Resumen

Hemos estudiado las distribuciones de frecuencia de longitudes y aquellos de longitudes de la correlación (d_c) de las secuencias de la codificación (exones) y de las secuencias de la no-codificación (intrones) para la DNA de seis especies de organismos vivos en el dominio del Eucarionte y obtuvimos los siguientes resultados:1) En el caso de los exones para cada especie de los organismos vivos todas las longitudes son distribuidas dentro del punto de ebullición limitando en la región < 8000 bp, mientras que las longitudes de la correlación de los exones se distribuyen sobre todo en distancias cortas con d_c < 10. Las longitudes de la correlación y las longitudes de los exones parecen no haber sido afectadas prácticamente en los procesos evolutivos. 2) En el caso de los intrones una gran diferencia se ha observado en las distribuciones de las longitudes reales, así como, en las que se presentan en las longitudes de la correlación dependiendo de la especie de los organismos vivos. Ambos se han encontrado para ser alargados de acuerdo con la orden del árbol evolutivo. Las longitudes de la correlación se distribuyen para arriba hasta una región de largo alcance que exceda de d_c ~ 500 uniformemente dependiendo de la especie de organismos vivos.

Abstract

We studied the distribution of the frequency of longitudes, that is, correlation longitudes (d_c) of the codification sequences (exons) and longitudes of the non-codification sequences (introns) for the DNA of six species of living organisms in the domain of the Eucarionte and we obtained the following results: 1) In the case of the exons, for each species of living organisms, all the longitudes are distributed within the point of boiling in the region of < 8000bp, while the correlation longitudes of exons are mostly distributed in short distances with d_c < 10. The correlation longitudes and the longitudes of exons appear not to have been affected by the evolutionary process. 2) In the case of the introns a great difference was observed in the distribution of the actual lengths, as well as those that are presented in the correlation lengths depending on the species of the living organisms. Both have been found to be elongated in accordance with the order of the evolutionary tree. The correlation lengths are distributed up till a long-range region which exceeds even d_c~500 depending on the species of living organisms.

Abstrait

On a étudié les distributions de fréquence de longitudes et ceux de longitudes de la corrélation (d_c) des séquences de la codification (exons) et des séquences de la non-codification (introns) de la DNA de six espèces d'organismes vivants dans le domaine de l'Eucarionte et on a obtenu les résultats suivants: 1) Dans le cas des exons pour chaque espèce des organismes vivants, toutes les longitudes sont distribuées à l'intérieur du point d'ébullition limitant dans la région < 8000 bp, alors que les longitudes de la corrélation des exons sont distribuées surtout en distances courtes avec d_c<10. Les longitudes de la corrélation et les longitudes des exons ne paraissent pratiquement pas avoir été affectées dans les processus évolutifs. 2) Dans le cas des introns, une grande différence a été observée dans les distributions des longitudes réelles, ainsi que dans celles qui se présentent dans les longitudes de la corrélation dépendant de l'espèce des organismes vivants. Les deux ont été mises en contact pour être allongées en accord avec l'ordre de l'arbre évolutif. Les longitudes de la corrélation sont distribuées vers le haut jusqu'à une région de longue portée qui dépasse d_c ~ 500 uniformément dépendant de l'espèce d'organismes vivants.

Keywords: Eucarionte, DNA, Base Sequence, Mutual Information Function, Correlation Length.

Palabras claves: Eucarionte, ADN, Secuencia Base, Función Mutua de la Información, Longitud de la Correlación.

* Yasuhiko Isohata
* Masaki Hayashi
**Honorio Vera Mendoza

* Tokyo University of Pharmacy and Life Science School of Life Science

** Universidad Tecnológica de Puebla

Sección 1. Introducción

Las correlaciones de largo alcance en los genes se basan en las secuencias [1]-[5] y en los patrones de la repetición, es decir, las periodicidades en secuencias de bases en las distancias de corto alcance [6] han sido estudiadas para varios genes usando el método del espectro de potencia basado en

el análisis de Fourier. Por ejemplo, las correlaciones de largo alcance en secuencias de bases del gen de S. Cerevisiae fueron comparadas con los del homo sapience. Se confirmó que las fuertes correlaciones de largo alcance existen en el caso del homo sapience comparado con aquellos del S. Cerevisiae. De tal análisis, uno podría en principio, conseguir información interesante sobre la evolución de los organismos vivos [6]. Funciones mutuas de la información (de aquí en adelante denotadas como MIF) si estadísticamente se analizó podríamos derivar nuevas estructuras interesantes de correlación en las secuencias de bases [7]. El MIF esta basado en la definición de la entropía de Shannon. Si los símbolos a y b (= A, G,C, T) aparecen en una secuencia base que tiene una distancia d entre ellos, entonces el MIF se define como:

$$M(d) = \sum_{a,b=A,G,C,T} P_{a,b}(d) \log_2 \left(P_{a,b}(d) / P_a P_b \right) \qquad (1)$$

