k-1}{2n^{(1)}\ln 2} + \frac{n^{(2)}}{N}\mathrm{H}\left(\mathcal{P}\right) - \frac{k-1}{2n^{(2)}\ln 2}\right) \Rightarrow$$

$$E\left[\mathrm{JS}_2\right] = \frac{k-1}{2N\ln 2} \tag{4.15}$$

Puesto que este valor se obtiene al segmentar una secuencia de variables aleatorias i.i.d. (que en principio no debería dividirse en dominios composicionales), se puede considerar como un "error de cero" de la medida.

Vemos que, en esta primera aproximación, $\mathrm{E}[\mathrm{JS}_2]$ es independiente de $n^{(1)}$, $n^{(2)}$ y $\mathcal{F}$, por lo que parece razonable suponer que la distribución de probabilidad de JS_2 también sea sólo función de N y k; esto es, que los cuantiles de esta distribución se puedan expresar, siempre suponiendo válida esta aproximación, como función únicamente de N, k y s:

$$D_s(N,k) = f\left(N,k,s\right) \tag{4.16}$$

donde $D_s(N,k)$ es el valor de JS_2 tal que, la probabilidad de que ocurran valores menores o iguales a él es s. Esto lo veremos más detenidamente en la sección siguiente.

4.3.2 Distribución de probabilidad de JS_2

No parece sencillo encontrar de forma analítica la distribución de probabilidad de JS_2. Para alfabetos pequeños se puede calcular numéricamente aunque con gran esfuerzo computacional para valores grandes de N. En esta sección vamos a ver una aproximación que, para N grande, permite expresar esta distribución en términos de la función gamma incompleta. En primer lugar lo haremos para alfabetos binarios y a continuación generalizaremos para número arbitrario de símbolos.

Alfabetos binarios

Consideremos ahora $k = 2$, si suponemos fijo el vector de frecuencias de S, $\mathcal{Q} = \{q_1, q_2\}$, y los tamaños de las subsecuencias, cada configuración (i.e. $\mathcal{Q}^{(1)}$, $\mathcal{Q}^{(2)}$) tendrá una probabilidad:

$$\text{Prob}\left\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)} \;/\; \mathcal{Q}\right\} = \frac{\binom{q_1}{q_1^{(1)}}\binom{q_2}{q_2^{(1)}}}{\binom{N}{n^{(1)}}} = \frac{\binom{n^{(1)}}{q_1^{(1)}}\binom{N-n^{(1)}}{q_1-q_1^{(1)}}}{\binom{N}{q_1}} \tag{4.17}$$

En este caso, esta probabilidad (distribución hipergeométrica univaluada) es función únicamente de $q_1^{(1)}$, ya que el resto de parámetros quedan fijados conociendo éste.

Si suponemos que las secuencias son suficientemente grandes, podemos usar la aproximación de Stirling para el factorial: $n! \approx \sqrt{2\pi} n^n e^{-n}$, con lo que los coeficientes binomiales quedan:

$$\begin{aligned}
\binom{n}{q} &= \frac{n!}{q!\,(n-q)!} \approx \frac{1}{\sqrt{2\pi}} \frac{n^n}{q^q\,(n-q)^{n-q}} = \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{n \ln n - q \ln q - (n-q) \ln (n-q)\right\} = \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{-q \ln\left(\frac{q}{n}\right) - (n-q) \ln\left(\frac{n-q}{n}\right)\right\} = \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{n\left(-\frac{q}{n} \ln\left(\frac{q}{n}\right) - \frac{n-q}{n} \ln\left(\frac{n-q}{n}\right)\right)\right\} = \\
&= \frac{1}{\sqrt{2\pi}} \exp\left\{n\mathcal{H}(q)\right\}
\end{aligned} \tag{4.18}$$

donde $\mathcal{H}(q)$ es la entropía, en nats, de la distribución binaria $\{\frac{q}{n}, \frac{n-q}{n}\}$.

Sustituyendo en (4.17) queda:

$$\text{Prob}\left\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)} \;/\; \mathcal{Q}\right\} \approx$$

$$\approx \frac{1}{\sqrt{2\pi}} \exp\left\{-N\mathcal{H}(Q) + n^{(1)}\mathcal{H}\left(q^{(1)}\right) + n^{(2)}\mathcal{H}\left(q^{(2)}\right)\right\} = \tag{4.19}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{-N\mathcal{JS}_2\right\}$$

donde $\mathcal{JS}_2$ indica la divergencia de Jensen-Shannon en nats.

La probabilidad de obtener un valor de divergencia menor o igual que uno dado será:

$$\begin{aligned} \text{Prob}\left\{\mathcal{JS}_2 \leq x\right\} &= \sum_{q_1^{(1)} / \mathcal{JS}_2 \leq x} \text{Prob}\left\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)} \;/\; \mathcal{Q}\right\} \approx \\ &\approx \sum_{\mathcal{JS}_2 \leq x} \frac{1}{\sqrt{2\pi}} \exp\left\{-N\mathcal{JS}_2\right\} \end{aligned} \tag{4.20}$$

Si $N, n^{(1)}$ y $n^{(2)}$ son suficientemente grandes, el número de valores posibles de $\mathcal{JS}_2$ será grande y podremos aproximar la sumatoria por una integral, pero hay que tener en cuenta que estos valores no están equiespaciados, sino que el número de valores en un intervalo es proporcional a $\mathcal{JS}_2^{-\frac{1}{2}}$, por tanto al pasar de la sumatoria a la integral en lugar de $d\mathcal{JS}_2$ tendremos que poner $A\mathcal{JS}_2^{-\frac{1}{2}} d\mathcal{JS}_2$, donde A es una constante que calcularemos normalizando la distribución de probabilidad.

$$\text{Prob}\left\{\mathcal{JS}_2 \leq x\right\} \approx \int_0^x \frac{1}{\sqrt{2\pi}} \exp\left\{-N\mathcal{JS}_2\right\} A\mathcal{JS}_2^{-\frac{1}{2}} d\mathcal{JS}_2 \tag{4.21}$$

Haciendo el cambio de variable $\xi = N\mathcal{JS}_2$ e imponiendo que la probabilidad total debe ser 1:

$$1 = \int_0^\infty \frac{A}{\sqrt{2\pi}} N^{-\frac{1}{2}} \xi^{-\frac{1}{2}} e^{-\xi} d\xi = \frac{A}{\sqrt{2\pi}} \frac{\Gamma\left(\frac{1}{2}\right)}{\sqrt{N}} \Longrightarrow \tag{4.22}$$

$$A = \frac{\sqrt{2\pi N}}{\Gamma\left(\frac{1}{2}\right)} = \frac{\sqrt{2\pi N}}{\sqrt{\pi}} = \sqrt{2N} \Longrightarrow \tag{4.23}$$

$$\text{Prob}\left\{\mathcal{JS}_2 \leq x\right\} \approx \frac{1}{\sqrt{\pi}} \int_0^{Nx} \xi^{-\frac{1}{2}} e^{-\xi} d\xi = \frac{\gamma\left(\frac{1}{2}, Nx\right)}{\Gamma\left(\frac{1}{2}\right)} \tag{4.24}$$

donde γ es la función gamma incompleta.

En vista de este resultado y teniendo en cuenta que el estadístico χ^2 con ν grados de libertad sigue la distribución de probabilidad:

$$\text{Prob}\left\{\chi^2 \leq x\right\} = \frac{\gamma\left(\frac{\nu}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \tag{4.25}$$

tenemos que, dentro de la aproximación con la que estamos tratando, la cantidad

$$d = 2N\mathcal{JS}_2 = 2N \ln 2 \ \text{JS}_2 \tag{4.26}$$

sigue la distribución χ^2 con 1 grado de libertad.

Generalización a otros alfabetos

En el caso de un alfabeto con número arbitrario de elementos k, la demostración anterior es algo más engorrosa, pero esencialmente, la misma. Ahora la probabilidad de una configuración determinada, que dependerá de $(k-1)$ parámetros, vendrá dada por la distribución hipergeométrica multivariada:

$$\text{Prob}\left\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)} \ / \ \mathcal{Q}\right\} = \frac{\binom{q_1}{q_1^{(1)}}\binom{q_2}{q_2^{(1)}}\cdots\binom{q_k}{q_k^{(1)}}}{\binom{N}{n^{(1)}}} \tag{4.27}$$

Si suponemos válida la aproximación de Stirling, y haciendo un cálculo similar al de (4.18) se tiene que:

$$\binom{q_i}{q_i^{(1)}} \approx \frac{1}{\sqrt{2\pi}} \exp\left\{q_i \ln q_i - q_i^{(1)} \ln q_i^{(1)} - q_i^{(2)} \ln q_i^{(2)}\right\} \tag{4.28}$$

$$\begin{aligned}\binom{N}{n^{(1)}} &\approx \frac{1}{\sqrt{2\pi}} \exp\left\{N \ln N - n^{(1)} \ln n^{(1)} - n^{(2)} \ln n^{(2)}\right\} = \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{\sum_{i=1}^{k} q_i \ln N - \sum_{i=1}^{k} q_i^{(1)} \ln n^{(1)} - \sum_{i=1}^{k} q_i^{(2)} \ln n^{(2)}\right\} \end{aligned} \tag{4.29}$$

y sustituyendo en (4.27) se tiene:

$$\text{Prob}\left\{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)} \;/\; \mathcal{Q}\right\} \approx$$

$$\approx \left(\frac{1}{\sqrt{2\pi}}\right)^{k-1} \exp\left\{\sum_{i=1}^{k} q_i \ln \frac{q_i}{N} - \sum_{i=1}^{k} q_i^{(1)} \ln \frac{q_i^{(1)}}{n^{(1)}} - \sum_{i=1}^{k} q_i^{(2)} \ln \frac{q_i^{(2)}}{n^{(2)}}\right\} \tag{4.30}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^{k-1} \exp\left\{-N\mathcal{JS}_2\right\}$$

que, salvo el factor que multiplica, es igual a lo que obteníamos para el caso binario. Numéricamente se ha comprobado que, en estos casos la densidad de valores de divergencia va como $\mathcal{JS}_2^{-\frac{k-1}{2}}$, con lo que haciendo cálculos similares a los que veíamos en la sección anterior tendremos:

$$\text{Prob}\left\{\mathcal{JS}_2 \leq x\right\} \approx \pi^{\frac{1-k}{2}} \int_0^{Nx} \xi^{-\frac{k-1}{2}} e^{-\xi} d\xi = \frac{\gamma\left(\frac{k-1}{2}, Nx\right)}{\Gamma\left(\frac{k-1}{2}\right)} \tag{4.31}$$

con lo que la cantidad

$$d = 2N\mathcal{JS}_2 = 2N \ln 2 \; \text{JS}_2 \tag{4.32}$$

sigue la distribución χ^2 con $k-1$ grados de libertad. Este resultado confirma la conjetura que hicimos en (4.16), ya que los cuantiles de esta distribución sólo dependen de N y k. Además podemos asegurar que, dentro de esta aproximación, la

dependencia con N es de proporcionalidad inversa, esto es, la cantidad $N \cdot D_s(N,k)$ es sólo función de k. Por otra parte esta aproximación nos lleva también al mismo resultado que obteníamos en la sección 4.3.1 para el valor medio de JS_2 (ec. 4.15):

$$\mathrm{E}\left[\mathrm{JS}_2\right] = \frac{1}{2N\ln 2}\mathrm{E}\left[d\right] = \pi^{\frac{1-k}{2}} \int_0^\infty x \cdot x^{-\frac{k-1}{2}} e^{-x} dx = \frac{k-1}{2N\ln 2} \tag{4.33}$$

4.3.3 Aproximación de JS_2 por χ^2

En el apartado anterior hemos visto que la distribución de probabilidad de JS_2 se puede aproximar por la distribución χ^2, veamos ahora que también JS_2 se puede aproximar por el estadístico χ^2, si la composición de las subsecuencias no es muy diferente. Esto es un resultado bastante lógico, ya que tanto JS_2 como χ^2 son medidas de distancia entre distribuciones ambas deben dar valores similares cuando la "distancia" entre las distribuciones es pequeña, de hecho relaciones como la que vamos a deducir aquí son ya conocidas para otras distancias [Bov 84].

En primer lugar tenemos, desarrollando por Taylor que:

$$x\ln x \approx x_0 \ln x_0 + (\ln x_0 + 1)(x - x_0) + \frac{1}{2x}(x - x_0)^2 \Longrightarrow \tag{4.34}$$

$$\mathcal{JS}_2 = -\sum_{i=1}^{k} f_i \ln f_i + \sum_{i=1}^{k} \frac{n^{(1)}}{N} f_i^{(1)} \ln f_i^{(1)} + \sum_{i=1}^{k} \frac{n^{(2)}}{N} f_i^{(2)} \ln f_i^{(2)} \approx$$

$$\approx -\sum_{i=1}^{k} f_i \ln f_i + \sum_{i=1}^{k} \frac{n^{(1)}}{N} \left[f_i \ln f_i + (\ln f_i + 1)\left(f_i^{(1)} - f_i\right) + \frac{1}{2f_i}\left(f_i^{(1)} - f_i\right)^2 \right] =$$

$$+ \sum_{i=1}^{k} \frac{n^{(2)}}{N} \left[f_i \ln f_i + (\ln f_i + 1)\left(f_i^{(2)} - f_i\right) + \frac{1}{2f_i}\left(f_i^{(2)} - f_i\right)^2 \right] \tag{4.35}$$

desarrollando y teniendo en cuenta que $f_i = \frac{n^{(1)}}{N} f_i^{(1)} + \frac{n^{(2)}}{N} f_i^{(2)}$, tenemos que:

$$\mathcal{JS}_2 \approx \sum_{i=1}^{k} \frac{1}{2Nf_i} \left[n^{(1)} \left(f_i^{(1)} - f_i \right)^2 + n^{(2)} \left(f_i^{(2)} - f_i \right)^2 \right] \tag{4.36}$$

y, tras algunos cálculos, sencillos aunque bastante engorrosos, se llega a:

$$\mathcal{JS}_2 \approx \frac{1}{2} \sum_{i=1}^{k} \frac{\left(n^{(1)} f_i^{(1)} - n^{(2)} f_i^{(2)} \right)^2}{Nf_i} + \frac{1}{2N} \sum_{i=1}^{k} \frac{\left(n^{(1)} - n^{(2)} \right) \left(n^{(2)} f_i^{(2)} - n^{(1)} f_i^{(1)} \right)}{Nf_i} \tag{4.37}$$

El primer término, salvo el factor $\frac{1}{2}$, no es otra cosa que test χ^2 entre las muestras $\mathcal{F}^{(1)}$ y $\mathcal{F}^{(2)}$, mientras que el segundo término se hace despreciable para N suficientemente grande, siempre que ni $n^{(1)}$ ni $n^{(2)}$ sean muy pequeños (nótese que, de hecho este término se anula si $n^{(1)} = n^{(2)}$). Esta aproximación tiene, más o menos, las mismas restricciones que la que obtuvimos en (4.18), ya que en aquel caso suponíamos que teníamos valores suficientemente grandes como para poder utilizar Stirling. Por último notar que este resultado está de acuerdo con lo que obteníamos en (4.32).

4.3.4 Validez de las aproximaciones

Las aproximaciones vistas en las secciones anteriores sólo son válidas para subsecuencias suficientemente grandes pero, con el método de segmentación que introduciremos en la sección siguiente, habrá muchos casos en los que, incluso con valores grandes de N, para $n^{(1)}$ ó $n^{(2)}$ serán pequeños. En estas situaciones, la probabilidad habrá que calcularla de forma exacta como:

$$\text{Prob}\left\{\mathcal{JS}_2 \leq x\right\} = \sum_{q_1^{(1)},\ldots,q_{k-1}^{(1)} / \mathcal{JS}_2 \leq x} \text{Prob}\left\{ \mathcal{Q}^{(1)}, \mathcal{Q}^{(2)} \;/\; \mathcal{Q} \right\} \tag{4.38}$$

donde las probabilidades que aparecen en la sumatoria vendrán dadas por (4.27).

Esta sumatoria, como ya se comentó anteriormente, no parece admitir una expresión analítica sencilla, y mucho menos una expresión analítica para los cuantiles. De todas formas, el cálculo numérico es bastante rápido para valores pequeños de N (justo donde no serán válidas la aproximación anteriores) y asequible para valores relativamente grandes, ya que, a pesar de lo que pueda parecer a primera vista, el número de sumandos de (4.38) no va a depender de N. Veamos esto.

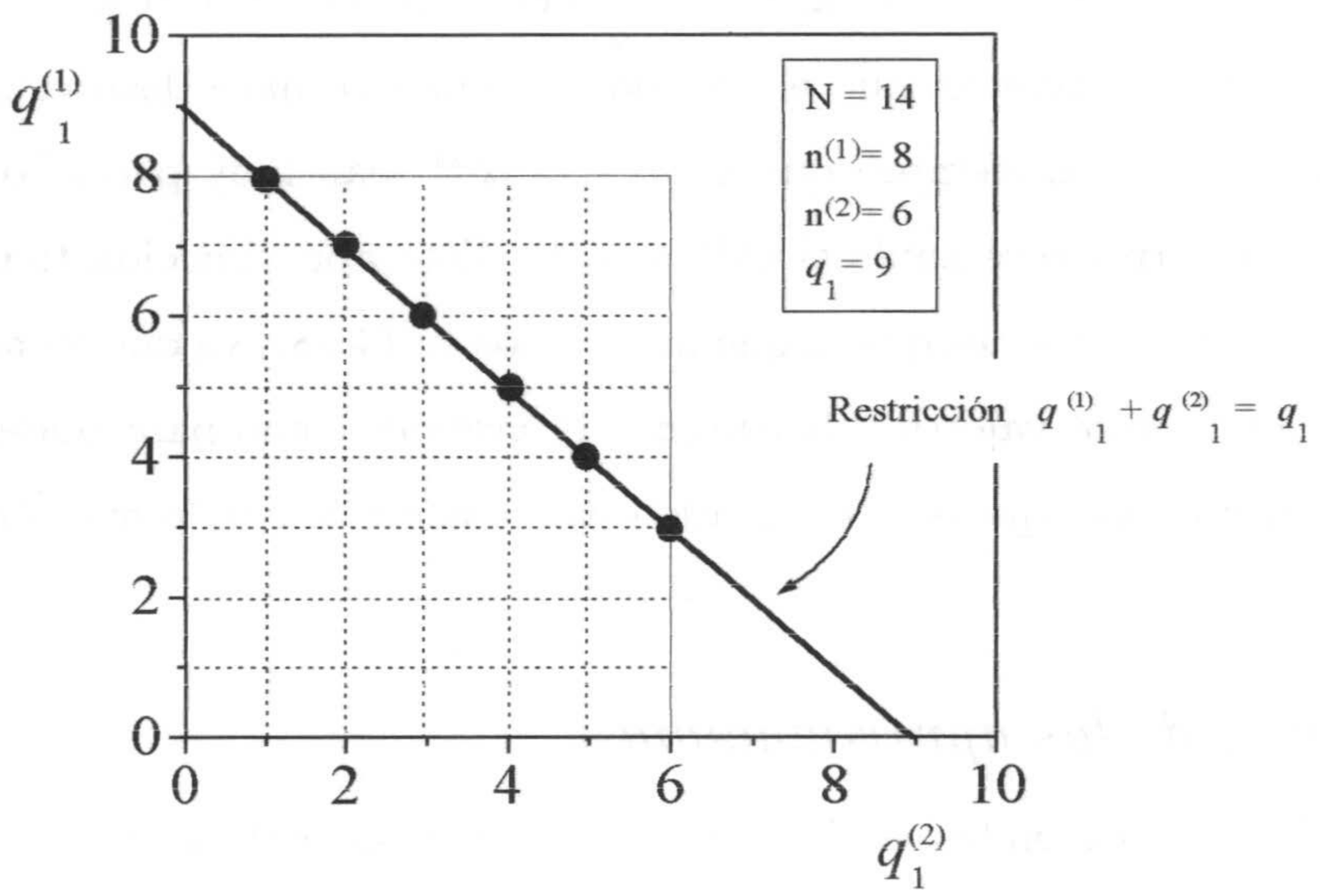

Figura 4.4: Cálculo gráfico del número de configuraciones de los histogramas de las subsecuencias compatibles la restricción de histograma total y tamaños fijos (alfabeto binario).

En la figura 4.4 se muestra una forma gráfica de calcular las configuraciones compatibles con las restricciones ($N, \mathcal{Q}, n^{(1)}$ y $n^{(2)}$ dados) para un caso sencillo con alfabeto binario [†]. Todos los valores permitidos por la restricción $q_1^{(1)} + q_1^{(2)} = q_1$ son

[†]En general habría que hacer una representación en un espacio de k dimensiones.

los que caen sobre la recta, además deben estar dentro del rectángulo ya que necesariamente $q_1^{(1)} \leq n^{(1)}$ y $q_1^{(2)} \leq n^{(2)}$. Según esto, a lo sumo, el número de términos de la sumatoria será $\min\left\{n^{(1)}, n^{(2)}\right\}$, o sea, el tiempo de cálculo depende del tamaño de la menor de las subsecuencias, no del tamaño total. En caso de alfabetos con k símbolos se puede ver que el número de sumandos crece como $\left(\min\left\{n^{(1)}, n^{(2)}\right\}\right)^{k-1}$.

Para verificar la validez de (4.31) se han calculado los cuantiles correspondientes al 99 y al 95% para diversos valores de los parámetros. En la figura 4.5.a) se representan $N \cdot D_{99}(N,2)$, y $N \cdot D_{95}(N,2)$ que según (4.32), deben ser constantes e independientes de todos los parámetros. Calculando numéricamente los percentiles de la distribución χ^2, se obtiene que, para la aproximación: $N \cdot D_{99}(N,2) = 2.39$, y $N \cdot D_{95}(N,2) = 1.38$. Como se ve en la figura, estos valores sólo coinciden con los reales para N suficientemente grande.

En la figura 4.5.b se han tomado 3 valores distintos de N, y se observa que, independientemente del valor de N, la aproximación se hace mejor a medida que crece $n^{(1)}$. Eso permite tomar un valor fijo de $n^{(1)}$, a partir del cual, se considera buena la aproximación (independientemente de la longitud total de la secuencia) de forma que, para valores menores se deba calcular la probabilidad de forma exacta. Por tanto, el tiempo de cálculo no va a crecer de forma ilimitada al aumentar la longitud de la secuencia.

Por último, se han calculado de forma exacta las distribuciones de probabilidad para algunos valores de los parámetros con alfabetos de tres y cuatro símbolos (figuras 4.6 a y b) y se han representado junto con las correspondientes distribuciones χ^2. Como se puede ver, el acuerdo entre los datos exactos y la aproximación es bastante bueno.

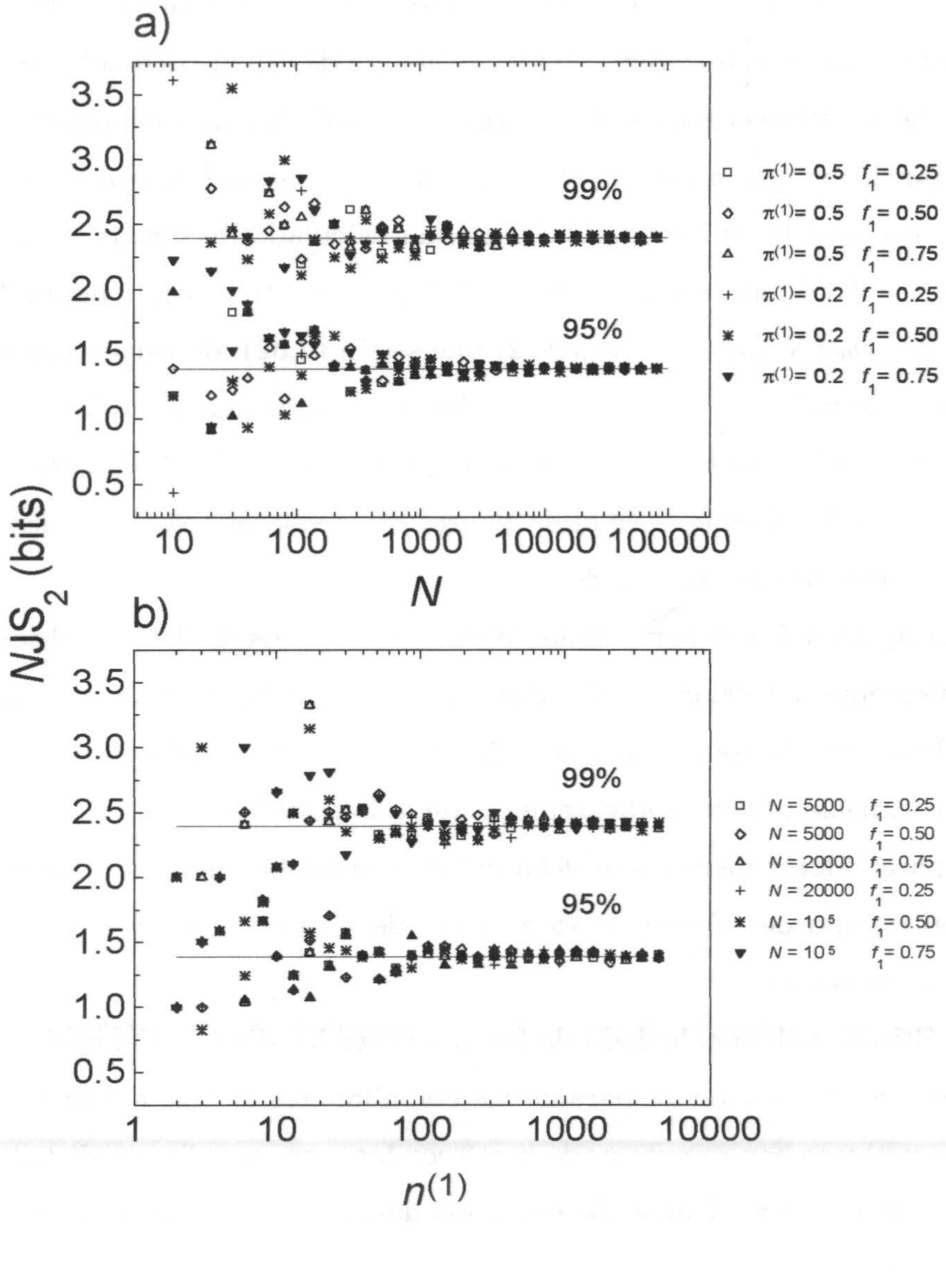

Figura 4.5: Percentiles 95 y 99 para la distribución de probabilidad de JS_2 y distintos valores de los parámetros.

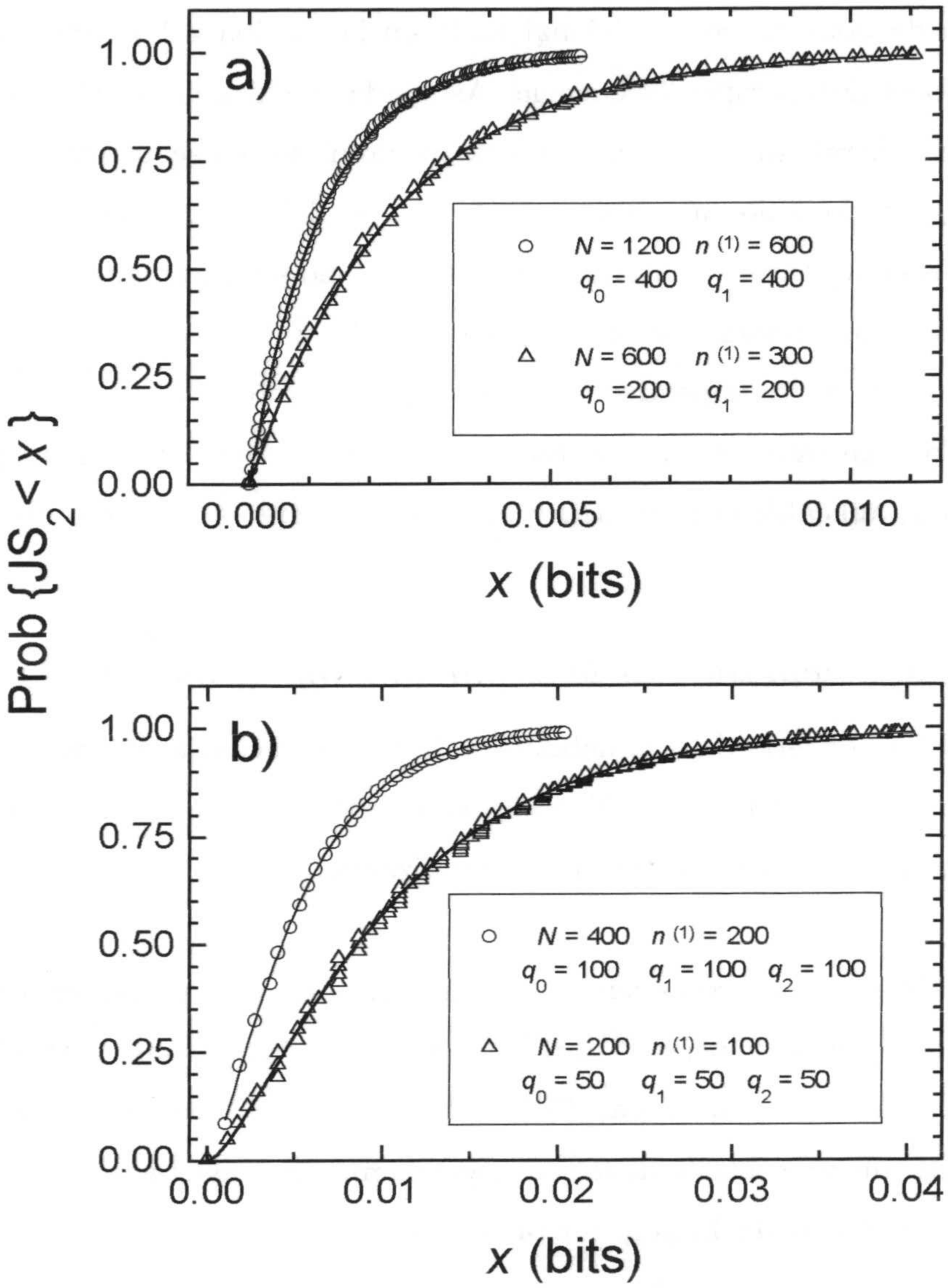

Figura 4.6: Distribuciones de probabilidad de JS_2 para: a) Alfabeto ternario. b) Alfabeto cuaternario. La línea representa en todos los casos la aproximación con χ^2.

4.4 Métodos de segmentación

La definición de dominio composicional dada en la sección 4.1 permite múltiples descomposiciones de una misma secuencia. Así, dada una descomposición cualquiera, es posible, en general, subdividir algunos de los dominios en segmentos más cortos que también son dominios al nivel especificado de fiabilidad estadística. Parece claro el interés en agotar esta subdivisión, hasta que no sea posible proseguir dentro del nivel dado. Sin embargo, la segmentación en dominios composicionales que se alcanza de este modo es dependiente del proceso seguido, por lo que el problema de alcanzar la segmentación óptima requiere un criterio de optimalidad preciso y global, es decir, aplicable al resultado y no en cada etapa del proceso, sea éste del tipo que sea.

4.4.1 Divergencia total de la segmentación. Segmentación óptima

Para establecer el criterio de optimalidad de la segmentación utilizaremos una medida derivada de la misma que utilizamos como distancia probabilística entre segmentos, de hecho, es simplemente una generalización de la divergencia de Jensen-Shannon.

Consideremos una secuencia S, que ha sido dividida, no importa mediante qué proceso, en m subsecuencias $S^{(1)}, S^{(2)}, ..., S^{(m)}$, y utilizando una notación similar a la que hemos usado hasta ahora, $\mathcal{F}^{(1)}$, $\mathcal{F}^{(2)}$,...,$\mathcal{F}^{(m)}$ y $\mathcal{F}$ son los vectores de frecuencias de las subsecuencias y de la secuencia completa respectivamente; definimos la **Divergencia Total de la segmentación** como:

$$\mathrm{JS}_m = \mathrm{H}(\mathcal{F}) - \sum_{i=1}^{m} \frac{n^{(i)}}{N} \mathrm{H}\left(\mathcal{F}^{(i)}\right) = \sum_{i=1}^{m} \frac{n^{(i)}}{N} \left[\mathrm{H}(\mathcal{F}) - \mathrm{H}\left(\mathcal{F}^{(i)}\right)\right] \tag{4.39}$$

donde, obviamente, se ha hecho la elección de pesos propuesta en (4.3).

Esta medida, que es aplicable al resultado de cualquier procedimiento de segmentación de una secuencia, es una medida global de la diversidad composicional entre los segmentos obtenidos. De hecho el último término de (4.39) es una media ponderada de la diferencia de la entropía de cada segmento con la entropía total de la secuencia. Nótese que, a pesar de ser aplicable a cualquier división de la secuencia en dominios, cuando es realmente relevante es cuando estos segmentos son dominios composicionales, esto es, cuando su diferencia de composición es significativa.

A igual que la versión para dos argumentos, JS_m es siempre positiva, salvo igualdad entre todos los argumentos, y simétrica con respecto a éstos. También se puede comprobar, por inducción sobre el número de segmentos, que verifica la propiedad de agrupamiento (ec. 4.10).

Además, es estrictamente creciente con la subdivisión de cualquier segmento en dos de distinto histograma, manteniendo los demás inalterados. En efecto, supongamos que el segmento j se divide en dos subsegmentos $S^{(j_1)}$ y $S^{(j_2)}$, con vectores de frecuencias distintos, ahora tendremos:

$$\mathrm{JS}_{m+1} = \mathrm{H}(\mathcal{F}) - \sum_{i\neq j} \frac{n^{(i)}}{N}\mathrm{H}\left(\mathcal{F}^{(i)}\right) - \frac{n^{(j_1)}}{N}\mathrm{H}\left(\mathcal{F}^{(j_1)}\right) - \frac{n^{(j_2)}}{N}\mathrm{H}\left(\mathcal{F}^{(j_2)}\right) =$$

$$= \underbrace{\mathrm{H}\left(\mathcal{F}\right) - \sum_{i=1}^{m} \frac{n^{(i)}}{N}\mathrm{H}\left(\mathcal{F}^{(i)}\right)}_{\mathrm{JS}_m} + \frac{n^{(j)}}{N}\mathrm{H}\left(\mathcal{F}^{(j)}\right) - \frac{n^{(j_1)}}{N}\mathrm{H}\left(\mathcal{F}^{(j_1)}\right) - \frac{n^{(j_2)}}{N}\mathrm{H}\left(\mathcal{F}^{(j_2)}\right) =$$

$$= \mathrm{JS}_m + \frac{n^{(j)}}{N}\left[\mathrm{H}\left(\mathcal{F}^{(j)}\right) - \frac{n^{(j_1)}}{n^{(j)}}\mathrm{H}\left(\mathcal{F}^{(j_1)}\right) - \frac{n^{(j_2)}}{n^{(j)}}\mathrm{H}\left(\mathcal{F}^{(j_2)}\right)\right] \Longrightarrow$$

$$\mathrm{JS}_{m+1} = \mathrm{JS}_m + \frac{n^{(j)}}{N}\mathrm{JS}_2\left(\mathcal{F}^{(j_1)}, \mathcal{F}^{(j_2)}\right) \tag{4.40}$$

con lo que, en virtud de las propiedades de JS_2, se tiene que $\mathrm{JS}_{m+1} > \mathrm{JS}_m$.

Además la última igualdad de (4.40) nos permite expresar JS_m como una suma ponderada de las contribuciones de cada uno de los cortes producidos en el

proceso:

$$\mathrm{JS}_m(s) = \sum_{j=1}^{m-1} \frac{n^{(j)}}{N} \mathrm{JS}_2\left(\mathcal{F}^{(j_1)}, \mathcal{F}^{(j_2)}\right) \tag{4.41}$$

donde los $n^{(j)}$ son los tamaños de las subsecuencias partidas en cada caso (téngase en cuenta que no coinciden en general con los tamaños de los dominios resultantes). Según esto JS_m se puede expresar en términos cada proceso particular de segmentación, aunque su valor debe ser igual para todos ellos en virtud de (4.39).

Por otra parte, a causa de la propiedad de simetría, la divergencia total no contiene información acerca de la ordenación de las subsecuencias en la secuencia total, ni por tanto, de la divergencia entre todos y cada uno de los pares de subsecuencias adyacentes. Por ello, el sólo uso de esta medida global no puede garantizar el cumplimiento de la condición de que las subsecuencias sean dominios composicionales, condición que deberá ser exigida necesariamente en todo caso a todos los pares de subsecuencias adyacentes.

4.4.2 Segmentación completa

Diremos que una secuencia S está **completamente segmentada** en dominios composicionales, a un nivel de significación estadística dado, si:

1. La divergencia JS_2 de todos los pares de segmentos adyacentes es mayor que la que cabría esperar a causa de las fluctuaciones aleatorias (ver sec. 4.3) y

2. No existe otra descomposición que, cumpliendo 1, tenga una JS_m mayor.

La condición 1 corresponde a la definición de dominio composicional dada en 4.1; mientras que la condición 2 trata de asegurar que no se pueden segregar más dominios, por lo que los que resultan de una segmentación completa pueden calificarse de **dominios composicionales estrictos**, en un sentido global que afecta

al conjunto de todos ellos. Este carácter global de la segmentación completa no garantiza que cada uno de los dominios sea indivisible; esto es, alguno puede ser escindido en dos que cumplen con 1, por lo que serían también dominios estrictos, pero necesariamente dejaría de serlo alguno de los adyacentes, pues el aumento de divergencia de la escisión (4.40) tiene que ser compensado con una disminución por "fusión", ya que JS_m era máxima por la hipótesis de segmentación completa.

Pero veamos que no va a ser posible obtener esta segmentación ideal. Como no existe ninguna forma de conocer a priori la divergencia total que tendría la secuencia completamente segmentada, la única forma de llevar a cabo una segmentación completa es probar con todas las posibles segmentaciones, esto es, tendríamos que chequear todas las posibles divisiones (número de cortes y posiciones), ya que ningún resultado permite predecir a priori otro por tratarse de una secuencia discreta y finita. Por tanto estamos ante un problema del tipo NP-completo, esto es, un problema en el que el número de cálculos crece forma más rápida que un polinomio (exponencial) con el tamaño de la secuencia en cuestión y no parece posible abordar el problema de forma exacta.

4.4.3 Algoritmos Heurísticos de segmentación por exceso y por defecto

Es habitual en problemas de este tipo, dada la alta complejidad, recurrir a métodos heurísticos que reducen fuertemente dicha complejidad a costa de una aproximación que se considera suficiente para la finalidad concreta perseguida [**?**]. El método de segmentación que publicamos en [Be 96] sigue un procedimiento de splitting o escisión rígida: el segmento en ensayo es recorrido, calculando en cada posición la JS_2 entre los dos subsegmentos resultantes, si en la posición de máxima divergencia, ésta es suficientemente grande como para considerar "distintos" los dos subsegmentos con la significación estadística que se considere, se produce el corte, el cual divide el

segmento en dos que son dominios composicionales (en el momento de producirse este corte), y se reitera el procedimiento hasta que no se obtienen nuevos dominios. Una vez producido un corte en una posición dada, ésta no se modifica en adelante, por lo que la condición 2 de la definición de segmentación completa, no queda garantizada, ya que los segmentos resultantes de un corte sí tendrán composiciones diferentes, pero no está asegurada la diferencia con los segmentos que los flanquean; aunque sí se sigue cumpliendo el hecho de que la JS_m es función creciente estricta con el número de cortes.

Segmentación heurística por exceso

Con el método heurístico anterior se obtiene un gran simplicidad ($m-1$ cortes para m segmentos), pero no se garantiza que todos los pares de segmentos adyacentes obtenidos (*adyacencias* en lo que sigue) sean dominios composicionales con la fiabilidad especificada. En efecto, algunos cortes (muchos, en general) se decidieron entre pares de segmentos que después quedaron modificados por cortes posteriores. Por este motivo, la divergencia total puede resultar mayor que la correspondiente a la segmentación completa de la secuencia a la fiabilidad dada. Se trata, pues, de un método de segmentación *heurística por exceso.*

Segmentación heurística por defecto

Para soslayar esta dificultad, en el trabajo presente se utilizará una modificación del método anterior, resultando un *heurístico por defecto.* Consiste en comprobar, antes de decidir cada nuevo corte, que éste y también los dos cortes contiguos una vez modificados por éste, satisfacen la condición de fiabilidad. Como los cortes son rígidos, esta comprobación en cada uno de ellos tiene carácter definitivo, y garantiza que la segmentación final cumple rigurosamente la condición 1 de la segmentación completa, por lo que todos los segmentos son dominios composicionales. La diver-

gencia total de la segmentación heurística por defecto no puede ser mayor que la teórica de la segmentación completa, siendo una cota inferior de ésta, puesto que, por definición es la máxima posible, cumpliendo la condición de que todos los segmentos sean dominios composicionales.

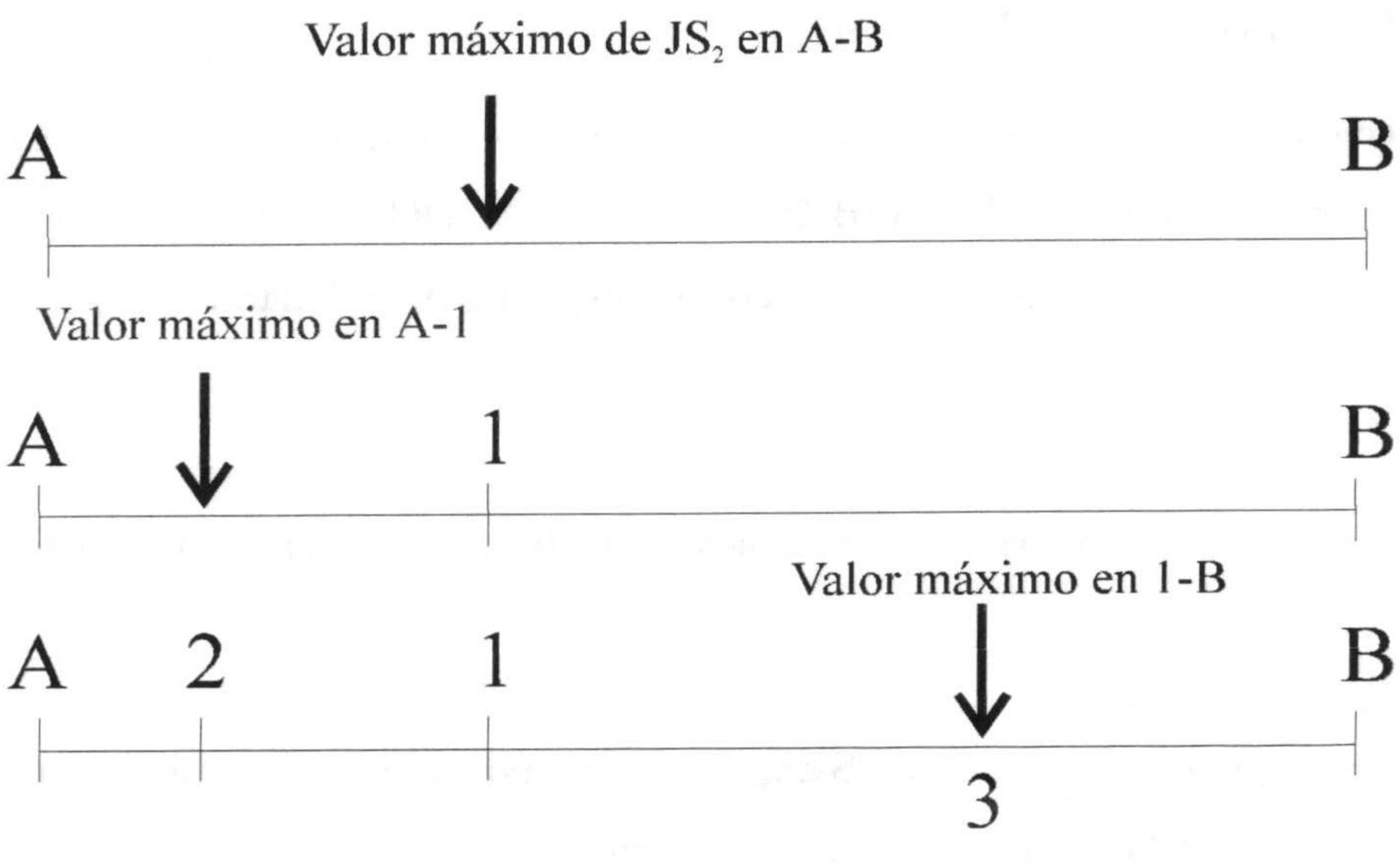

Figura 4.7: Ejemplo de proceso de segmentación heurístico.

Con ayuda del ejemplo de la figura 4.7, veamos cómo influiría el uso de un algoritmo u otro: Supongamos la secuencia, representada como un segmento, A-B y supongamos que se encuentra un punto de división (1) acorde con la fiabilidad estadística especificada. Los dos segmentos resultantes (A-1 y 1-B) son, obviamente, dominios composicionales. Proseguimos la segmentación buscando ahora cortes en A-1 y supongamos que también se encuentra un punto de corte (2). Ahora aparece la primera diferencia entre los dos métodos: el heurístico por exceso establece como nuevos dominios los segmentos A-2 y 2-1, verificando únicamente que la divergencia entre ambos es suficientemente grande, sin embargo, no verifica la divergencia entre

2-1 y 1-B, que también son, por el momento segmentos adyacentes. Por el contrario, el heurístico por defecto sólo admitiría el nuevo corte después de haber verificado la divergencia entre A-2 y 2-1 y entre 2-1 y 1-B. De la misma forma el heurístico por exceso admitiría el corte 3 sin verificar la divergencia entre 2-1 y 1-3.

4.4.4 Descripción del algoritmo

En esta sección describiremos brevemente la implementación del algoritmo. En concreto, nosotros hemos usado **Borland Pascal 7.0**, bajo DOS, aunque la descripción es válida para cualquier lenguaje que pueda manejar listas enlazadas.

Variables

Aparte de los contadores y resto de variables secundarias, las estructuras principales son:

- Un array que denominaremos **SEQ** y que, como indica su nombre, almacena los símbolos de la secuencia en formato entero.

- Una lista enlazada de registros (**CUTS**) que almacena los cortes que se han realizado, además de información útil para los cálculos:

  Estos registros contienen, la posición sobre la secuencia en la que se ha producido el corte, el vector de frecuencias desde el inicio de la secuencia hasta el punto de corte y la divergencia que añade el corte a la divergencia de la segmentación (4.41).

Procedimiento

1. Inicialmente **CUTS** contiene únicamente la posición 0 (que no corresponde a la secuencia) y la posición final de la secuencia.

2. El algoritmo recorre **CUTS** repitiendo el siguiente proceso entre un punto de corte y el siguiente:

   Un contador recorre todas las posiciones de la subsecuencia calculando los vectores de frecuencias de la parte que queda a la derecha y a la izquierda. Para acelerar este proceso no se recalculan los vectores de frecuencias cada vez que se avanza una posición, sino que cada vez que se lee un nuevo elemento de **SEQ** se resta uno al elemento correspondiente del vector de frecuencias de la parte izquierda y se añade al de la derecha.

   Se calcula la divergencia entre estos dos vectores y se almacena el valor máximo encontrado a lo largo de la subsecuencia.

   Una vez recorrida la subsecuencia se comprueba que este valor máximo es significativo, en caso contrario se desecha el corte y se marca la secuencia como indivisible. En caso de ser significativo se verifica que la divergencia entre estos segmentos y los que los flanquean es significativa, si lo es, se inserta este nuevo corte entre los dos anteriores, en caso contrario se desecha el corte (nótese que en este caso no se marca la secuencia como indivisible, ya que en una ronda posterior, al modificarse los segmentos que lo flanquean podría ser aceptado el corte).

3. Se vuelve a recorrer **CUTS**, realizando el mismo proceso hasta que en un recorrido completo no se obtienen nuevos cortes.

4. Se recorre finalmente **CUTS** para volcar al fichero de salida las posiciones de inicio y final de los dominios encontrados. También se aprovecha para obtener la divergencia de la segmentación sumando las contribuciones de cada corte, almacenadas en los registros de **CUTS**.

Cálculo de la significación

Para calcular la significación estadística de los valores de divergencia obtenidos, se ha tomado en todos los programas un valor umbral a partir del cual se consideran buenas las aproximaciones de la sección 4.3. En la elección de este valor se ha buscado un balance entre la precisión y el tiempo de cálculo. En concreto hemos usado el cálculo exacto (4.38) para $\min\left\{n^{(1)}, n^{(2)}\right\} \leq 2000$ y la aproximación a la distribución χ^2 (4.31) en el resto de los casos, usando las rutinas de [Pr 94] para el cálculo de la función gamma incompleta.

Capítulo 5

Segmentación de secuencias de ADN

5.1 Introducción

En este capítulo veremos los resultados de la aplicación del algoritmo de segmentación a secuencias de ADN. En primer lugar justificaremos el uso que se hará en la mayoría de los ejemplos que veremos de aquí en adelante el alfabeto {R,Y}, ya que, al igual que ocurre con las técnicas comentadas en el capítulo 2, con el algoritmo de segmentación también aparece un mayor poder de discriminación al utilizar este alfabeto (sección 5.2). A continuación veremos las distribuciones de tamaños de los dominios obtenidos en las segmentaciones. Prestando especial interés a las diferencias que ponen de manifiesto entre secuencias con y sin correlaciones de largo alcance y codificadoras y no codificadoras. También se dedica un epígrafe a la comparación de las distribuciones de dominios obtenidos con los 16 cromosomas de la levadura de la cerveza (sección 5.3). En la sección 5.4 veremos la relación que hay

entre la organización de los dominios composicionales obtenidos con el método de segmentación y las correlaciones de largo alcance. Finalmente, en las secciones 5.5 y 5.6 estudiaremos la composición interna de los dominios. Veremos lo que ocurre al segmentar con niveles de significación inferiores los dominios obtenidos a un nivel dado (segmentación recursiva) y cómo dependen los resultados de la presencia o no de correlaciones de largo alcance en la secuencia de partida.

5.2 Elección del alfabeto

El algoritmo de segmentación presentado en el capítulo anterior es válido con cualquier alfabeto, por lo que, en principio, sería aplicable al análisis de todo tipo de información almacenada secuencialmente en forma simbólica, aunque por una parte, como se comentó en la sección 4.2.2, el poder de discriminación de la divergencia es mayor para alfabetos pequeños y por otra, el tiempo de cómputo puede crecer considerablemente para alfabetos grandes en aquellas subsecuencias en las que no es válida la aproximación deducida en la sección 4.3. Aquí nos limitaremos a analizar secuencias de ADN, utilizando alfabetos binarios por simplicidad y para tener la posibilidad de comparar con otros métodos de análisis. En concreto usaremos el alfabeto {R,Y}, ya que la estructura composicional más relevante encontrada en estas secuencias parece residir en la alternancia de purinas-pirimidinas [Pe 92], [Pe 94].

Para justificar el uso que haremos, en lo sucesivo, del alfabeto {R,Y} veamos que con esta elección se encuentran las mayores diferencias entre las dos secuencias consideradas en la bibliografía como paradigmas de la presencia de correlación de largo alcance (HUMTCRAD, Humana, Número de acceso en GenBank: M94081, con 97634 bp) y de ausencia de ella (ECO110K, *Escherichia coli,* Número de acceso: D10483, con 111401 bp). La primera presenta un exponente de DFA de 0.61 con un ajuste muy bueno en el rango de 10 a 10^4 bp, mientras que la segunda presenta un

exponente de 0.51 [Pe 94].

Tabla I.

	Nivel de significación	nº de dominios	longitud media	desviación standard
HUMTCRAD {R,Y}	99	125	781	2293
	95	538	181	513
	90	1412	69	194
HUMTCRAD {S,W}	99	95	1027	1167
	95	256	381	752
	90	550	178	405
ECO110K {R,Y}	99	10	11140	25219
	95	91	1224	2591
	90	294	379	1297
ECO110K {S,W}	99	44	2532	3664
	95	167	667	1884
	90	368	303	1112

En la Tabla I aparecen los estadísticos básicos de las distribuciones de dominios que se obtienen al segmentar estas secuencias a distintos niveles de significación con los alfabetos {R,Y} y {S,W}. En términos generales cabe destacar la gran dispersión de las distribuciones (desviación estándard $\gg$ media). En cuanto a la diferencia entre la secuencia humana y la bacteriana, vemos que, a pesar de tener longitudes similares, el número de dominios obtenidos en HUMTCRAD es superior al obtenido en ECO110K para los dos alfabetos. Este mayor número de dominios, parece indicar una mayor heterogeneidad composicional, al menos una mayor variación de la composición a lo largo de la secuencia, y es justamente con el alfabeto {R,Y} con el que aparecen mayores diferencias entre las dos secuencias, lo que estaría de acuerdo con el modelo de Inserción-delección propuesto en [Bu 93a] (ver sec. 2.4.4).

En la figura 5.1 se han representado las distribuciones de dominios de estas segmentaciones. A causa del amplio rango de valores, se representa como variable

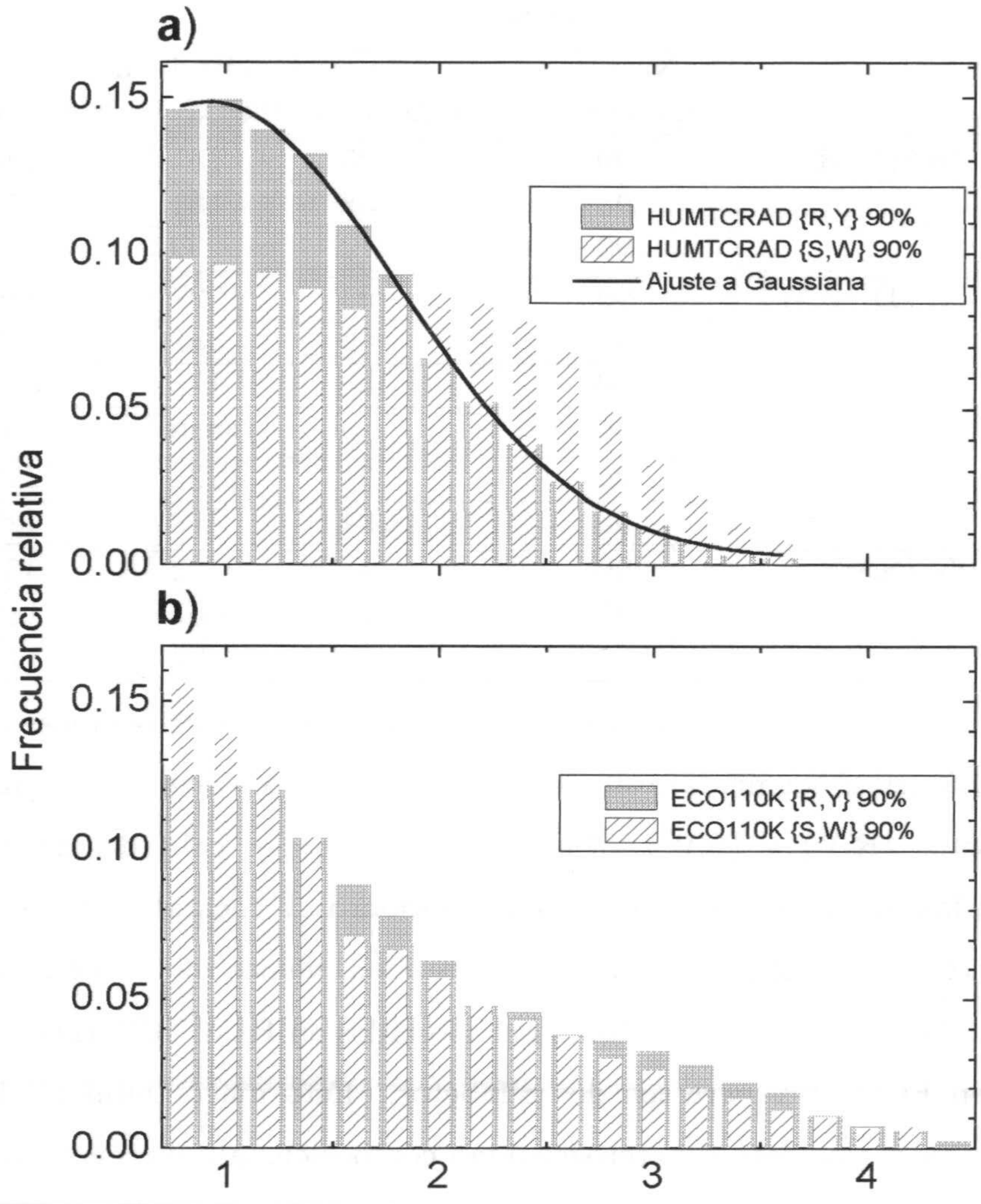

Figura 5.1: Distribución de longitudes de dominios para las segmentaciones con alfabetos {R,Y} y {S,W} al 90%. a) HUMTCRAD. b) ECO110K.

independiente el logaritmo decimal de la longitud, en lugar de la longitud. En la figura 5.1.a) se aprecia una marcada diferencia entre la distribución para el alfabeto {R,Y} y la correspondiente al alfabeto {S,W}. Sin embargo para ECO110K las diferencias entre los dos alfabetos no son tan evidentes, lo que concuerda con los resultados obtenidos a partir del análisis con DFA.

5.3 Distribución de las longitudes de dominios

5.3.1 Secuencias con y sin correlaciones de largo alcance

Centrándonos en el alfabeto {R,Y}, la diferencia más notoria entre la secuencia humana y la bacteriana está en el número de dominios que se obtienen, téngase en cuenta que tienen una longitud similar ($\sim$ 100 kbases) y la humana tiene de 5 a 10 veces más dominios. Esta mayor heterogeneidad composicional de la secuencia humana puede estar ligada a la presencia de correlaciones de largo alcance observadas en ella y no en la bacteriana. De hecho ésta no es una situación excepcional, hemos comprobado con otras secuencias que existe cierta relación entre el valor del exponente del DFA y el número de dominios que se detectan con el algoritmo de segmentación, a todos los niveles de significación. Por ejemplo, en la figura 5.2 se han representado, para secuencias de ADN de diversos organismos y niveles de significación del 99, 95 y 90%, en ordenadas, el número de dominios encontrados por cada 1000 bases y en abcisas, el exponente de escala calculado con el DFA en el intervalo de mejor ajuste. Como se puede ver, aunque no existe una dependencia claramente definida se aprecia una tendencia evidente al aumento del número de dominios con el aumento del exponente del DFA. De hecho el ajuste por mínimos cuadrados a una recta da un coeficiente de correlación del orden de 0.8, más que suficiente para suponer que existe dependencia entre las dos variables.

Centrándonos en el alfabeto {R,Y}, aunque, salvo en el número de dominios,

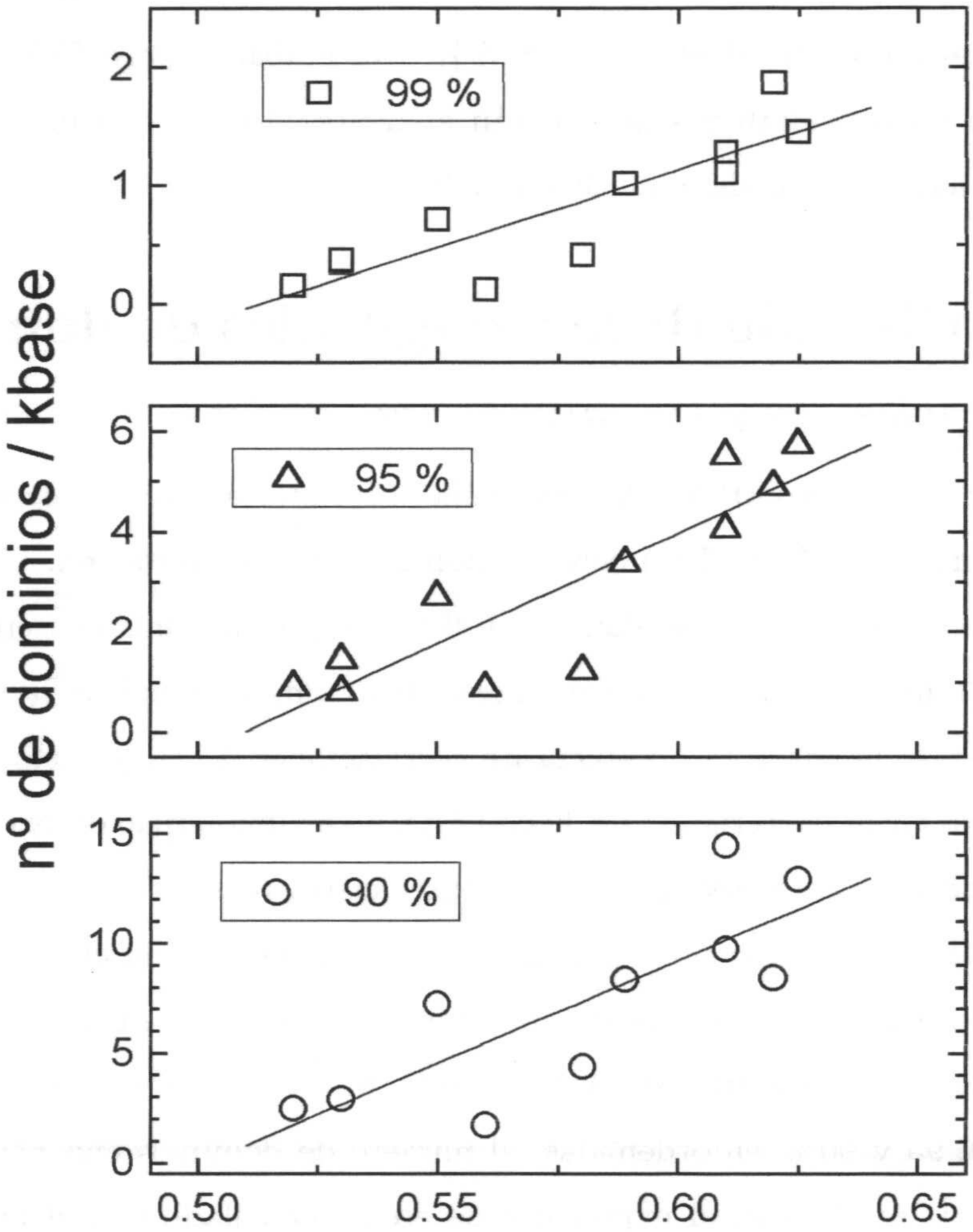

Figura 5.2: Número de dominos por kbase frente al exponente DFA para algunas secuencias de ADN.

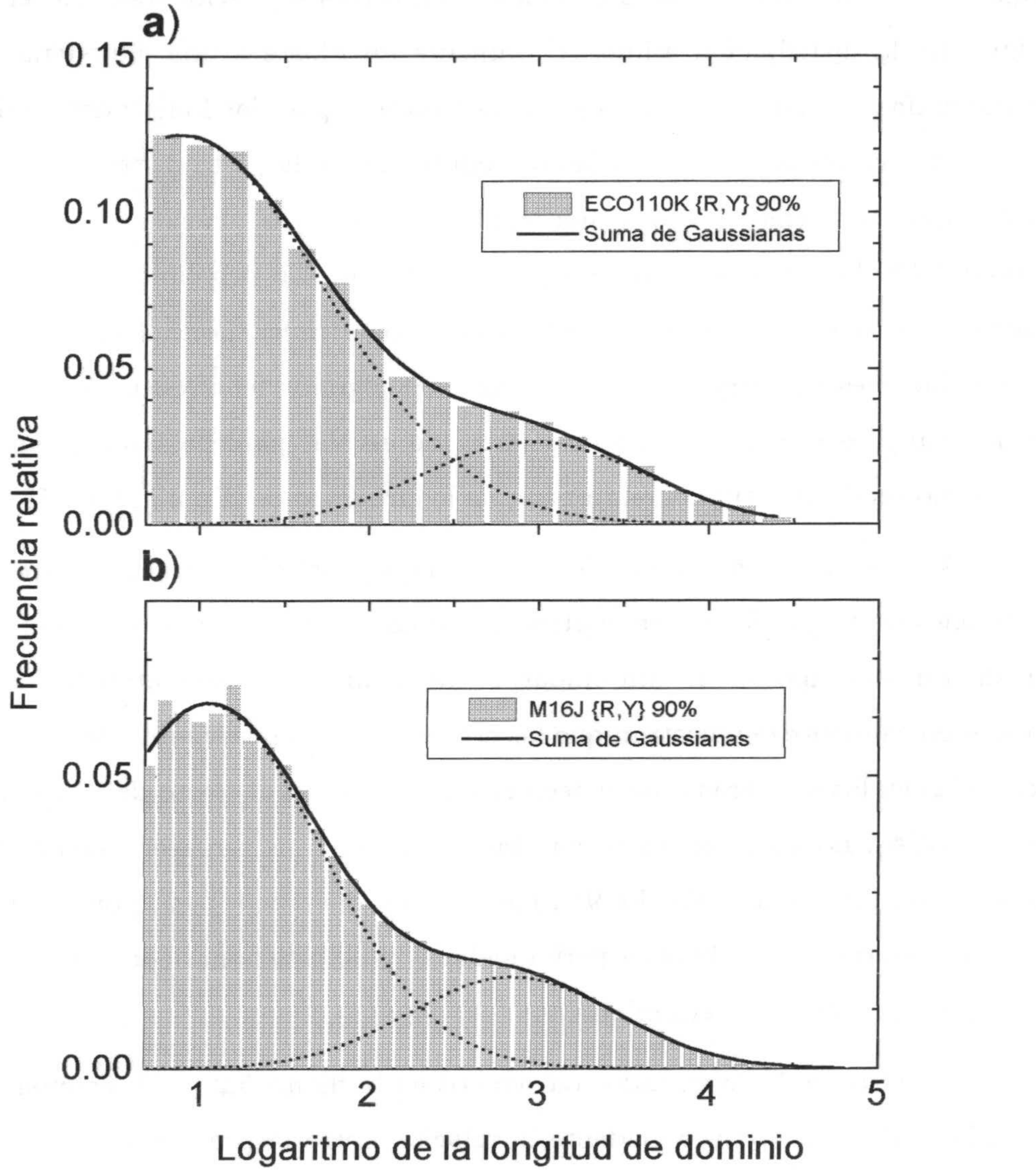

Figura 5.3: Distribución de longitudes de dominios para las segmentaciones con alfabeto {R,Y} al 90%. a) ECO110K. b) M16J.

no aparecen grandes diferencias cuantitativas entre las distribuciones de la secuencia humana y la bacteriana, las diferencias cualitativas son evidentes: En el caso de la humana la distribución admite claramente un ajuste a una gaussiana, que correspondería a una distribución logarítmica normal para las longitudes de dominios (fig. 5.1.a), mientras que para la bacteriana no, de hecho, esta distribución se puede ajustar muy bien (fig. 5.3.a) a una suma de gaussianas, una de ellas centrada en torno a 1 (10 bp) y la otra en torno a 3 (1000 bp). En principio no existe ningún modelo teórico que relacione la distribución logarítmica normal con la presencia de correlaciones de largo alcance, aunque, en algunos fenómenos de fractura, en los que subyace el concepto de invarianza de escala, parecen haberse encontrado distribuciones de este tipo en el tamaño de los fragmentos [Ep 47], [Sch 75], [Fu 86].

Algunos autores [Ka 93], [La 93], [Cha 94] sostienen que la presencia de correlaciones de largo alcance se explica de forma trivial por la alternancia de zonas con diferente composición. Sin embargo, los análisis que presentan son algo ingenuos: esencialmente demuestran que las correlaciones son trivialmente explicadas en secuencias en las que, tras aplicar técnicas que eliminan el efecto de la no estacionariedad (DFA), no aparecen tales correlaciones. Hay que tener en cuenta que estos trabajos son previos al DFA [Pe 94], método que diferencia bien, por una parte, las correlaciones triviales debidas a periodicidades y no estacionariedad de la secuencia y por otra, aquellas más complejas.

En vista de los resultados mostrados en la figura 5.2 y las distribuciones de dominios obtenidas, parece evidente la relación entre la presencia de correlaciones de largo alcance (*long-range)* y la alternancia de zonas con diferente composición. Lo que no está tan claro es que cualquier distribución, como sostienen los autores citados arriba, pueda llevar a correlaciones de largo alcance no triviales.

Otros autores sostienen también que el origen de estas correlaciones está

en la presencia de dominios de diferente composición (*patches)*, pero con tamaños distribuidos según una ley de potencia con un exponente cercano a -2 ([Bu 93a], [Bu 93b]). Esta última afirmación no estaría en completo desacuerdo con los resultados que aquí aparecen, ya que las distribuciones en forma de ley de potencia que proponen estos autores no cubren todo el rango de longitudes, sino que empiezan en una longitud *umbral* en torno a 10 bp, por debajo de la cual, según ellos [Bu 93b] aparecen los efectos de las correlaciones a corto alcance debidas a la longitud 3 de los codones. Esto se traduciría en una caída de la distribución para longitudes menores o del orden de 10 bp (como ocurre en la figura 5.1.a). Además, el comportamiento asintótico de la distribución logarítmica normal puede confundirse con una ley de potencia de exponente próximo a -2, o viceversa.

Lo cierto es que se ha observado que el ajuste a esta distribución es tanto mejor cuanto mejor es el ajuste a una ley de potencia en el análisis DFA (lo que indicaría una mayor consistencia de la hipótesis de la presencia de estructura fractal), además la posición de la "joroba" que presenta la distribución de longitudes para ECO110K (1000 bp) coincide con el cambio de pendiente que se observa en el DFA (fig. 5.4) en torno a $L = 10^3$ bp (que correspondería a las longitudes características de los genes [Pe 94]). Todo lo dicho para esta secuencia, se puede observar también en otra mucho más larga del mismo organismo, M16J con 1.6 Mbases (figs. 5.3.b y 5.4), con lo que la presencia de esta longitud característica no parece depender del tamaño total de la secuencia. Nótese, por último que a pesar de las grandes diferencias de tamaño entre las secuencias, las distribuciones de longitudes son muy similares, lo cual es un dato a favor del método.

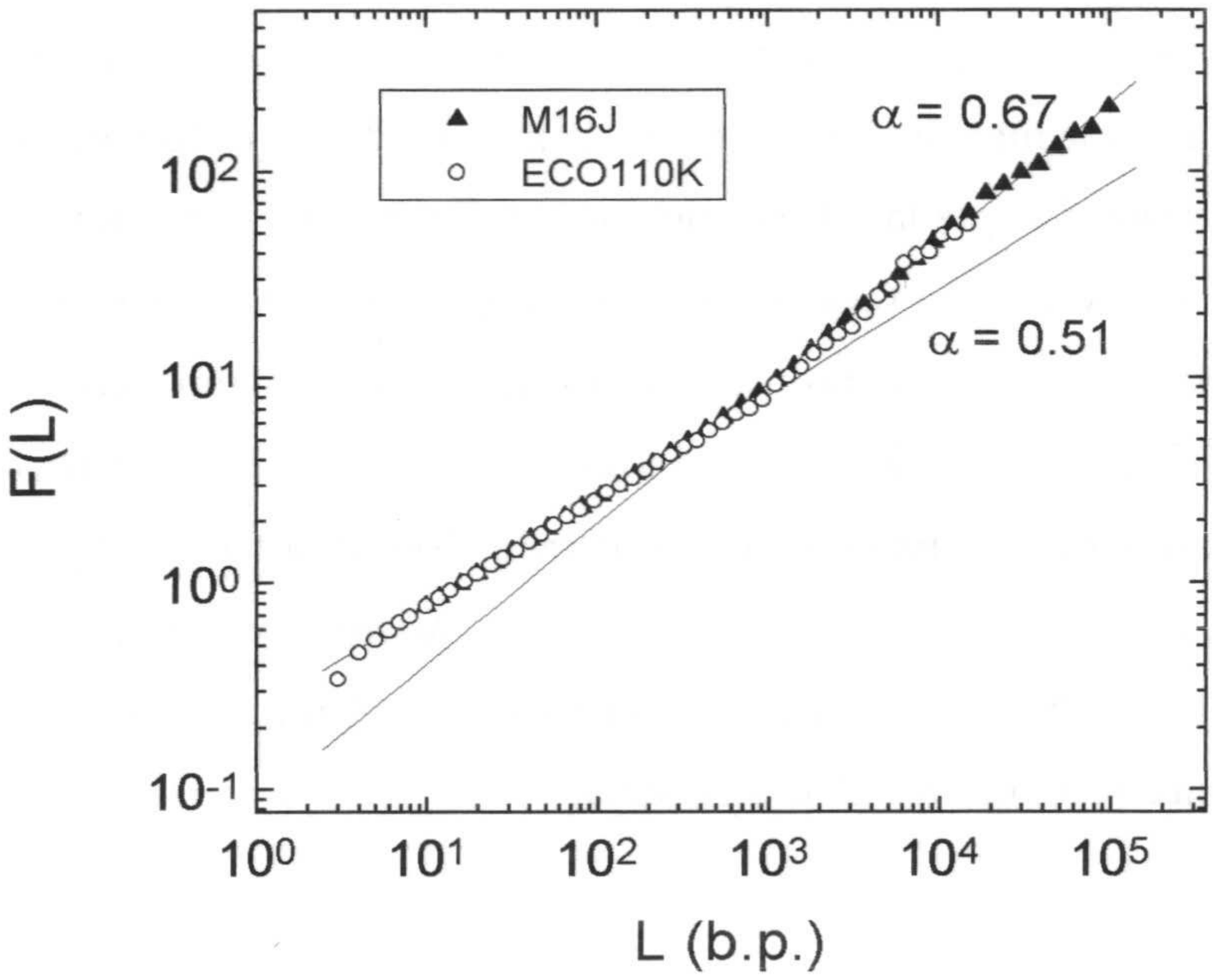

Figura 5.4: F(L) frente a L (DFA) para ECO110K y M16J con ajustes a ley de potencia para $L < 10^3$ b.p. y $L > 10^3$ b.p.

5.3.2 Cromosomas de la levadura

Recientemente se ha secuenciado, por primera vez, el genoma completo de un organismo eucariota, se trata de la levadura de la cerveza (*Saccharomyces cerevisiæ*)*, con un total de 12.136.020 bp distribuidos en 16 cromosomas. Una cuestión que se está planteando en la actualidad es comprobar si todos los cromosomas tienen o no propiedades estadísticas similares. Según parece hay argumentos a priori tanto a favor como en contra de la similitud entre ellos. Los argumentos en contra se basan, por lo general, en estudios locales [Lio 96], pero todos los estudios globales (en sen-

*Las secuencias se pueden obtener en formato ASCII, a través de FTP anónimo en: ftp.ebi.ac.uk.

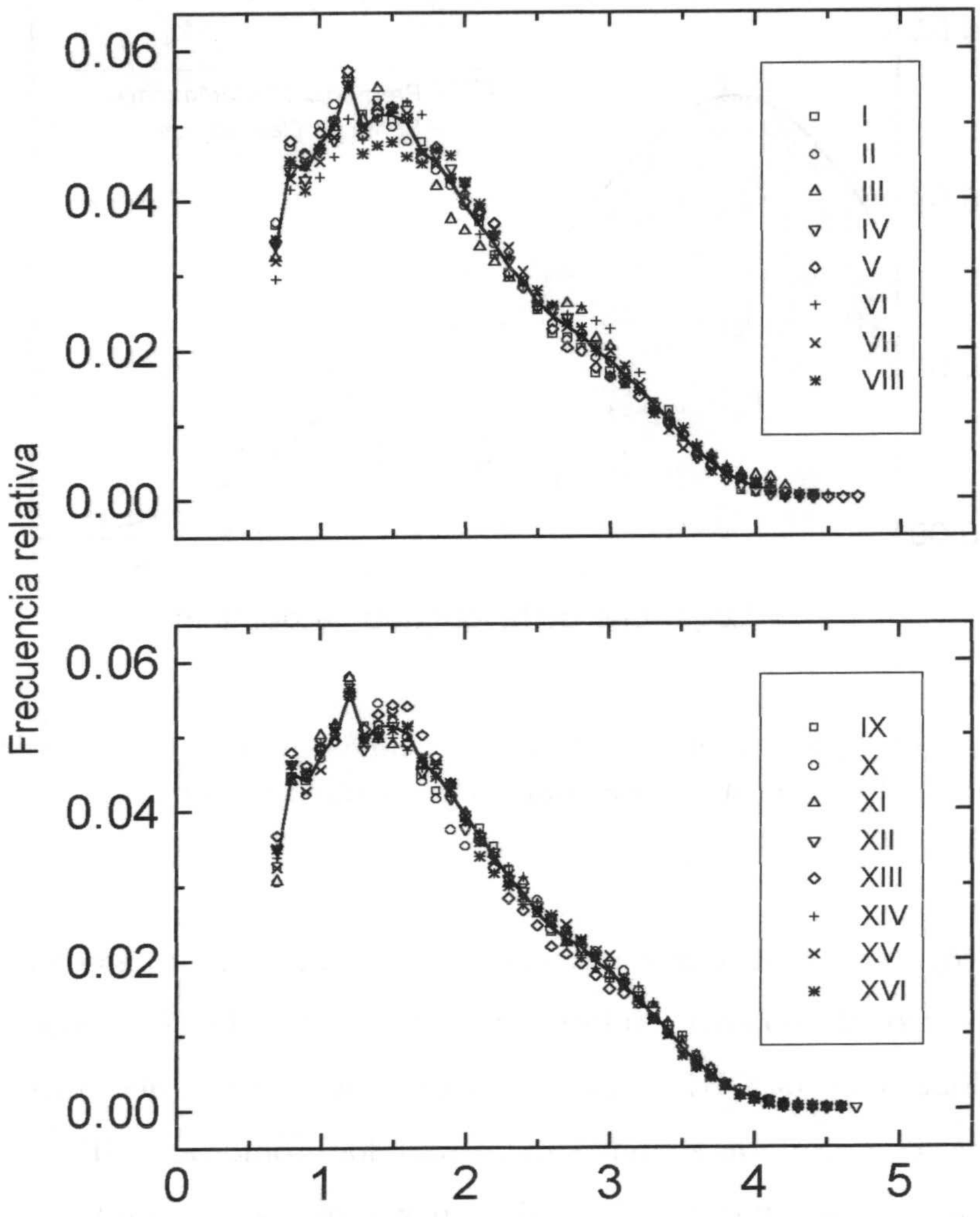

Figura 5.5: Distribuciones de dominios para los 16 cromosomas de la levadura para una significación del 95%. La línea continua corresponde a la distribución promedio de los 16.

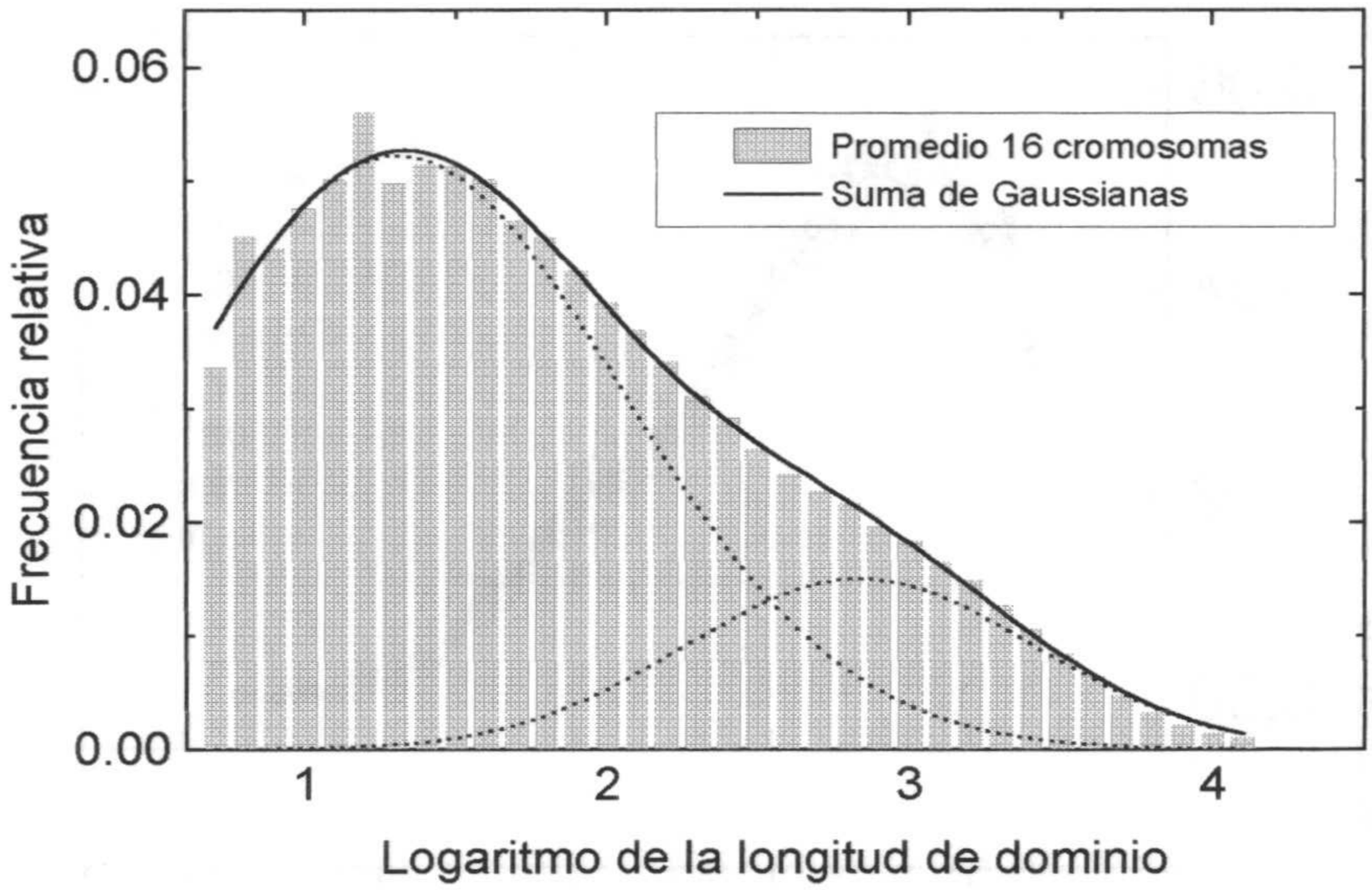

Figura 5.6: Ajuste a una suma de gaussianas para la distribución de dominios de los 16 cromosomas para una significación del 95%.

tido estadístico) dan propiedades muy similares para los 16 cromosomas: Análisis con DFA, espectros de potencia, información muta, etc. [Li 97c]. Además, la distribución de longitudes de ORFs (marcos abiertos de lectura), que corresponderían a posibles zonas codificadoras es muy similar para los cromosomas III, VI, IX y XII, los únicos para los que hay datos disponibles por el momento [Duj 96].

Por otra parte parece haber acuerdo en cuanto a la existencia y relevancia de la heterogeneidad composicional dentro de cada cromosoma para los organismos eucariotas en general [Ber 89], [Ber 95], y en particular para la levadura [Li 97a], [Li 97c], por lo que la comparación de las distribuciones de dominios de los cromosomas puede ser interesante como otro método para caracterizar la similitud entre ellos.

Tabla II.

	Longitud (bp)	χ^2	ν	**Probabilidad (%)**
I	230.195	7.9	26	99.9
II	813.137	10.4	29	99.8
III	315.344	11.6	26	98.9
IV	1.522.191	19.3	30	91.2
V	574.860	11.2	27	99.5
VI	270.148	12.2	26	98.5
VII	1.090.936	13.8	29	98.9
VIII	562.638	15.5	28	96.2
IX	439.885	4.9	27	99.9
X	745.443	12.3	28	99.3
XI	666.448	8.1	27	99.9
XII	1.078.171	13.7	29	98.9
XIII	924.430	15.7	29	96.9
XIV	784.328	14.4	29	98.4
XV	1.091.282	13.8	29	98.9
XVI	948.061	12.8	29	99.4

Veamos qué ocurre al aplicar a estas secuencias nuestro algoritmo de segmentación. En la figura 5.5 se representan las distribuciones de longitudes de dominios para los 16 cromosomas al segmentarlos para una significación del 95 %, y como comparación se ha incluido en las gráficas la distribución media de longitudes. Como puede verse, los resultados de la segmentación están de acuerdo con los resultados obtenidos por los métodos estadísticos que antes hemos comentado: todas las distribuciones son bastante parecidas a simple vista, además el test χ^2, da probabilidades bastante altas de que las distribuciones de cada cromosoma sean las mismas que la total (tabla II).

De nuevo nótese que la considerable diferencia de tamaños entre los cromosomas no afecta a la distribución de longitudes. Por tanto parece razonable hablar de la distribución de dominios en el ADN de la levadura, sin hacer referencia a un cromosoma en concreto, y esto puede ser interesante para estudios posteriores, ya que

abre la posibilidad de un análisis estadístico bastante preciso dado el gran número de datos disponibles (por ejemplo, en la segmentación al 95% tenemos un total de 36.539 dominios).