Donde P_a y P_b denoten una probabilidad de la ocurrencia de símbolos a y b (= A, G, C, T)) en la secuencia, respectivamente, mientras que $P_{a,b}$ (b) denota una probabilidad condicional para los símbolos a y b parecen que tiene una distancia d en la secuencia. P_a P_b , y $P_{a,b}$ (d) son calculados según las estadísticas a lo largo de la secuencia dadas. Por las aplicaciones prácticas de la Eq. (1) para calcular el MIF M(d), es más conveniente reescribirlo como:

$$M(d) = \sum_{a,b=A,G,C,T} P_{a,b}(d) \log_2 \left(P_{a,b}(d) / P_a P_b \right) - \sum_{a=A,G,C,T} P_a(d) \log_2 P_a(d) \left(\sum_{b=A,G,C,T} P_b \log_2 P_b \right) - \sum_{b=A,G,C,T} P_b \log_2 P_b \Big). \qquad (2)$$

En lo sucesivo como un ejemplo para calcular el MIF M(d) consideramos el caso de gen supresor tumoral Von Hippel-Lindau del homo sapiens (VHL) (la longitud o el número de la secuencia base: 14543 bp). Exhibimos el resultado del cálculo para el MIF de esta muestra por una gráfica en Fig. 1. Para la comparación también exhibimos en la misma figura una gráfica del MIF para una secuencia aleatoria que posee la misma longitud y la misma velocidad como la secuencia baja de la muestra.

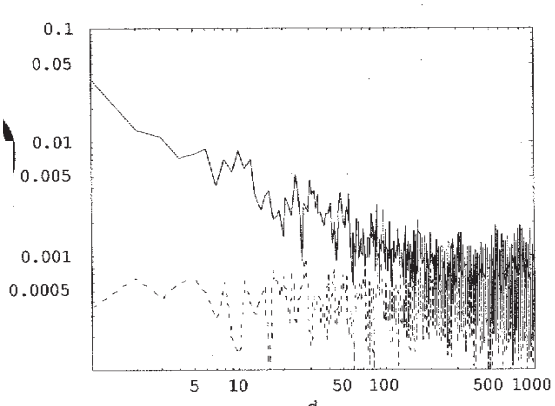


Fig. 1. El MIF para el caso de la muestra del gen supresor tumoral Von Hippel-Lindau del homo sapiens (VHL) (el número de bases: 14543 bp). La ordenada muestra que el MIF (d) y la abscisa muestra la distancia d entre las dos bases relevantes. La línea sólida corresponde al MIF de la secuencia de la muestra mientras que la línea discontinua corresponde al MIF de una secuencia aleatoria.

Encontramos una diferencia notable entre el MIF de las secuencias de la base del gen M_{Gene}(d) y el de la secuencia aleatoria M_{Random} (d) en las distancias con d<200. El MIF M_{Random} (d) no depende sobre una distancia d y fluctúa alrededor de cero. Por otra parte, el MIF M_{Gene}(d) gradualmente decrece con un incremento de d dentro de una región donde d< 200. En la otra región donde d>200 tenemos