Como se ha dicho arriba, el análisis con DFA y espectros de potencia también da resultados muy similares para todos los cromosomas, y al igual que ocurre con las secuencias de *E. coli* (M16J y ECO110K), también se observa un cambio de tendencia muy acusado. Pero en este caso, el cambio se produce en torno a una longitud de 600 a 800 bp, que no correspondería al tamaño característico de las zonas codificadoras (en torno a 1400 bp [Duj 96]). Hay que tener en cuenta que, en el genoma de la levadura, la proporción de zona no codificadora es considerable, por lo que la longitud característica de los cambios de composición estará relacionada con el tamaño conjunto de los genes y zonas intergénicas, que son de tamaño comparable. De hecho el valor medio de las longitudes de zonas codificadoras y no codificadoras es algo menor, en torno a 1000 bp y además, teniendo en cuenta la semejanza de las distribuciones con la logarítmica normal, quizá sea más significativa como longitud característica la media geométrica (algo menor) que la aritmética.

De nuevo podemos ver que este cambio de tendencia observado con otros métodos se refleja en la distribución de dominios: en la figura 5.6 se ha representado la distribución total de segmentos para los 16 cromosomas y un ajuste a una suma de gaussianas, una de las cuales está centrada en torno 2.8 (lo que corresponde a una longitud de unas 630 bp). Nótese que esta posición coincide con el centro de la distribución de logaritmos de longitudes de zona codificadora y no codificadora (figura 5.7).

5.3.3 Diferencias entre zonas codificadoras y no codificadoras

Prácticamente desde el descubrimiento de la presencia de correlaciones de largo alcance en secuencias de ADN, apareció la controversia entre los que afirmaban que

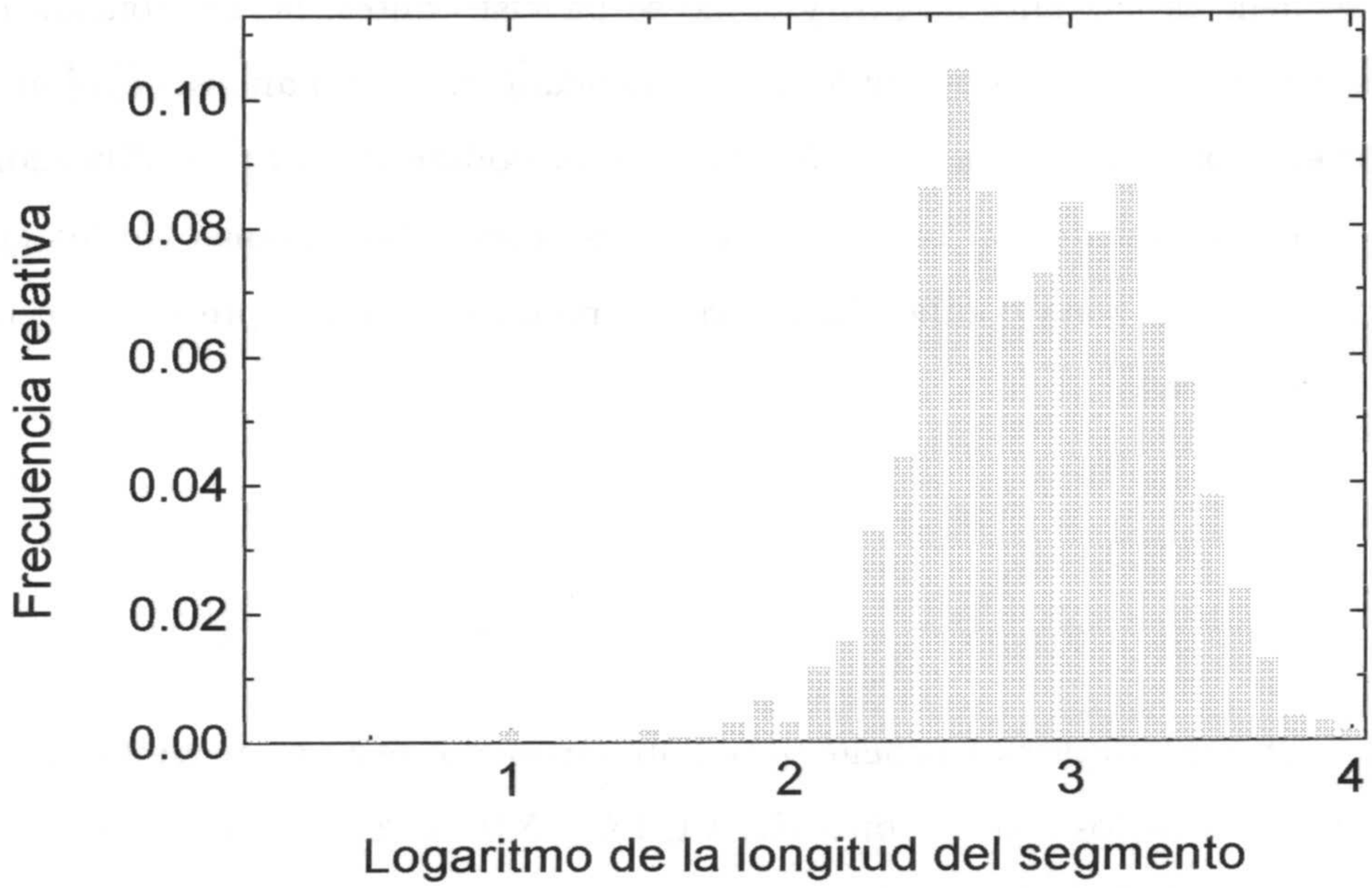

Figura 5.7: Distribución conjunta de los logaritmos de las longitudes de zonas codificadoras y nocodificadoras para los cromosomas III, VI, IX y XII de la levadura.

éstas se encontraban tanto en las zonas codificadoras como en las no codificadoras [Vo 92] y los que sostenían que aparecían sólo en las no codificadoras [Bu 93c]. Esta controversia aún no está resuelta de forma definitiva, aunque tras los trabajos de Buldyrev et al. ([Bu 93b], [Bu 95]) son muchos los que sostienen que el origen de las correlaciones de largo alcance está en las zonas no codificadoras.

Veamos qué aspecto tienen las distribuciones de dominios obtenidos al segmentar zonas codificadoras y no codificadoras. Para ello utilizaremos los cromosomas de la levadura que, aunque no son las secuencias en las que se han encontrado mejores

ajustes a ley de potencia en el análisis con DFA y otros métodos, tienen varias ventajas: son secuencias bastante largas, y como se ha visto antes, la distribución de dominios es similar en todos ellos con lo que es razonable hacer un análisis conjunto. Además, tienen proporciones más equilibradas de zona codificadora y no codificadora (70 y 30% respectivamente), al contrario de las secuencias bacterianas, en las que prácticamente todo es zona codificadora y las humanas en las que predominan las zonas no codificadoras.

Para obtener estadística suficiente se han agrupado, por una parte todas las zonas codificadoras de los cromosomas III, VI, IX y XII (1.295.100 bp) y por otra las no codificadoras (590.844) y se han segmentado. Los resultados aparecen en la figura 5.8. En el primer caso puede verse como el escalón que se observaba ya en la distribución de dominios de los cromosomas completos (en torno a unas 600-800 bp) aparece de forma mucho más notoria, y algo más desplazado a la derecha. Nótese que esta distribución es muy similar a la que veíamos en la figura 5.3 para M16J (*Escherichia coli)*, mientras que en el caso de la zona no codificadora, el escalón es mucho menos notorio y la distribución se asemeja a la que obteníamos para la secuencia humana (figura 5.1.a). Este resultado, tiene una explicación inmediata, y a primera vista, trivial: la secuencia bacteriana tiene un gran porcentaje de zona codificadora y la humana, por el contrario, es casi todo zona no codificadora. Pero en cualquier caso confirma la diferente organización de dominios composicionales en cada zona y parece un argumento bastante razonable a favor de la tesis de que es, justamente, esta distinta organización de dominios composicionales la responsable de la aparición de correlaciones de largo alcance en unas sí y en otras no.

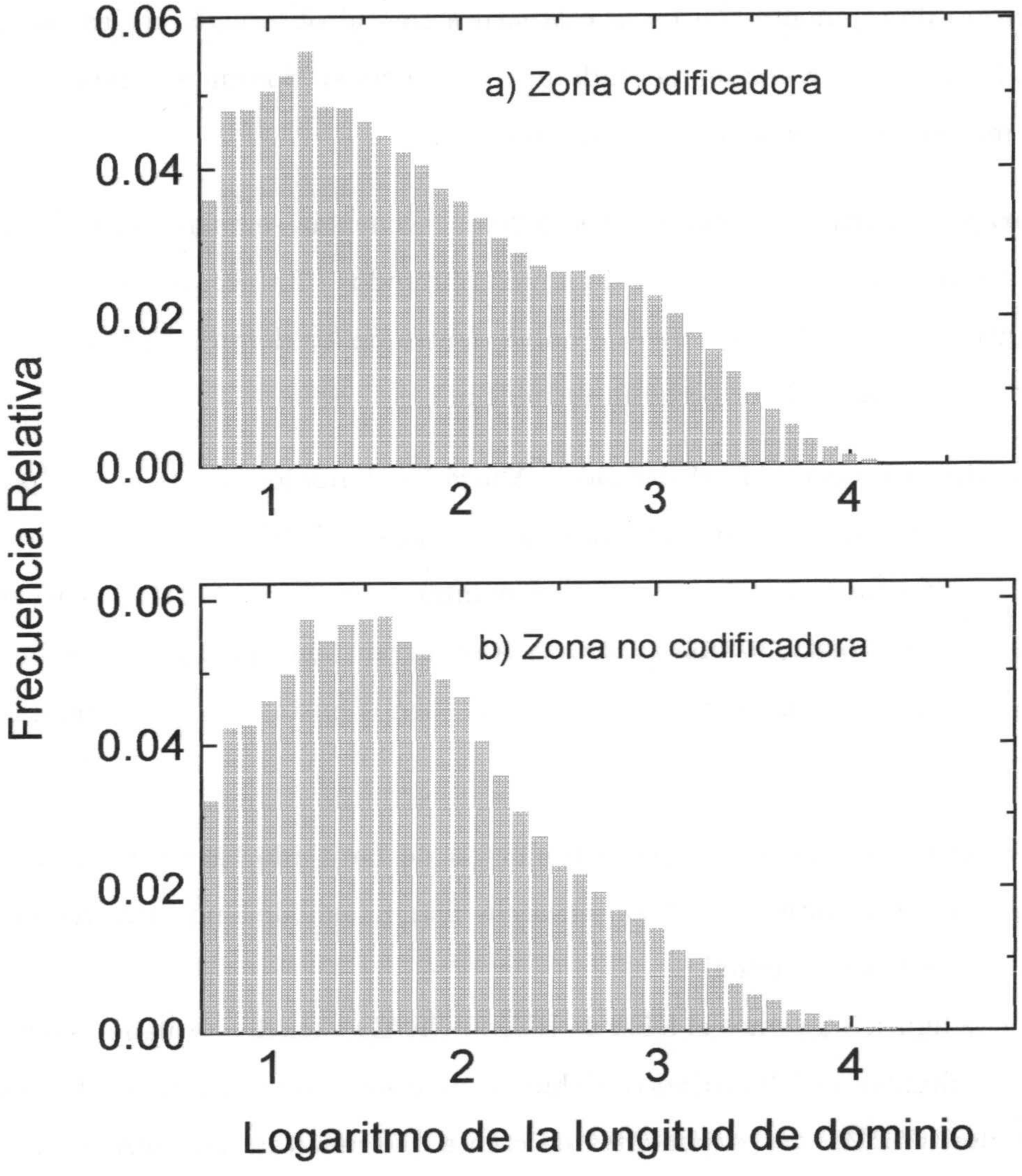

Figura 5.8: Segmentación para un 95% de significación de la zona codificadora y no codificadora de los cromosomas III,VI,IX y XII de la levadura.

5.4 Organización de los dominios

Veamos ahora hasta qué punto pueden dar cuenta los dominios encontrados con nuestro algoritmo de segmentación de la estructura fractal observada en las secuencias de ADN. Para ello, una vez segmentada una secuencia en dominios composicionales, vamos a manipularlos de dos formas diferentes:

a) **Barajado entre dominios**. Tomamos la secuencia segmentada y obtenemos otra reordenando al azar los dominios obtenidos. La secuencia así generada mantiene toda la estructura que había dentro de los dominios pero no la ordenación de éstos a lo largo de la secuencia.

b) **Barajado interno de dominios.** Ahora los dominios mantienen su posición en la secuencia pero se reordenan al azar los nucleótidos que los componen. De esta forma la ordenación de los dominios queda inalterada, pero se destruye toda la organización interna que pudieran tener (téngase en cuenta que esta organización debe ser irrelevante, al menos hasta el nivel de significación considerado).

Sin embargo, estos dos procedimientos de barajado tienen en común una característica importante: ambos conservan la distribución de longitudes de dominios que tenía la secuencia original.

En la figura 5.9 vemos el DFA de HUMTCRAD y las dos secuencias obtenidas por los procedimientos de barajado que hemos descrito arriba, utilizando los segmentos obtenidos al 95%. Los exponentes de escala que resultan del DFA para las dos secuencias artificiales ($\alpha = 0.62$ en ambos casos) son muy similares al observado para la secuencia natural ($\alpha = 0.61$). Esto indica que nuestro algoritmo de segmentación permite descomponer secuencias complejas en dominios, composicionalmente homogéneos hasta cierto nivel de significación, de forma que las correlaciones

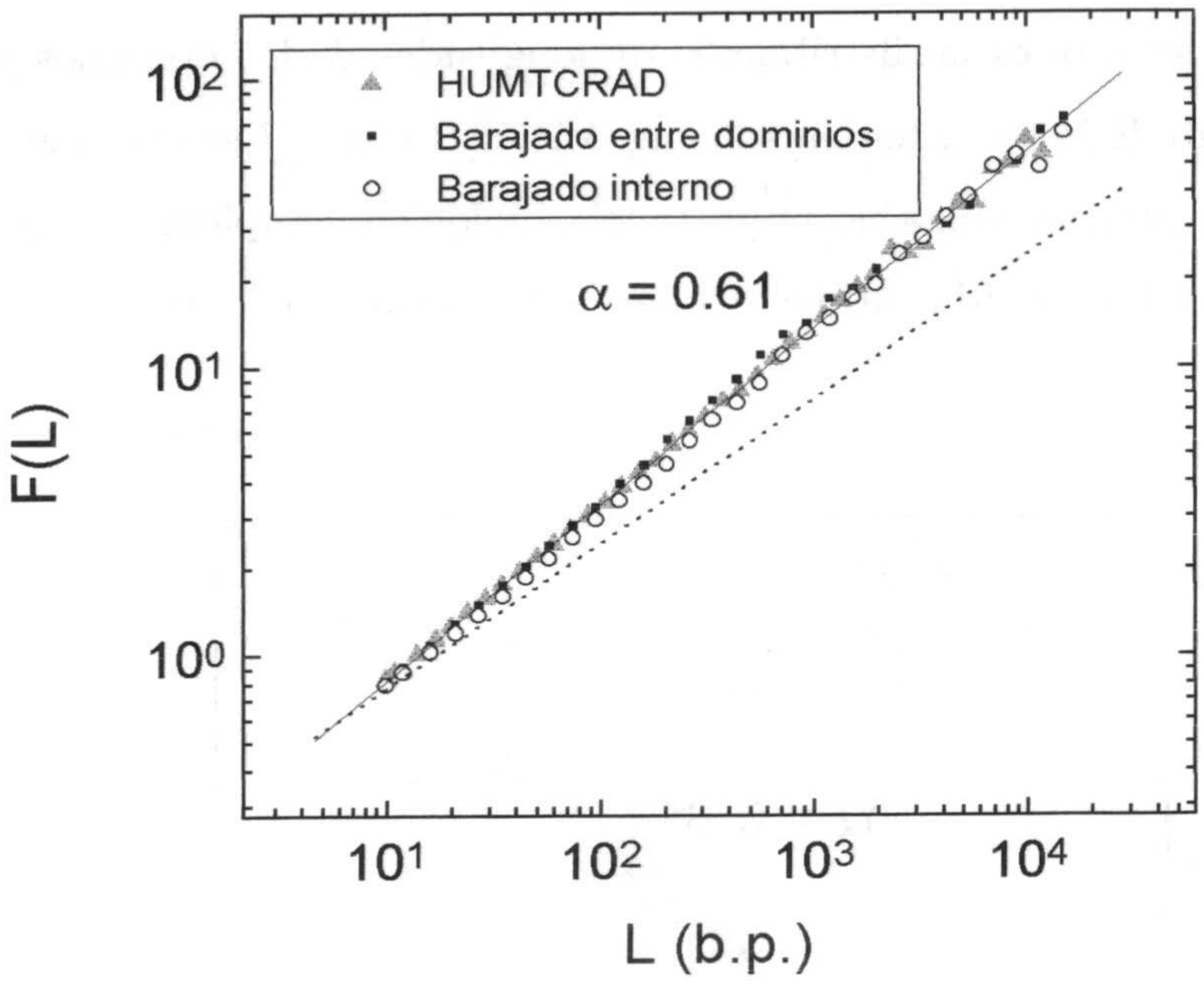

Figura 5.9: DFA para HUMTCRAD y las secuencias artificiales obtenidas barajando los dominios obtenidos para una significación del 95%. Como comparación se incluye (línea punteada) la recta con pendiente 0.5, que corresponde a la ausencia de correlaciones de largo alcance.

originales presentes en la secuencia se conservan.

Estos resultados muestran, en primer lugar, que nuestro algoritmo estima razonablemente bien la distribución de longitudes presente en las secuencias con correlaciones de largo alcance; una vez identificados, los dominios se pueden reorganizar al azar sin alterar el exponente de escala de la secuencia completa, al menos tal y como lo mide el DFA, y esto indica que el proceso de reordenación no altera la distribución de longitudes. En segundo lugar, estos resultados también demuestran que la estructura interna de los dominios no parece ser muy relevante a la hora de

estimar el exponente de escala, al menos calculado con el DFA. Por tanto parece que lo realmente relevante es la distribución de longitudes de los dominios y no su ordenación a lo largo de la secuencia ni su composición interna. Esto no quiere decir que la organización interna tenga no importancia biológica, simplemente que no es tenida en cuenta con las medidas utilizadas para estimar la estructura fractal de las secuencias.

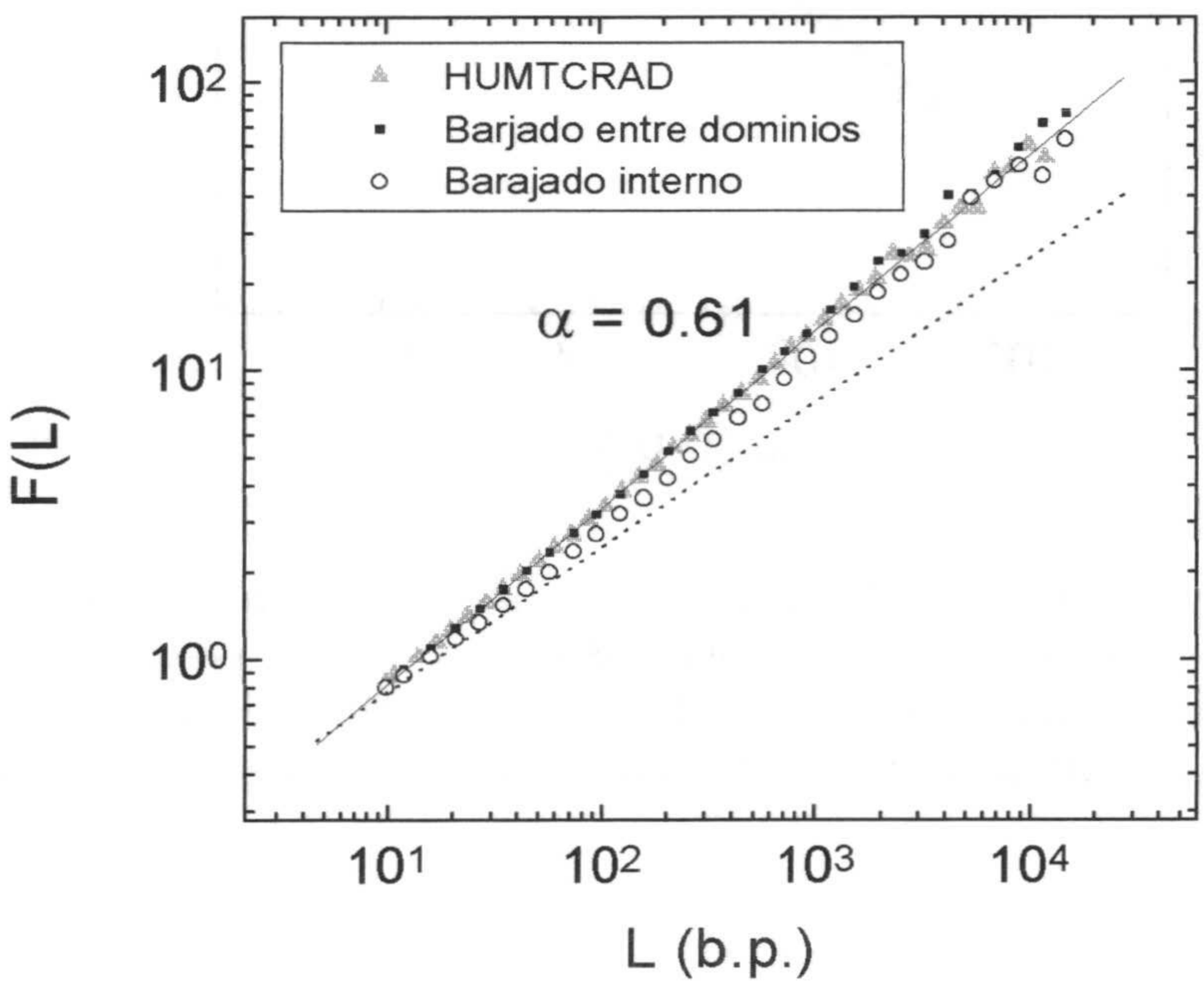

Figura 5.10: DFA para HUMTCRAD y las secuencias artificiales obtenidas barajando los dominios obtenidos para una significación del 99%. Como comparación se incluye (línea punteada) la recta con pendiente 0.5, que corresponde a la ausencia de correlaciones de largo alcance.

En este punto, el nivel de significación utilizado para la segmentación no parece ser un factor determinante, ya que se obtienen resultados similares repitiendo

el experimento con distintos niveles de significación desde 80 al 99% (por ejemplo, en la figura 5.10 vemos los resultados obtenidos con la segmentación al 99%). Sin embargo, hay que notar que los resultados obtenidos barajando internamente, aunque no de forma muy significativa, se apartan más del comportamiento de la secuencia original que los obtenidos con el barajado entre dominios. Esta desviación, como puede verse comparando las figuras 5.9 y 5.10, es más notoria para la segmentación al 99%. Teniendo en cuenta que la segmentación al 99% es más estricta que al 95%, en el sentido de que requiere una mayor divergencia entre los segmentos, los dominios obtenidos al 99% pueden contener estructura interna que no es aceptada como significativa, pero que sí lo es al disminuir el nivel de significación. De esto se deduce que habría que puntualizar la afirmación anterior: la organización de la secuencia en dominios composicionales justifica, *a grandes rasgos*, las correlaciones de largo alcance observadas, pero que no es suficiente para justificarla por completo, ya que la estructura interna de los dominios también parece influir en los resultados (aunque no de forma decisiva).

5.5 Segmentación recursiva.

Otra diferencia importante que se observa en la segmentación de secuencias con y sin long-range, es la forma en que se subdividen los dominios que se obtuvieron a un nivel de significación cuando se vuelven a segmentar a un nivel inferior. De nuevo recurrimos a HUMTCRAD y ECO110K para ilustrar este comportamiento. Para ello, el mayor de los segmentos obtenidos a un nivel de significación, es segmentado de nuevo a un nivel más bajo, repitiendo este proceso varias veces. En la figura 5.11 se presenta el resultado de este proceso aplicado a HUMTCRAD y en la figura 5.12 a ECO110K. Como puede verse, aparecen claras diferencias entre ambas: mientras que en la secuencia bacteriana pronto aparecen dominios homogéneos y relativa-

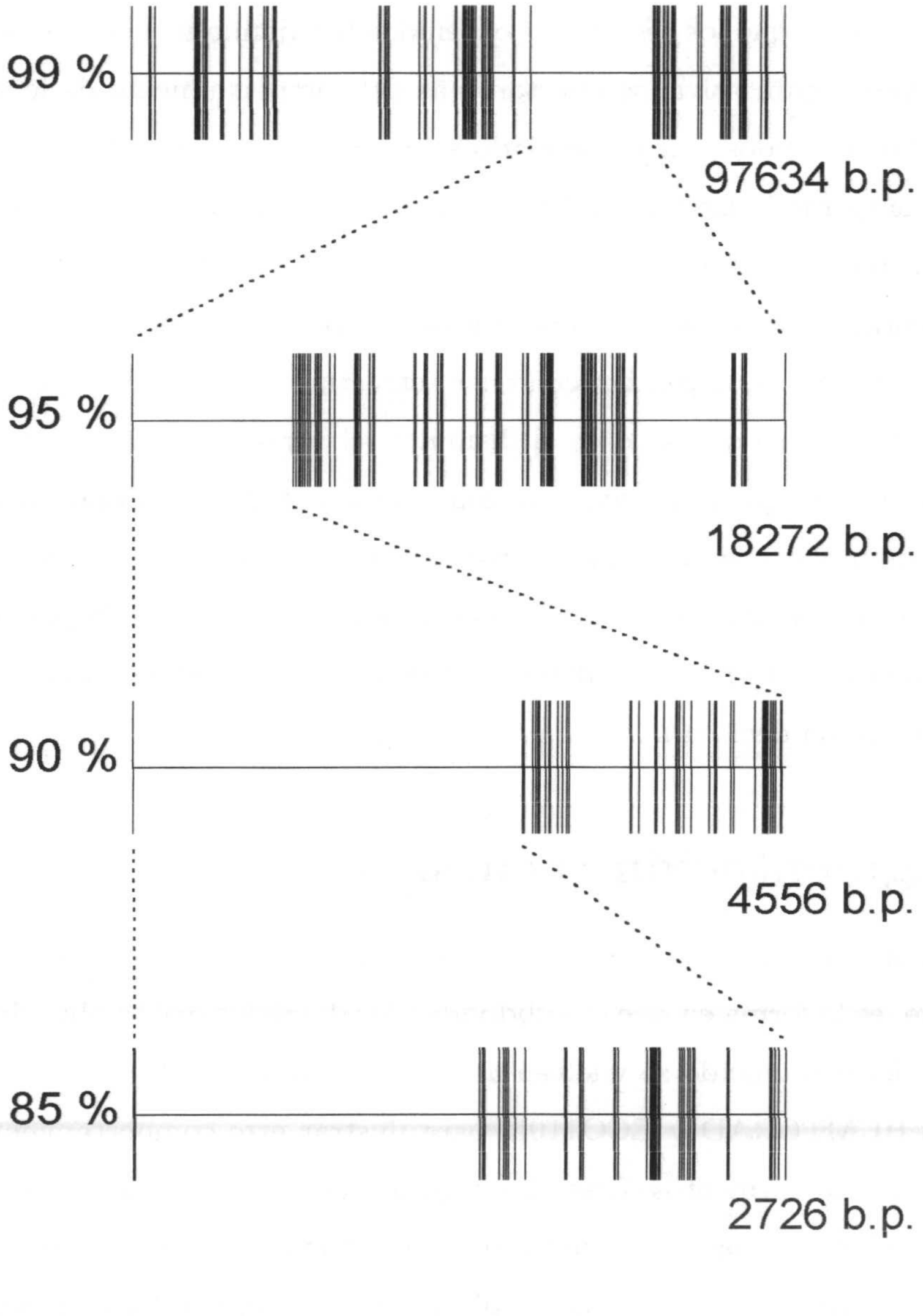

Figura 5.11: Segmentación recursiva de HUMTCRAD

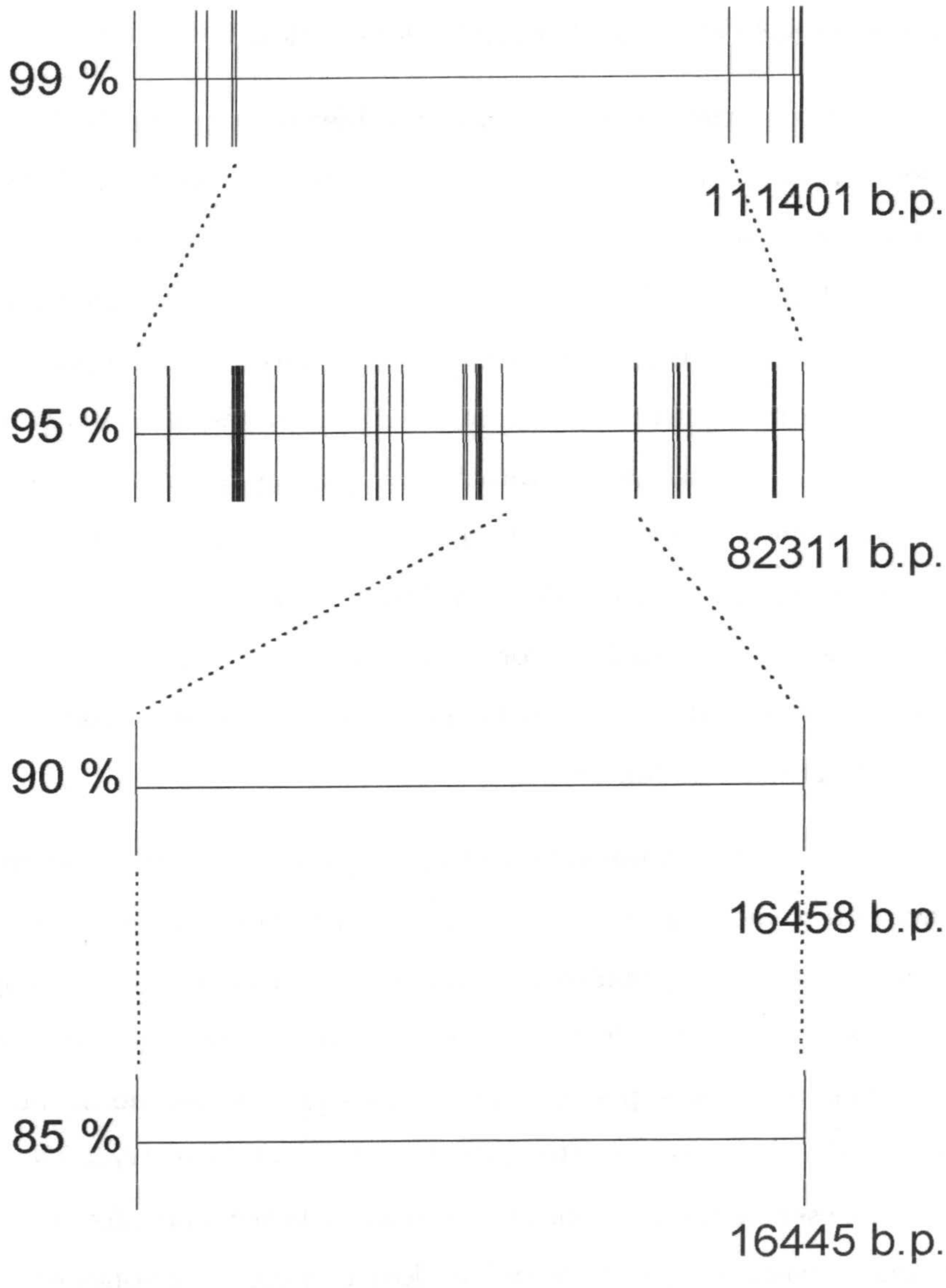

Figura 5.12: Segmentación recursiva de ECO110K

mente grandes, que no siguen subdividiéndose, 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.

Este resultado es interesante para apoyar la idea de la presencia de estructura fractal en las secuencias de ADN, y en concreto, la importancia de la alternancia de zonas de diferente composición para justificarla, esto es, su relación con la complejidad composicional. La figura 5.11 recuerda el aspecto de los conjuntos fractales en una dimensión con autosemejanza estadística [Sr 91], 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 un "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. Por otra parte, esta figura recuerda también el aspecto de los espectros atómicos, en los que, recientemente se han encontrado propiedades multifractales [Sa 96].

En la sección anterior vimos que la tesis de que el origen de las correlaciones de largo alcance son debidas a la alternancia de zonas con diferente composición, pero con una distribución de tamaños determinada, podría justificar el aspecto de las distribuciones de longitudes de dominios obtenidas, pero veamos que tal suposición no es suficiente para explicar lo que hemos encontrado aquí. Los autores que sostienen esta idea propusieron [Bu 93b] el modelo del *Lévy Walk Generalizado* (ver sec. 2.4.3) que, esencialmente, consiste en suponer la secuencia formada por una sucesión de tramos aleatorios (no correlacionados) con composiciones alternadas y cuyos tamaños siguen una ley de potencia. Estos tramos, por construcción, serían homogéneos, y por tanto, no seguirían subdividiéndose al reducir el nivel de significación. En la figura 5.13 vemos la segmentación recursiva de una secuencia gener-

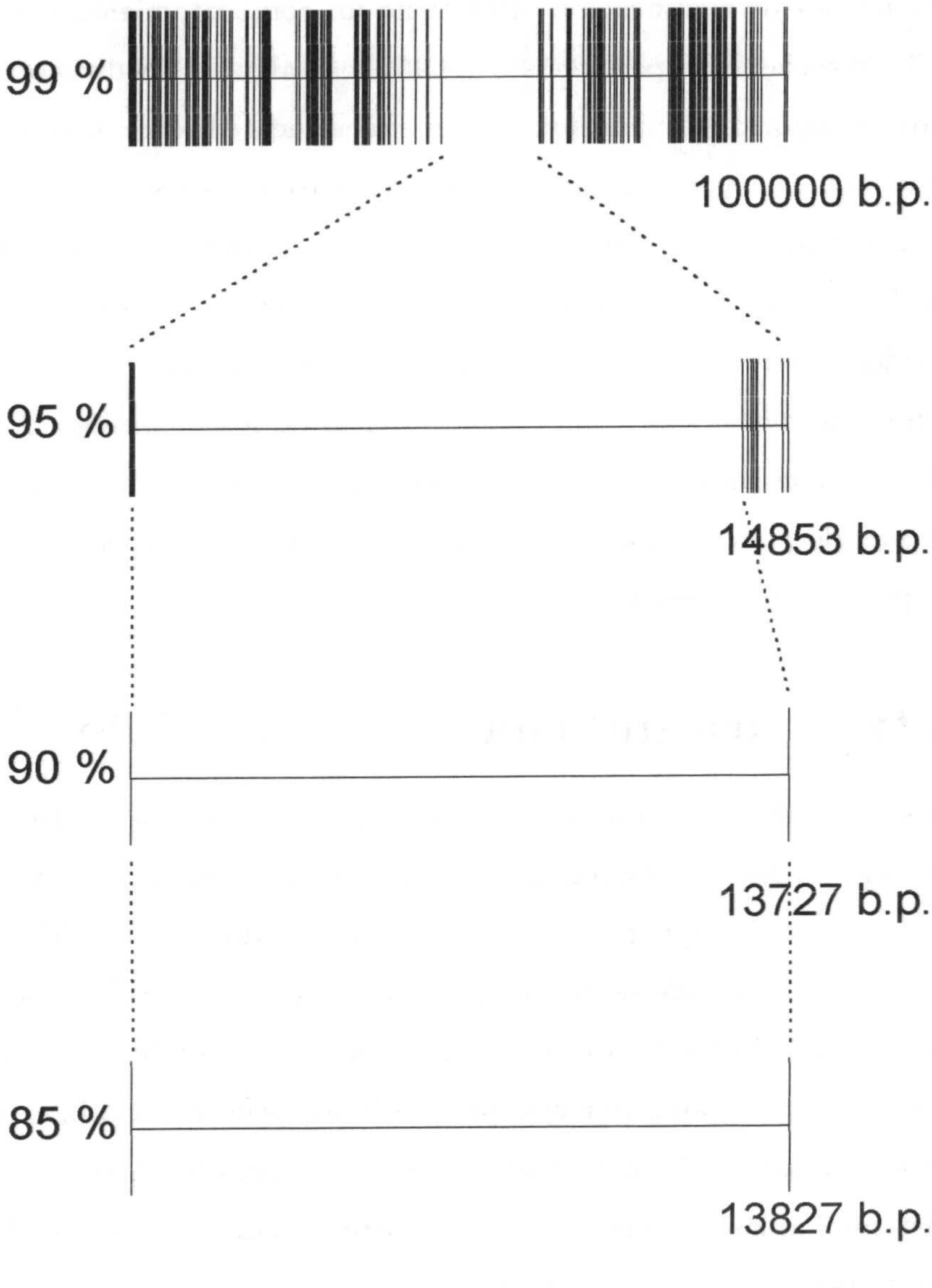

Figura 5.13: Segmentación recursiva de una secuencia generada a partir del modelo del Lévy-walk generalizado.

ada con este modelo (los parámetros son (sec. 2.5) $\mu = 2.2$, $\epsilon = 0.3$ y $l_0 = 10$). Vemos que, a niveles de significación altos tiene un comportamiento similar al de HUMTCRAD, de hecho, en el paso del 99% al 95% hay algunas subdivisiones, lo cual es lógico, ya que el algoritmo puede no detectar exactamente los dominios con los que fue generada la secuencia y agrupar varios de ellos, pero pronto se obtienen dominios homogéneos de tamaño considerable, algo parecido a lo que ocurre en ECO110K. Nótese, sin embargo, la similitud de la segmentación al 99% con la de HUMTCRAD, esto es, el modelo "imita a simple vista" bastante bien el aspecto de la secuencia humana, aunque no captura la complejidad interna de los dominios. A pesar de esto, el modelo aparece indistinguible de las secuencias con long-range en el análisis con DFA, lo que, de nuevo, parece indicar que este análisis tampoco captura toda la estructura presente en la secuencia.

5.6 Estructura interna de los dominios.

En vista de estos resultados, no sólo parece relevante la presencia de dominios y su organización, sino también su estructura interna, de hecho en secuencias con long-range, casi podríamos decir que pierde por completo sentido el hablar del número de dominios, ya que éste depende de forma muy importante del nivel de significación que utilicemos (siguiendo las analogías con los conjuntos fractales, algo parecido a lo que ocurre cuando se habla, por ejemplo, de la longitud de una línea costera o la superficie de una nube [Vo 88]). Por ello, en esta sección vamos a analizar de forma más sistemática las diferencias internas entre los dominios que se obtienen al segmentar las secuencias con y sin long-range.

Para esto, calcularemos el DFA a los dominios obtenidos en uno y otro caso. Este método, aunque no parece poner de manifiesto toda la complejidad de la secuencia, es bueno al menos para diferenciar entre secuencias completamente homogéneas

y secuencias con algún tipo de estructura interna.

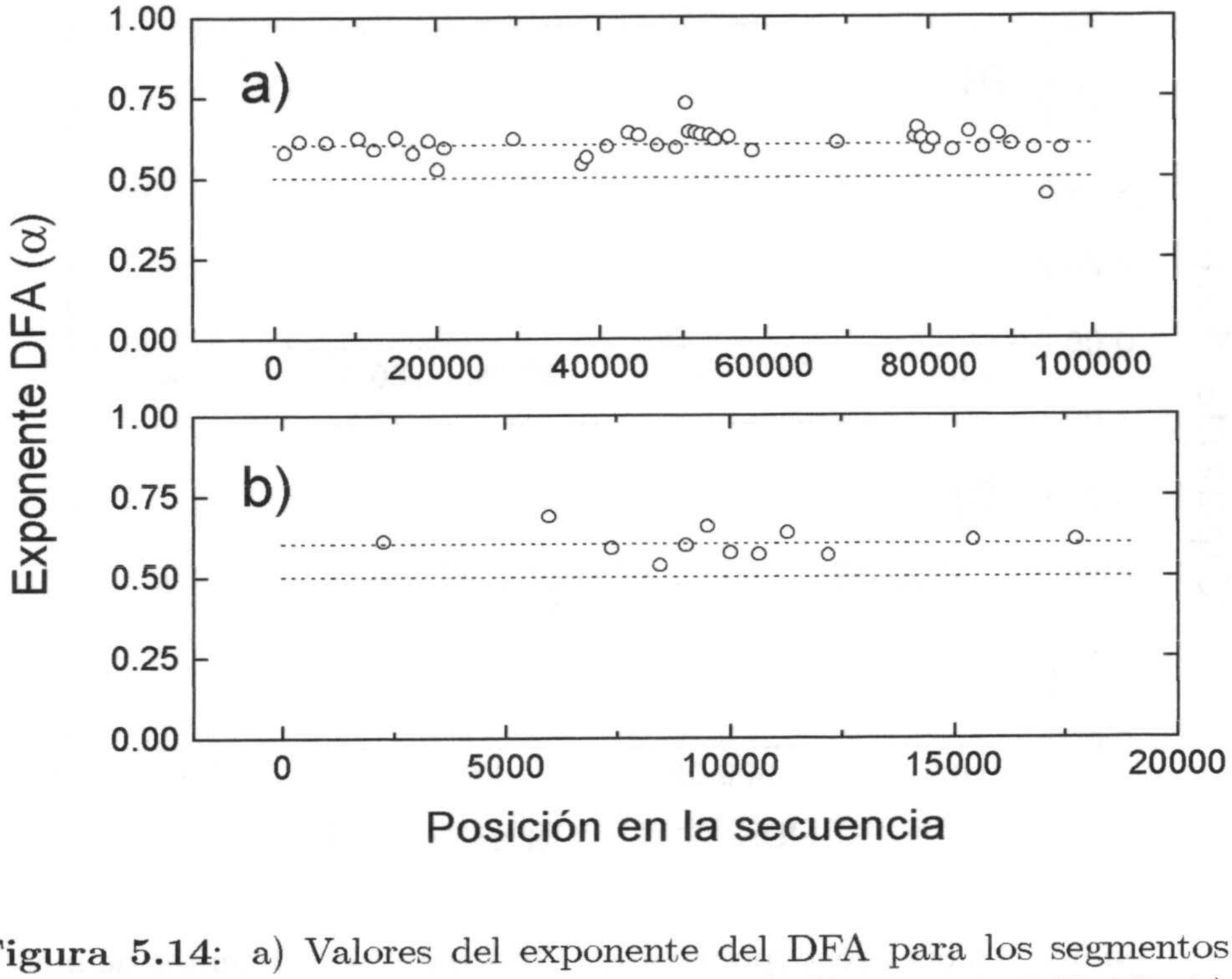

Figura 5.14: a) Valores del exponente del DFA para los segmentos de más de 200 b.p. de la segmentación al 99% de HUMTCRAD. b) El mayor de ellos (18272 b.p.) se segmenta de nuevo al 95% y se le calculan también los exponentes del DFA. Las líneas punteadas marcan los valores de $\alpha = 0.6$ (secuencia completa) y $\alpha = 0.5$ (ausencia de long-range)

En la figura 5.14 vemos los valores del exponente del DFA para los dominios de más de 200 bp obtenidos al segmentar HUMTCRAD al 99% (a) y los obtenidos al segmentar el mayor de éstos (18272 bp) al 95% (b). Como puede apreciarse, la mayoría tienen exponentes próximos a 0.6 (exponente de la secuencia completa), incluso mayores. También aparecen algunos valores cercanos a 0.5 (ausencia de

long-range), pero en casi todos los casos, en dominios de poca longitud.

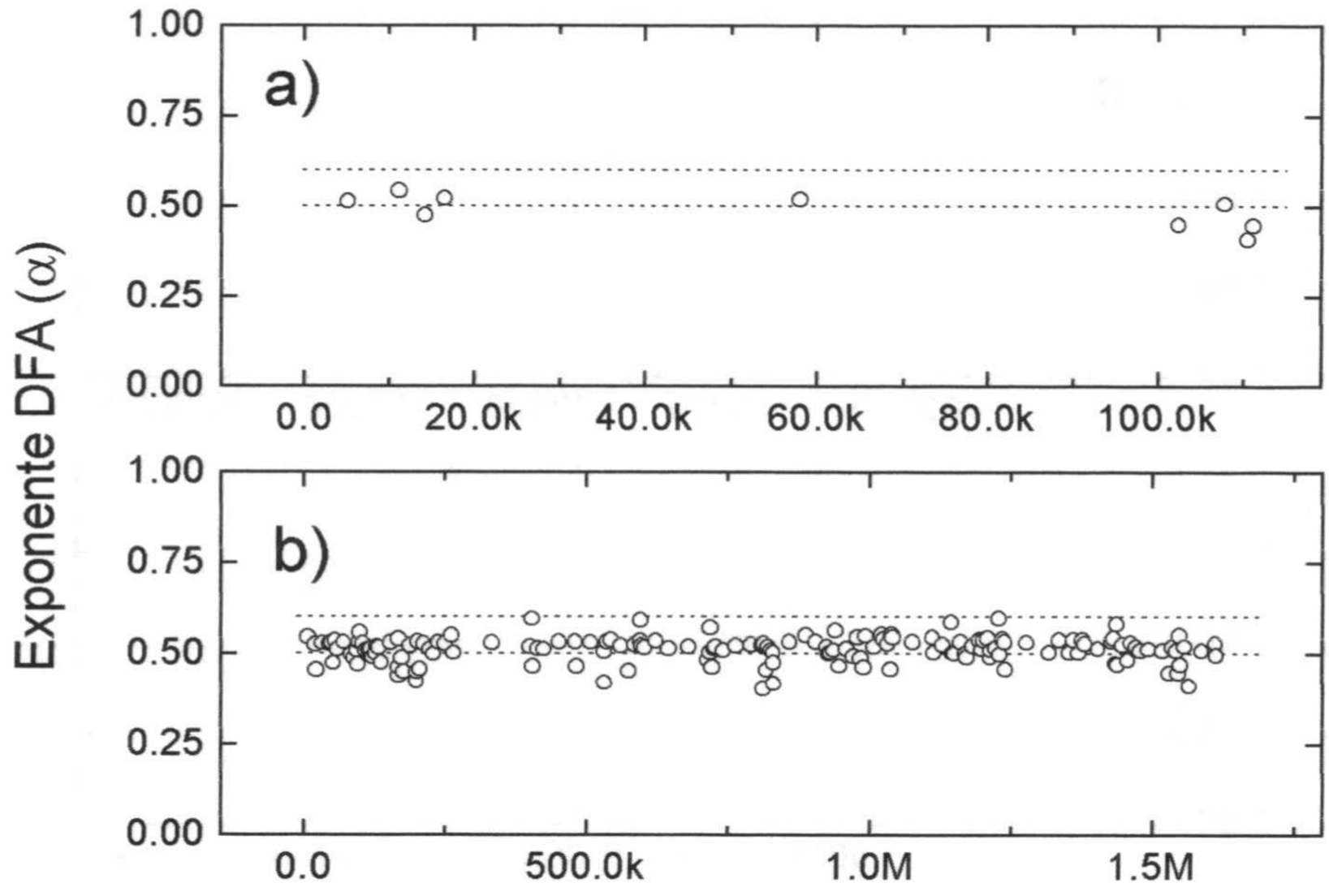

Figura 5.15: Valores del exponente del DFA para los segmentos de más de 200 b.p. de las segmentaciones al 99% de: a) ECO110K y b) M16J. De nuevo las líneas punteadas marcan los valores de $\alpha = 0.6$ y 0.5

Este resultado está completamente de acuerdo con lo que veíamos en la sección anterior, los dominios resultantes de la segmentación de HUMTCRAD no son homogéneos, como resulta del análisis con DFA, es más, tienen exponentes de escala α similares a los de la secuencia completa, como cabría esperar en un conjunto fractal con invarianza ante cambio de escala. Por el contrario, aplicando un análisis similar a las secuencias bacterianas M16J y ECO110K, el resultado es bien distinto (fig. 5.15): ahora los exponentes son próximos a 0.5 para la mayoría de los dominios,

al igual que ocurre con la secuencia completa, lo cual indica que ni ésta, ni las partes que la componen tienen una organización interna tan compleja como la de la secuencia humana.

Capítulo 6

Medida de la complejidad composicional

6.1 Introducción

Las secuencias de ADN almacenan la información genética completa de los organismos biológicos, considerados por todos como complejos, aunque no haya un acuerdo común en cuanto a la definición de lo que se entiende por complejidad. Algunos autores sostienen que esta complejidad debe reflejarse en la información genética, ya que ésta se puede considerar de alguna forma como el mecanismo que genera el organismo [Llo 88], aunque hay algunas objeciones a la correlación entre la complejidad de un sistema y la del mecanismo que lo genera [Rom 74]. En definitiva el objetivo final del proyecto Genoma Humano es justamente la comprensión del "lenguaje genético" que contienen las secuencias de ADN. En cierto sentido, este lenguaje es más simple que, por ejemplo, el castellano ya que su alfabeto está compuesto sólo por cuatro letras A,T,C y G que representan los cuatro nucleótidos o

cuatro bases. Pero, por otra parte, las secuencias de ADN son mucho más complejas que el castellano a causa de su gran longitud, que permite más combinaciones de letras y, por tanto, más "palabras".

La cuestión que se plantea es ¿cómo de complejas son las secuencias de ADN? Ha habido algunos intentos de relacionar la complejidad global del genoma con la del correspondiente organismo. En los primeros estudios se partía de la hipótesis de que el genoma de los organismos más avanzados debía ser más complejo, en cierto sentido no muy bien especificado. En general, la idea era que para organismos más complejos se requerirían más genes y genomas más grandes [Brt 69], [Sp 72]. En la mayoría de los casos [CS 85] [Sz 95] los resultados no fueron muy buenos, la correlación entre el tamaño del genoma y el grado de evolución del organismo era bastante pobre. La correlación con el número de genes da resultados mejores, pero las últimas estimaciones colocan al hombre al mismo nivel que los peces pulmonados [Sz 95].

Pero, aparte de la complejidad global del genoma, otra cuestión es cómo de complicado es el texto genético en sí. Esta última pregunta requiere la medida de las características estadísticas de la ordenación de los nucleótidos a lo largo de la secuencia. En la actualidad es admitido por la mayoría de los autores que la composición de la secuencia y las correlaciones de corto alcance no son factores decisivos. Parece que la heterogeneidad espacial en la composición de bases [Ber 89], [Ber 95] y la presencia de correlaciones de largo alcance [St 94], [Li 94] son las que dan cuenta de las propiedades estadísticas más relevantes del genoma.

En el capítulo anterior hemos visto que estos dos factores no parecen ser independientes y que, en gran medida, las distribuciones de longitudes de dominios están relacionadas con la presencia o no de correlaciones de largo alcance. Según esto, la heterogeneidad presente a todas las escalas en algunas secuencias de ADN,

parece ser muy relevante a la hora de explicar la complejidad de su estructura. La mayoría de los métodos de análisis propuestos hasta el momento suelen plantear problemas cuando la secuencia estudiada no es estacionaria [Ka 93], aunque algunos de ellos han sido mejorados con las técnicas estándard para el tratamiento de series no estacionarias [Man 96]. Pero en cualquier caso estas técnicas no abordan de lleno la cuestión, ya que sólo intentan modificar la secuencia o los estadísticos derivados de ella para que le sean aplicables resultados, en principio válidos únicamente para secuencias estacionarias. Se plantea, por tanto, la necesidad de una medida capaz de cuantificar la complejidad presente en una secuencia a causa de la heterogeneidad composicional, medida que podría ser útil tanto en el modelado de secuencias [Li 97a], como en estudios de evolución molecular o genómica comparada. Por el momento, no existe un acuerdo en la Comunidad Científica sobre las definiciones y medidas de la complejidad, ya que éstas deben ser objetivas y matemáticamente tratables además de contener la idea intuitiva de complejidad [Ben 90]. Ni la complejidad algorítmica, que es máxima para la aleatoriedad [Ch 87], ni otras medidas derivadas como la complejidad efectiva y la información total [Ge 96] o la información mutua, que mide la dependencia estadística, son apropiadas aquí. En su lugar, parece mucho más adecuada una medida que tenga en cuenta tanto el número de dominios como su heterogeneidad composicional. En este capítulo presentamos una medida , basada en el algoritmo de segmentación que vimos en el capítulo 4 y que, como veremos tiene muchas de las propiedades deseables para una medida de complejidad [Ro 97], [Li 97b].

6.2 Divergencia total de la segmentación

En el capítulo anterior vimos que, fijado el nivel de significación, la principal diferencia que aparece entre las secuencias consideradas en la bibliografía como complejas

(p.e. HUMTCRAD) y las consideradas como simples (p.e. ECO110K) es el número de dominios en los que se dividen. Pero esta cantidad, por sí sola, aunque bastante relevante, no da cuenta por completo de la heterogeneidad de la secuencia. Por ejemplo, una secuencia en la que aparezcan unos cuantos dominios muy bien definidos (con una divergencia muy alta) no sería distinguible de otra aleatoria en la que, a causa de las fluctuaciones estadísticas, aparezca el mismo número de dominios con la divergencia justa para ser considerados como significativos. En la figura 6.1 se han representado los random-walks de dos secuencias que ilustran esta situación: una secuencia binaria con 23 zonas alternadas de ceros y unos y una secuencia aleatoria. Si segmentamos las dos secuencias con una significación del 90%, en ambos casos obtenemos 23 dominios composicionales, por lo que, a este nivel de significación, serían indistinguibles si tenemos en cuenta sólo el número de dominios. Es por esto, que se requiere una medida que también tenga en cuenta, de alguna forma, cómo de diferente es la composición de los distintos dominios encontrados.

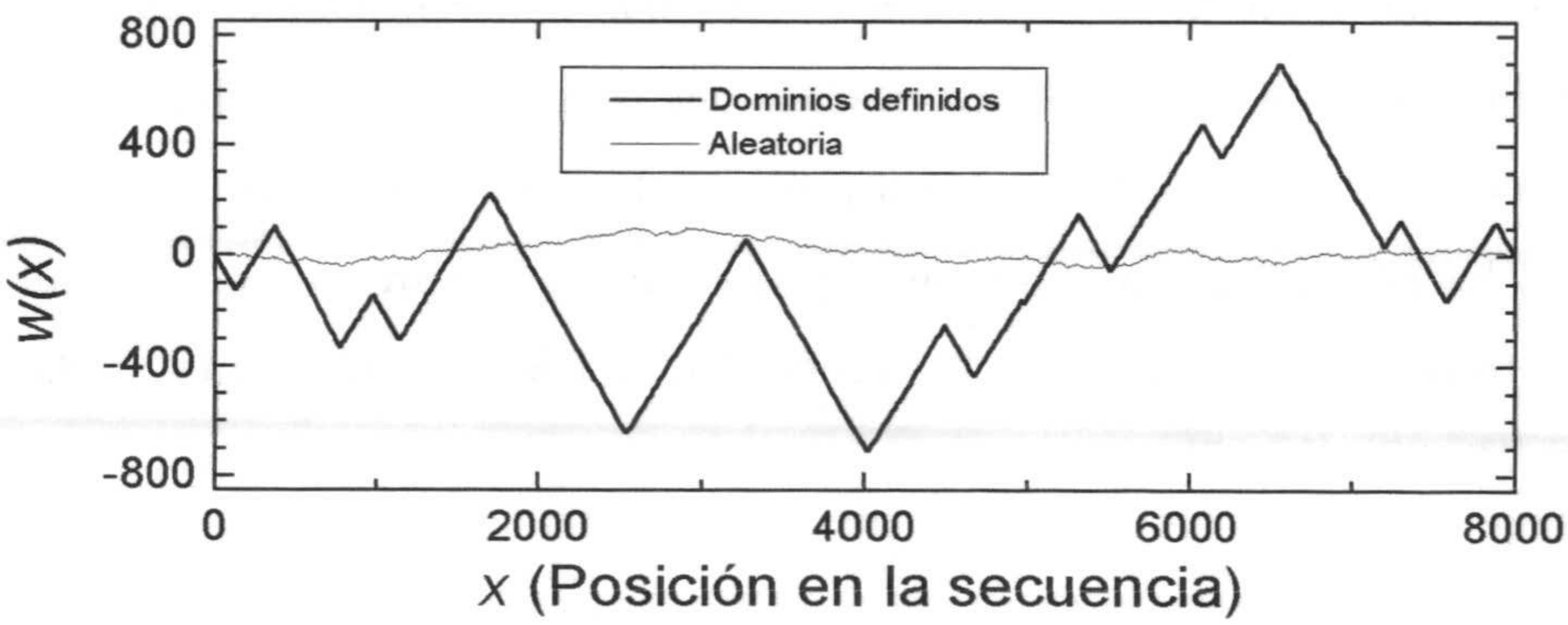

Figura 6.1: Random-walk para una secuencia artificial con dominios bien definidos y una aleatoria generada por ordenador

Aquí proponemos para tal fin la **Divergencia Total de la segmentación,** la misma medida utilizada como criterio de optimización en el proceso de segmentación:

$$\mathrm{JS}_m = \mathrm{H}(\mathcal{F}) - \sum_{i=1}^{m} \frac{n^{(i)}}{N} \mathrm{H}\left(\mathcal{F}^{(i)}\right) = \sum_{i=1}^{m} \frac{n^{(i)}}{N} \left[\mathrm{H}(\mathcal{F}) - \mathrm{H}\left(\mathcal{F}^{(i)}\right)\right] \tag{6.1}$$

Es evidente que JS_m reúne los requisitos como medida de la heterogeneidad composicional de la secuencia. En efecto, nótese como el tercer término de (6.1) es la suma, pesada, de las desviaciones de los histogramas de cada dominio con respecto al histograma medio de la secuencia. Por tanto, si las diferencias de composición entre los dominios son importantes, la divergencia total tomará un valor más alto que si los segmentos tienen una composición parecida y son únicamente resultado de las fluctuaciones estadísticas. En concreto, en el ejemplo que veíamos antes, la divergencia total de la segmentación vale 1 bit para la secuencia con dominios bien definidos, mientras que para la aleatoria vale 0.01 bits.

Por otra parte, esta medida crece al aumentar el número de dominios, por lo que también tiene en cuenta este factor. Esta propiedad no es evidente a partir de (6.1), pero se deduce de forma inmediata de (4.41).

Otras propiedades interesantes son:

- Esta divergencia, al ser la misma que establece el criterio de optimalidad de la segmentación, daría el valor máximo, fijado el nivel de significación, para la segmentación óptima, lo cual es bastante razonable.

- Puede ser utilizada para secuencias no estacionarias. De hecho podríamos decir que no es otra cosa que una forma de medir la no estacionariedad de la secuencia: una secuencia estacionaria, en teoría, tendría una divergencia total nula, ya que no se dividiría en dominios composicionales.

- Es independiente de la longitud total de la secuencia.

Esta última propiedad aparece a consecuencia de los factores proporcionales a las longitudes que multiplican a cada sumando. De todas formas es una propiedad controvertida ya que para verificarla habría que tener dos secuencias de longitud distinta con la misma heterogeneidad composicional y verificar que ambas tienen una misma divergencia total al ser segmentadas; pero ¿cómo podemos afirmar a priori que tienen una misma heterogeneidad composicional? Una prueba, aunque no definitiva, que puede convencernos de la independencia de la medida con la longitud total de la secuencia, es la siguiente: consideremos una secuencia dividida en m dominios y supongamos que obtenemos otra duplicando cada símbolo. Esta secuencia tendría una longitud doble a la inicial y se dividiría en los mismos dominios, que además tendrían los mismos histogramas y los mismos pesos en relación a la longitud total, por lo que su divergencia total sería la misma. Y en este caso parece bastante razonable asignar a priori la misma heterogeneidad a las dos secuencias.

Por último, comentar que esta medida es aplicable a cualquier división de la secuencia en segmentos, aunque cuando realmente tiene sentido como medida de la heterogeneidad composicional es cuando estos segmentos son dominios composicionales; esto es, cuando la división de la secuencia es el resultado de un proceso de segmentación como el descrito en el capítulo 4. Tal y como se planteó entonces la estrategia de segmentación y la definición de dominios, está claro que la divergencia total de la segmentación depende del nivel de significación (s) considerado, por ello en lo sucesivo la notaremos como función de s, $\mathrm{JS}_m(s)$.

6.3 Perfiles de complejidad

Puesto que la medida de la heterogeneidad composicional que aquí proponemos depende del nivel de significación, parece interesante explorar cómo depende de ella

y además, vistos los resultados obtenidos al final del capítulo anterior, en los que se pone de manifiesto la importancia de comparar las segmentaciones a distintos niveles para el análisis de la compleja heterogeneidad de algunas secuencias de ADN, parece que una representación de JS_m en función de s puede dar una información bastante más interesante que la de un único valor. En lo sucesivo nos referiremos a este tipo de representaciones como **Perfiles de Complejidad**. Un perfil de complejidad estará definido en el intervalo $[0, 100]$. Su valor mínimo es 0, es fácil comprobar que este valor sólo se alcanza si la secuencia contiene un único dominio (no ha habido ninguna segmentación). Este valor se alcanzará necesariamente para $s = 100\%$, puesto que nunca podremos afirmar que dos segmentos tienen diferente composición con total seguridad, y por tanto a este nivel de significación la secuencia no debe segmentarse.

En cuanto al valor máximo, depende de la secuencia y se alcanza para la segmentación en la que todos los dominios son cadenas de símbolos iguales (*runs*), que coincidiría con la idea intuitiva de secuencia segmentada al máximo. Además, veamos que este valor máximo es justamente $\mathrm{H}(\mathcal{F})$, esto es, la entropía del histograma de la secuencia completa.

En efecto: basta con tener en cuenta que $\mathrm{JS}_m \geq 0$ y que en el segundo miembro de (6.1) aparece como diferencia de dos cantidades positivas, por tanto el valor máximo lo alcanzará cuando el término que resta sea nulo. En esta situación $\mathrm{H}\left(\mathcal{F}^{(i)}\right) = 0 \ \ \forall i$ y los histogramas de los dominios serán degenerados. En la práctica nuestro algoritmo no alcanza por lo general este máximo ya que, para dividir una subsecuencia, ésta debe tener al menos 3 símbolos; por lo que incluso bajando el nivel de significación a cero, no llegaríamos a deshacer la secuencia en símbolos. Por otra parte el criterio de significación puede plantear paradojas cuando intentamos dividir secuencias muy cortas. Consideremos por ejemplo la secuencia binaria:

$$S = \{1, 1, 0, 1\}$$

Si intentamos segmentarla, obtenemos que el máximo valor de divergencia ($\mathrm{JS}_2 = 0.311$ bits) es el que aparece al tomar $S^{(1)} = \{1,1\}$ y $S^{(2)} = \{0,1\}$; y ahora para ver si la división en dominios es significativa nos preguntamos por la probabilidad de que al generar al azar una secuencia con $N = 4$ y $\mathcal{F} = \left\{\frac{1}{4}, \frac{3}{4}\right\}$ y dividirla de forma que $n^{(1)} = n^{(2)} = 2$ obtengamos valores menores o iguales de divergencia. Pero hay un problema, en estas condiciones sólo hay dos posibles configuraciones $\{1,1\}\{0,1\}$ y $\{0,1\}\{1,1\}$ que además tienen la misma divergencia y, por tanto, no parece razonable decir que la divergencia entre los dos segmentos sea debida a una diferencia significativa cuando es la única posible compatible con las restricciones.

De todas formas los valores próximos al máximo se alcanzan para niveles de significación muy bajos que, como veremos, no serán importantes ya que a estos niveles prácticamente sólo se detectan fluctuaciones estadísticas que no son muy relevantes a la hora de considerar la complejidad de la secuencia.

Por otra parte, también es fácil comprobar que $\mathrm{JS}_m(s)$ es una función decreciente. Consideremos dos niveles de significación s y s' con $s > s'$, obviamente todos los segmentos que son dominios composicionales para s lo son también para s', por lo que, al menos $\mathrm{JS}_m(s') = \mathrm{JS}_m(s)$; pero en general al disminuir el nivel de significación algunos de estos dominios se escindirán en dos o más, y teniendo en cuenta que esta escisión sólo puede aumentar el valor de la divergencia (4.40) tenemos que $\mathrm{JS}_m(s') \geq \mathrm{JS}_m(s)$. Nótese que la expresión de JS_m en términos de las divergencias de los cortes sucesivos que nos llevan a la segmentación final (4.41) nos asegura que esta propiedad también se cumple para el resultado de segmentaciones realizadas con nuestro algoritmo heurístico.

6.3.1 Comparación con otras medidas de complejidad

Desde el punto de vista más genérico, la complejidad de algo se puede definir como alguna medida que caracterice la dificultad necesaria para llevarlo a cabo [Li 91b]. Intentando dar una definición amplia se han propuesto diversas medidas de la complejidad: La longitud de la descripción mínima y completa del sistema, usando cierto lenguaje [Lof 77], [Pa 80], [Pa 82], o bien la longitud del menor algoritmo que lo puede generar [Ko 65], [Ch 75], [Ch 87]. Estas medidas reciben la denominación genérica de complejidad algorítmica. Además de no ser calculables [Kam 91], sólo tienen en cuenta la dificultad para predecir una parte del sistema conocidas otras, pero de ninguna forma contemplan la posible dificultad para generar el sistema o la presencia de patrones más o menos complejos. Por ejemplo para una secuencia binaria, la máxima complejidad algorítmica se alcanzaría para una secuencia aleatoria.

Teniendo el cuenta la complejidad del proceso que genera el sistema a partir de su descripción, se ha propuesto la profundidad lógica [Ben 88], y recientemente [Ge 96], la complejidad efectiva, similar a la complejidad algorítmica pero, en lugar de tratar de describir por completo el sistema busca la descripción mínima de las regularidades de éste, esto es, elimina la componente aleatoria.

En cualquier caso lo que parece claro es que resulta difícil dar una definición en sentido amplio, que sea aplicable a todos los sistemas. Desde luego, en los sistemas biológicos, una definición de complejidad que se hace máxima para la aleatoriedad, como ocurre con la complejidad algorítmica, no parece muy adecuada. En algunos estudios de evolución se adopta un punto de vista más limitado a la hora de hablar de complejidad: se considera un sistema más complejo cuanto más diferenciado es. Más exactamente, la complejidad de un sistema debe ser una función creciente con el número de partes diferentes que lo componen y el número de diferentes interacciones

entre ellas. McShea [Mc 96], siguiendo este punto de vista, considera la complejidad compuesta de cuatro factores o cuatro tipos: (1) Complejidad no jerárquica de las partes, (2) complejidad no jerárquica de las interacciones, (3) complejidad jerárquica de las partes y (4) complejidad jerárquica de las interacciones.

Atendiendo a la clasificación de McSea, nuestra medida de complejidad estaría enmarcada dentro del tipo (1): Las partes que consideramos son los dominios composicionales que contiene la secuencia de ADN sin tener en cuenta las interacciones entre ellas. La complejidad jerárquica, también se tiene en cuenta cuando llevamos a cabo una segmentación recursiva (ver capítulo 5), aunque el perfil de complejidad, al promediar sobre todos los dominios, pierde un poco de vista la relación jerárquica entre dominios concretos, aunque sí nos puede dar una idea en términos estadísticos. Por otra parte nuestra medida también guarda cierta relación con la complejidad algorítmica y la complejidad efectiva. La primera trataría de describir la secuencia base a base, mientras que la segunda tan sólo se ocuparía de describir las regularidades. En nuestro caso la regularidad que buscamos es la homogeneidad composicional de un dominio: tratamos de describir la secuencia en términos del número de dominios que contiene y su composición. Pero no hay que perder de vista el nivel de significación, para un nivel muy alto, sólo las regularidades más relevantes entrarían a formar parte de la descripción de la secuencia, y a medida que este baja regularidades menos relevantes se tienen en cuenta en la descripción, llegando al extremo de que para resolución cero la descripción tendría en cuenta cada símbolo de la secuencia. De alguna forma nuestra medida, al variar el nivel de significación, se "mueve" de forma continua entre la complejidad algorítmica y la efectiva, aunque si se quiere enmarcar dentro de una de ellas estaría más cerca de la complejidad efectiva [Li 97b].

Nosotros estamos interesados en tres aspectos de la descripción de los domi-

nios: a) ¿Cuántos dominios homogéneos tiene una secuencia de ADN? b)¿Cómo de grandes son las diferencias entre esos dominios? y c)¿A qué nivel de detalle estamos describiendo esa heterogeneidad?

Con algunas excepciones, a mayor número de dominios tendremos una descripción más larga y $\mathrm{JS}_m(s)$ es mayor cuanto mayor sea el número de dominios, por tanto esto es un dato a favor de la medida. En concreto una de las excepciones es el caso de una secuencia con una estructura de dominios perfectamente periódica, cosa que rara vez ocurre en las secuencias de ADN.

El que haya una diferencia mayor con los dominios adyacentes no hace que la descripción sea más complicada. Por tanto $\mathrm{JS}_m(s)$ contiene información que, realmente, no está relacionada con la dificultad de la descripción de la heterogeneidad de la secuencia. Pero hay que tener en cuenta que una mayor diferencia de composición es una forma de asegurar que la división en dominios es más fiable.

Quizá uno de los aspectos más interesantes de $\mathrm{JS}_m(s)$ sea el tercero de los mencionados arriba. Que una secuencia sea homogénea o no depende el nivel de detalle con el que la estamos describiendo: detalles que son relevantes a un nivel de significación pueden no serlo al aumentar éste. De hecho, para menores niveles de significación, mayor es la heterogeneidad que vemos y por consiguiente mayor la complejidad. Esta es una propiedad fundamental para una medida de la complejidad desde el punto de vista de la heterogeneidad [Li 91b], [Ge 94]. Las secuencias aleatorias requieren una descripción muy larga si se tienen que describir todos los detalles (de hecho, salvo términos constantes la misma longitud que la de la propia secuencia), pero la descripción es muy simple si sólo lo hacemos a grandes rasgos.

6.3.2 Interpretación de los perfiles

Consideremos en primer lugar secuencias binarias formadas por tramos de la misma longitud con composiciones complementarias, esto es, la frecuencia de aparición de

ceros en un tramo es igual a la frecuencia de aparición de unos en las adyacentes (representando este tipo de secuencias mediante un random walk aparecería como una sucesión de dientes de sierra con una inclinación que dependerá de la composición de los tramos).

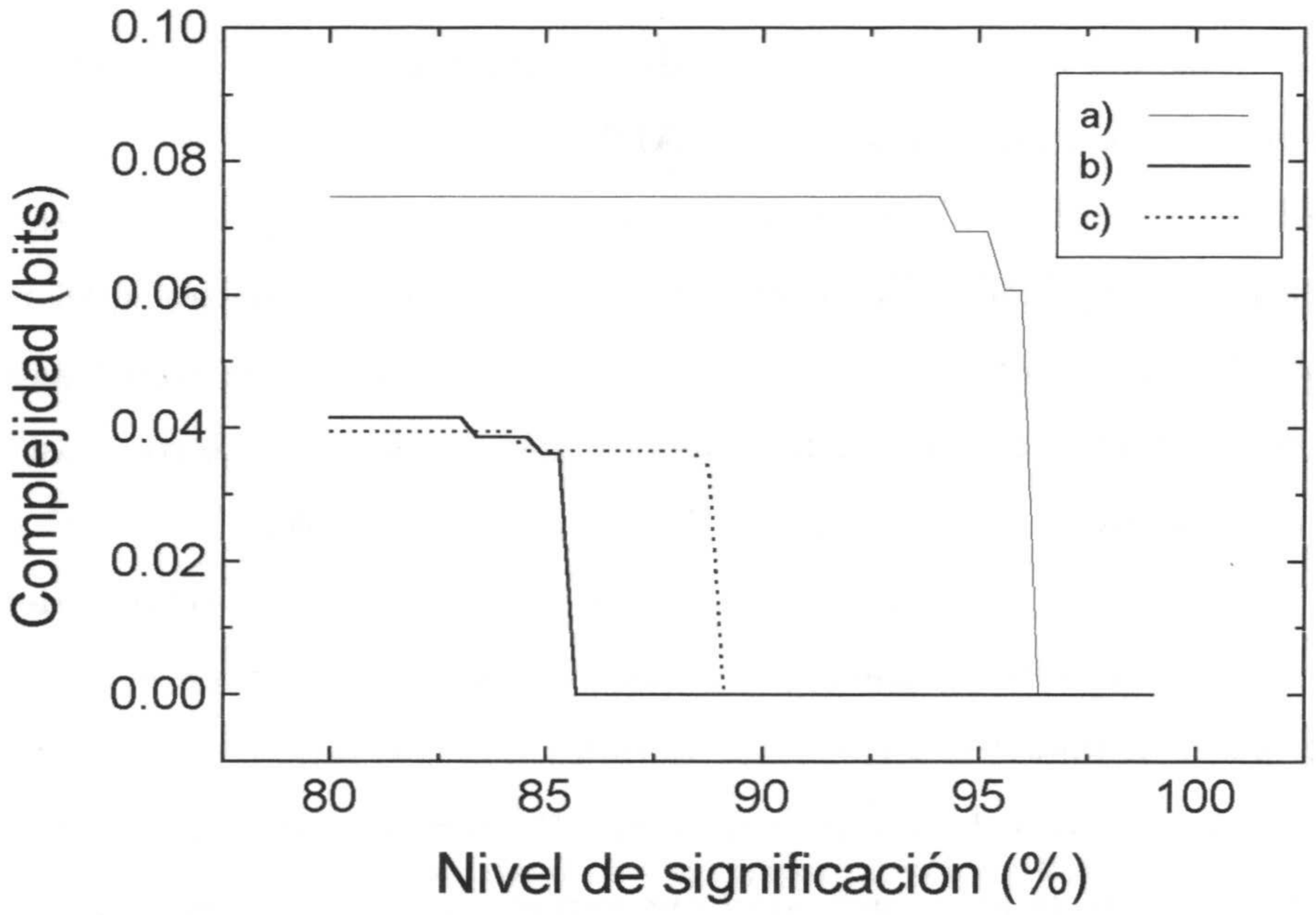

Figura 6.2: Perfiles de complejidad de varias secuencias con dominios de igual longitud y sesgo alternado.

En la figura 6.2 aparecen los perfiles de complejidad de tres secuencias de este tipo. Las secuencias se han construido de la forma siguiente:

a) Se han unido 36 tramos de 50 símbolos en los que aparecen 33 unos y 17 ceros o viceversa.

b) También 36 tramos de 50 símbolos pero ahora con 31 unos y 19 ceros o viceversa.

c) También 36 tramos pero en este caso con 60 símbolos y una composición de 37 unos y 23 ceros o viceversa.

En todas ellas el aspecto del perfil es similar: Inicialmente toma el valor cero durante cierto intervalo de niveles de significación, esto indica que las divergencias entre los tramos de diferente composición no son suficientemente significativas para que se produzcan divisiones. Llegado a un nivel de significación "crítico" se produce un estallido y aparecen bruscamente muchos dominios durante un intervalo relativamente pequeño y de aquí en adelante no aparecen más dominios dado que la composición interna de los tramos es homogénea (si se prosiguiese hasta niveles más bajos aparecerían más dominios cuando el algoritmo comenzase a segmentar repeticiones de dos o tres símbolos). La aparición brusca de dominios se explica de forma inmediata teniendo en cuenta que, al tener todos la misma longitud y composiciones alternadas, la divergencia entre tramos adyacentes es la misma y además será significativa para el mismo valor de s. Lo que no se explica únicamente en términos de niveles de significación es el hecho de que el escalón no sea perfecto, esto es, que se extienda a lo largo de un intervalo de valores de s; de hecho esto es consecuencia del algoritmo heurístico de segmentación: en teoría todos los dominios deberían aparecer a un mismo nivel de significación, concretamente aquél para el cual la divergencia entre tramos adyacentes (que es igual para todos) coincida con el percentil correspondiente. Esto no es así porque el algoritmo, cuando intenta la primera división, lo hace a partir de la secuencia completa y no "conoce" a priori la existencia de los tramos, para él dos tramos consecutivos de composición alternada son indistinguibles de un tramo con la composición media. El único punto por el que puede empezar a segmentar es por la separación entre el primer o último tramo

con el resto de la secuencia. Se puede comprobar que en este punto la divergencia (y la significación de ésta) es menor que la que habría entre dos tramos consecutivos y, por tanto el algoritmo empezará a segmentar a un valor de s menor del que cabría esperar en teoría*. Esta situación se repetiría, una vez escindido el primer tramo, cuando el algoritmo pasase a intentar dividir los restantes, y como en general la divergencia entre un tramo y un número variable de ellos no va a ser siempre la misma, no todos los cortes se dan al mismo nivel de significación. Tal vez la forma de evitar este problema sea introduciendo en el algoritmo alguna forma de explorar un segmento aunque el valor máximo de divergencia no se encuentre en él. En las simulaciones Montecarlo en las que el algoritmo busca el equilibrio dirigiendo el estado del sistema hacia el mínimo de energía, ocurre un problema similar con los estados metaestables (Ref. Montecarlo). En este caso el problema se resuelve asignando una pequeña probabilidad, pero no nula, a los pasos que suponen una aumento de energía.

De todas formas aquí el problema es distinto, ni siquiera está claro que esto deba ser un defecto del método. El algoritmo detecta los dominios repetidos a un nivel de significación siempre menor que el teórico y, además se puede comprobar que este nivel será menor cuanto menor sea el tamaño de los dominios en relación con la longitud total de la secuencia. Este comportamiento podría ser razonable: aunque los dominios sean del mismo tamaño y tengan la misma diferencia de composición, si son menores en proporción al tamaño de la secuencia completa son menos relevantes a la hora de considerar la heterogeneidad composicional de la secuencia.

En cuanto al valor de s para el que aparece el escalón, como se ha dicho, este debería ser el valor cuyo correspondiente cuantil es la divergencia entre tramos consecutivos, pero el algoritmo heurístico hace que sea algo menor. De todas formas

*Esto está de acuerdo con el hecho de que la divergencia total que se obtiene por nuestro procedimeinto heurístico es una cota inferior del valor teórico de la segmentación completa.

veamos cómo se relaciona este valor con las características de los tramos. Para la secuencia a) el estallido se produce en torno al 96% mientras que para la b) en torno al 86%; esto nos dice que a pesar de tener la misma longitud los dominios de la primera son más significativos que los de la segunda, esto es bastante razonable ya que la diferencia de composición entre tramos consecutivos es mayor en el primer caso y, a igualdad de longitud, una mayor diferencia en la composición nos da una divergencia más significativa. Por otra parte los tramos de la secuencia c) tienen una diferencia de composición muy similar a los de la b) (algo menor, de hecho) y sin embargo aparecen para un valor mayor de s, esto es debido a que tienen una mayor longitud: a igualdad en la diferencia de composición es más significativa la diferencia entre los tramos al ser de mayor longitud, por así decirlo, es menos probable que la diferencia haya sido generada al azar.

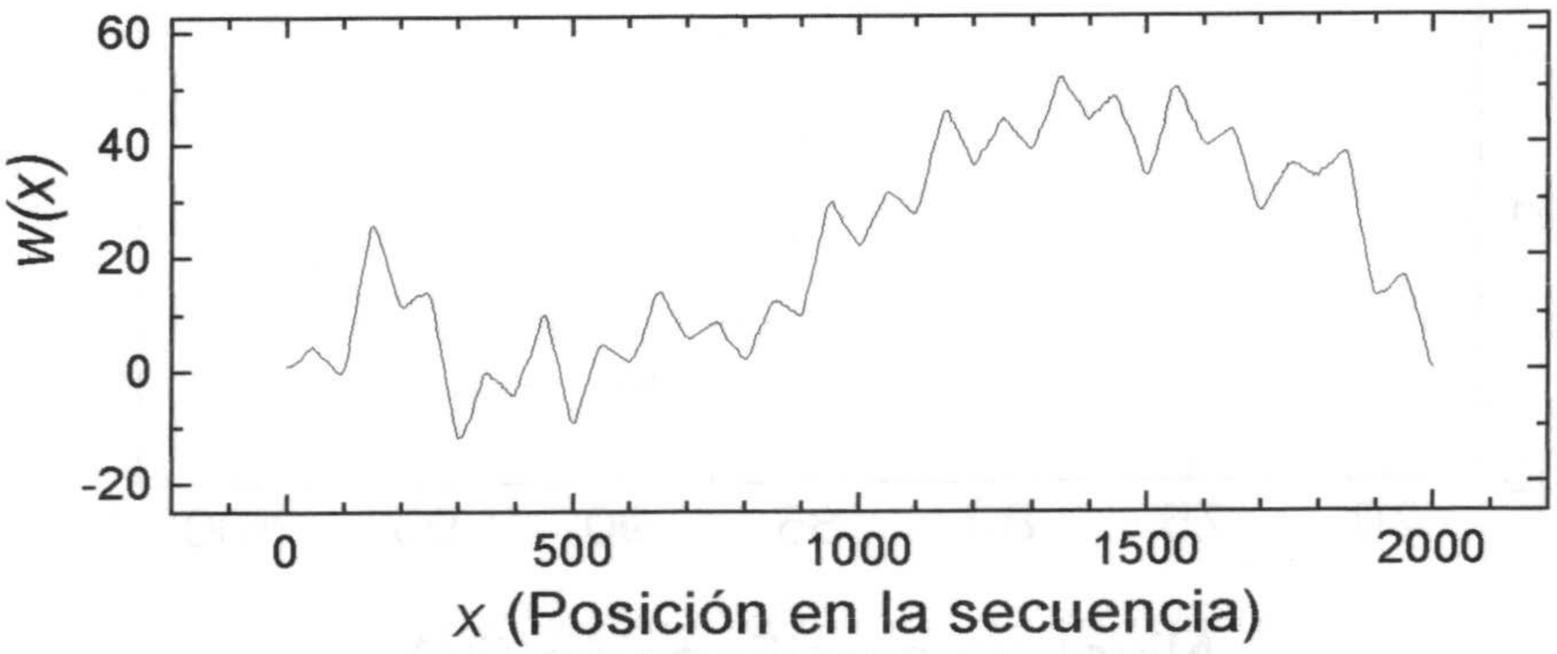

Figura 6.3: Random walk de una secuencia construida con tramos de longitud 50 pero con diferentes composiciones.

En vista de esto, la presencia de un escalón en un perfil no se deberá necesariamente a la presencia de muchos dominios con la misma longitud y diferencia

de composición sino que puede deberse a la existencia de dominios con diferentes longitudes y varias diferencias de composición pero dando lugar a divergencias con la misma significación. Si suponemos válida la aproximación de la distribución de JS_2 por la distribución χ^2 (4.31) tenemos que las parejas de dominios cuya divergencia y suma de longitudes (N) verifiquen:

$$N \ JS_2 = cte. \tag{6.2}$$

tendrán una misma significación (véase ec. 4.32).