$$M_{Gene}(d) \sqcup M_{Random}(d). \qquad (3)$$

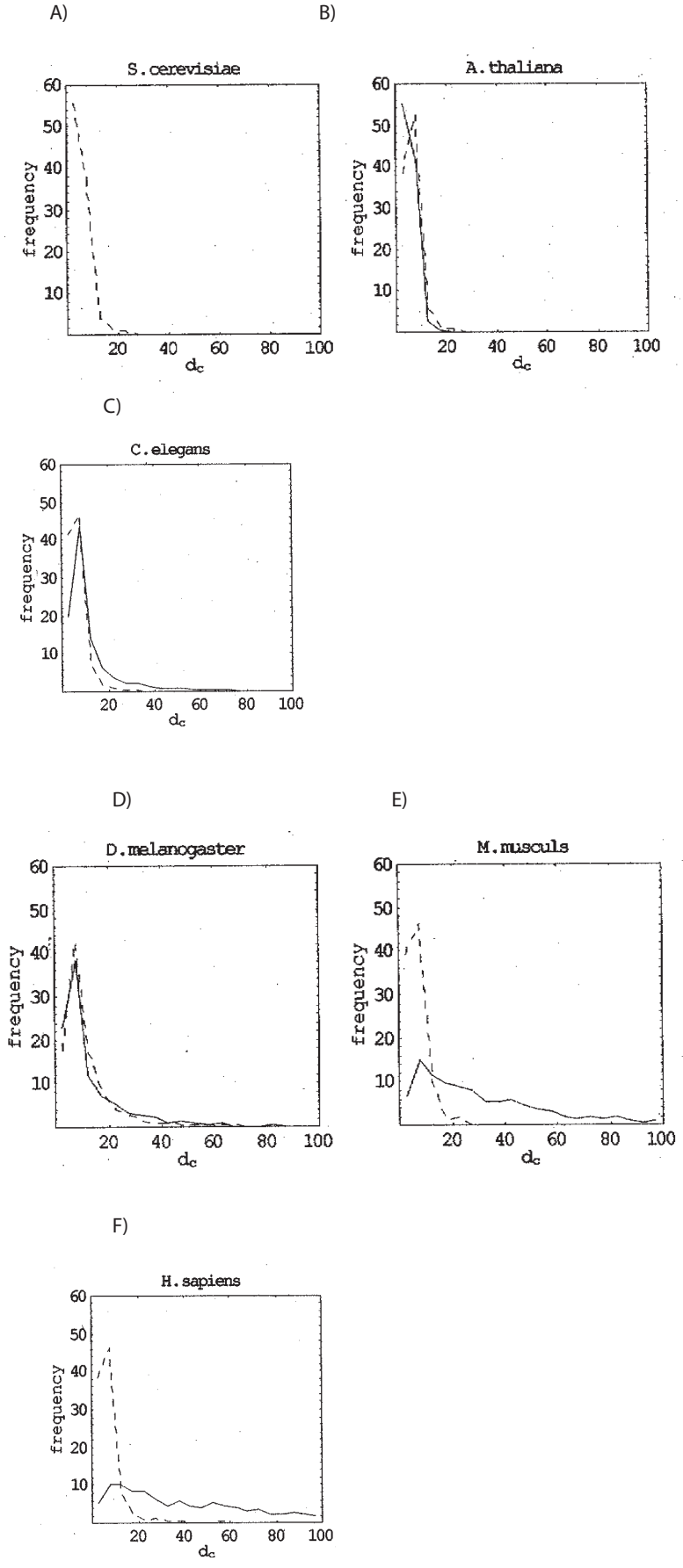
Consecuentemente definimos en este caso una longitud de la correlación que caracteriza la estructura de la correlación en el MIF como d_c 200. Un valor finito diferente de cero de d_c es una reflexión de la cantidad de información correlacionada contenida en las secuencias dadas de la base del gen.

Sección 2. Las distribuciones de la longitud y las distribuciones de la longitud de la correlación de los exones y los intrones de las secuencias bases del gen en el dominio del Eucarionte

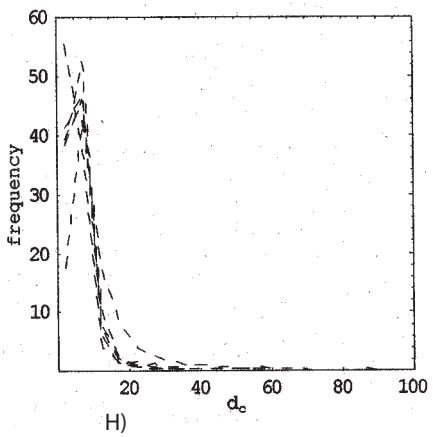
Adoptando la técnica descrita en la sección 1 hemos calculado y comparado las longitudes de la correlación d_c de los exones y los intrones de las secuencias

bases del gen para 6 especies de los organismos vivos en el dominio del Eucarionte (células con núcleos), es decir A) S.cerevisiae: 1798, B) A.thaliana: 2582, C) C.elegans: 7743, D) D.melanogaster: 2564, E) M.musculs: 903, y F) H.sapiens:1184 (los números refieren a sus genes respectivos). Las distribuciones de frecuencia respectivas se muestran en Fig. 2. Hemos seleccionado las muestras de la base de datos del ADN de NCBI GenBank [8] de una manera tal que las longitudes de las secuencias bases del gen sean más largas que 1000 bp. Con el fin de mejorar el nivel de la confiabilidad en la determinación de las longitudes de la correlación, hemos elegido cinco secuencias aleatorias para cada secuencia del gen y hemos definido la longitud de la correlación de la secuencia del gen como el promedio sobre el valor calculado (cinco veces). Para la comparación, hemos exhibido también los gráficos de los exones y los intrones para varias especies de los organismos vivos compilándolos en conjunto en las mismas gráficas de fig. 2 G y de fig. 2 H, respectivamente. En el caso de los exones no vemos una diferencia clara en las distribuciones de sus longitudes de correlación en las gráficas de la Fig. 2 G. Muchos de ellos están concentrados en distancias cortas dentro de d_c <10. (sin embargo, una excepción es el caso de D. Melanogaster en el cual la distribución se extiende hasta d_c~60.

La mediana de las longitudes de la correlación es también comparativamente más grande que la de otras especies como se muestra en la Tabla 1.) Por otra parte, en cuanto a los intrones una diferencia grande se ha observado en las distribuciones de las longitudes de la correlación d_c para cada especie de los organismos vivos en el dominio del Eucarionte como lo demuestra en la gráfica de la Fig. 2 H. Las distribuciones de las longitudes de la correlación para las diversas especies se han encontrado para ser alargadas de acuerdo con la orden conocida del árbol evolutivo. Sin embargo se distribuyen en una región de largo alcance dependiendo sobre una especie de los organismos vivos. En particular ésto excede hasta ~500 en el caso de H.Sapiens



G)



H)

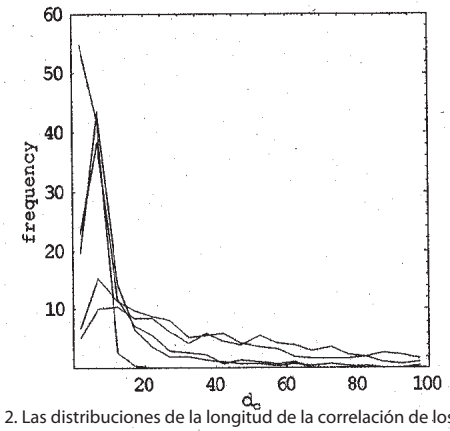


Fig. 2. Las distribuciones de la longitud de la correlación de los exones y los intrones para los organismos vivos: A)~F) (A) S.cerevisiae, B) A.thaliana, C) C.elegans, D) D.melanogaster, E) M.musculs, y F) H.sapiens). La ordenada demuestra que la frecuencia (la razón en % de los números de genes con respecto al número total de los genes respectivamente analizados) y la abscisa corresponde a la longitud de la correlación d_c. Las líneas sólidas corresponden a las distribuciones de la longitud de la correlación de los intrones mientras que las líneas discontinuas a las de los exones. Las gráficas de los exones y los intrones de varias especies de los organismos vivos se compilan también en las mismas figuras de fig. 2 G y de fig. 2 H, respectivamente.

	Exones	Intrones
S.cerevisiae	4.6	----
A.thaliana	5.6	4.6
C.elegans	5.4	7.8
D.melanogaster	8.4	7.8
M.musculs	5.4	26.2
H.sapiens	5.4	46.2

Tabla 1. Las medianas de las longitudes de la correlación d_c para varios organismos vivos en el dominio del Eucarionte.

Sería interesante examinar la relación entre las longitudes reales de las secuencias bases para varias especies de los organismos vivos en el dominio del Eucarionte y las longitudes correspondientes de la correlación d_c. Para este propósito hemos exhibido en la Fig. 3 los valores promedios (bp) de las longitudes reales de los exones y de los intrones, respectivamente. En esta figura observamos un aumento notable de las longitudes medias de los intrones en la etapa evolutiva de D. Melanogaster M. Musculs H. Sapiens. Esto corresponde exactamente al comportamiento similar observado en la Fig. 4 en donde hemos mostrado los gráficos para las medianas de las longitudes de la correlación de los intrones.