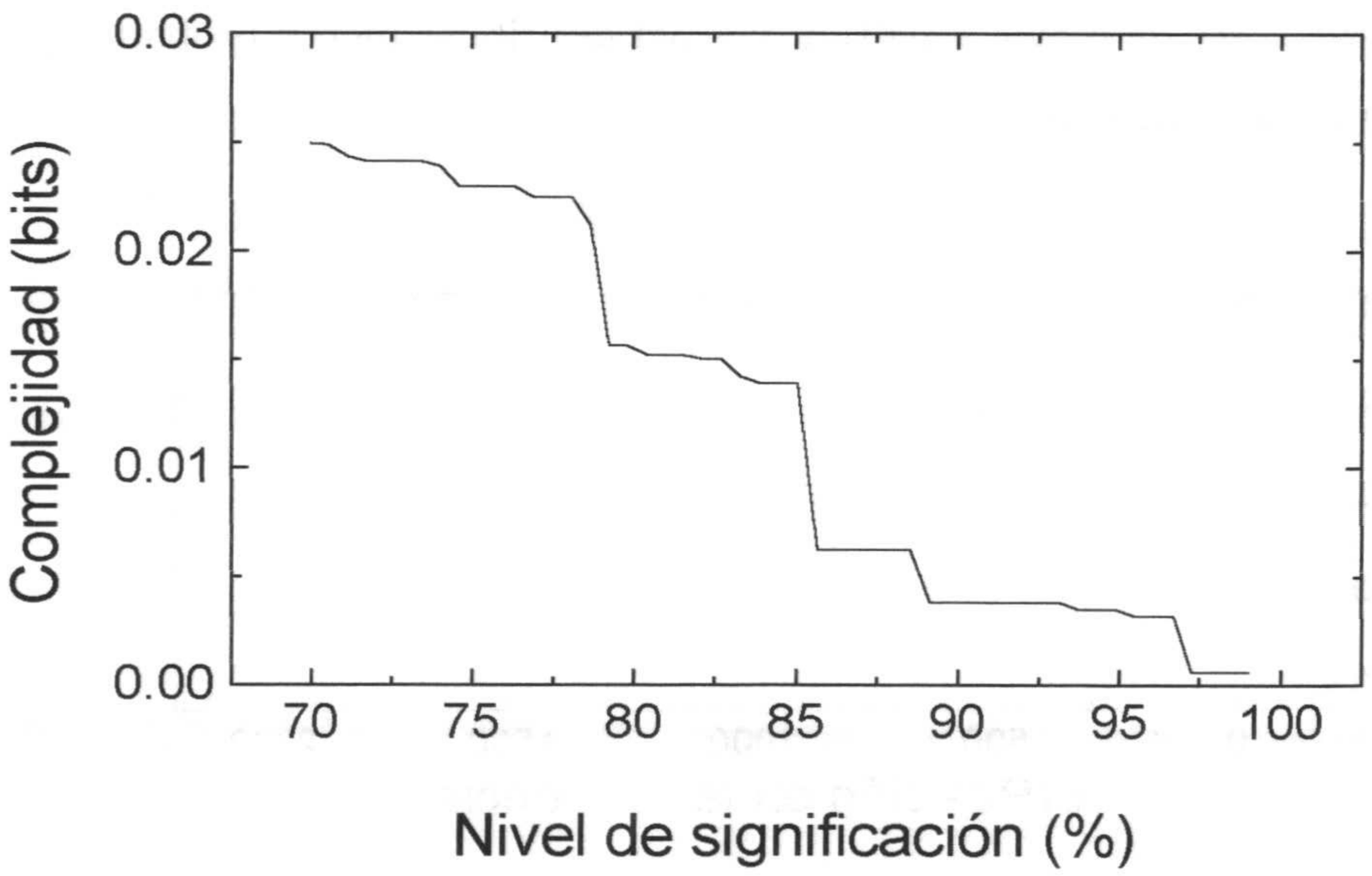

Figura 6.4: Perfil de complejidad de una secuencia construida con tramos de longitud 50 pero con diferentes composiciones.

Por el contrario a los ejemplos que hemos visto aquí, un perfil en el que

aparezca una variación más o menos continua indicará la presencia de dominios a todos los niveles de significación. Esto puede indicar la presencia de dominios de múltiples longitudes, múltiples composiciones o una combinación de ambos. Veamos algunos ejemplos.

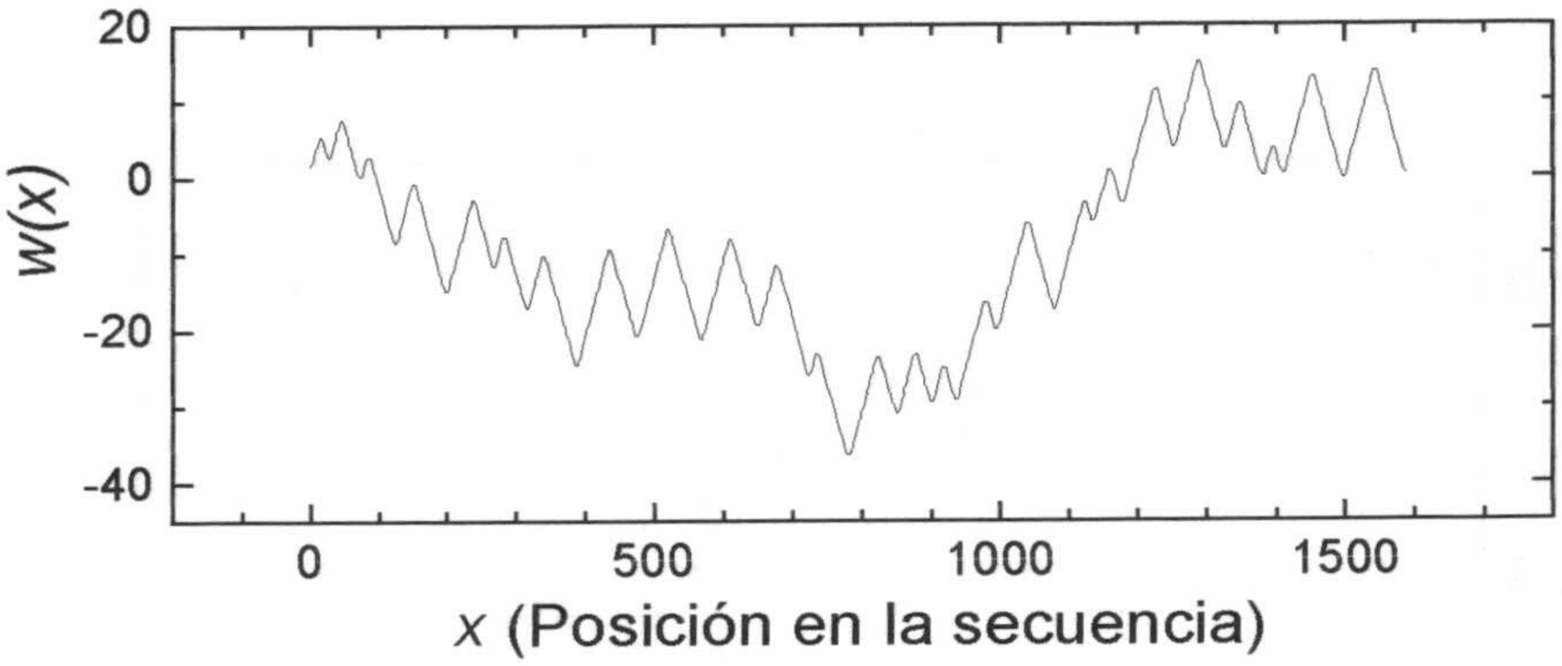

Figura 6.5: Random walk de una secuencia construida con tramos con 1/3 de unos alternados con tramos con 2/3 de unos. Las longitudes de los tramos van desde 12 a 60.

En la figura 6.3 vemos el random walk de una secuencia binaria que ha sido construida uniendo tramos de longitud 50, pero al contrario que en los ejemplos anteriores, ahora las diferencias de composición entre tramos adyacentes no es siempre la misma (en la figura esto se refleja en una diferente agudeza de los picos del walk). Al calcular el perfil de complejidad de esta secuencia (figura 6.4) en lugar de aparecer un escalón abrupto como teníamos antes, el perfil crece de forma más o menos suave a medida que disminuye el nivel de significación, ya que no todas las diferencias entre dominios son igualmente significativas.

En el siguiente ejemplo (figura 6.5) construimos la secuencia uniendo tramos de diferente longitud (desde 12 hasta 60 símbolos) pero ahora todos los tramos

adyacentes tienen una misma diferencia de composición, se alternan tramos con 1/3 de unos con tramos con 2/3 de unos. De nuevo el perfil de esta secuencia (figura 6.6), en lugar de presentar un escalón abrupto presenta una subida progresiva, ya que ahora, a pesar de tener una misma diferencia de composición, las longitudes de los dominios son diferentes.

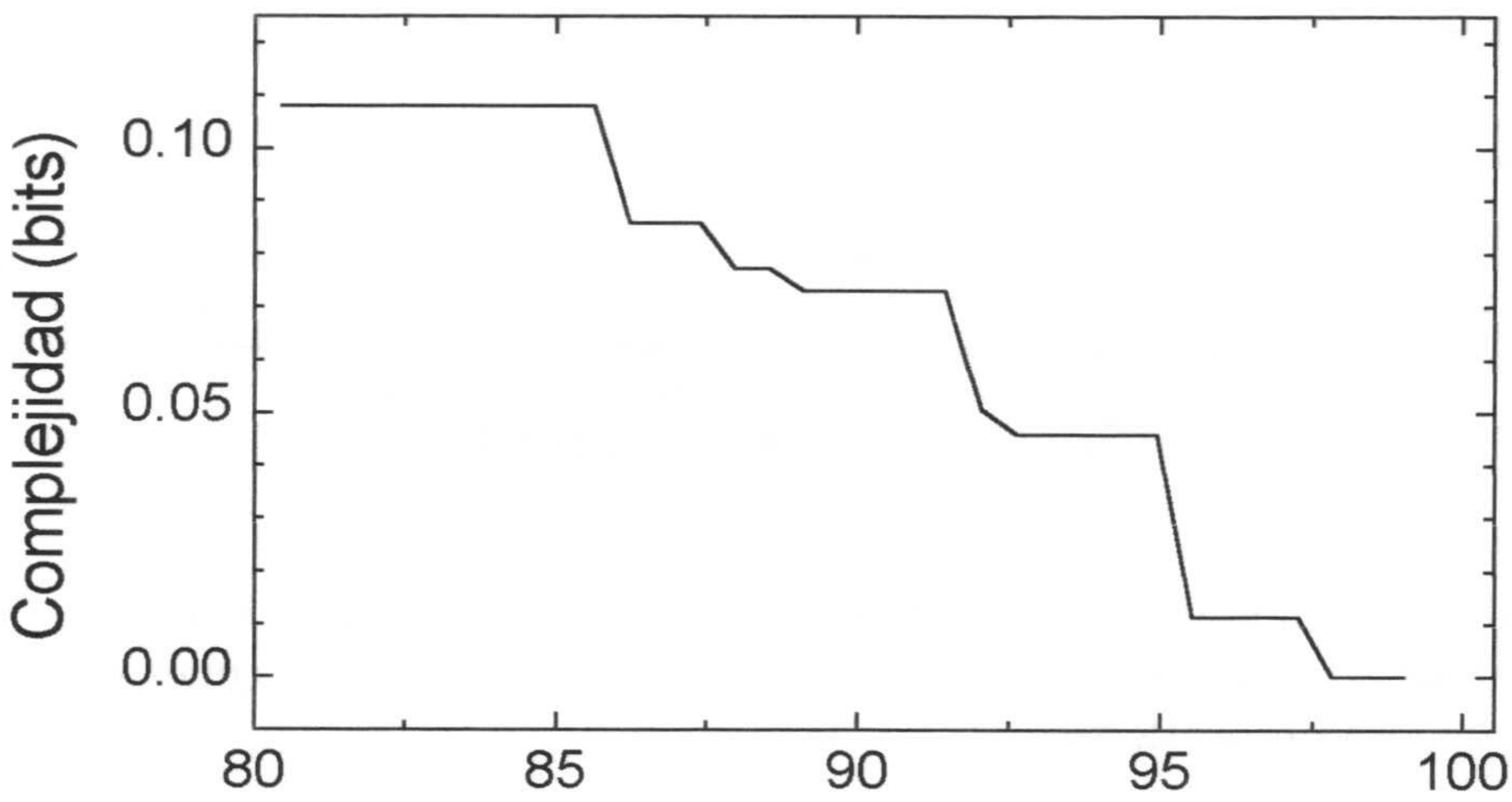

Figura 6.6: Perfil de complejidad de una secuencia construida con tramos con 1/3 de unos alternados con tramos con 2/3 de unos. Las longitudes de los tramos van desde 12 a 60.

6.3.3 Dominios dentro de dominios

Otra cuestión para la que resulta interesante el análisis a distintos niveles de resolución es la cuestión de la homogeneidad relativa de los dominios. Los dominios

resultantes del proceso de segmentación son homogéneos hasta el nivel de significación considerado†, pero puede que a resoluciones inferiores aparezca una estructura que quedaba oculta (o no era considerada suficientemente significativa), por así decirlo, los detalles que pasan inadvertidos a simple vista, se aprecian al mirarlos con una lupa. Además, como se vio al final del capítulo anterior, la forma en que los dominios, que a un nivel de resolución se dieron como homogéneos, se subdividen al reducir el nivel parece ser una diferencia importante entre las secuencias consideradas en la literatura como complejas y no complejas; heterogeneidad simple y compleja como lo define W. Li ([Li 97b]).

Consideremos el siguiente ejemplo. Se ha generado una secuencia binaria en la que hay tramos de longitud 10 con distinta composición: 3,4,6 y 7 unos respectivamente y el resto de ceros; y estos tramos se colocan alternando primero 15 tramos con 6 unos y 15 tramos con 3 unos, a continuación 15 tramos con 7 unos y 15 con 4 unos y así sucesivamente. De esta forma la secuencia, además de tener la estructura debida a los tramos de longitud 10 tiene zonas de longitud 300 con composición definida. El random walk de esta secuencia se ha representado en la figura 6.7 y, como se puede apreciar, a simple vista son mucho más relevantes estos tramos que los de longitud 10. En la figura 6.8 tenemos el perfil de complejidad correspondiente a esta secuencia. El primer escalón que aparece se debe a la presencia de los dominios de tamaño 300, que son detectados a un nivel de significación relativamente alto (en torno al 90%), a partir de aquí el perfil continúa horizontal hasta que, para un nivel de significación en torno al 75% se detectan súbitamente todos los dominios de tamaño 10. Como comparación se ha incluido en esta gráfica el perfil correspondiente a una secuencia generada con los mismos tramos de longitud

†Téngase en cuenta que aquí el concepto de homogeneidad es global, en el sentido de que el dominio de forma aislada podría continuar dividiéndose, pero si el algoritmo no lo divide es porque los subdominios resultantes pueden no tener composición significativamente diferente de los que flanqueaban al dominio original.

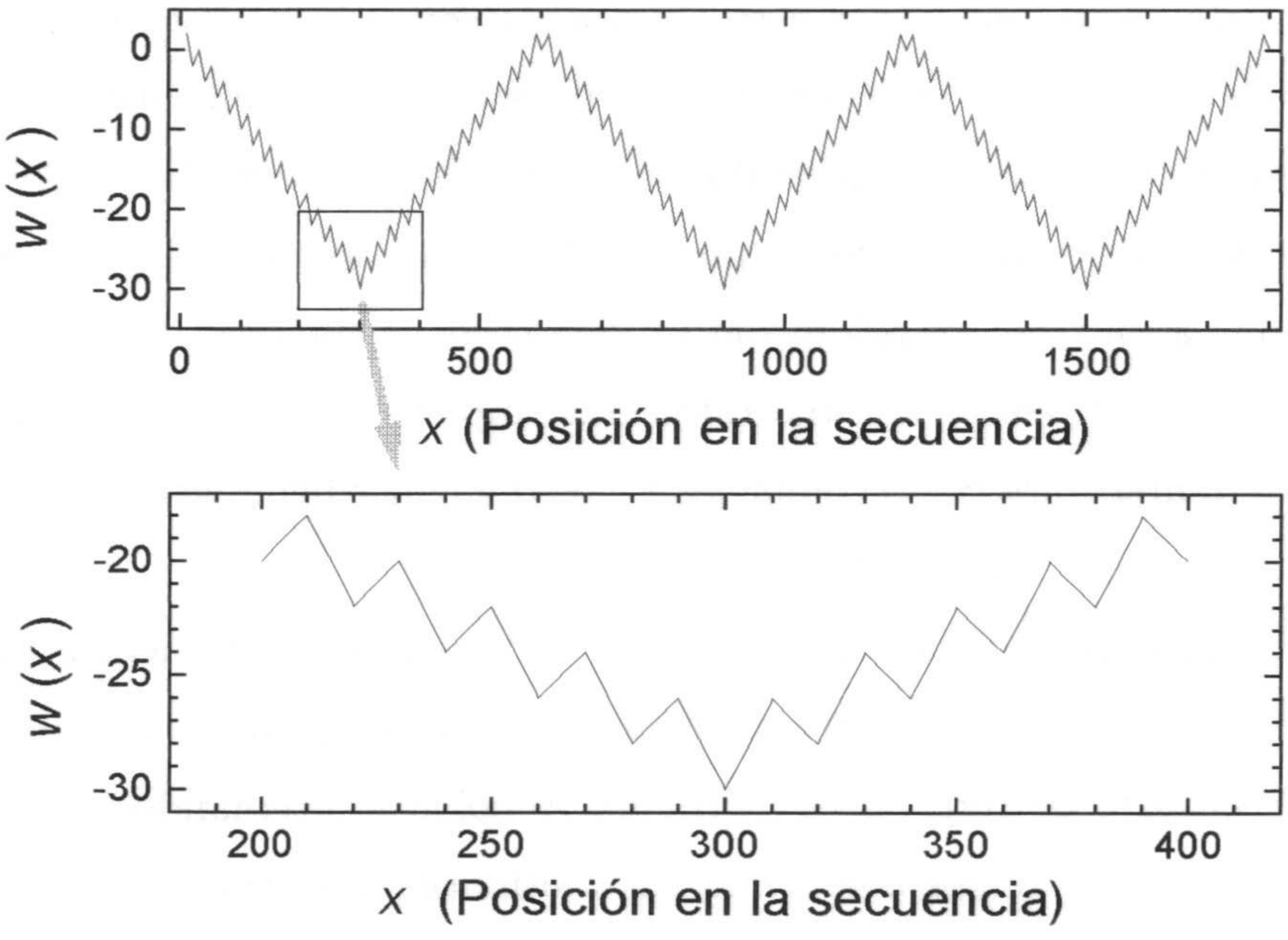

Figura 6.7: Random walk de una secuencia con tramos de longitud 10 organizados en dominios de mayor tamaño.

10 pero colocados al azar. Ahora el perfil alcanza valores mucho menores y además la subida se produce de forma más suave. Esto último es debido a que al estar distribuidos al azar las diferencias entre tramos consecutivos no tiene por qué ser la misma, además pueden aparecer varios tramos de igual composición seguidos, con lo cual no existirían tales tramos.

6.3.4 Influencia de las fluctuaciones

En los ejemplos del apartado anterior las secuencias que hemos utilizado habían sido generadas de forma determinista, esto es, todos los tramos eran iguales y tenían una

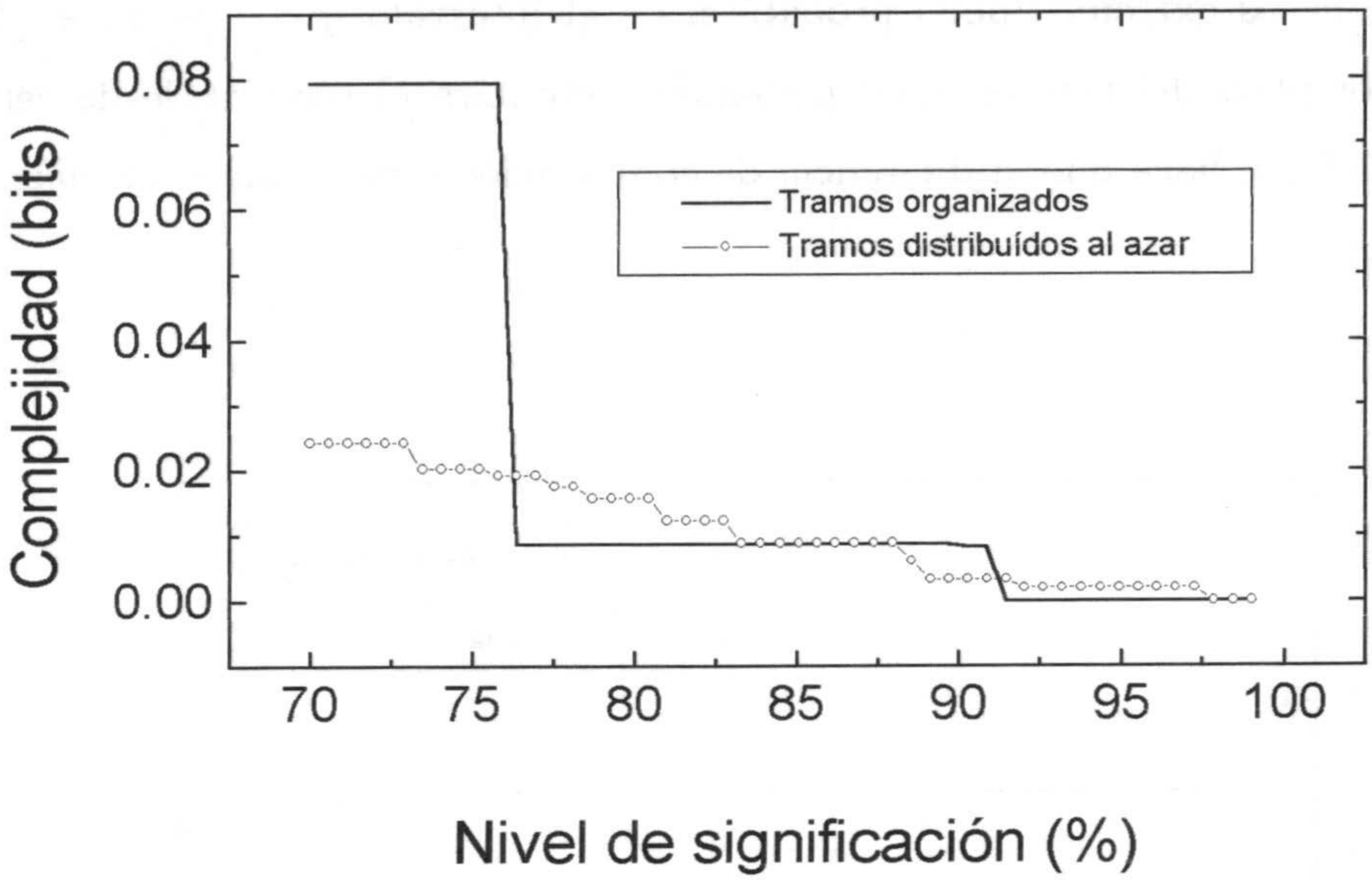

Figura 6.8: Perfil de complejidad de una secuencia con tramos de longitud 10 organizados en dominios de mayor tamaño y de la secuencia en la que los mismos tramos se han dispuesto al azar.

composición perfectamente determinada. Veamos ahora qué ocurre si construimos secuencias con tramos generados con distribuciones de probabilidad determinadas pero cuya composición final está sujeta a fluctuaciones estadísticas. En términos generales podemos decir que se presentan dos efectos: en primer lugar los tramos ya no son estrictamente homogéneos ya que, a causa de las fluctuaciones estadísticas pueden aparecer, dominios dentro de ellos y, por otra parte, la composición de los tramos no va a ser exactamente la correspondiente a las probabilidades asignadas a cada símbolo. Si llamamos p_0 y p_1 a las probabilidades con que la fuente genera

1s y 0s respectivamente, una realización concreta tendrá un porcentaje de unos que variará, salvo casos extremos poco probables, en el intervalo $[p_1 - \frac{1}{\sqrt{L}}, p_1 + \frac{1}{\sqrt{L}}]$, siendo L la longitud del tramo; igual podemos decir para el porcentaje de ceros. Este segundo efecto hace que la diferencia de composición entre tramos adyacentes no sea la esperada.

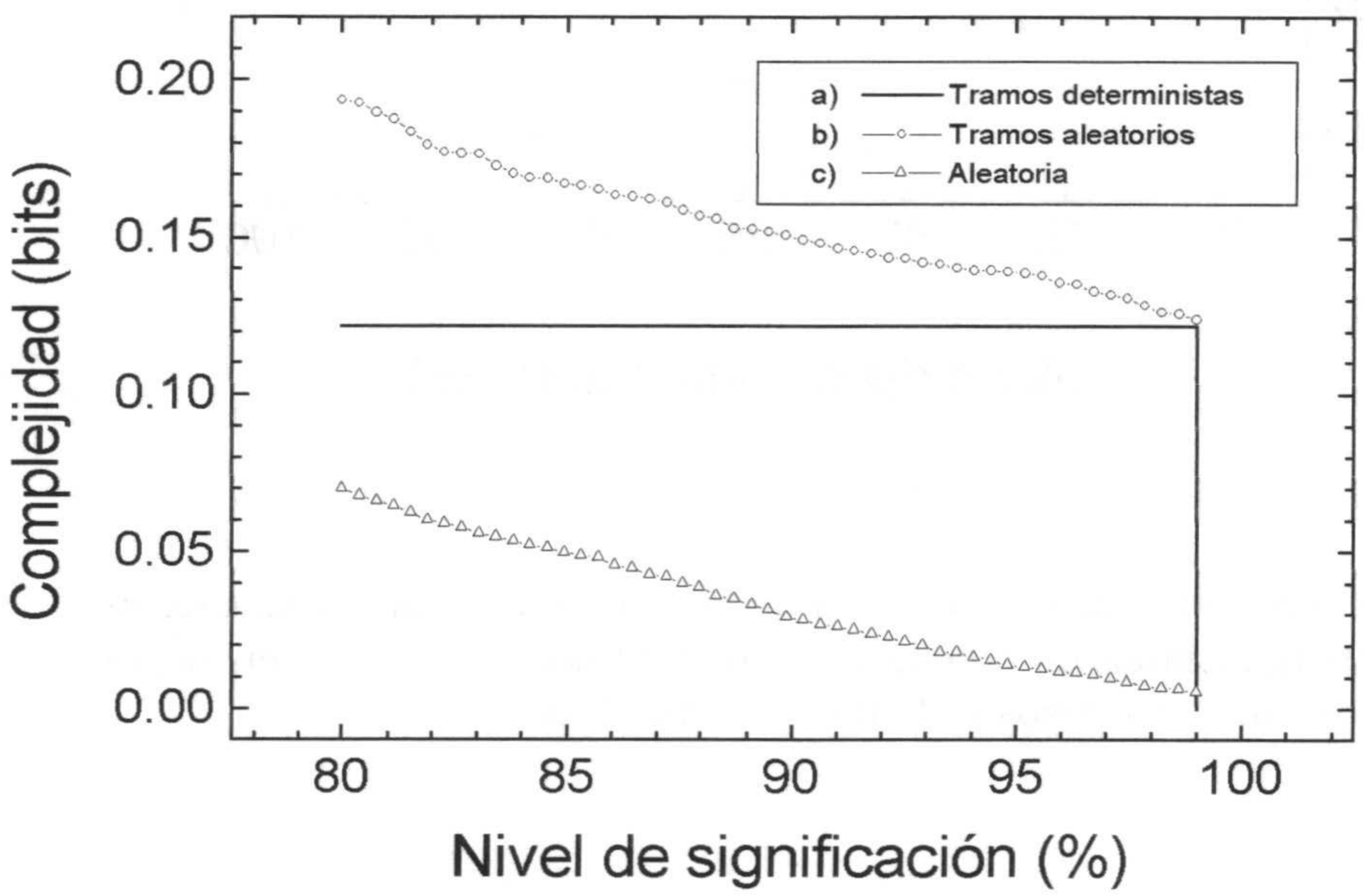

Figura 6.9: Perfiles de complejidad de: a) una secuencia con de 64 tramos deterministas de 400 símbolos con composiciones composiciones alternadas (280 unos y 120 ceros y viceversa); b) secuencia con los mismos tramos pero generados al azar con probabilidades las probabilidades correspondientes (0.7 para unos y 0.3 para ceros y viceversa); c) promedio de 100 secuencias aleatorias de 400 símbolos con las mismas probabilidades.

Para ver cómo afecta esta situación a los perfiles se ha generado, en primer

lugar, una secuencia binaria con 64 tramos de longitud 400, de forma que un tramo tiene 280 unos y 120 ceros y el adyacente 120 unos y 280 ceros, de forma determinista; esto es, cada tramo tiene *exactamente* la composición complementaria a los adyacentes. Por otra parte, se ha generado una secuencia, también con 64 tramos de 400 símbolos pero en este caso los tramos se han obtenido a partir de un generador pseudoaleatorio asignando las probabilidades $p_1 = 0.7 = \frac{280}{400}$ y $p_0 = 0.3 = \frac{120}{400}$, y las complementarias. Como comparación también se han generado secuencias pseudoaleatorias de 400 símbolos con la misma composición de los tramos.

En la figura 6.9 vemos los perfiles de estas secuencias. Para los tramos deterministas tenemos, como era de esperar, un escalón (que en este caso aparece a una resolución muy alta, ya que los dominios están muy bien definidos) y a partir de aquí el perfil es plano, reflejando la homogeneidad interna de los dominios. Para los tramos generados al azar vemos que también aparece el escalón (los tramos siguen apareciendo), pero el perfil no se mantiene plano a causa de la aparición de dominios dentro de los tramos. En este caso, dado que los tramos tienen un tamaño considerable y unas diferencias de composición grandes, las fluctuaciones no modifican la posición en la que se sitúa el escalón (realmente se produce a una significación tan alta que la diferencia prácticamente imperceptible). Nótese también cómo la subida del perfil por encima del de la secuencia con tramos deterministas es similar al perfil de la secuencia aleatoria[‡].

En la figura 6.10 se ha representado el número de dominios obtenidos para estas mismas secuencias en función del nivel de resolución (en el caso de las aleatorias, el número de dominios corresponde a las 64 secuencias). Nótese como ahora la diferencia es mucho mayor, el número de dominios crece tremendamente a causa de las fluctuaciones estadísticas. Mientras que en la secuencia de tramos deterministas

[‡]Para suavizar el aspecto del perfil, éste se ha promediado sobre 64 secuencias pseudoaleatorias.

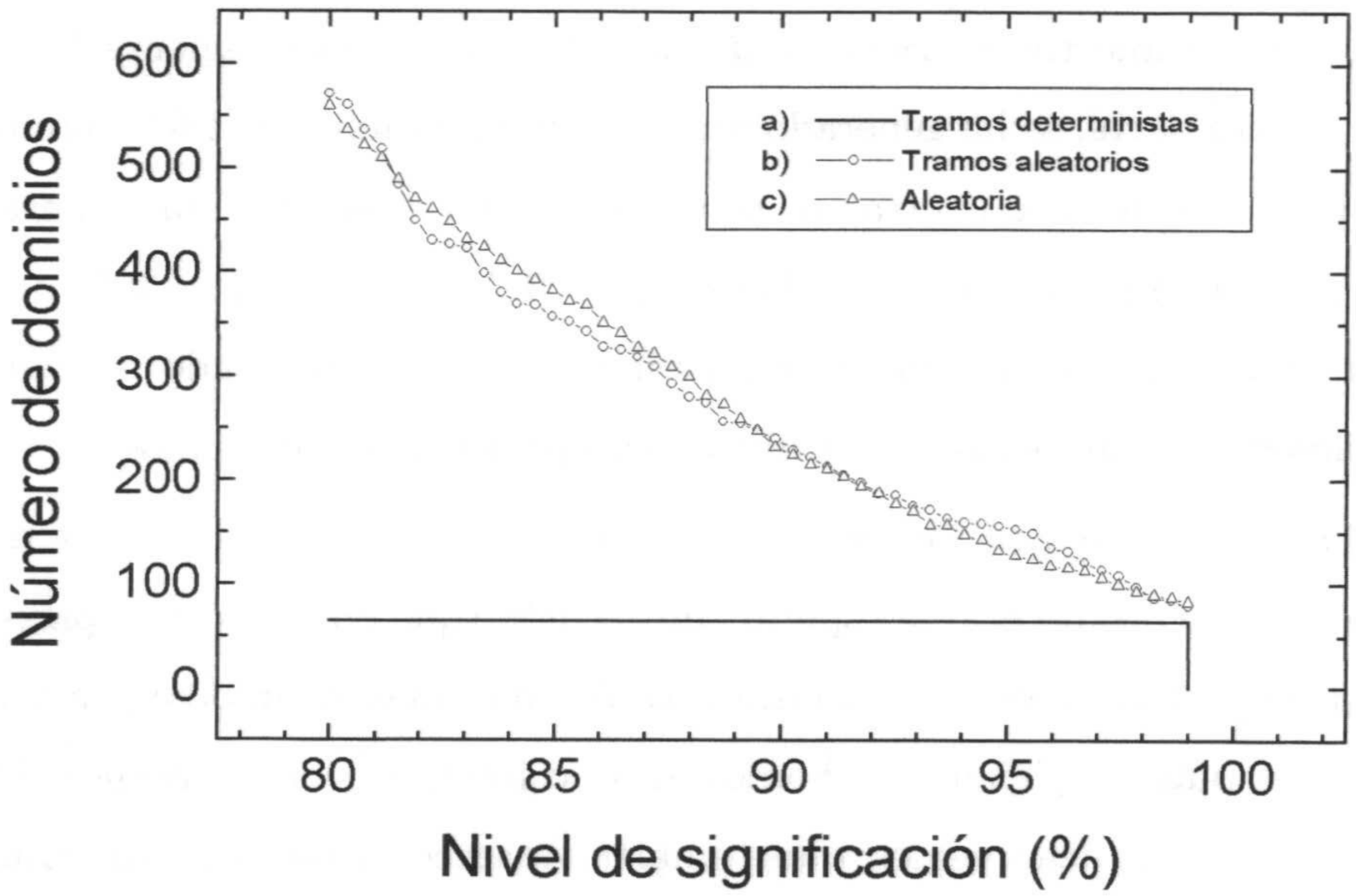

Figura 6.10: Número de dominios para las mismas secuencias de la figura anterior.

sólo hay 64, en la secuencia de tramos aleatorios llega hasta 600 para los niveles de resolución más bajos. Es importante destacar cómo esta gran diferencia en el número de dominios queda amortiguada cuando representamos el perfil: en la aleatoria hay muchos más dominios pero éstos aparecen a niveles de resolución más bajos por lo que su contribución a la divergencia total de la segmentación es menor. También es importante destacar que el "perfil" del número de dominios no diferencia entre la secuencia aleatoria y la de tramos aleatorios: realmente, una vez segmentadas, las dos tienen el mismo numero de dominios, sin tener en cuenta de ninguna forma los dominios que en una fueron encontrados a resoluciones superiores y en la otra no.

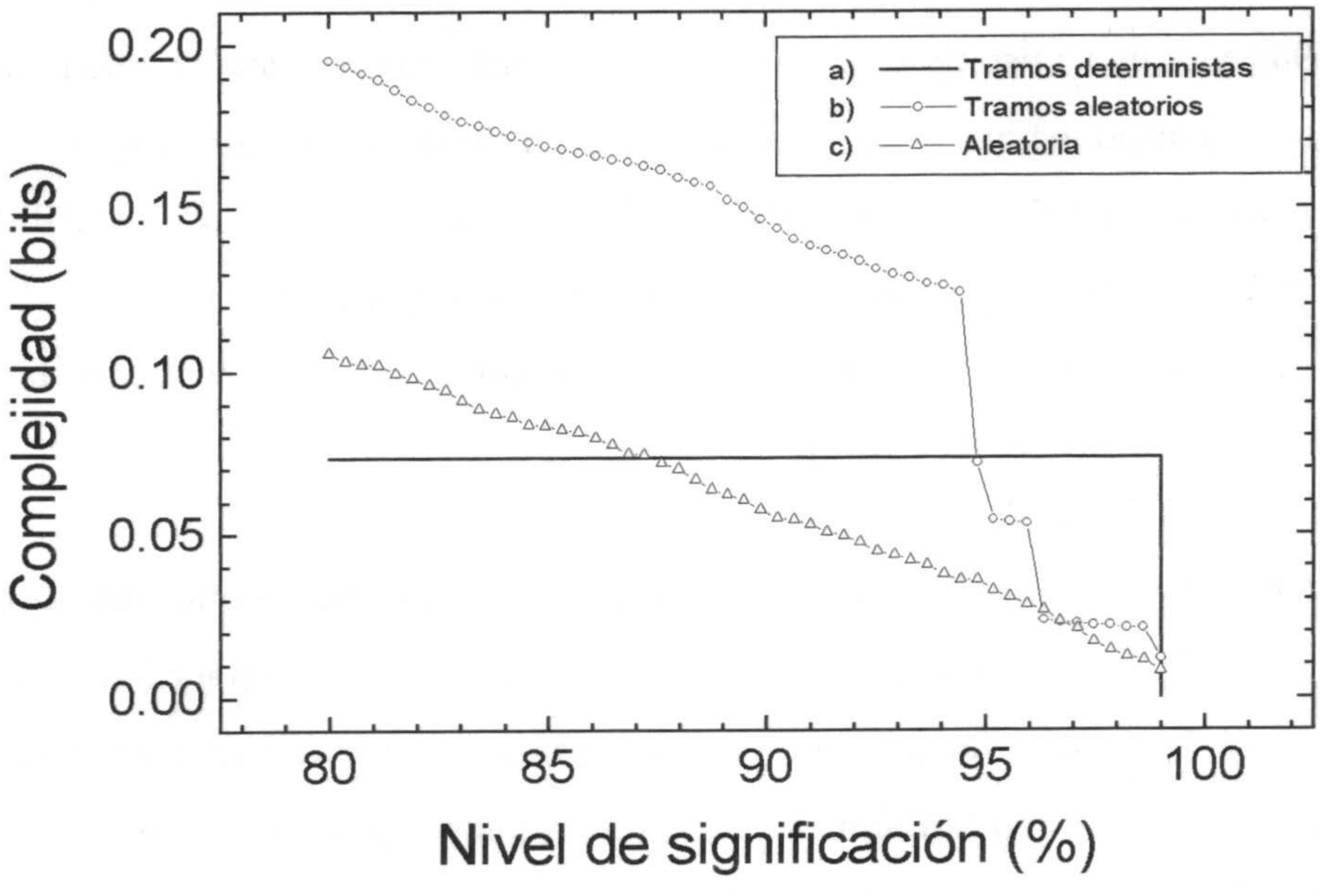

Figura 6.11:

En la figura 6.11 vemos los perfiles de secuencias similares a las del ejemplo anterior pero ahora la longitud de los tramos es de 120 símbolos (y la diferencia de composición algo menor 0.65 y 0.35). De nuevo aparecen en el perfil zonas inclinadas donde debía aparecer un tramo plano, a causa de la existencia de dominios dentro de los tramos. En este caso, además, se puede ver cómo el escalón se encuentra situado algo más a la derecha y no está tan definido como consecuencia de las fluctuaciones en la composición de los tramos que ya no son exactamente las complementarias en tramos adyacentes.

Estos resultados pueden ser útiles para descontar de un perfil los posibles efectos del "ruido" aleatorio. De hecho técnicas para la eliminación del ruido blanco,

muy utilizadas en teoría de la señal, han sido propuestas para el estudio de secuencias de ADN con espectros de potencia [Vo 92], técnicas que han provocado controversia [Bu 93c], [Vo 93], ya que su uso no está del todo justificado en este caso ya que no hay ningún motivo para suponer que tenga que aparecer ruido en las secuencias de ADN. Sin hacer suposiciones a priori sobre la existencia o no de este "ruido", la utilidad de estos resultados al menos está en la posibilidad de comparar la heterogeneidad presente en una secuencia con la que tendría una generada al azar y juzgar hasta qué punto se puede considerar relevante.

Por último veamos qué ocurre al segmentar secuencias pseudo-aleatorias hasta niveles de significación bajos. Como ya se comentó antes, independientemente de la estructura de la secuencia, para niveles de significación suficientemente bajos se debe alcanzar una complejidad similar para todas las secuencias con una misma composición global. Vimos que el valor máximo de la complejidad viene dado por la entropía Shannon del histograma de la secuencia completa, aunque este valor no se alcanza nunca con nuestro algoritmo de segmentación que, como mínimo, necesita una secuencia de tres símbolos para actuar. Veamos una estimación de este valor máximo efectivo que haremos, por simplicidad, para secuencias binarias.

Consideremos una secuencia binaria de símbolos independientes e idénticamente distribuidos (i.i.d.), y sean p_0 y p_1 las probabilidades de aparición de ceros y unos respectivamente. Si suponemos que al disminuir hasta cero el nivel de significación el algoritmo divide la secuencia en dominios de longitud 3 tendremos las siguientes posibilidades: $\{0,0,0\}$ con probabilidad p_0^3, $\{1,0,0\}$ y sus permutaciones con probabilidad $3p_0^2p_1$, $\{1,1,0\}$ y sus permutaciones con probabilidad $3p_0p_1^2$ y $\{1,1,1\}$ con probabilidad p_1^3. Por tanto, en valor medio la divergencia total de la segmentación vendrá dada por:

$$\begin{aligned} \mathrm{JS}_m &= \mathrm{H}(p_0, p_1) - 3p_0^2 p_1 \mathrm{H}\left(\frac{2}{3}, \frac{1}{3}\right) - 3p_0 p_1^2 \mathrm{H}\left(\frac{1}{3}, \frac{2}{3}\right) = \\ &= \mathrm{H}(p_0, p_1) - 3p_0 p_1 \mathrm{H}\left(\frac{2}{3}, \frac{1}{3}\right) \end{aligned} \tag{6.3}$$

En la figura 6.12 se han representado los perfiles de complejidad de algunas secuencias pseudo-aleatorias obtenidas con tres generadores aceptables. Las probabilidades se han tomado $p_0 = p_1 = \frac{1}{2}$, con lo que 6.3 nos da un valor máximo de 0.311 bits, bastante próximo a lo que se obtiene en la gráfica.

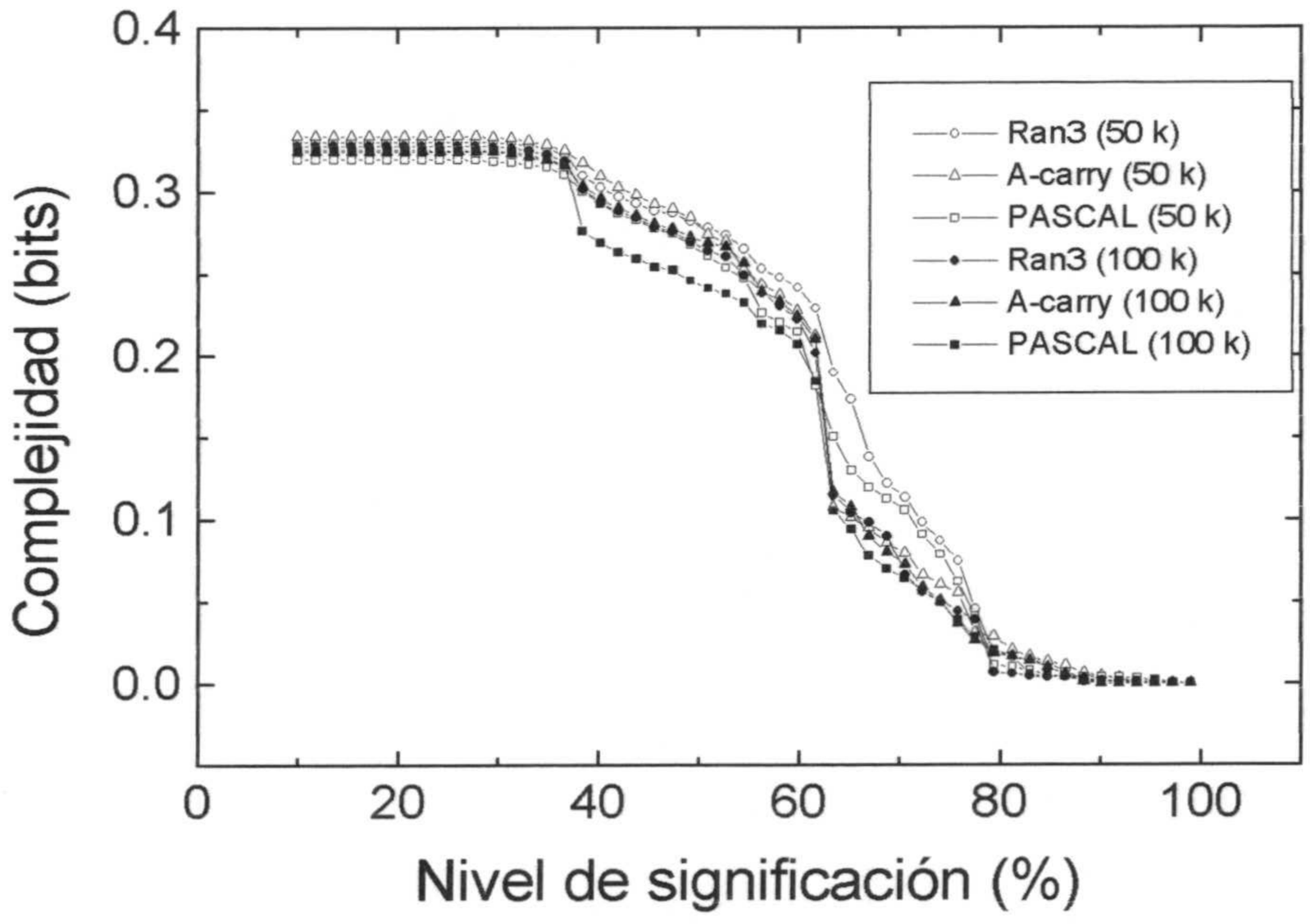

Figura 6.12: Perfiles de complejidad hasta niveles de significación bajos para tres generadores pseudo-aleatorios aceptables: Ran3 [Pr 94], A-carry [Ja 90] y el generador interno del Borland PASCAL 7.0.