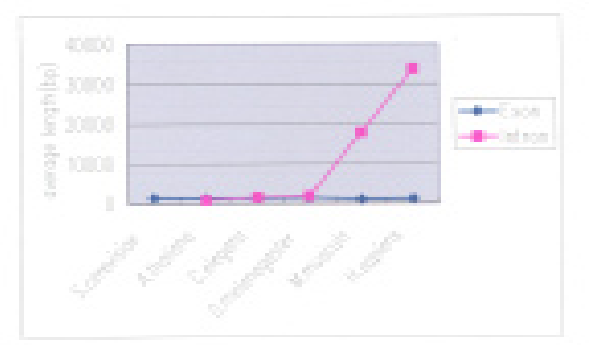


Fig. 3. Las longitudes promedio (bp) de las secuencias de bases para varios organismos vivos en el dominio del Eucarionte.

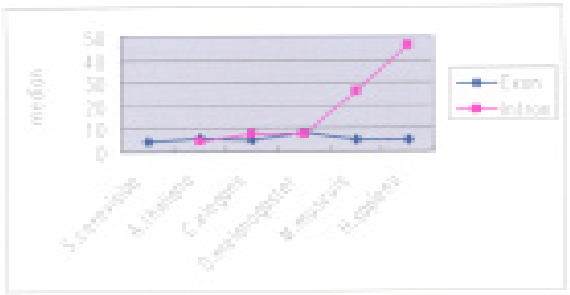
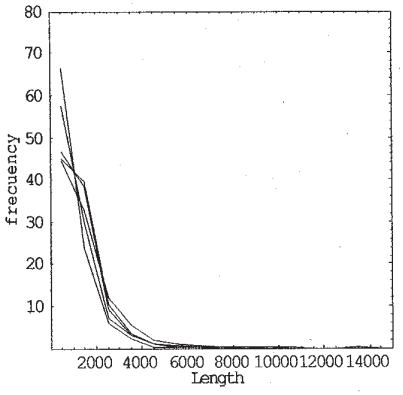


Fig. 4. Las medianas de las longitudes de la correlación para varios organismos vivos en el dominio del Eucarionte.

Además, hemos demostrado la longitud de distribuciones de los exones y los intrones para varios organismos vivos en Fig. 5 A y fig. 5 B, respectivamente. Observamos que los comportamientos de estas distribuciones y en algunas de las longitudes de correlación para los exones y los intrones se exhibieron en fig. 2 G y Fig. 2 H, son respectivamente similares el uno al otro, aunque las escalas son diferentes. En el caso de los exones las longitudes de la correlación y sus longitudes actuales tienen la tendencia a ser distribuidas dentro

de un rango más corto, cerca de las secuencias aleatorias que indican que no se han afectado mucho en los procesos evolutivos. Por otra parte en el caso de los intrones las longitudes de correlación, así como, sus longitudes actuales se distribuyen a lo largo de un rango mucho más largo. Se distribuyen más largos de acuerdo con la orden del árbol evolutivo. Uno puede concluir que los alargamientos de las longitudes de la correlación de las secuencias de bases del gen han sido afectados grandemente por los alargamientos de los intrones en los procesos evolutivos. Esto debe darnos la información importante sobre el mecanismo del alargamiento de las secuencias bases de la ADN en los procesos evolutivos.

A)



B)

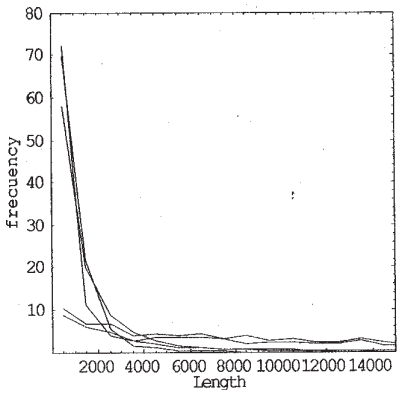


Fig. 5. Las distribuciones de frecuencia en la longitud de los exones y de los intrones de las secuencias bases para varias especies de los organismos vivos. Las gráficas en Fig. 5 A corresponden a los casos de los exones, mientras que las gráficas en Fig. 5 B a los intrones. La ordenada muestra que la frecuencia (la razón en % del número de genes con respecto al número total de genes analizado) y la abscisa en los gráficos en Fig. 5 A corresponde a las longitudes de las secuencias de bases del gen y en las gráficas en Fig. 5 B para sus longitudes de la correlación, respectivamente.