Otra cuestión interesante que se puede apreciar en esta gráfica es que, para todos los generadores se produce una subida considerable en el perfil a partir de, más o menos, el 80% de significación. Según esto, la complejidad que se detecta a partir de este nivel de significación no parece ser muy relevante ya que aparece incluso en estas secuencias pseudo-aleatorias; y es por este motivo que en lo sucesivo todos los perfiles de secuencias naturales que veremos en el capítulo siguiente se limitarán al rango 80-100%.

Capítulo 7

Perfiles de complejidad para secuencias de ADN

7.1 Introducción

En este capítulo aplicaremos la medida de la complejidad composicional propuesta en el capítulo anterior a secuencias de ADN. En primer lugar veremos cómo nuestra medida de complejidad discrimina entre secuencias con y sin correlaciones de largo alcance (sección 7.2), y entre secuencias codificadoras y no codificadoras (sección 7.3), comparando en ambos casos con los resultados obtenidos por otros métodos. En la sección 7.4 estudiaremos la relación entre la complejidad composicional y el grado de complejidad del organismo al que pertenece la secuencia, también comparando con los resultados previos que relacionan la complejidad del organismo con lo que ellos llaman "complejidad fractal". Para terminar, en la sección 7.5 veremos los perfiles de complejidad de secuencias generadas con algunos de los modelos propuestos en la bibliografía y comprobaremos hasta qué punto dan cuenta de la complejidad

composicional de las secuencias naturales.

7.2 Correlaciones de largo alcance

Al igual que se hizo en el capítulo 5, aquí nos centraremos en los resultados obtenidos codificando la secuencia con el alfabeto binario {R,Y}. Tan sólo veamos antes un ejemplo del aspecto que tienen los perfiles de las secuencias HUMTCRAD y ECO110K que, como ya se ha dicho, son los mejores ejemplos de secuencias con y sin correlaciones de largo alcance.

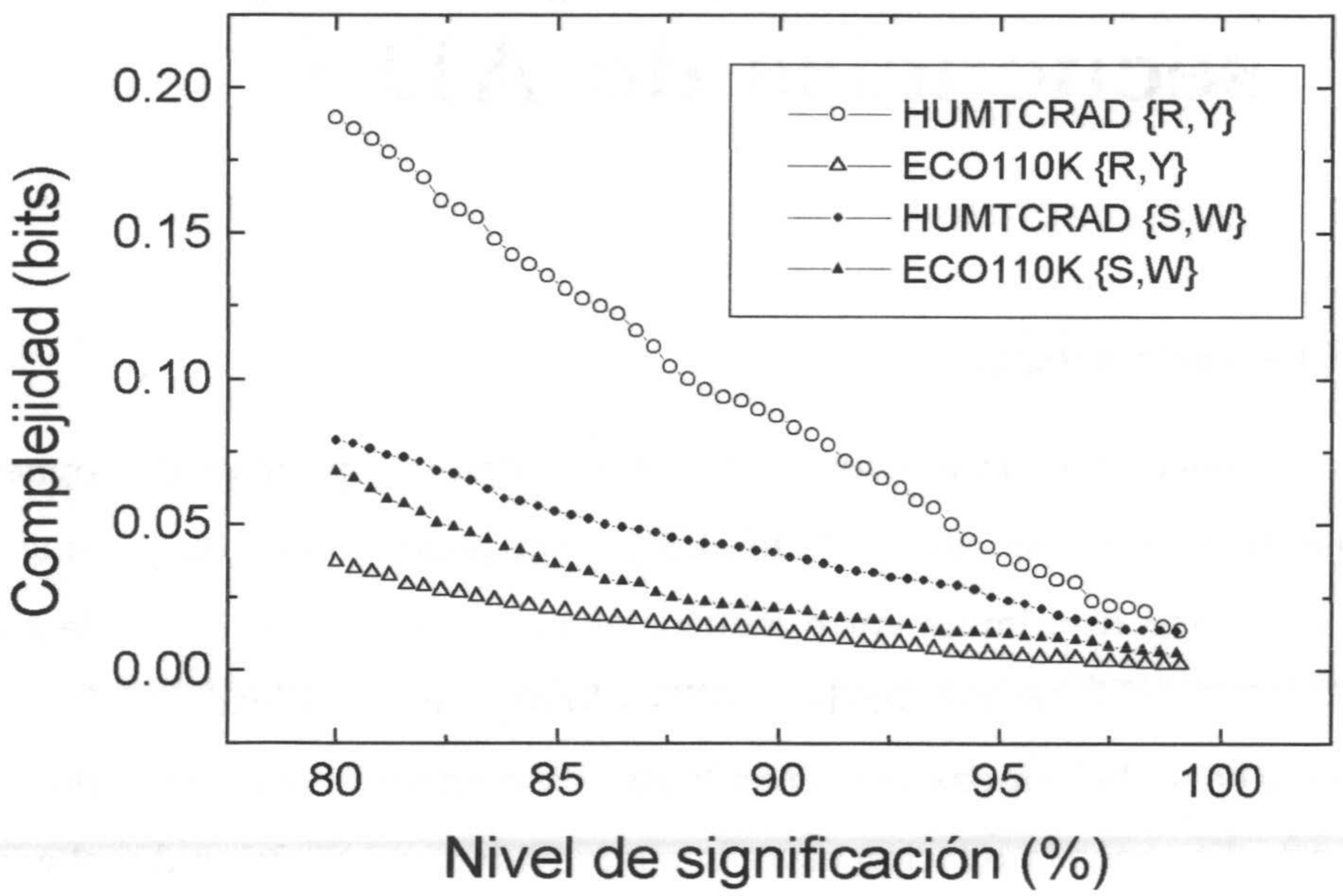

Figura 7.1: Perfiles de complejidad de HUMTCRAD y ECO110K con alfabetos binarios {R,Y} y {S,W}.

En la figura 7.1 vemos los perfiles de complejidad de ambas secuencias para

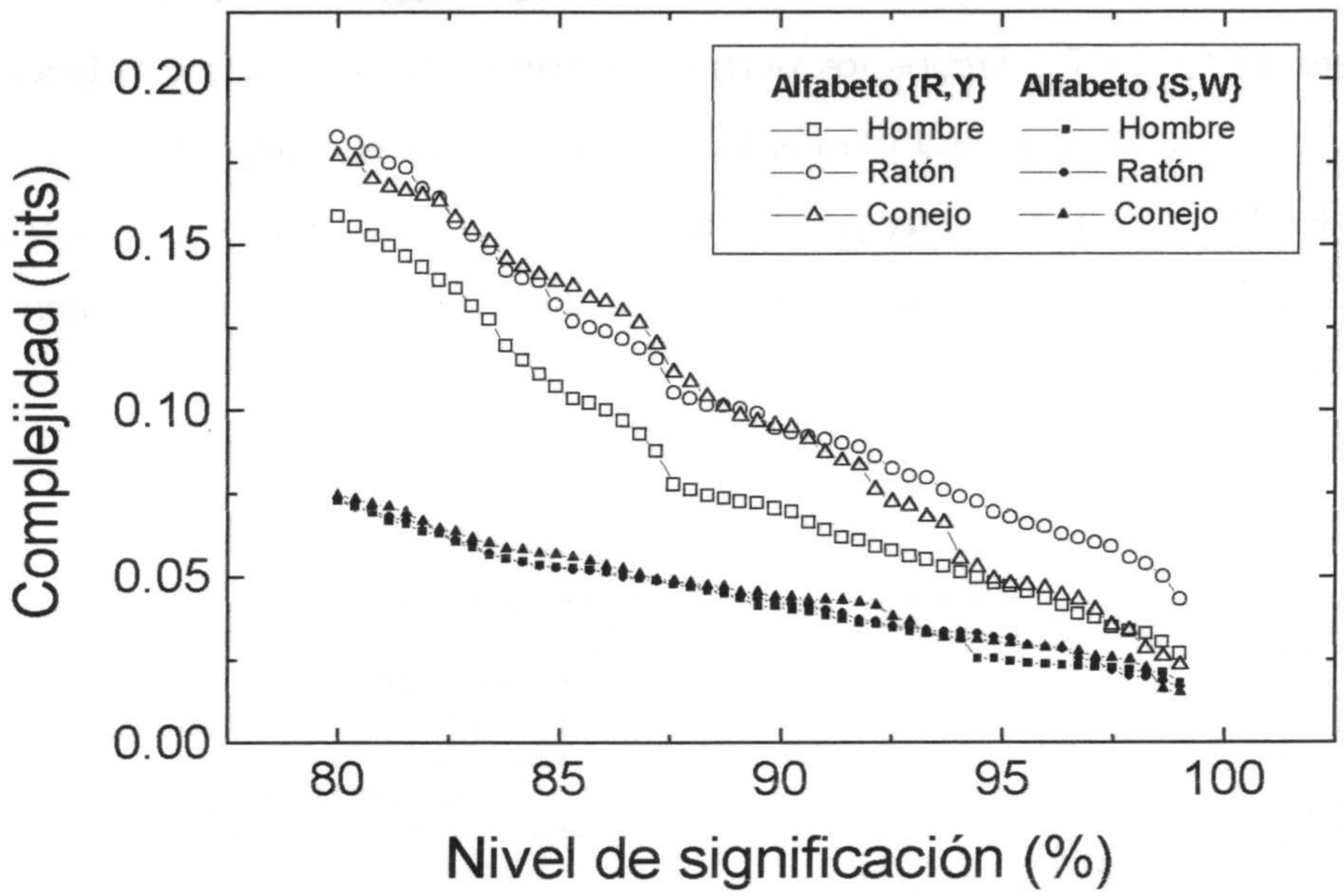

Figura 7.2: Perfiles de complejidad con alfabeto {R,Y} y {S,W} para tres secuencias que contienen la zona codificadora de la betaglobina.

las segmentaciones con los dos alfabetos binarios. Al igual que ocurría con el número de dominios, en este caso tenemos una diferencia de complejidad mucho mayor al usar el alfabeto {R,Y}: el perfil de la secuencia bacteriana apenas se levanta significativamente de cero, mientras que la humana alcanza valores de complejidad realmente altos. Por el contrario, con el alfabeto {S,W} los perfiles están mucho más próximos. Otra cuestión interesante es el hecho de que el perfil con {S,W} es más bajo que el {R,Y} para la secuencia humana, mientras que en la bacteriana ocurre al contrario.

En términos generales los perfiles {S,W} son inferiores para prácticamente

todas las secuencias naturales analizadas, siendo las únicas excepciones observadas algunas pertenecientes a *E. coli* y al bacteriófago lambda (parásito de esta bacteria). Por ejemplo, en la figura 7.2 vemos los perfiles de tres secuencias que contienen la zona codificadora de la betaglobina para el hombre, el ratón y el conejo. En todos los casos los perfiles {S,W} quedan muy por debajo de los {R,Y}, además curiosamente en este caso los primeros son prácticamente iguales para las tres secuencias mientras que los otros no.

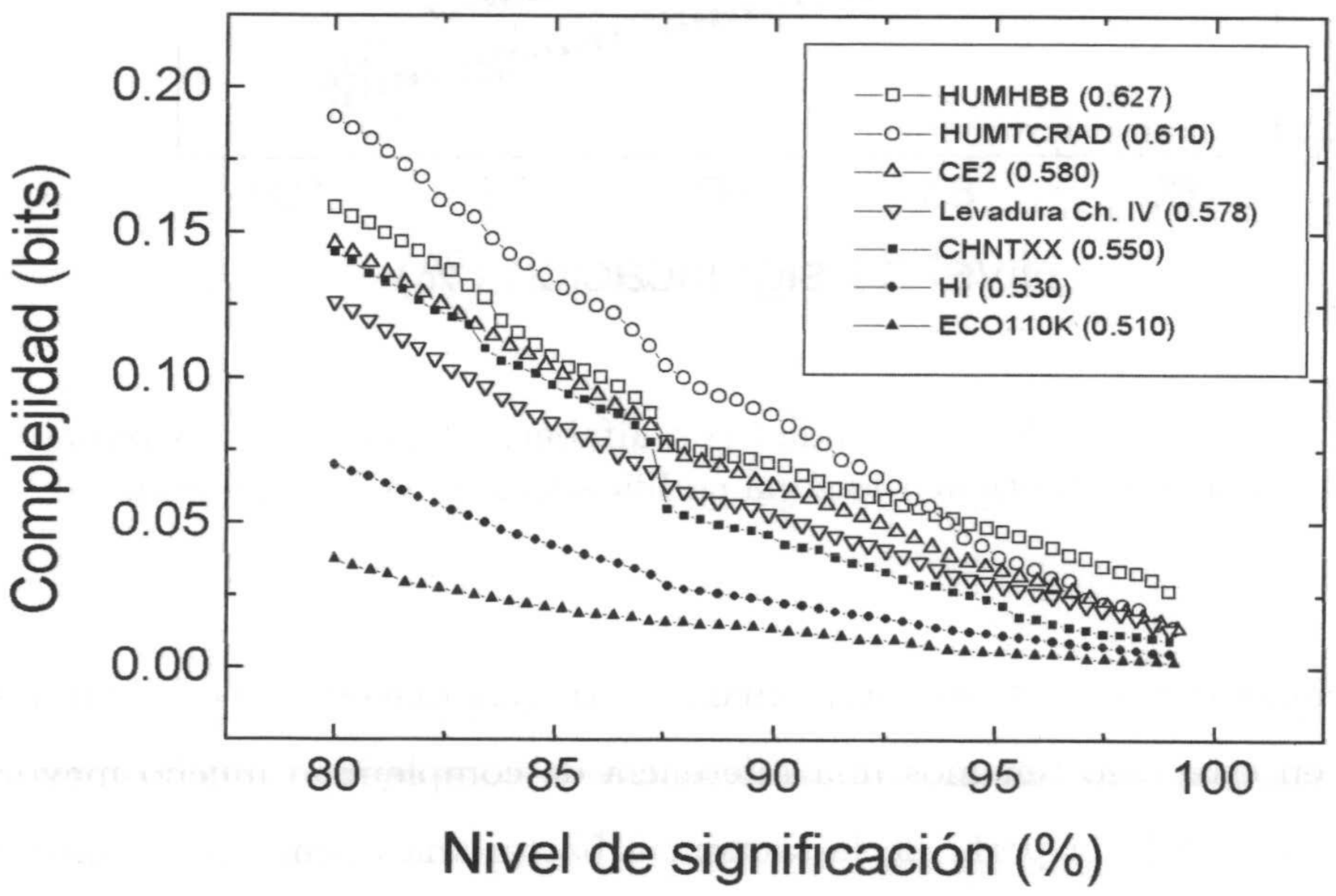

Figura 7.3: Pefiles de complejidad para secuencias naturales con distintos exponentes de escala.

El primer ejemplo pone de manifiesto una diferencia muy considerable entre los perfiles de la secuencia con correlaciones de largo alcance y el de la que no

las tiene. En la figura 7.3 se presentan los perfiles de algunas secuencias más con distintos exponentes DFA, como se puede ver no existe una correlación clara entre la altura del perfil y el exponente de escala, aunque preferentemente las secuencias con exponentes mayores aparecen más arriba en la gráfica. De hecho, como veremos en apartados siguientes, los perfiles no miden exactamente lo mismo que el DFA, es más distinguen entre secuencias con un mismo exponente pero con una estructura de dominios diferente. Algo que sí se ha observado es que la correlación entre el nivel de complejidad, tal y como lo miden nuestros perfiles, y el exponente del DFA es mucho más claro entre secuencias naturales que entre secuencias artificiales generadas con diversos modelos. Esto podría indicar que todas las secuencias naturales, independientemente de su exponente de escala, tienen un tipo "especial" de heterogeneidad que cualquier modelo que reproduce el exponente de escala no es capaz de reproducir.

7.2.1 Perfiles autosemejantes

Una de las características más relevantes de las estructuras fractales es la autosemejanza, o en términos más generales la autoafinidad. En vista de las propiedades fractales encontradas en algunas secuencias de ADN, cabe preguntarse si los perfiles de complejidad pueden reflejar de alguna manera esta semejanza de las partes con el todo. En el capítulo 5 ya vimos un ejemplo (fig. 5.11) en el que el mayor de los dominios obtenidos en la segmentación de HUMTCRAD se segmentaba, al reducir el nivel de significación, de forma similar a como lo hacía la secuencia completa.

En la figura 7.4 vemos los perfiles de complejidad de HUMTCRAD, la subsecuencia antes citada y otros dos dominios obtenidos en la segmentación al 99% de significación. Los dos primeros dominios, salvo para las significaciones más altas en las que casi no se dividen, presentan un perfil de complejidad muy similar al de la secuencia completa. De hecho esta similitud también se refleja en el análisis con

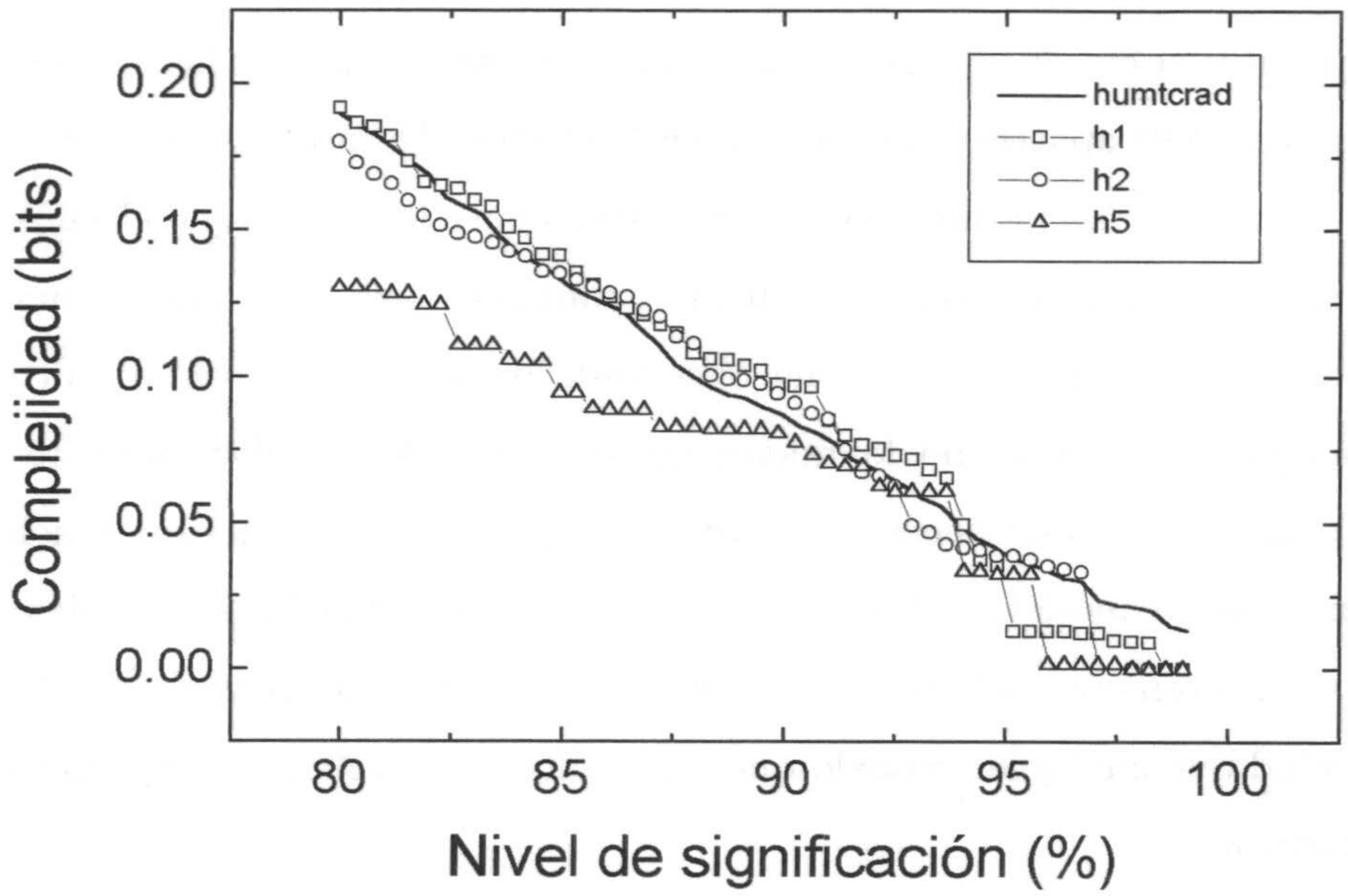

Figura 7.4:

DFA: el primer segmento tiene un exponente $\alpha = 0.594$ y el segundo $\alpha = 0.606$, muy cercanos al exponente que presenta la secuencia completa ($\alpha = 0.61$). Estos perfiles muestran de forma mucho más clara el fenómeno de dominios dentro de dominios que comentábamos en el capítulo 5, allí veíamos que los dominios se subdividían de forma similar a la secuencia completa, pero no teníamos ninguna forma de cuantificar esta similitud; la igualdad en los perfiles nos asegura que la heterogeneidad presente en los dominios es similar a la de la secuencia completa. Téngase en cuenta que este resultado no es trivial, para secuencias con heterogeneidad simple (por ejemplo, formadas por tramos homogéneos de diferente composición) esto no ocurre así, el perfil de la secuencia estará por encima de los perfiles de los dominios. Esta situación la

ilustra bastante bien la gráfica 6.9 del capítulo anterior.

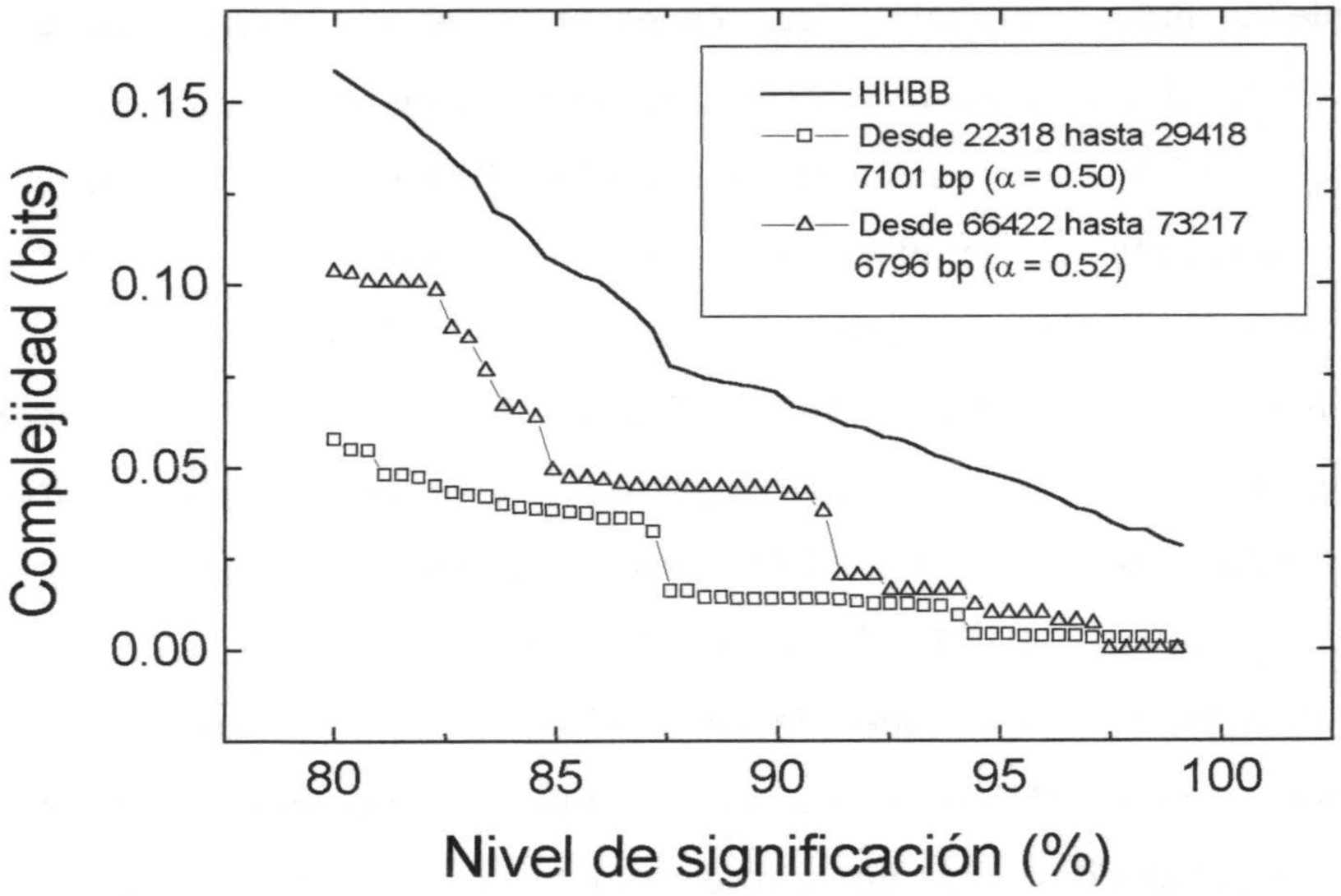

Figura 7.5:

Por otra parte, el perfil del tercer dominio difiere notablemente del resto pero, de nuevo, el resultado concuerda con lo obtenido mediante el análisis con DFA ya que este dominio tiene un exponente de escala de $\alpha = 0.561$, significativamente inferior al del resto de la secuencia. Este resultado no entra en contradicción con la existencia de estructura fractal ya que, en primer lugar, la autosemejanza que cabe esperar en una secuencia natural no sería exacta sino estadística, esto es, las partes tienen un comportamiento estadístico similar al de la secuencia completo, no son copias reducidas *exactas* de la secuencia completa. Además tampoco parece muy probable que exista un exponente de escala uniforme a lo largo de toda la

secuencia dado el gran número de factores que pueden influir en la evolución de éstas y las diferentes funciones (la mayor parte de ellas desconocidas) que pueden llevar a cabo distintas zonas del genoma. Sin ir más lejos, la presencia de zonas repetidas puede dar lugar a dominios homogéneos incluso en secuencias en las que el exponente de escala global indica la presencia de correlaciones de largo alcance. Por ejemplo, en la figura 7.5 tenemos los perfiles de HUMHBB y dos dominios obtenidos en la segmentación al 99% de significación. Estos dos dominios concuerdan bastante bien con dos zonas de ADN repetido descritas en la bibliografía [Ju 89], [Hat 85]: las familias Line-1c (22896-29407) y KnpI (67089-73213). Como se puede ver los perfiles de estos dominios quedan bastante más bajos que los de la secuencia completa. Nótese como también los exponentes DFA de estos dominios son inferiores a los de la secuencia completa ($\alpha = 0.627$), de hecho, el primero presenta $\alpha = 0.5$, lo cual indica total ausencia de correlaciones de largo alcance. De aquí parece bastante razonable suponer que el ADN repetido no introduce heterogeneidades suficientes para provocar los perfiles de complejidad observados en las secuencias eucariotas que hemos visto.

Por último comentar que fue, justamente esta secuencia, la que se utilizó como ejemplo en el artículo de Buldyrev et. al ([Bu 93b]) en el que se propuso el modelo del Lévy-walk generalizado, ya que se trata de una secuencia con un exponente de escala considerablemente alto y que, prácticamente a simple vista, presenta zonas grandes de composición homogénea.

7.3 Diferencias entre zonas codificadoras y no codificadoras.

Vimos en el capítulo 5 que las distribuciones de dominios obtenidas segmentando zona codificadora y no codificadora eran diferentes, tanto cualitativa (abundancia

relativa de dominios en torno a 1000 bp) como cuantitativa (número de dominios por kbase), y ya se comentaron entonces las opiniones de algunos autores justificando la presencia de correlaciones de largo alcance sólo en las codificadoras. Dada la relación existente entre la heterogeneidad composicional y este tipo de correlaciones, los perfiles de complejidad pueden servir para cuantificar estas diferencias.

Para ver esto hemos tomado la secuencia humana HUMTCRAD, el cromosoma III de la levadura de la cerveza y el genoma completo de *E. coli*, y a partir de ellas hemos construido secuencias uniendo, por una parte, toda la zona codificadora y por otra toda la zona no codificadora. Los perfiles de estas secuencias se muestran en la figura 7.6.

Independientemente de la proporción de zonas codificadoras, se observa que en todos los casos el perfil de la zona no codificadora queda por encima del de la secuencia completa y el de la codificadora, por debajo. Según esto parece evidente que la presencia de esta última reduce la complejidad composicional global de la secuencia y que, en gran medida, esta complejidad se debe a la zona no codificadora. Nótese que esto es cierto tanto para los intrones de la secuencia humana, las zonas intergénicas de la levadura y las de la secuencia bacteriana. Hay que tener en cuenta que incluso los valores de complejidad que resultan para las zonas codificadoras pueden haberse incrementado algo al construir una secuencia "pegando" de forma artificial todos los genes, ya que éstos, dependiendo de la dirección de transcripción pueden tener composiciones más o menos alternadas de purinas-pirimidinas causadas por el uso no uniforme de codones. Sin embargo no hay motivo para suponer esta alternancia de composición en la zona no codificadora por lo que, en principio, estos perfiles se verían menos afectados por la manipulación a la que hemos sometido a las secuencias. Este efecto quizá sea más importante para la levadura y *E. coli* que para la secuencia humana.

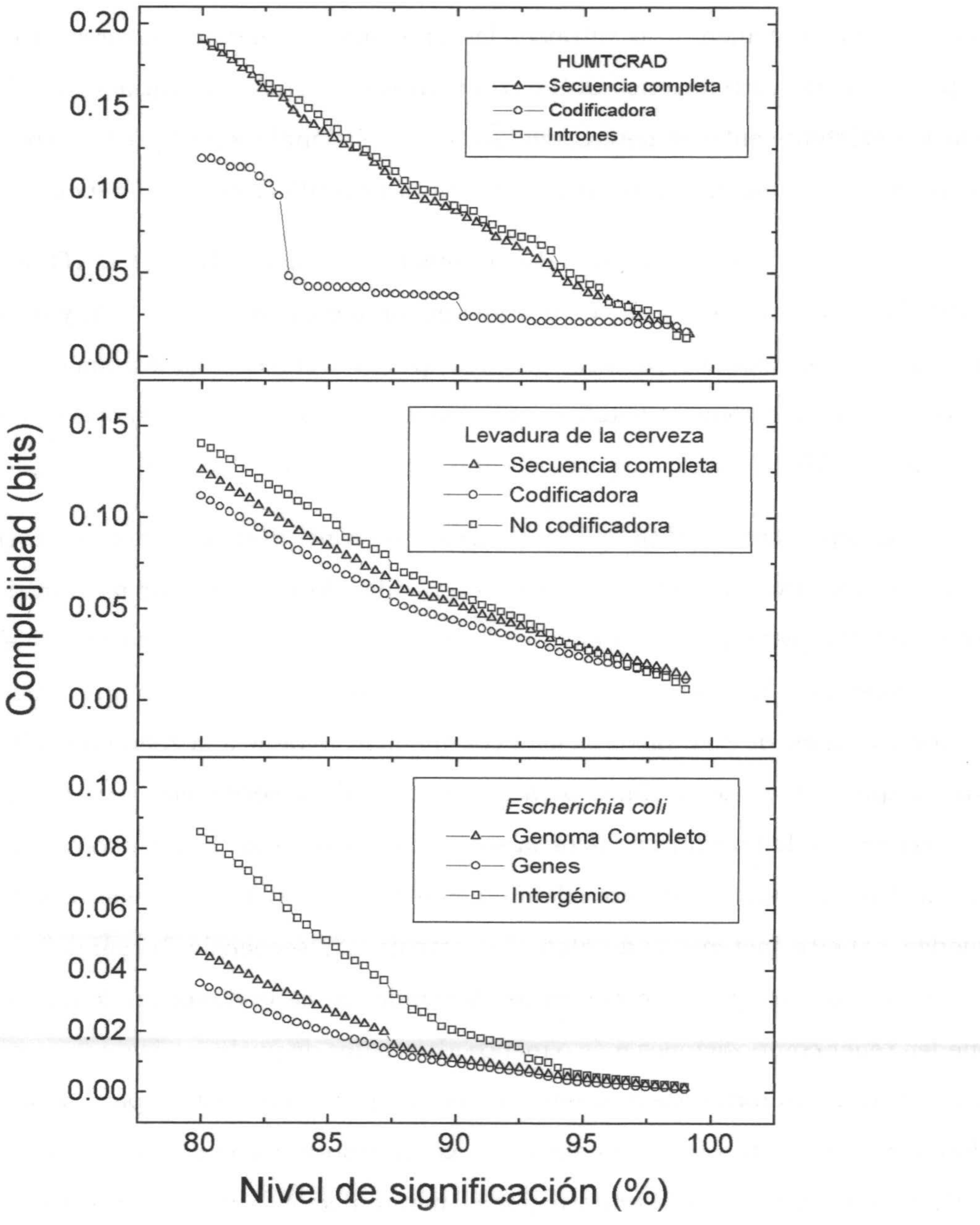
HUMTCRAD
Secuencia completa
Codificadora
Intrones
Levadura de la cerveza
Secuencia completa
Codificadora
No codificadora
Escherichia coli
Genoma Completo
Genes
Intergénico
Complejidad (bits)
0.20
0.15
0.10
0.05
0.00
0.15
0.10
0.05
0.00
0.10
0.08
0.06
0.04
0.02
0.00
80
85
90
95
100
Nivel de significación (%)

Figura 7.6:

Otro experimento que puede servir para poner de manifiesto la contribución a la complejidad total de la secuencia de zonas codificadoras y no codificadoras sería barajarlas internamente y ver como afecta al perfil de la secuencia. De esta forma no se altera la estructura global de la secuencia y además se puede ver la contribución a la complejidad de la alternancia de estas zonas. De hecho se pueden plantear tres experimentos:

a) Barajar internamente los genes y respetar las zonas intergénicas.

b) Barajar internamente las zonas intergénicas y respetar los genes.

c) Barajar internamente ambos.

En la figura 7.7 vemos los perfiles de las secuencias que se obtienen al llevar a cabo estos experimentos con el cromosoma III de la levadura. El perfil más bajo corresponde al experimento c). Esta secuencia sólo conserva (salvo las fluctuaciones estadísticas) la heterogeneidad debida a la alternancia de zonas codificadoras y no codificadoras. Como se puede ver, para todos los niveles de significación, alcanza alrededor de la mitad del perfil de la secuencia completa, por este motivo, aunque supone una contribución importante, no se puede achacar a esta alternancia la heterogeneidad presente en la secuencia como proponen algunos autores [Du 95], [Ce 95], [Az 95]. Aquí es importante destacar que el análisis con DFA apenas muestra diferencias entre el cromosoma III y la secuencia obtenida con el experimento c).

Continuando con los experimentos, el perfil a) es algo superior al c). Esto indica que las heterogeneidades de la zona codificadora también añaden complejidad a la secuencia, aunque menos que la zona no codificadora ya que el perfil correspondiente al experimento b) queda aún más arriba, indicando un mayor aumento de la complejidad.

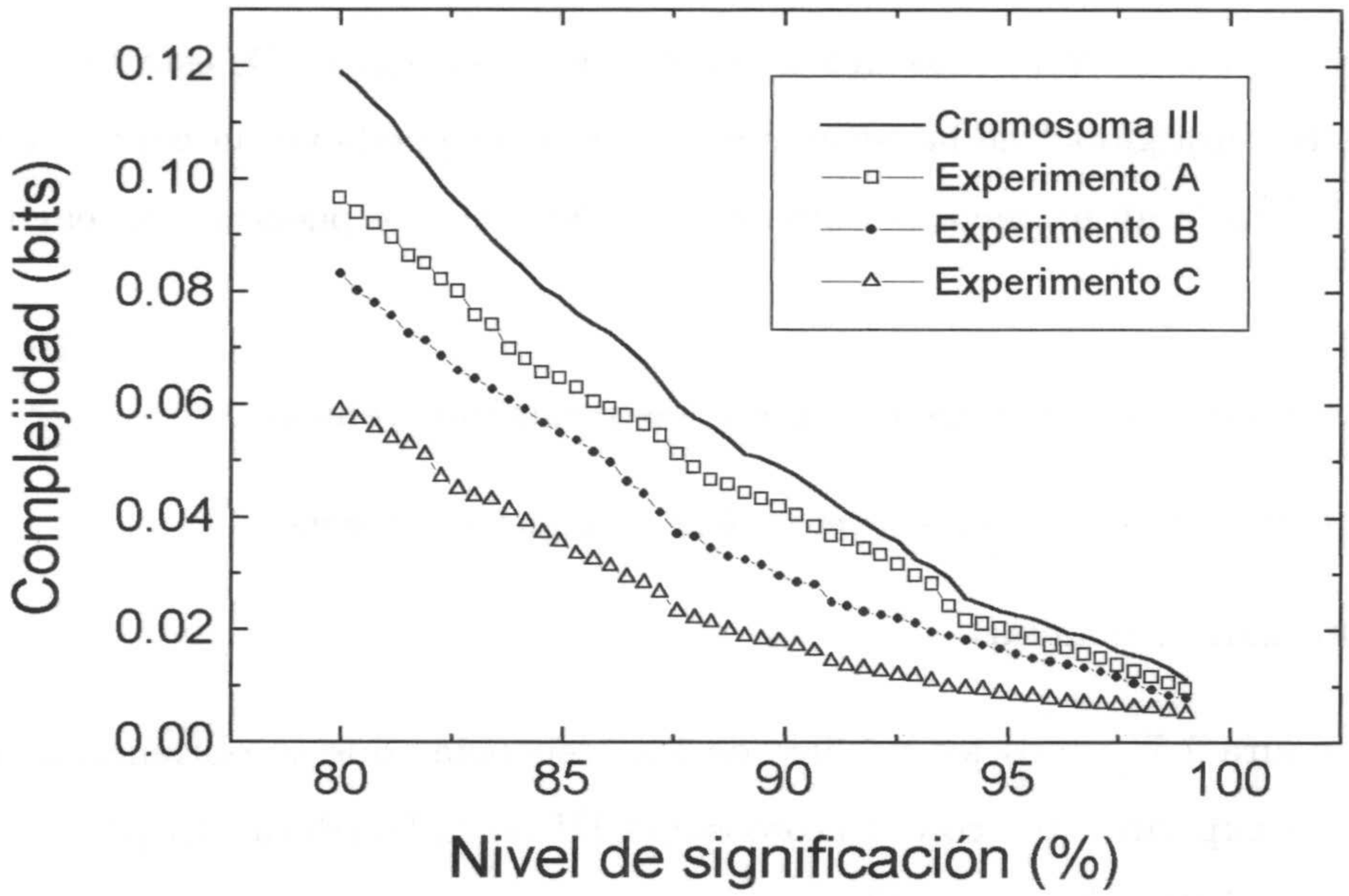

Figura 7.7: Perfiles de complejidad de secuencias obtenidas barajando internamente la zona codificadora del cromosoma III de la levadura, la no codificadora y ambas.

7.4 Complejidad composicional y evolución

Otra cuestión interesante, relacionada con los perfiles de complejidad, es la forma en que varían éstos al aplicarlos a secuencias pertenecientes a organismos con distintos grados de evolución. Aunque, a primera vista, parece muy razonable suponer que la evolución debe llevar consigo un aumento de la complejidad de los organismos (sea cual sea la definición de esta), no es algo aceptado por todos los autores y sigue siendo hoy un tema controvertido [Mc 96].

Si no hay acuerdo en esto, mucho menos en que la complejidad del genoma

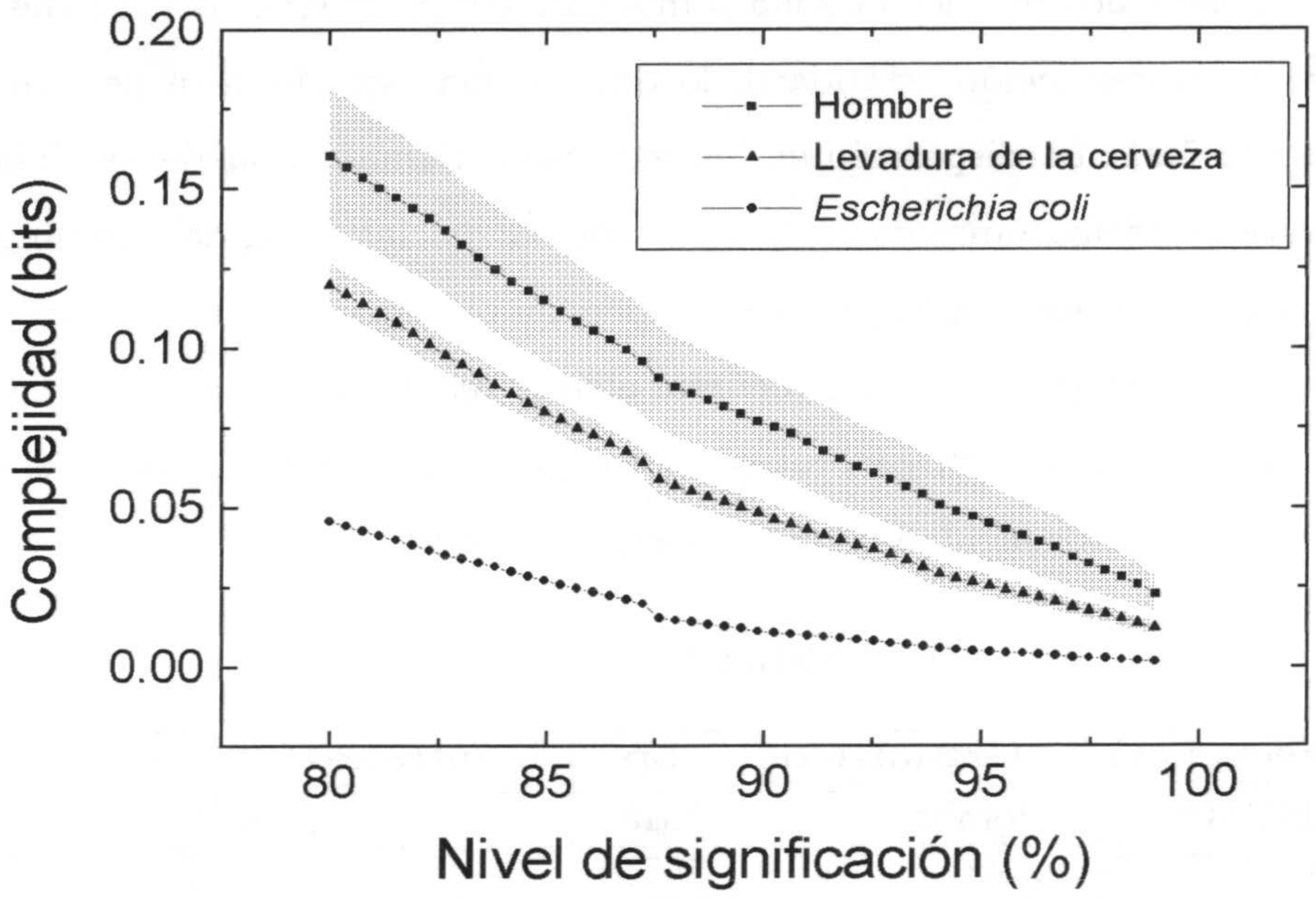

Figura 7.8: Perfiles de complejidad de secuencias humanas, de la levadura de la cerveza y del *Escherichia coli*.

deba aumentar con la evolución. En el capítulo anterior se comentaron algunos intentos de definir la complejidad global del genoma, en base al número de nucleótidos o al número de genes y, como se dijo entonces, los resultados no fueron muy buenos. En los ejemplos que hemos visto hasta ahora, aunque no se ha hecho un estudio exhaustivo, se puede ver que los perfiles de organismos, a priori más complejos, suelen aparecer más altos. Como ejemplo vamos a tomar los tres organismos de los que disponemos de un mayor número de datos: la bacteria *Escherichia coli (*procariota), la levadura de la cerveza (eucariota) y el hombre. Para *E. coli* representamos el perfil del genoma completo, para la levadura el perfil promedio de los 16 cromosomas

y para el hombre el promedio de los perfiles de todas las secuencias de las que disponemos. En estos dos últimos la zona sombreada centrada en el perfil tiene un grosor de 4 veces la desviación estándard, lo que da una idea de la dispersión de los perfiles (figura 7.8). La dispersión en las secuencias humanas puede ser debida a que no se trata de cromosomas completos sino de secuencias escogidas sin ningún criterio explícito. Cabe destacar la pequeña dispersión que aparece entre los 16 cromosomas de la levadura (ya vimos que ocurría algo parecido con las distribuciones de longitudes de dominios, y en general con gran número de medidas estadísticas que caracterizan la heterogeneidad composicional [Li 97c]).

Tabla I.

Secuencia	**Organismo**	**bp**	**% intrones**	α
HSBMYH7	Hombre	28438	74	0.59
RATMHCG	Rata	25759	77	0.57
GGMYHE	Pollo	31111	74	0.56
DROMHC	*Drosophila*	22663	66	0.52
BMMYOH	*Brugia Malayi*	11766	32	0.53
CEMYO3	*Caenorhabditis*	11621	19	0.53
ACMHC	*Acanthamoeba*	5894	10	0.51
SCMYO1G	Levadura	6108	0	0.51

Como se puede ver en la figura, en estos tres casos, digamos extremos, (un organismo procariota, un eucariota unicelular y el hombre) sí aparece un acuerdo entre complejidad composicional y evolución. Pero para hacer un análisis más riguroso veamos como se comportan los perfiles de complejidad para secuencias correspondientes a un mismo gen para diversos organismos. Lo haremos para las secuencias que contienen la zona codificadora de la cadena pesada de la miosina, que fueron analizadas por Buldyrev et al. en 1993 [Bu 93a] obteniendo un buen acuerdo entre complejidad fractal (medida con el exponente DFA) y evolución. En la tabla I se muestran los resultados obtenidos por este autor, junto con las longitudes de cada

secuencia y el porcentaje de intrones.

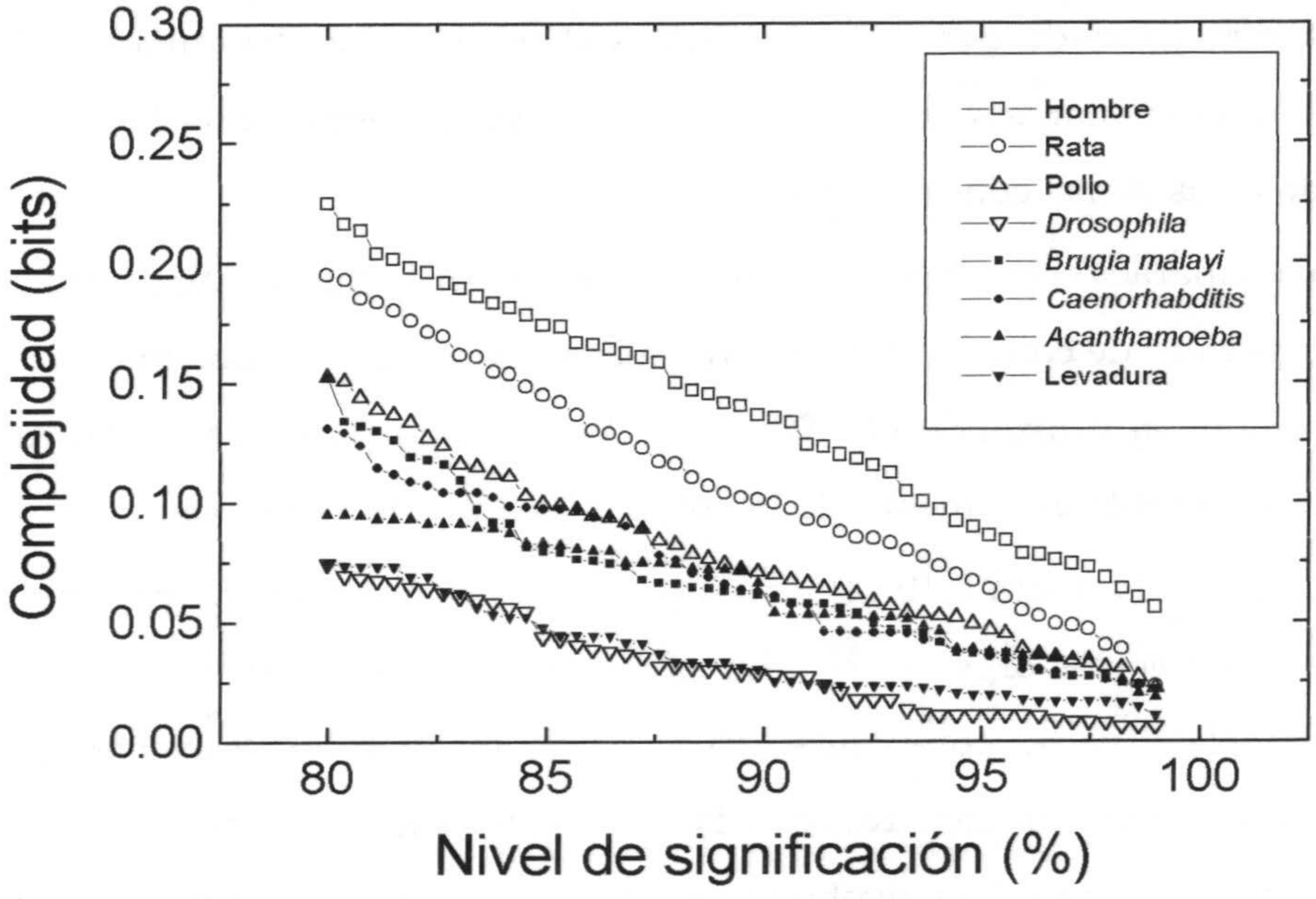

Figura 7.9: Pefiles de complejidad para las secuencias que contienen el gen de la cadena pesada de la miosina de diversos organismos.

En la figura 7.9 se muestran los perfiles de estas secuencias. Como se puede ver los resultados son bastante buenos, la única secuencia que se aparta de la tendencia es la correspondiente a la *Drosophila* que aparece al mismo nivel que la levadura. Nótese, sin embargo, que esta secuencia también da un valor algo más bajo de lo que cabría esperar con el exponente del DFA.

También cabe destacar la relación que existe entre la complejidad composicional de las secuencias y el porcentaje de intrones presentes en ellas, lo que estaría de acuerdo con lo visto en la sección anterior sobre la mayor complejidad de las zonas

no codificadoras. Sin embargo las secuencias correspondientes al hombre, la rata y el pollo tienen un porcentaje de intrones muy similar y presentan perfiles claramente diferentes. Es por este motivo que no sólo el porcentaje de zona no codificadora (que suele estar relacionado con la evolución) es el responsable del aumento de la complejidad composicional, sino que también se observa un incremento de la propia complejidad de estas zonas con la evolución.

Como comparación con los resultados expuestos en la figura 7.9, en la figura 7.10 vemos los perfiles de complejidad de estas mismas secuencias pero, en este caso, tras una segmentación utilizando el alfabeto {S,W}. Como puede verse, no existe una tendencia apreciable al aumento de la complejidad con la evolución.

Por último en la figura 7.10 vemos los perfiles de complejidad de los genomas mitocondriales de diversos organismos. En este caso cabe destacar, ante todo, las pequeñas diferencias que se observan entre los distintos perfiles, a excepción del paramecio, lo que estaría de acuerdo con la idea, generalmente aceptada, de que las mitocondrias tienen un origen evolutivo común (endosimbiosis de una bacteria). Por otra parte, la notable diferencia del perfil correspondiente al paramecio también estaría de acuerdo con el hecho de que este microorganismo, y en general todos los ciliados, presenta grandes divergencias con el resto de los organismos; hasta el punto de que su origen evolutivo se data con una antigüedad superior a la de la Tierra, lo que hace considerar seriamente su origen extraterrestre.

7.5 Modelos de secuencias

Los modelos que vimos en la sección 2.4 intentan reproducir alguna característica determinada de las secuencias reales: la composición a nivel de di o trinucleótidos (cadenas de Markov), dependencia de la información mutua con la longitud (pseudocromosomas) y, principalmente, las correlaciones de largo alcance (replicación-

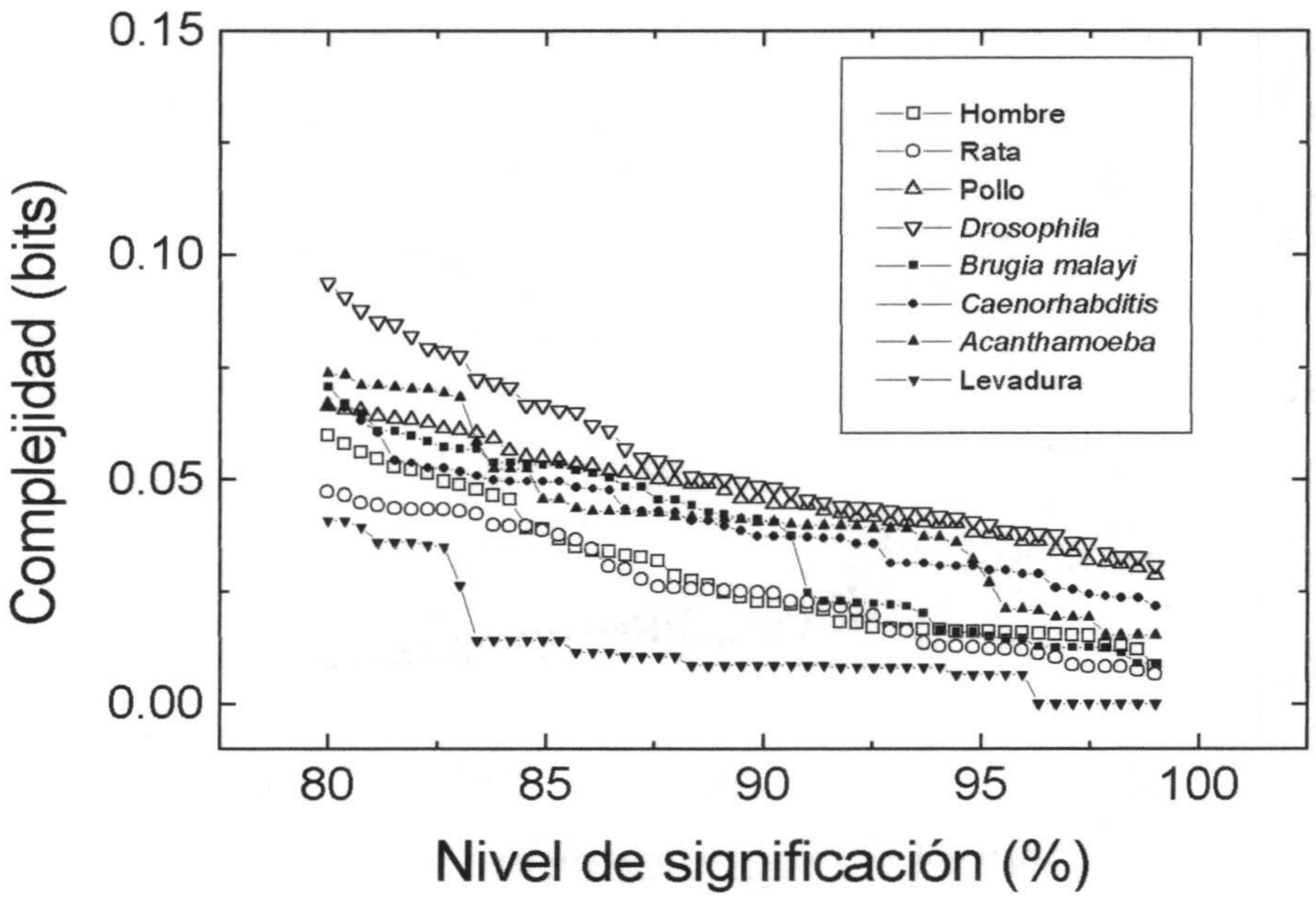

Figura 7.10: Perfiles de complejidad con alfabeto {S,W} para las secuencias de la figura anterior.

mutación, Lévy-walk e inserción-delección). Veamos ahora cómo reproducen estos modelos la heterogeneidad composicional de las secuencias, tal y como la mide nuestro perfil de complejidad. En lo que sigue incluiremos en las gráficas los perfiles de HUMTCRAD y ECO110K como referencia.

7.5.1 Cadenas de Markov

En la figura 7.12 vemos los perfiles de dos secuencias generadas con un modelo de Markov de orden 1. Se han utilizado las matrices de transición obtenidas a partir de HUMTCRAD y ECO110K:

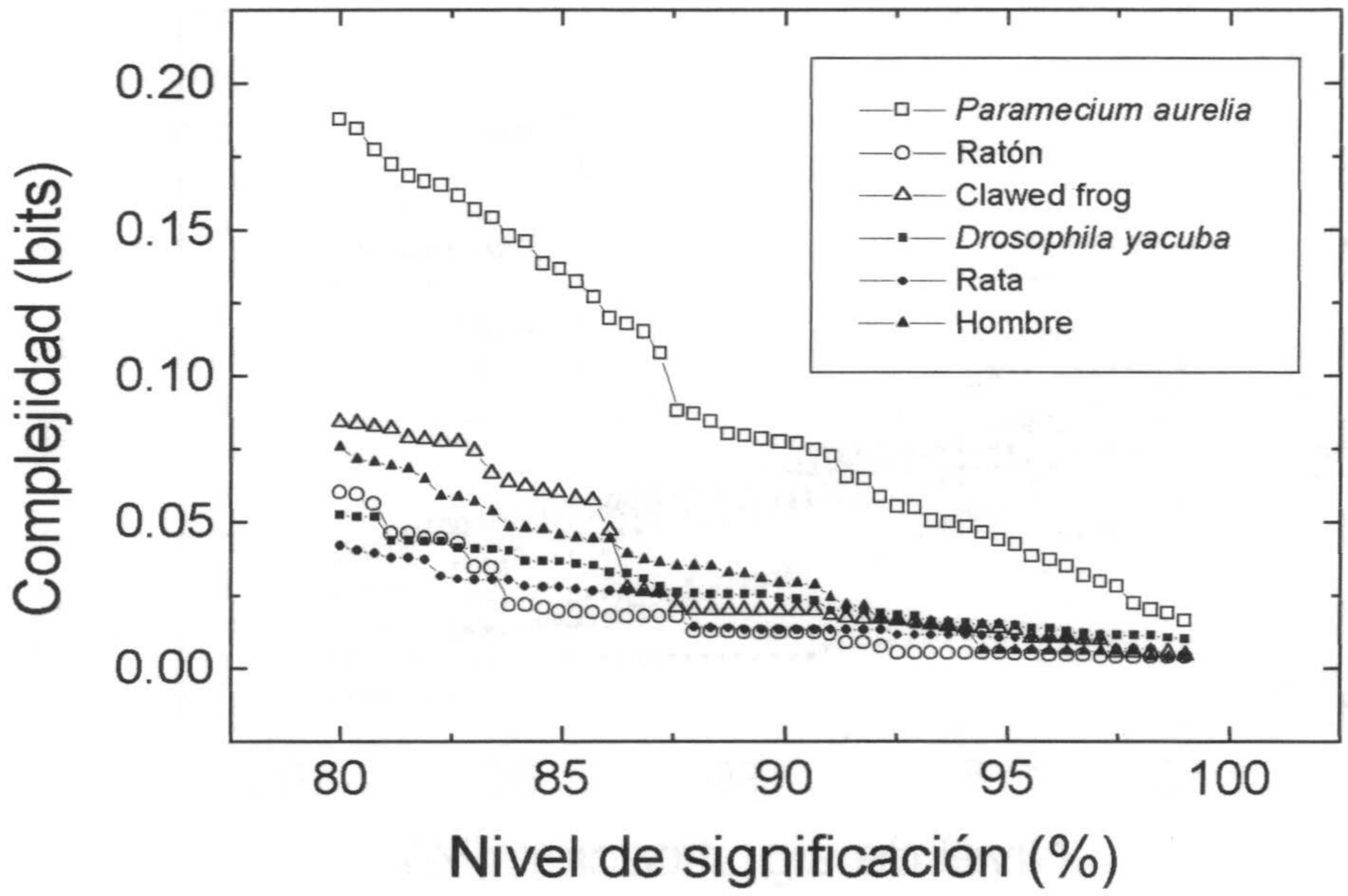

Figura 7.11: Perfiles de complejidad del genoma mitocondrial de algunos organismos.

$$\text{HUMTCRAD} \begin{pmatrix} 0.5894 & 0.4363 \\ 0.4106 & 0.5637 \end{pmatrix}$$

$$\text{ECO110K} \begin{pmatrix} 0.4854 & 0.5316 \\ 0.5146 & 0.4684 \end{pmatrix}$$

El resultado era bastante previsible, como es sabido, las cadenas de Markov son estacionarias y, por tanto, tienen una composición homogénea; esto se traduce en un perfil casi plano (no se producen apenas divisiones). Tan sólo aparecen algunos dominios para resoluciones bajas, causados por las correlaciones de corto alcance que presentan estas secuencias [Ta 94] y que producen, para estas resoluciones, perfiles notablemente superiores a los de las secuencias aleatorias. Cabe destacar que el

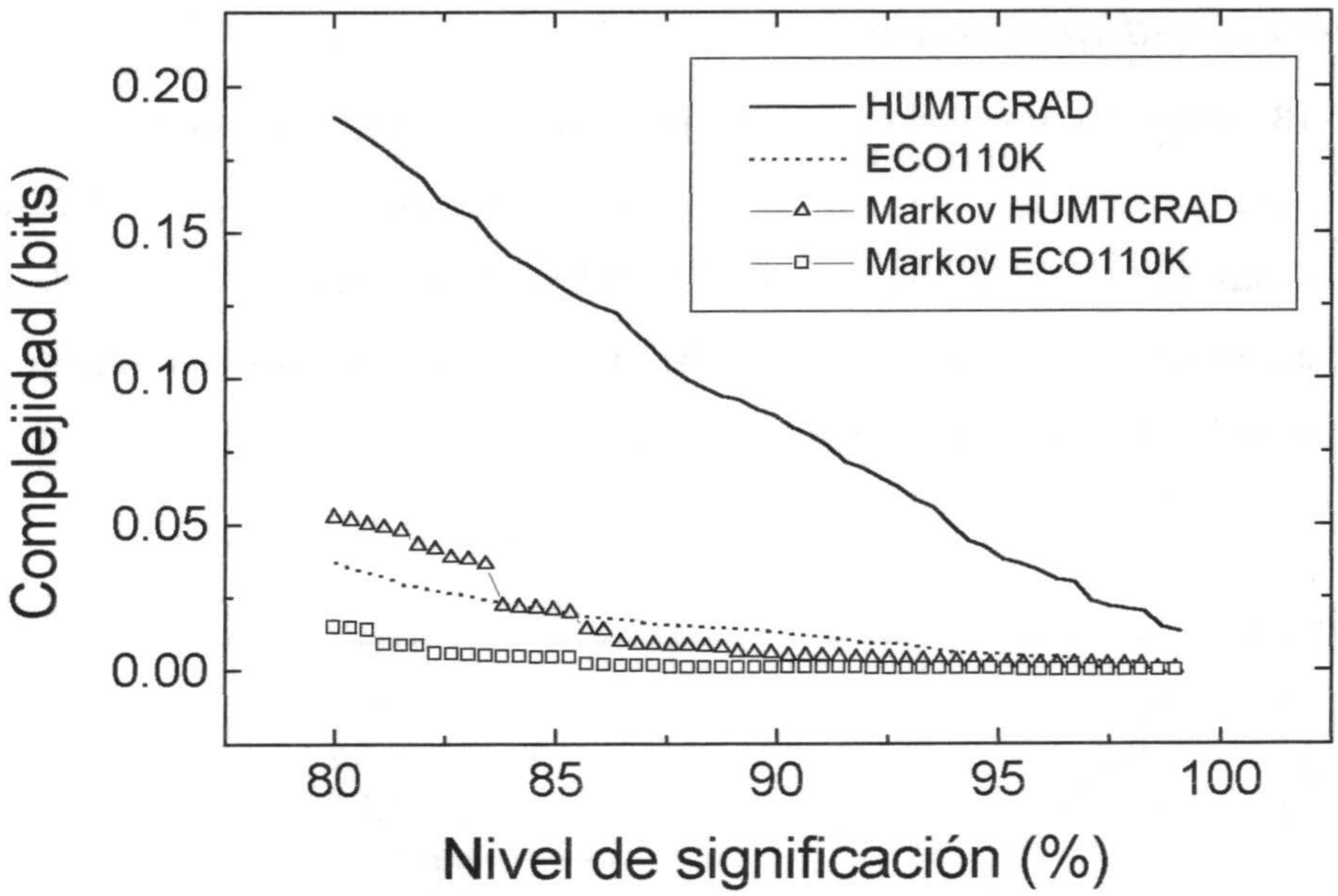

Figura 7.12: Perfiles de complejidad de secuencias generadas con el modelo de Markov con las probabilidades de transición de HUMTCRAD y ECO110K.

perfil de la cadena de Markov generada con las probabilidades de transición de HUMTCRAD se levanta, a resoluciones bajas, por encima del perfil de ECO110K. De nuevo este resultado relaciona los perfiles de complejidad con la presencia de correlaciones: para la matriz de transición de HUMTCRAD el exponente de escala para longitudes pequeñas es de $\alpha_0 = 0.6$ [Bu 93b], lo que indica correlaciones debidas a la persistencia de un símbolo, sin embargo para ECO110K este exponente es de $\alpha_0 = 0.46$, lo que indica anticorrelaciones, esto es, alternancia de símbolos y por tanto ausencia de dominios composicionales. Esto explicaría el perfil casi plano de

la cadena de Markov obtenida a partir de ECO110K.

7.5.2 Mutación-duplicación

En la figura 7.13 vemos los perfiles de complejidad de dos secuencias generadas con el modelo de mutación-duplicación con los valores del parámetro p (sec. 2.4.2) que dan resultados más parecidos a los de HUMTCRAD. Nótese como los perfiles no reproducen el comportamiento de HUMTCRAD para valores altos de significación, aunque sí se ásemejan bastante para significaciones bajas.

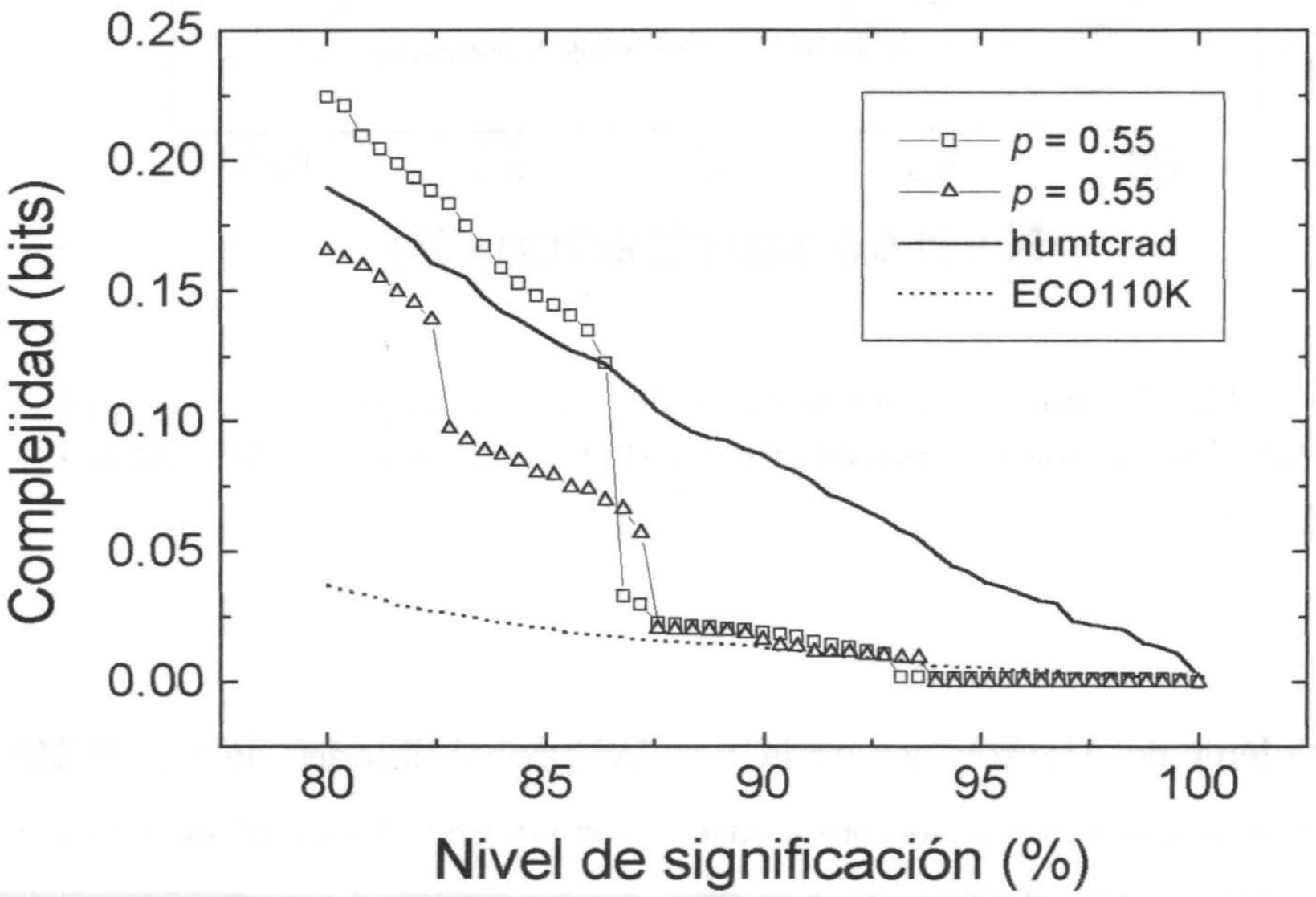

Figura 7.13: Perfiles de complejidad de secuencias generadas con el modelo de mutación-duplicación para los valores del parámetro p con los que más se acerca al perfil de HUMTCRAD.

Como ya se comentó, este modelo, da cuenta de algunos procesos que tienen

lugar en la evolución de las secuencias de ADN y, que en general, justificaría la presencia del ADN repetido. Pero hemos visto que la mera presencia de ADN repetido no es suficiente para generar la complejidad composicional que se observa en algunas secuencias eucariotas. Un dato interesante es la similitud de los perfiles de las secuencias generadas con este modelo y los perfiles de subsecuencias de ADN repetido que veíamos en la figura 7.5.

Este resultado concuerda con la idea de que, para significaciones más bajas suelen aparecer detalles de menor tamaño y, como es sabido, este modelo aunque genera correlaciones de largo alcance, no produce heterogeneidades de gran tamaño.

Nótese por último que, aunque estos perfiles difieren claramente de los de la secuencia humana, su análisis con espectros de potencia da exponentes similares [Li 92b]. Esto corrobora lo que comentábamos al principio del capítulo: aunque hay bastante relación entre los resultados obtenidos con los perfiles de complejidad y los basados en análisis fractales, no se puede decir que sean equivalentes.

7.5.3 Inserción-delección

Veamos ahora los resultados que se obtienen con el modelo de inserción-delección que, como se comentó en el capítulo 2, justificaba bastante bien la existencia de correlaciones de largo alcance en secuencias no codificadoras y que, según los autores, además reproducía bastante bien la "evolución" de estas correlaciones al observar secuencias de organismos con diferente grado complejidad biológica.

En la figura 7.14 vemos los perfiles de tres secuencias generadas con este modelo para distinto número de iteraciones. Los parámetros que se han escogido son los propuestos por los autores del modelo [Bu 93b]. En primer lugar cabe destacar cómo los perfiles aumentan su ordenada con el número de iteraciones (también se puede comprobar que aumenta el exponente de escala calculado con el DFA). Esta figura tiene cierta similitud con la figura 7.9 en la que se representan los perfiles de

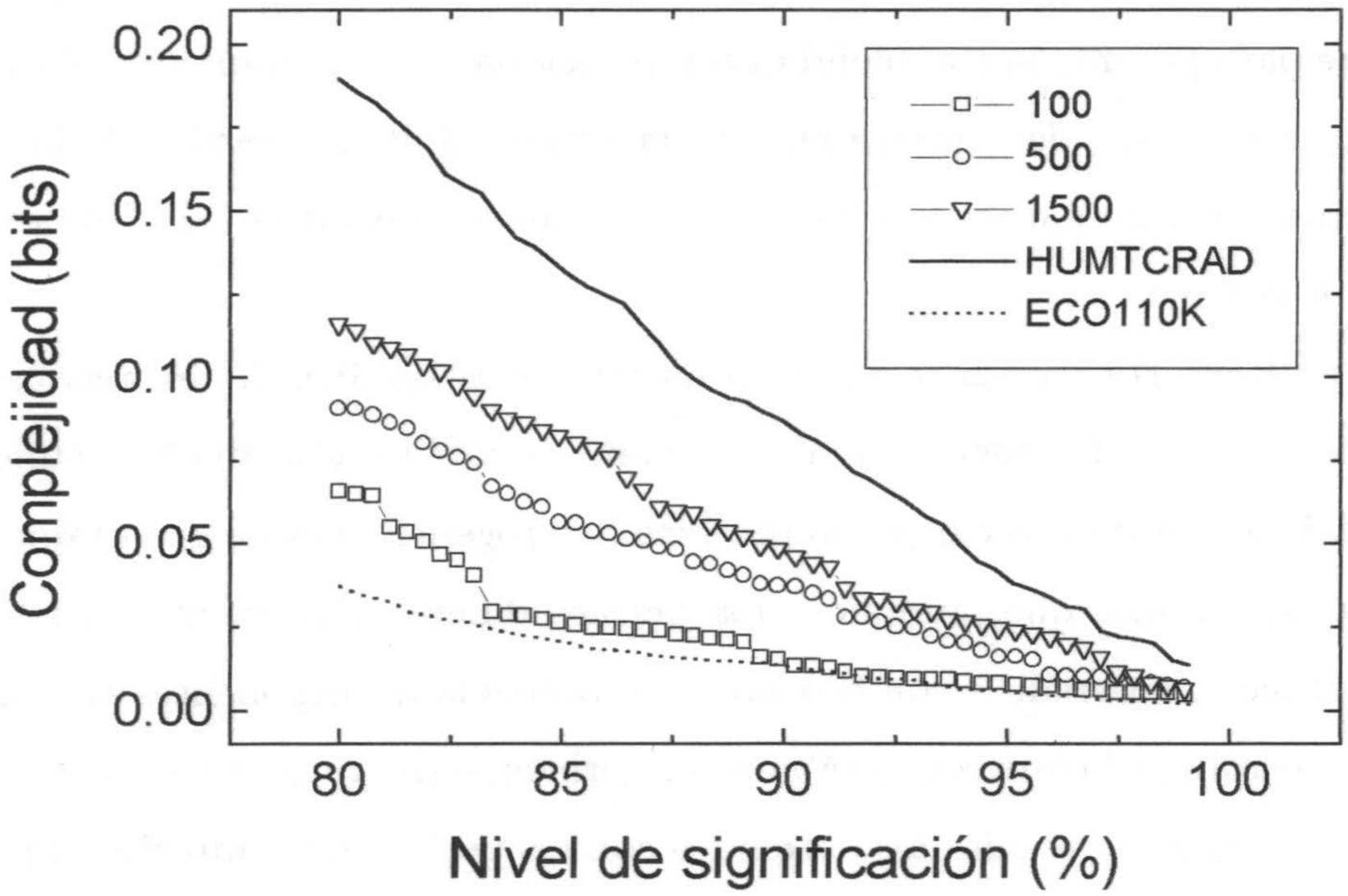

Figura 7.14: Perfiles de complejidad para distinto número de iteraciones del modelo de inserción-delección. Los parámetros son: $\mu = 2.1$, $p_i = 0.2$ y $l_0 = 20$ y un 60% de purinas en la secuencia de partida.

complejidad de las secuencias que contienen el gen de la cadena pesada de la miosina para organismos con distinto grado de complejidad biológica.

En la figura 7.15 se han representado la secuencias límite que se obtienen con este modelo después de un número suficientemente grande de iteraciones. Se han utilizado, para una de ellas los mismos parámetros que en la figura 7.14, y para la otra se ha aumentado la proporción de purinas en la secuencia de partida. La zona sombreada tiene un grosor de 4 veces la desviación estándard de los perfiles obtenidos desde 1000 hasta 10.000 iteraciones, para indicar las fluctuaciones que

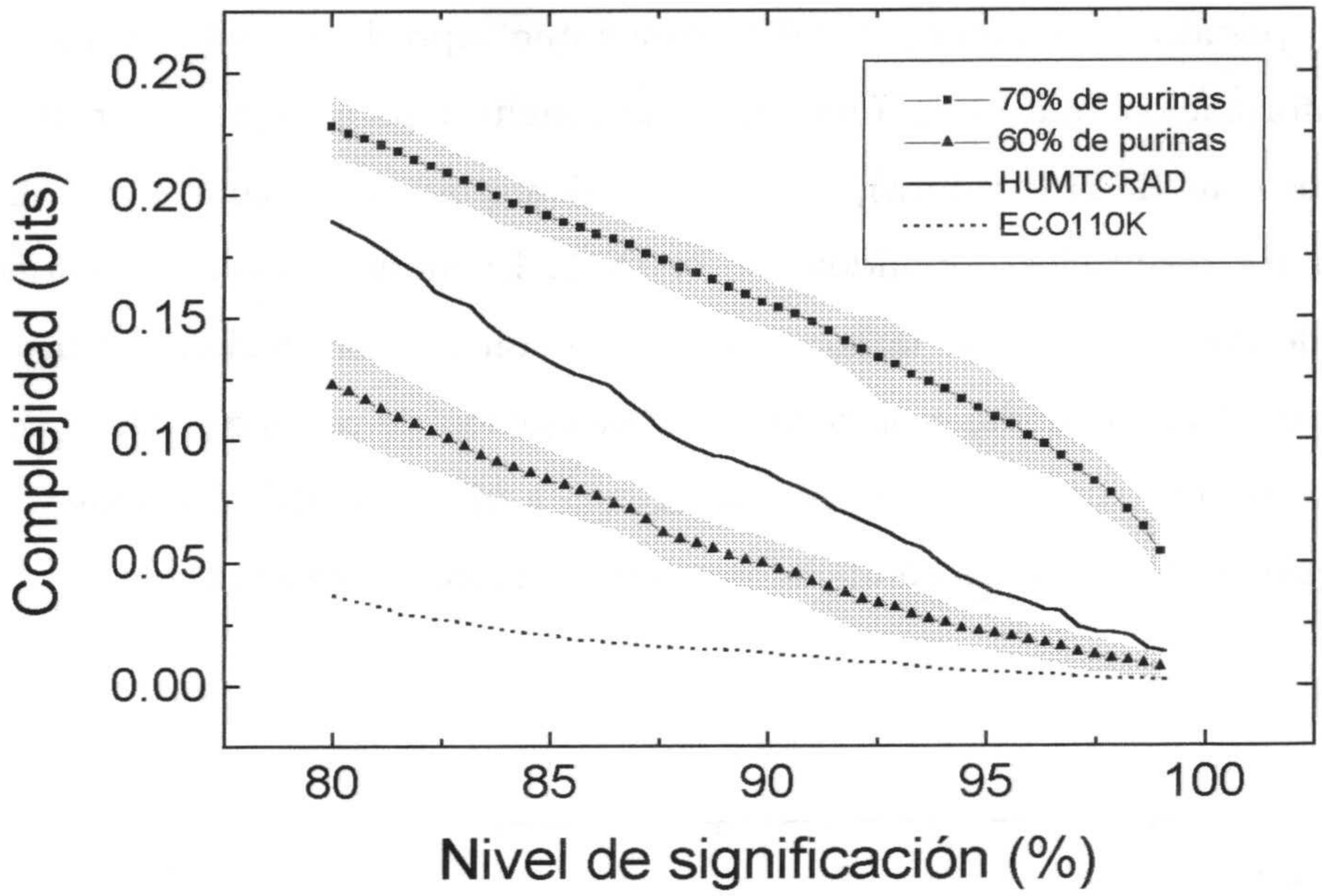

Figura 7.15:

se producen en el perfil, una vez alcanzado un estado estacionario. Vemos que la secuencia generada con los parámetros más realistas (60 % de purinas), tiene un perfil que queda por debajo del de HUMTCRAD, aunque su exponente de escala coincide exactamente con el de la secuencia humana. De todas formas, como se puede ver en la gráfica, con este modelo se pueden generar secuencias con perfiles de complejidad iguales e incluso superiores a los de la secuencia humana pero, en este caso, se utiliza un sesgo inicial de purinas que es algo superior a los encontrados en secuencias codificadoras reales, además esta secuencia presenta un exponente de escala, medido con el DFA, notablemente superior al de HUMTCRAD (0.7 y 0.6 respectivamente).

7.5.4 Pseudo-cromosomas

El modelo de pseudo-cromosomas [He 97], vimos que reproduce bastante bien los perfiles de información mutua en función de la longitud para algunas secuencias (ej. los cromosomas de la levadura), pero no era demasiado satisfactorio a la hora de reproducir los resultados obtenidos con el DFA. El modelo pretende justificar la presencia de correlaciones de largo alcance sólo con las correlaciones a distancias múltiplo de 3 introducidas por el uso no homogéneo de codones. Tal y como está propuesto supone una composición homogénea a nivel de trinucleótidos y una composición global de las zonas codificadoras igual a la las no codificadoras.