Sección 3. Resumen y Discusiones

En este artículo hemos comparado las distribuciones de la longitud de correlación de los exones y de los intrones y las longitudes de distribuciones correspondientes para las 6 especies de organismos vivos en el dominio del Eucarionte, es decir A) S.cerevisiae, B) A.thaliana, C) C.elegans, D) D.melanogaster, E) M.musculs, y F) H.sapiens, que hemos exhibido en Fig. 2 G y Fig. 2 H, y en Fig. 5 A y Fig. 5 B, respectivamente. Los resultados del análisis pueden ser resumidos como sigue:

1) En el caso de los exones para varias especies de organismos vivientes analizados no se podía observar una diferencia clara en las distribuciones de las longitudes de correlación. Se concentran de preferencia en distancias cortas con d_c =10 y están cerca con las de una secuencia aleatoria (véase Fig. 2 G). Una excepción aparece en el caso del D. Melanogaster en el cual la mediana de las longitudes de correlación del exon es comparativamente más alto que los de otras especies de organismos vivos y de la distribución, se ha demostrado que se extiende hasta d_c~60 (véase Fig. 2 D).

2) En el caso de los intrones, sin embargo, hemos observado una diferencia clara en las distribuciones de frecuencia de las longitudes de correlación d_c que se encuentran para ser alargadas de acuerdo con la orden conocida del árbol evolutivo. Por otra parte las longitudes de correlación de los intrones se distribuyen en regiones de rango largo dependiendo de la especie de los organismos vivos (véase Fig. 2 H). En el caso del H. Sapiens se extienden hasta el d_c~500. La comparación de las medianas de las longitudes de correlación d_c para varias especies también apoya esta observación (véase Fig. 3 y la Tabla 1).

3) En cuanto a los valores promedios de la longitud base de la secuencia de los intrones de las respectivas especies de los organismos vivos, hemos encontrado una tendencia de incremento de acuerdo con la orden conocida de los procesos evolutivos. Observamos un incremento particular en la etapa de D.melanogaster, M.musculs, H.sapiens (véase Fig. 3). Esto corresponde al comportamiento de un gráfico para los puntos medios de las longitudes de correlación de los intrones demostrados en Fig. 4.

4) Aunque las escalas son diferentes, los comportamientos funcionales de las distribuciones de la longitud de los exones y los intrones mostrados por las

gráficas en Fig. 5 A y Fig. 5 B son similares a otro de distribuciones de longitudes de correlación demostradas por las gráficas en Fig. 2 G y Fig. 2 H, respectivamente. Encontramos que las longitudes de correlación de los exones igualmente como sus longitudes actuales para todos los organismos vivos analizados están distribuidas dentro de un rango corto cerca de los de las secuencias aleatorias que indican que no se han afectado mucho en los procesos evolutivos. Por otra parte las longitudes de correlación de los intrones, así como, sus longitudes actuales se distribuyen a lo largo de una gama mucho más larga. Se distribuyen más de largo de acuerdo con la orden del árbol evolutivo.

Nuestros análisis nos conducen a la siguiente hipótesis:los genes de los organismos vivos han sido alargados principalmente por los intrones mientras que adquieren la estructura de correlación en secuencias de bases del ADN en los procesos evolutivos. Sin embargo, puesto que los números de las especies de los organismos vivos tratados en nuestro análisis se han restringido a solamente seis casos, un análisis más sistemático aumentando los números de datos y combinando el análisis tales como los que están basados en el espectro de potencia, permitiría deducir más información convincente sobre las evoluciones de los organismos vivos [9].

Referencias

[1] W. Li.
1989 Europhys. Lett. 10 395-400
W. Li
1991 Phys. Rev. A43 5240-5260.
[2] W. Li and K. Kaneko
1992 Europhys. Lett. 17 655-660.
[3] C. K. Peng, et al.
1992 Nature 356 168-170.
[4] R. Voss.
1992 Phys. Rev. Lett. 68 3805-3808.
[5] E. Takushi, M. Yogi, and C. Yamada
1998 Bulletin of the Faculty of Science, University of the Ryukyus 65 31-38
E. Takushi and H. Miyagi
2000 ibid. 70 43-46.
[6] Y. Isohata and M. Hayashi
2003 J. Phys. Soc. Jpn, 72, No.3 en prensa.
[7] S. Guharay,B. R. Hunt,J. A. Yorke and O. R.
2000 White,Physica D146 388-396.
[8] Dirección de NCBI (Centro Nacional de Información Biotecnológica)ver en http: //www.ddbj. nig.ac.jp/ Welcome-j.html.