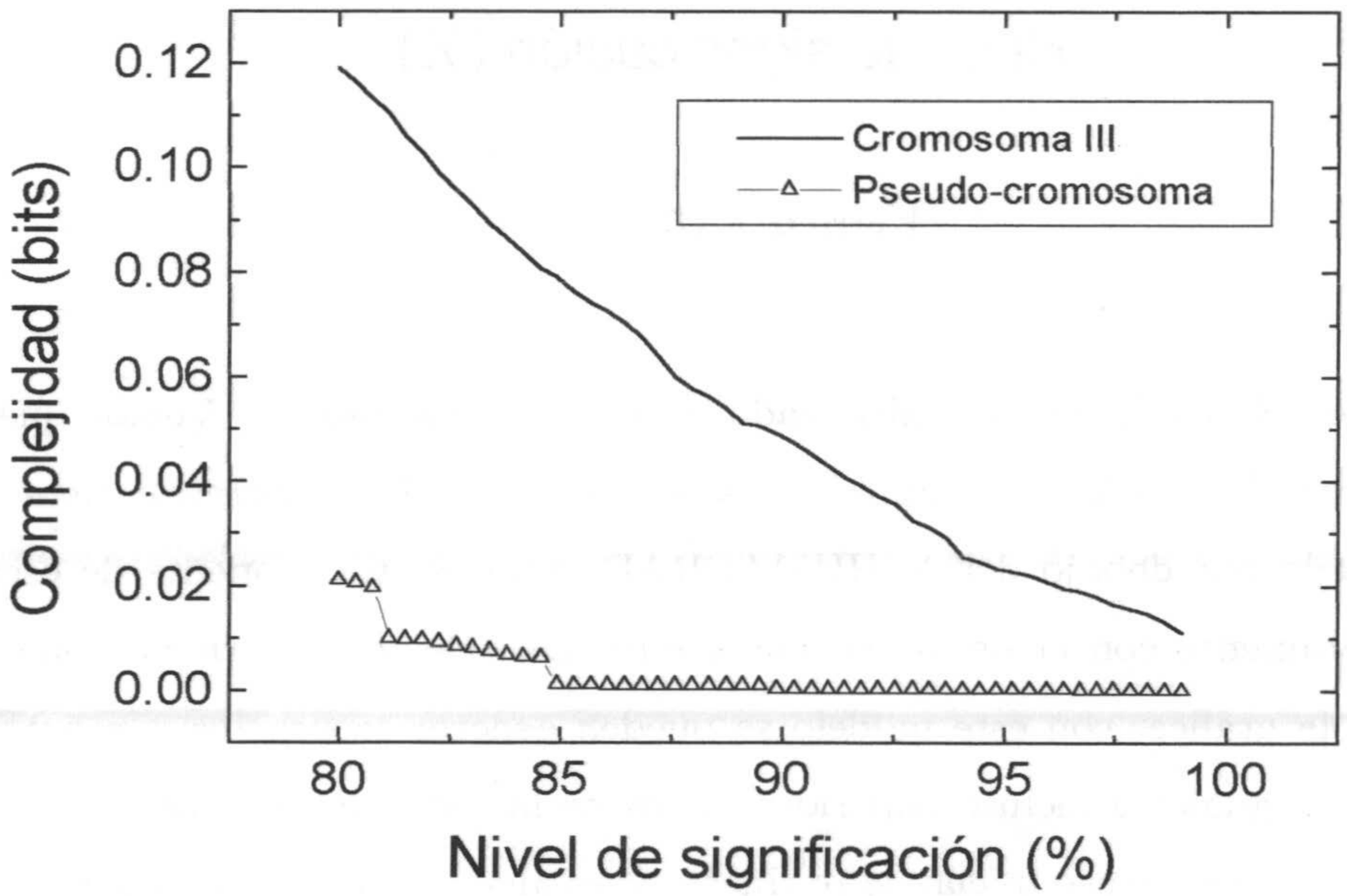

Figura 7.16: Perfil de complejidad de una secuencia generada con el modelo de Pseudo-cromosomas.

En la figura 7.16 vemos el perfil de una secuencia simulada con este modelo junto con el perfil del cromosoma III de la levadura de la cerveza. Este modelo, desde luego, no reproduce la complejidad composicional de la secuencia natural pero, en principio podría achacarse esto a la excesiva simplificación: no se tiene en cuenta la distinta composición de diferentes zonas codificadoras, la alternancia de zonas transcritas en un sentido y en otro y la diferente composición de zonas codificadoras y no codificadoras. Pero aún así, si realmente el modelo diese cuenta de la complejidad de las zonas codificadoras, el perfil debería tomar valores del orden de la diferencia entre los perfiles del experimento b) y el c) de la figura 7.7, cosa que no ocurre.

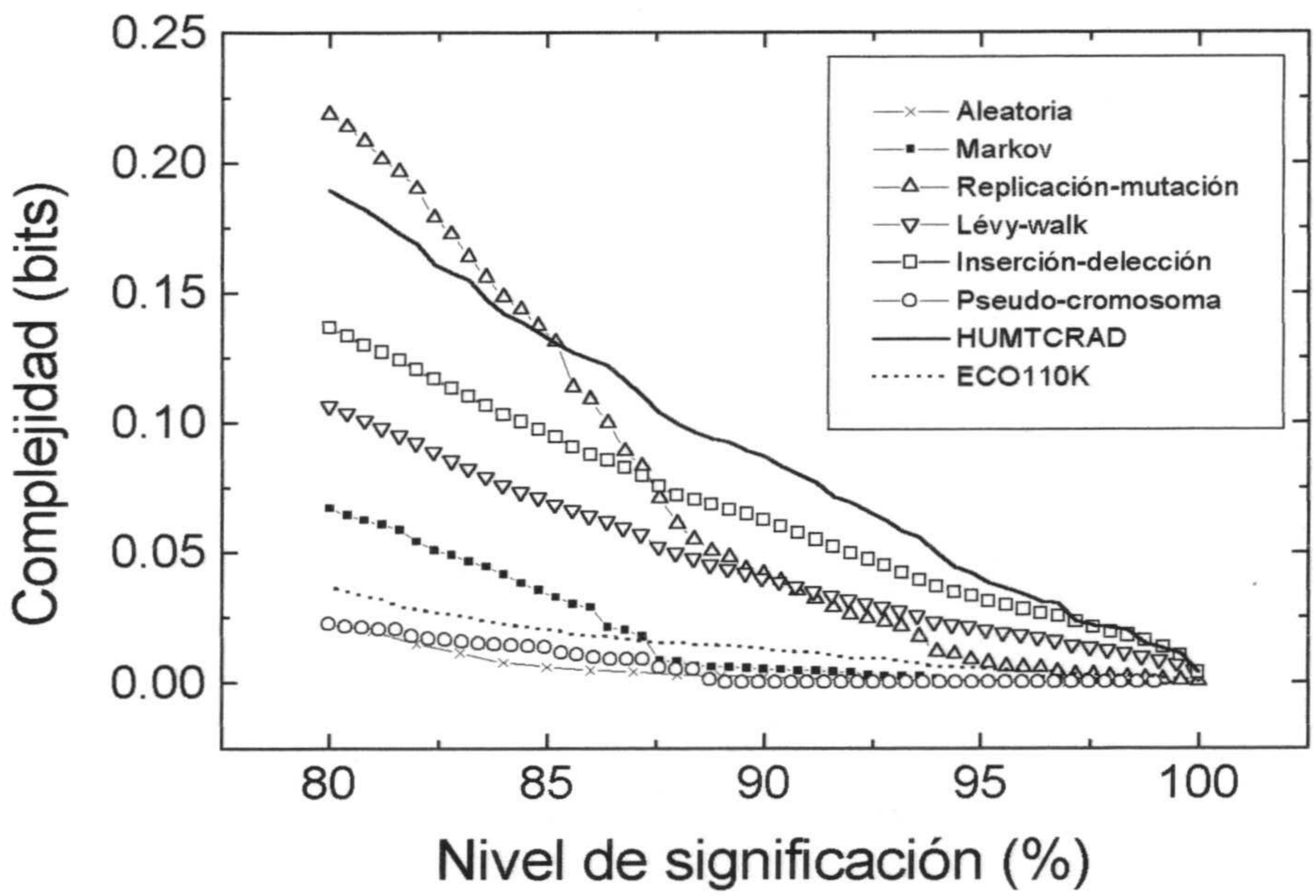

Figura 7.17: Perfiles de complejidad para secuencias artificiales generadas con diversos modelos y con parámetros realistas.

Por último, vemos en la figura 7.17 agrupados los perfiles de secuencias generadas con estos modelos. Para el modelo de Markov de orden 1 se ha tomado la matriz de transición de HUMTCRAD y para el resto se han tomado los parámetros que, según sus autores, son óptimos.

Capítulo 8

Conclusiones y problemas abiertos

8.1 Conclusiones

Dedicaremos esta sección a resumir los resultados más importantes que aquí se presentan. En primer lugar enumeraremos los métodos originales de análisis propuestos y, a continuación, describiremos las conclusiones obtenidas con estos métodos.

8.1.1 Métodos

1. Se ha propuesto un algoritmo de segmentación (capítulo 4) capaz de determinar zonas de composición homogénea en secuencias simbólicas en general y de ADN en particular. Como principales novedades del algoritmo están el uso de la divergencia de Jensen-Shannon y el procedimiento de segmentación, que parte de la secuencia completa y la va escindiendo, en lugar de utilizar una ventana móvil para localizar los cortes.

2. Para estimar las significación de los cortes producidos por el algoritmo se ha obtenido una aproximación asintótica de la distribución de probabilidad que sigue la divergencia de Jensen-Shannon para el caso de una secuencia aleatoria i.i.d.

3. A partir del algoritmo de segmentación se ha propuesto en el capítulo 6 una medida de la complejidad composicional en secuencias simbólicas que, como se discute en este capítulo, tiene muchas propiedades que la hacen adecuada como medida de complejidad en general. En particular, parece idónea para el estudio de la compleja estructura de las secuencias de ADN, en vista de la importancia de la heterogeneidad composicional de estas secuencias.

   Cabe destacar que esta medida es la única, de las propuestas hasta el momento, que presenta la complejidad como una función del nivel de detalle con el que se observa.

8.1.2 Resultados

1. En primer lugar se obtienen resultados que sostienen de forma definitiva, me atrevería a decir, 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. Esta cuestión, como se comentó, es uno de los debates abiertos entre los especialistas, principalmente a causa de la controversia sobre el uso de métodos pensados para el análisis de secuencias estacionarias.

2. La posibilidad de segmentar secuencias a distintos niveles de significación ha puesto de manifiesto, no sólo la mayor heterogeneidad composicional en las secuencias no codificadoras sino también la existencia de una heterogeneidad compleja, como la denominan algunos autores [Li 97d]. Esta estructura

jerárquica de los dominios se aprecia claramente en las segmentaciones recursivas que se muestran en el capítulo 5 (sección 5.5). También cabe destacar cómo los perfiles de complejidad ponen de manifiesto la existencia de autosemejanza, en lo que a complejidad composicional se refiere, en las secuencias con correlaciones de largo alcance (sección 7.2.1)

3. Por último, aunque no menos importante, cabría citar el buen acuerdo que se obtiene entre los niveles de complejidad composicional y complejidad biológica (sección 7.4). Este resultado indica la relevancia de la gran heterogeneidad composicional observada en las secuencias de organismos superiores y podría abrir el camino a futuros estudios de evolución molecular.

8.2 Problemas abiertos

Como uno de los objetivos a más corto plazo de nuestra línea de investigación estaría la obtención de modelos teóricos que relacionen las distribuciones de dominios obtenidos en el proceso de segmentación y los perfiles de complejidad con las propiedades estadísticas de la secuencia.

En concreto, una cuestión prioritaria sería la obtención de una caracterización de los perfiles correspondientes a secuencias con propiedades fractales.

Otra cuestión, también relevante, es el estudio de la influencia de las heterogeneidades de cada nucleótido, haciendo uso de las propiedades de agrupamiento vistas en la sección 4.3.2.; y, en general, aprovechar las posibilidades que brinda el hecho de que la medida utilizada para segmentar puede utilizarse con cualquier alfabeto (no tiene las limitaciones de las medidas basadas en la teoría de la señal).

Por último citar que en la actualidad y en colaboración con otros grupos estamos planteando la posibilidad de obtener métodos de segmentación, basados en el uso de algoritmos genéticos, que consigan resultados más próximos a la segmentación completa.

Referencias

[Al 95] Allegrini, P., Barbi, M., Grigolini, P. and West, B.J. *Dynamical model for DNA sequences* Physical Review E (1995), vol. **52**(5), pp. 5281-5296.

[An 96] Anteneodo, C. and de Souza, M. C. *Prototype for time correlation effects: analytical results.* Journal of Physics A (1996), vol. **29** pp. 6151-6160.

[Ar 95] Arneodo, A., Bacry, E., Graves, P.V. and Muzy, J.F. *Characterizing Long-Range Correlations in DNA Sequences from Wavelets Analysis.* Physical Review Letters (1995), vol.**74**(16) pp. 3293-3296.

[Ar 96] Arneodo, A., d'Aubenton-Carafa, Y., Bacry, E., Graves, P.V., Muzy, J.F. and Thermes, C. *Wavelet based fractal analysis of DNA sequences.* Physica D (1996) vol. **96**, pp. 291-320.

[At 96] Attard, G.S., Hurworth, A.C. and Jack, J.P. *Language-like features in DNA: transposable elements footprints in the genome.* Europhysics Letters (1996) vol. **36**(5), pp. 391-396.

[Az 95] Azbel, M.Y., *Universality in a DNA statistical structure* Physical Review Letters (1995), vol. **75**, pp. 168-171.

[Ba 95a] Barranco-López, V., Luque-Escamilla, P., Martínez- Aroza, J. and Román-Roldán, R. *Entropic texture-edge detection for image segmentation,* Electronics Letters (1995), vol. **31**(11), pp. 867-869.

[Ba 95b] Barranco-López, V., Luque-Escamilla, P., Martínez- Aroza, J. and Román-Roldán, R. *Texture segmentation based on Information-Theoretic edge detection method.* Proceedings of the VI Spanish Symposium on Pattern Recognition and Image Analysis (1995), pp. 58-64.

[Bar 88] Barnsley, M. *Fractals Everywhere.* Academic Press, London (1988).

[Be 94] Bernaola-Galván, P., *Contribución al análisis de secuencias mediante medidas de perfiles entrópicos.* Tesis de Licenciatura. Universidad de Granada (1994).

[Be 96] Bernaola-Galván, P., Román-Roldán, R. and Oliver, J.L., *Compositional segmentation and long-range fractal correlations in DNA sequences.* Physical Review E (1996), vol **53**(5), pp. 5181-5189.

[Ben 88] Benneth, C.H., *Dissipation, information, computational complexity and the definition of organization.* pp. 215-233 en: *Emergin syntesis in science: Proceedings of the founding workshops of the Santa Fe Institute.* Ed. Pines, D., vol. **1**, Addison-Wesley, Redwood City, C.A. (1988).

[Ben 90] Benneth, C.H., *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, (1990).

[Ber 89] Bernardi, G. *The isochore organization of the human genome*, Annual Review of Genetics (1989), vol. **23**, pp. 637-661.

[Ber 95] Bernardi, G. *The human genome: organization and evolutionary history*, Annual Review of Genetics (1995), vol. **29**, pp. 445-476.

[Bo 87a] Borodovskii, M.Y., Sprinzhitskii, Y.A., Golovanov, E.J., Aleksandrov, A.A., *Statistical patterns in the primary structures of functional regions of the genome in Escherichia coli I. Frequency characterization.* Molecular Biology (1987), vol. **20**, pp. 826-833.

[Bo 87b] Borodovskii, M.Y., Sprinzhitskii, Y.A., Golovanov, E.J., Aleksandrov, A.A., *Statistical patterns in the primary structures of functional regions of the genome in Escherichia coli II. Nonuniform Markov models.* Molecular Biology (1987), vol. **20**, pp. 833-840.

[Bog 94] Bogobubskaya, A.A. and Bogolubsky, I.L. *Two-components localized solutions in a nonlinear DNA model.* Physics Letters A (1994), vol. **192**, pp. 239-246.

[Bog 96] Bogolubsky I.L. and Bogolubskaya, A.A. *Reply to the comment on "Two-component localized solutions in a nonlinear DNA model".* Physics Letters A (1996), vol. **213**, pp. 311-312.

[Bon 96] Bonhoeffer, S., Herz, A.V.M., Boerlijst, M.C., Nee, S., Nowak, M.A. and May, R.M. *No signs of hidden language in noncoding DNA.* Physical Review Letters (1996), vol. **76**(11), p. 1977.

[Bor 93] Borŝtnik, B., Pumpernik, D. and Lukman, D. *Analysis of Apparent $1/f^{\alpha}$ Spectrum in DNA Sequences.* Europhysics Letters (1993), vol. **23**(6), pp. 389-394.

[Bov 84] Borovkov, A.A., *Estadística Matemática.* Ed. Mir (1984) (versión en castellano 1988).

[Br 88] Brooks, D.R. and Willey, E.O., *Evolution as Entropy: Toward a Unified Theory of Bioilogy* (1988) 2nd edn. University of Chicago.

[Bri 62] Brillouin, L. *Science and Information Theory.* Academic Press, New York (1962).

[Bu 93a] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Stanley, H.E., Stanley, M.H.R. and Simons, M. *Fractal Landscapes and Molecular Evolution: Modeling the Myosin Heavy Chain Gene Family.* Biophysical Journal (1993), vol.**65**, pp. 2673-2679.

[Brt 69] Britten, R. and Davidson, E.H., *Gene regulation for higher cells: A theory.* Science (1969), vol. **165**, pp. 349-357.

[Bu 93b] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simmons, M. and Stanley, H. *Generalized Lèvy-walk model for DNA Nucleotide Sequences.* Physical Review E

(1993), vol. **47**(6), pp. 4514-4523.

[Bu 93c] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simmons, M., Scortino, F. and Stanley, H.E. *Long-range power-law correlations in DNA*. Physical Review Letters (1993), vol. **71**(11), p. 1776.

[Bu 94] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K. and Stanley, H.E., *Fractals in Biology and Medicine: From DNA to the Heartbeat.* pp. 49-83 en: *Fractals in Science.* Ed. Brunde, A., Havlin, S., Springer-Verlag (1994).

[Bu 95] Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Matsa, M.E., Peng, C.K., Simmons, M. and Stanley, H. *Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis.* Physical Review E (1995), vol. **51**(5), pp. 5084-5091.

[Bur 93] Burlando, B. *The Fractal Geometry of Evolution.* Journal of Theoretical Biology (1993), vol. **163**, pp. 162-172.

[Ca 97] Castro e Silva, A. and Moreira, J.G. *Roughness exponents to calculate multi-affine fractal exponents.* Physica A (1997), vol. **235**, pp. 327-333.

[Cav 92] Caves, C.M, *Information and Entropy* Proceedings of Physics of Computation Workshop.(1990).

[Ce 95] Cebrat, S. and Dudek, M.R. *Coding rhythm of DNA strands.* Preprint (1995).

[Ce 96] Cebrat, S. and Dudek, M.R. *Symetry in chromosome fractal organization and DNA domain structure.* Preprint (1996).

[Ch 75] Chaitin, G.J., *Randomness adn Mathematical Proof.* Scientific American (1975) vol. **232**, pp. 47-52.

[Ch 87] Chaitin, G.J., *Algorithmic Information Theory*, Cambridge University Press, (1987).

[Cha 94] Chatzidimitriou-Dreismann, C.A., Friedrich Streffer, R.M. and Larharmmar, D. *Variations in base pair composition and associed long-range correlations in DNA sequences:*

computer simulation results. Biochimica et Biophysica Acta (1994), vol. **1217**, pp. 181-187.

[Cha 96] Chatzidimitriou-Dreismann, C.A., Streffer, R.M.F. and Larharmar, D., ... Nucleic Acid Research (1996), vol.**24**, pp. 1676-.

[Che 95] Chechetkin, V.R. and Turygin, A.Y., *Size-dependence of three-periodicity and long-range correlations in DNA sequences.* Physics Letters A (1995), vol. **199**, pp.75-80.

[Che 96] Chechetkin, V.R. and Lobzin, V.V., *Levels of ordering in coding and noncoding regions of DNA sequences.* Physics Letters A (1996), vol. **222**, pp. 354-360.

[Chg 90] Cheng, Z., Savit, R., *Structure factor of substitutional sequences.* Journal of Statistical Physics (1990), vol.**60**, pp. 383-393.

[Cho 56] Chomsky, N., *Three models for the description of language.* IRE Transactions on Information Theory (1956), vol.**2**, pp. 113-129.

[Chu 89] Churchill, G.A., *Stochastic models for heterogeneous DNA sequences.* Bulletin of Mathematical Biology (1989), vol.**51**(1), pp. 79-94.

[Chu 92] Churchill, G.A., *Hidden Markov chains and the analysis of genome structure.* Computer and Chemistry (1992), vol.**16**(2), pp. 107-115.

[Co 90] Cosmi, C., Cuomo, V., Ragosta, M. and Macchiato, M.F., *Characterization of nucleotide sequences using maximum entropy techniques,* Journal of Theoretical Biology (1990) vol. **147**, pp. 423-432.

[Col 92] Collins, J.J., Fanciulli, M. and Hohlfeld, R.G. *A random number generator based on the transform of the logistic variable.* Computer in Physics (1992), vol. **6**(6), pp. 630-632.

[Cov 91] Cover, T. and Thomas, J. *Elements of Information Theory.* John Wiley & Sons, Inc., New York (1991).

[Cp 91] Compagner, A. *Definitions of Randomness.* American Journal of Physics (1991) vol.

58(8), pp. 700-705.

[CS 85] Cavalier-Smith, T., *Eukaryotic gene numbers, non-coding DNA and genome size.* pp. 69-103 en: Cavalier-Smith, T. ed. *The evolution of genome size.* Wiley (1985).

[Cz 95] Czirók, A., Mantegna, R.N., Havlin, S. and Stanley, H.E. *Correlations in binary sequences and generalized Zipf analysis.* Physical Review E (1995), vol. **52**(1), pp. 446-452.

[Cz 96] Czirók, A., Stanley, H.E. and Vicsek, T. *Posible origin of power-law behavior in n-tuple Zipf analysis.* Physical Review E (1996), vol. **53**(6), pp. 6371-6375.

[Da 97] Dandliker, P.J., Holmlin, R.E. and Barton, J.K. *Oxidative Thymine Dimer Repair in the DNA Helix,* Science (1997) vol. **275**, pp. 1465-1468.

[Dau 88] Daubechies, I., *Orthonormal bases of compactly supported wavelets.* Communications on Pure and Applied Mathematics (1988), vol.**41**, pp. 909-996.

[Dau 92] Daubechies, I., *Ten Lectures on Wavelets.*, SIAM (1992).

[Des 80] Des Cloizeaux, J. *Short range correlations between elements of a long polimer in a good solvent.* Journal de Physique (1980), vol. **41**, pp. 223-238.

[Du 95] Dudek, M.R. and Cebrat, S. *Stochastic DNA in the presence of the coding bias.* Preprint (1995).

[Duj 96] Dujon, B., *The yeast genome proyect: what did we learn?* T.I.G. (1996), vol. **12**(7), pp. 263-270.

[Eb 90] Ebeling, W. and Volkenstein, M.V., *Entropy and the evolution of biological information,* Physica A (1990) vol. **163**, pp.398-402.

[El 74] Elton, R.A., *Theoretical models for heterogeneity of base composition in DNA.* Journal of Theoretical Biology (1974) vol. **45**, pp. 533-553.

[Ep 47] Epstein, B., *The mathematical description of certain breakage mechanisms leading to*

the logarithmico-normal distribution. Journal of Franklin Institute (1947), vol. **244**, pp. 275-288.

[Fr 96] Freund, J., Ebeling, W. and Rateitschak, K. *Self-Similar sequences and universal scaling of dynamical entropies.* Physical Review E (1996), vol. **54**(5), pp. 5561-5566.

[Fu 86] Furukawa, H., *Universal spectra of quasirandom objects produced by off-equilibrium space divisions.* Physical Review A (1986), vol. **34**(3), pp. 2323-2315.

[Ga 66] Gatlin, L. *The information content of DNA.* Journal of Theoretical Biology (1966), vol.**10**, pp. 281-300.

[Ga 72] Gatlin, L. *Information Theory and the Living System.* (1972) Columbia University Press, New York.

[Ge 94] Gell-Mann, M. *The Quark and the Jaguar*, W.H. Freeman and Co. (1994).

[Ge 96] Gell-Mann, M. and Lloyd, S. *Information Measures, Efective Complexity and Total Information.* Complexity (1996), vol. **2**. pp. 44-52.

[Go 87] Gonzalez, R.C., *Digital Image Processing.* Addison-Wesley (1987).

[Gu 95] Gu, X. and Li, W.H., *The size distribution of insertions and deletions in human and rodent pseudogenes suggest the logaritmic gap penalty for sequence alignement.* Journal of Molecular Evolution (1995) vol. **40** pp. 464-473.

[Gui 77] Guiasu, S. *Information Theory With Applications.* McGraw-Hill, New York (1977).

[Ha 85] Hamori, E. *Novel DNA sequence representations.* Nature (1985), vol. **314**, p. 585.

[Har 88] Hariri, A., Weber, B. and Olmsted III, J., *On the validity of Shannon-information calculations for molecular biológical sequences.* Journal of Theoretical Biology (1988), vol.**147**, pp. 235-254.

[Hat 85] Hattori M., Hidaka S., Sakaki Y., *Sequence analysis of a KpnI family member near the 3' end of human beta-globin gene.* Nucleic Acids Research (1985) vol. **13**, pp.7813-

7827.

[Hau 96] Haussdorff, J.M. and Peng, C.K. *Multiscaled randomness: A posible source of 1/f noise in biology.* Physical Review E (1996), vol. **54**(2), pp. 2154-2157.

[He 94a] Herzel, H., Ebeling, W. and Schmitt, O., *Entropies of Biosequences: The role of repeats.* Physical Review E (1994), vol **50**(6), pp. 5061-5071.

[He 94b] Herzel, H., Schmitt, A.O. and Ebeling, W. *Finite Sample Effects in Sequence Analysis.* Chaos, Solitons and Fractals (1994), vol. **4**(1), pp. 97-113.

[He 95] Herzel, H. and Groβe, I. *Measuring correlations in symbol sequences.* Physica A (1995), vol. **216**, pp. 518-542.

[He 97] Herzel, H. and Groβe, I. *Correlations in DNA Sequences: the Role of Protein Coding Segments.* Physical Review E, (1997) vol. **55**(1), pp. 800-810.

[Is 96] Israeloff, N.E., Kagalenko, M. and Chan, K. *Can Zipf Distinguish Language From Noise in Noncoding DNA?* Physical Review Letters (1996), vol. **76**(11), p. 1976.

[Ja 90] James, F. *A Review of Pseudo-Random Number Generators.* Computer Physics Communications (1990), vol. **60**(3), pp. 329-344.

[Je 90] Jeffrey, H.J. *Chaos game representation of gene structure.* Nucleic Acid Research (1990), vol. **18**, pp. 2163-2170.

[Ju 89] Jurka, J., ... Journal of Molecular Evolution (1989), vol.**29** pp.496- .

[Ka 93] Karlin, S. and Brendel, V. *Patchiness and Correlations in DNA Sequences.* Science (1993), vol. **259**, pp. 677-680.

[Kam 91] Kampis, G. *Self-Modifying Systems in Biology and Cognitive Science.* Pergamon Press, Oxford (1991).

[Kat 92] Kathlen, A.H., Nicholas, J.S. and Singh, S.M., *Chaos Game Representation of Coding Regions of Human Globin Genes and Alcohol Dehydrogenase Genes of Phylogenetically*

Divergent Species. Journal of Molecular Evolution (1992) vol.**35**, pp. 261-269.

[Ko 65] Kolmogorov, A.N., *Three approaches to the concept of the amount of information,* IEEE Prob. Info. Trans. (1965) vol. **1**, pp. 1-7.

[La 93] Larharmmar, D. and Chatzidimitriou-Dreismann, C.A., *Biological origins of long-range correlations and compositional variations in DNA.* Nucleic Acids Research (1993), vol. **21**(22) pp. 5167-5170.

[Le 95] Leong, P.M. and Morgenthaler, S., *Random walk and gap plots of DNA sequences.* CABIOS (1995), vol. **11**(5), pp. 503-507.

[Li 89] Li, W. *Spatial 1/f Spectra in Open Dynamical Systems.* Europhysics Letters (1989), vol. **10**(5), pp. 395-400.

[Li 90] Li, W. *Mutual Information Functions versus Correlation Functions.* Journal of Statistical Physics (1990), vol.**60**(5/6), pp. 823-837.

[Li 91a] Li, W., *Expansion-modification systems: A model for spatial 1/f spectra.* Physical Review A (1991), vol. **43**(10), pp. 5240-5260.

[Li 91b] Li, W., *On the relationship between complexity and entropy for Markov chains and regular laguages*, Complex Systems, (1991) vol. **5**(4), pp. 381-399.

[Li 92a] Li, W. and Kaneko, K. *DNA correlations.* Nature (1992), vol. **360**, pp. 635-636.

[Li 92b] Li, W., *Generating nontrivial long-range correlations and 1/f spectra by replication and mutation.* International Journal of Bifurcation and Chaos (1992), vol. **2**(1), pp. 137-154.

[Li 92c] Li, W. and Kaneko, K., *Long-range Correlations and Partial $1/f^{\alpha}$ Spectrum in a Noncoding DNA Sequence.* Europhysics Letters (1992), vol. **17**(7), pp. 555-660.

[Li 94] Li, W., Marr, T.G. and Kaneko, K., *Understanding Long-range Correlations in DNA Sequences.* Physica D (1994), vol. **75**, pp. 392-416.

[Li 96] Li, W., ... Complexity (1996), vol.**1**(6), pp.6-.

[Li 97a] Li, W. *Characterizing Correlation Structures of DNA Sequences by Spectral Analysis.* No publicado.

[Li 97b] Li, W. *The Measure of Compositional Heterogeneity in DNA Sequences Is Also A Measure of Complexity* Enviado a Complexity.

[Li 97c] Li, W., Bernaola Galván, P., Stolovitzky, G., Oliver, J.L. *Compositional Heterogeneity Within, and Uniformity Between, DNA Sequences of Yeast Chromosomes* En preparación.

[Li 97d] Li, W., *The Study of Correlation Structures of DNA: A Critical Review.* Computer and Chemistry (1997). En prensa.

[Ld 68] Lindenmayer, A., *Mathematical models for cellular interactions in development. I and II.* Journal of Theoretical Biology (1968), vol.**18**, pp. 280-299,300-315.

[Lin 91] Lin, J. *Divergence measures based on the Shannon Entropy.* IEEE Transactions on Information Theory (1991), vol. **37**(1) pp. 145-151.

[Lio 96] Liò, P., Politi, A., Buiatti, M., Ruffo, S., *High Statistics Block Entropy Measures of DNA-Sequences* Journal of Theoretical Biology (1996), vol. **180**(2), pp. 151-160.

[Lis 96] Lisý, V. and Miškovský P. *Comment on "Two-component localized solutions in a nonlinear DNA model".* Physics Letters A (1996), vol. **213**, pp. 308-310.

[Llo 88] Lloyd, S. and Pagels, H. *Complexity as thermodydamic depth* Annals of Physics (1988), vol. **188**, pp. 186-213.

[Lo 93] Lowen, S.B. and Teich, M.C. *Fractal renewal processes generate 1/f noise.* Physical Review E (1993), vol. **47**(2), pp. 992-1001.

[Lof 77] Löfgren, L., *Complexity of description of systems: A foundational study.* Int. J. Gen. Sys. (1977), vol. **3**, pp. 197-214.

[Ma 94] Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M. and Stanley, H.E. *Linguistic Features of Noncoding DNA Sequences.* Physical Review Letters (1994), vol. **73**(23), pp. 3169-3172.

[Mad 92] Maddox, J. Nature (1992) (Scientific Correspondence) vol. **358** p. 103.

[Mak 96] Makse, H.A., Havlin, S., Shwartz, M. and Stanley H.E. *Method for generating long-range correlations for large systems.* Physical Review E (1996), vol. **53**(5), pp. 5445-5449.

[Man 96] Manuca, R. and Savit, R. *Stationarity and nonstationarity in time series analysis.* Physica D (1996), vol. **99**, pp. 134-161.

[Mas 87] Mansuripur, M., *Introduction to Information Theory.* Prentice-Hall, Inc., Englewood Cliffs, N.J.(1987).

[Mc 96] McShea, D.W., *Metazoan complexity and evolution: Is there a trend?* Evolution (1996), vol. **50**(2), pp. 477-492.

[Mun 92] Munson, P.J., Taylor, R.C. and Michaels, G.S. *Long range DNA correlations extends over entire cromosome.* Nature (1992), vol. **360**, p. 636.

[Nee 92] Nee, S., Nature (1992), (Scientific Correspondence), vol.**357**, pp. 450.

[Ol 93] Oliver, J.L., Bernaola-Galván, P., Guerrero-García, J. and Román-Roldán, R. *Entropic Profiles of DNA Sequences Through Chaos-game-derived Images.* Journal of Theoretical Biology (1993), vol. **160**, pp. 457-470.

[Os 94] Ossadnik, S.M., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N. and Peng, C.-K., *Correlation Approach to Identify Coding Regions in DNA Sequences.* Biophysical Journal (1994), vol.**67**, pp. 64-70.

[Pa 80] Papentin, F., *On order and complexity I: General considerations.* Journal of Theoretical Biology (1980), vol. **87**, pp. 421-456.

[Pa 82] Papentin, F., *On order and complexity II: Applications to chemical an biochemical*

structures. Journal of Theoretical Biology (1982), vol. **95**, pp. 225-245.

[Pe 92] Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Scortino, F., Simons, M. and Stantley, H.E. *Long-range correlations in nucleotide sequences.* Nature (1992), vol. **356** pp. 168-170.

[Pe 93a] Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Simons, M. and Stanley, H.E. *Finite-size effects on long-range correlations: Implications for analyzing DNA sequences.* Physical Review E (1993), vol. **47**(5), pp. 3730-3733.

[Pe 93b] Peng, C.K., Mietus, J., Hausdorf, J.M., Havlin, S., Stanley, H.E. and Goldberger, A.L. *Long-range Anticorrelations and Non-Gaussian Behavior of the Heartbeat.* Phsyscal Review Letters (1993), vol. **70**(9), pp. 1343-1346.

[Pe 94] Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stantley, H.E. and Goldberger, A.L. *Mosaic organization of DNA nucleotides.* Physical Review E (1994), vol. **49**(2), pp. 1685-1689.

[Pei 92] Peitge, H.O., Jurgens, H. and Saupe, D. *Fractals for the Classroom* (1992).

[Per 93] Percival, D.B. and Walden, A.T., *Spectral Analysis for Physical Applications.* Cambridge University Press (1993).

[Pr 94] Press, W.H., Flannery, B.P., Teuloksy, S.A. and Vetterling, W.T. *Numerical Recipes in Pascal.* Cambridge University Press, Cambridge (1994).

[Ra 96] Rama Kant, *Statistics of approximately self-affine fractals: Random corrugated surface and time series.* Physical Review E (1996) vol. **56**(6), pp. 5749-5763.

[Raf 94] Raffery, A., Tavaré, S., *Estimation and modeling repeated patterns in high order Markov chains with the mixture transition distribution model.* Applied Statistics (1994), vol.**43**(1), pp. 179-199.

[Re 73] Reichert, T.A., Cohen, D.N., Wong, A.K.C. *An Application of Information Theory to Genetic Mutations and the Matching of Polypeptide Sequences.* Journal of Theoretical

Biology (1973) vol. **42**, pp. 245-261.

[Rei 65] Reif, F., *Fundamentals of Statistical and Thermal Physics* McGraw-Hill, New York (1965).

[Ro 91] Román-Roldán, R., Quesada-Molina, J.J. and Martínez-Aroza, J., *Multirresolultion-Information Analysis for Images.* Signal Processing (1991), vol.**24**, pp. 77-91.

[Ro 93a] Román Roldán, R. *Análisis de secuencias de ADN mediante Teoría de la Información.* En: *Homenaje a Safwan al Khouri Ibrahim.* Universidad de Granada, 1993.

[Ro 93b] Román Roldán, R., Bernaola Galván, P. and Luque Escamilla, P. *Entropic Vector Feature of Sequences.* Proceedings of the 1993 Workshop on Information Theory.

[Ro 94] Román-Roldán, R. Bernaola-Galván, P. and Oliver, J.L. *Entropic feature for sequence pattern through iterated function systems.* Pattern Recognition Letters (1994), vol. **15**, pp. 567-573.

[Ro 96] Román Roldán, R, Bernaola Galván, P., Oliver J.L. *Application of Information Theory to DNA sequence analysis: a review.* Pattern Recognition (1996) vol. **29**(7), pp. 1187-1194.

[Ro 97] Román Roldán, R, Bernaola Galván, P., Oliver J.L. *DNA compositional complexity through an entropic segmentation algorithm.* Enviado a Physical Review Letters (1997).

[RP 95] Robles-Pérez, A., Ben-Hamza, Martínez-Aroza, J. and Román Roldán, R. *The Maximum Value of the Jensen-Shannon Divergence.* No publicado (1995).

[Rob 74] Robinson, F.N.H., *Noise and fluctuations* Clarendon, Oxford (1974).

[Rom 74] Rombauer, I.S. and Becker, M.R. *Joy of cooking* Signet Books (1974).

[Ros 93] Rosu, H. and Canessa, E. *Solitons and 1/f noise in molecular chains.* Physical Review E (1993), vol. **47**(6), pp. R3818-R3821.

[Roz 80] Rozenberg, G., Salomaa, A., *The Mathematical Theory of L Systems.* Academic Press (1980).

[Sa 96] Sáiz, A. and Martínez, V.J. *Multifractal analysis of the atomic spectral line series.* Physical Review E (1996), vol. **54**(3), pp. 2431-2437.

[Sc 96] Schmitt, A.O., Ebeling, W. and Herzel, H. *The modular structure of informational sequences.* Biosystems (1996), vol. **37**, pp. 199-210.

[Sr 91] Schroeder, M., *Fractals, Chaos, Power Laws. Minutes from an infinite paradise.* Ed. W.H. Freeman adn Co., New York (1991).

[Sch 75] Schultz, D.M., *Mass-size distributions: a review and a proposed new model.* en *Statistical distributions in scientific work.* editores: Patil, Kotz & Ord, D. Reidel Publishing (1975), pp. 275-288.

[Se 92]

[Sh 94] Shnerb, N. and Eisenberg, E. *Analyzing long-range correlations in finite sequences.* Phsysical Review E (1994), vol. **49**(2), pp. R1005-R1008.

[Sha 90] Shakhnovich, E.I. and Gutin, A.M., ... Nature (1990), vol.**346**, pp. 773-.

[Shl 86] Shlesinger, M.F. and Klafter, ... en *On Growth and Form: Fractal and Non-Fractal Pattern in Physics.* Ed. Stanley H.E. and Ostrowsky N., Martinus Nijhoff (1986).

[Si 89] Sibbald, P.R., Banerjee, S. and Maze, J., *Calculating higher order DNA sequence information measures,* Journal of Theoretical Biology (1988) vol. **136**, pp. 457-482.

[Sn 96] Snarskii, A.A., Morozovsky, A.E., Kolek, A. and Kusy, A. *1/f noise in percolation and percolation-like systems.* Physical Review E (1996), vol. **53**(6), pp. 5596-5605.

[So 96] Sousa Vieira, M. and Herrmann, H.J. *A growth model for DNA evolution.* Europhysics Letters (1996), vol. **33**(5), pp. 409-414.

[Sp 72] Sparrow, A.H., Price, H.J., and Underbrink, A.G. *A survey of DNA content per cell*

and per chromosome of prokaryotic and eukaryotic organism: Some evolutionary considerations. pp. 451-494 en: Smith, H.H. ed. *Evolution of genetic systems.* vol. **23**, Gordon and Breach (1972).

[St 94] Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Goldberger, Z.D., Havlin, S., Mantegna, R.N., Ossadnick, S.M., Peng, C.K. and Simons, M. *Statistical mechanics in biology: how ubiquitous are long-range correlations?* Physica A (1994), vol. **205**, pp. 214-253.

[Str 95] Strait, B.J. and Dewey, T.G. *Multifractals and decoded walks: Applications to protein sequence correlations.* Physical Review E (1995), vol. **52**(6), pp. 6588-6592.

[Sz 95] Szathmáry, E., and Maynard Smith, J. *The major evolutionary transitions.* Nature (1995), vol. **374**, pp. 227-232.

[Ta 94] Taylor, H.M. and Karlin, S., *An Introduction To Stochastic Modeling.* Academic Press Ltd. (1994).

[Tav 89] Tavaré, S., Song, B., *Codon preference and primary sequence structure of nucleotide sequences.* Bulletin of Mathematical Biology (1989), vol.**51** pp. 95-115.

[Te 96] Teitelman, M. and Eeckman, F.H., *Principal Component Analysis and Large-Scale Correlations in Non-Coding Sequences of Human DNA.* Journal of Computational Biology (1996), vol. **3**(4), pp. 573-576.

[Ts 95] Tsonis, A.A. and Elsner, J.B. *Testing for Scaling in Natural Forms and observables.* Journal of Statistical Physics (1995), vol. **81**(5/6), pp. 869-880.

[Ts 96] Tsonis, A.A., Kumar, P., Elsner, J.B. and Tsonis, P.A. *Wavelet analysis of DNA sequences.* Physical Review E (1996), vol. **53**(2), pp. 1828-1834.

[Vi 97] Viswanathan, G.M., Buldyrev, S.V., Havlin, S. and Stanley, H.E. *Quantification of DNA Patchiness Using Long-Range Correlation Measures.* Biophysical Journal (1997) vol. **72**, pp. 866-875.

[Vo 75] Voss, R.F. and Clarke, J. *1/f noise in music and speech.* Nature (1975), vol. **257**(27), pp. 317-318.

[Vo 88] Voss, R.F. *Fractals in nature: From characterization to simulation* En: *The Science of Fractals Images* Ed. Peitgen, H.O. and Saupe, D., Springer-Verlag (1988).

[Vo 92] Voss, R.F. *Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences.* Physical Review Letters (1992), vol. **68**(25), pp. 3805-3808.

[Vo 93] Voss, R.F., Réplica al artículo [Bu 93c]. Physical Review Letters (1993), vol. **71**(11), p. 1777.

[Vo 94] Voss, R.F., *Long-range fractal correlations in DNA introns and exons.*, Fractals (1994), vol. **2**(1), pp. 1-6.

[Vo 96] Voss, R.F. *Comment on "Linguistic Features of Noncoding DNA Sequences".* Physical Review Letters (1996), vol. **76**(11), p. 1978.

[We 88] Weber, B.H., Depew, D.J. and Smith J.D. (ed.), *Entropy, Information and Evolution.*, (1988), M.I.T. Press, Massachusetts.

[Weg 90] Wegrzyn, S., Gille, J.C. and Vidal, P., *Development Systems.* Springer-Verlag (1990).

[Wei 88] Weissman, M.B. *1/f noise and other slow, nonexponential kinetics in condensed matter.* Review of Modern Physics (1988), vol. **60**(2), pp. 537-570.

[Wes 90] West, B.J., *Fractal Physiology and Chaos in Medicine.* (1990) Word Scientific, Singapore.

[Wes 95] West, B.J. *Fractal Statistics in Biology.* Physica D (1995), vol. **86**, pp. 12-18.

[Wi 87] Wicken, J.S., *Evolution, Thermodynamics, and Information.* (1987) Oxford University Press, New York.

[Wil 97] Wilson, E.K., *DNA - Insulator or Wire.* Chemical & Engineering News (1997), vol. **75**(8), pp. 33-34.

[Ya 92] Yam, P., P. Sci. Am. (1992) vol. **267** p. 23.

[Zha 95] Zhang, *Exploratory analysis of long genomic DNA sequences using the wavelet transform: examples using polymavirus genomes.* en Genome Sequencing and Analisys Conference VI (1995), Mary Ann Liebert.