BIOLOGIA COMPUTACIONAL

Andrés Moya (editor)

Universitat de València (Estudi General)

Uso de codones y preferencias de dinucleótidos: un estudio de las diferencias de composición dentro del genomio humano.

A. Marln, G. Gutiérrez y J. L. Oliver *

Departamento de Genética. Universidad de Sevilla.
* Departamento de Genética. Universidad de Granada.

Introducción

El genomio humano está formado por unos tres mil millones de nucleótidos distribuidos en 24 cromosomas: 22 autosomas y los cromosomas sexuales X e Y. El modelo de doble hélice del ADN requiere cantidades equimolares de nucleótidos de citosina (C) y de guanina (G), y también de nucleótidos de adenina (A) y timina (T), el ADN humano tiene un 37'5 % de G+C. Las primeras determinaciones de composición de los ácidos nucléicos pusieron de manifiesto que el contenido de G+C podía diferir mucho de unas especies a otras, pero se pensaba que dentro de cada genomio el contenido de G+C era constante. Ahora sabemos que en los genomios de los vertebrados homeotermos existen grandes diferencias de composición entre diferentes segmentos del mismo genomio. En este articulo examinaremos las variaciones de composición dentro del genomio humano relacionadas con el uso de codones sinónimos y con la frecuencia de dinucleótidos.

Las mejoras en las técnicas de tinción de los cromosomas introducidas a partir de los años 70, permitieron observar que los cromosomas humanos no son homogéneos por lo que respecta a sus propiedades de tinción. Después de una serie de tratamientos más o **menos** drásticos es posible observar un patrón de bandas que es característico de cada cromosoma. Cuando se utiliza el método de bandas G, el cariotipo humano se puede dividir en dos compartimentos: el ocupado por las bandas G y el complementario ocupado por las bandas R.

El material genético incluido en esas dos clases de bandas tiene características diferentes: 1) Las bandas G contienen secuencias de ADN con contenidos de G+C más bajo que las bandas R. 2) los genes **que sólo se expresan en algunos tejidos o los** que sólo se expresan en algunos estadios del desarrollo tienden a localizarse en bandas G, mientras que los de expresión más generalizada suelen estar en las

bandas R 3) Los tiempos de replicación de las bandas parecen ser tambien diferentes (temprana para las R y tardía para las G), además hay otras diferencias en la clase de secuencias repetidas predominantes en uno y otro tipo de bandas (el lector interesado puede consultar la referencia 1).

Con métodos físico químicos (centrifugación diferencial en gradiente), el grupo de Bernardi puso de manifiesto en 1985, que el genomio nuclear humano y en general el de los vertebrados homeotermos, tiene una estructura en mosaico en el que las piezas son segmentos de ADN de unas 300 kilobases. La composición nucleotídica dentro de esos segmentos, a los que se les ha llamado isócoras, es homogénea (2). Aunque hay varias clases de isócoras definibles por sus densidades de flotación, se les puede agrupar en dos grupos: isócoras H (pesadas) con un contenido alto de G+C, e isócoras L (ligeras) con un contenido más bajo de G+C. Como era de esperar, los genes localizados en bandas R pertenecen a isócoras H y los localizados en bandas G a isócoras L.

La estructura genómica en compartimentos determina que dentro del mismo genomio haya genes con diferentes composiciones nucleotídicas, en términos generales se puede hablar de genes ricos en G+C y de genes pobres en G+C, por ejemplo el gen de la globina beta humana tiene 56'30 % de G+C, y el el gen de la globina zeta tiene 65'03. Las diferencias de composición afectan tanto a las secuencias codificadoras de proteínas como a las zonas no codificadoras, de manera que en un segmento de ADN rico en G+C encontramos genes cuyos exones, intrones y flancos también son ricos en G+C y viceversa (2,3).

Dentro de los exones las relaciones de composición son diferentes según las posiciones del codón. El contenido de G+C de un exón medido en las tres posiciones de sus codones está también correlacionado positivamente con el contenido de G+C del segmento genómico que lo contiene, pero la dependencia aumenta en el sentido II - I - III. En otras palabras, en un gen rico en G+C el contenido de G+C es mayor en las terceras posiciones de los codones, después en las primeras y finalmente en las segundas posiciones. Siguiendo con el ejemplo de la globina zeta humana, el contenido de G+C en las posiciones I, II y III vale 57'34 %, 43'35 % y 94'40 %, respectivamente.

Esta circunstancia es una consecuencia de la naturaleza redundante de la clave genética, sólo dos aminoácidos (metionina y triptófano) son codificados por un único codón, de los 18 restantes algunos son codificados por dos codones (duetos), uno por tres codones, otros por cuatro codones (cuartetos), y hay tres aminoácidos que son codificados por seis codones (sextetos) (Tabla 1). Una peculiaridad de la clave genética es que hay sólo dos clases de duetos, los que llevan en terceras posiciones una pirimidina (C o T) y los que llevan una purina (A o G). La tercera posición de cada uno de los cuartetos puede ser cualquiera de los nucleótidos. Los tres sextetos se pueden descomponer en un cuarteto y en un dueto, que difieren en el primer nucleótido (arginina y leucina) o en el primero y segundo (serina). De esta manera, para la mayoría de los aminoácidos existe un catálogo de codones para elegir el tercer nucleótido, y en el caso de arginina, leucina y serina la elección se puede extender también al primer nucleótido. Las segundas posiciones de los codones son fijas y no existen dos codones sinónimos que difieran sólo en ellas.

Así se explican las relaciones de dependencia antes mencionadas, los genes ricos en G+C tienen una estrategia codificadora diferente a la de los genes pobres en G+C, mientras los primeros utilizan preferentemente los codones cuyas terceras posiciones son C o G, los segundos utilizan más los codones acabados en A o T.

En el caso del genomio humano no existen pruebas de que las diferencias en el uso de codones sinónimos se deban a fenómenos de selección relacionados con la eficacia de la traducción, el fenómeno más bien se puede considerar como una táctica de acomodación a las circunstancias de composición de la región genómica en cuestión. Como la relación composicional afecta también, aunque mucho menos, a las segundas posiciones, la composición de las proteínas varía también con el contenido en G+C de los genes que las codifican.

Uso de codones sinónimos y contenidos de G+C.

En nuestro grupo de investigación hemos estudiado cómo afectan los cambios en el contenido de G+C a la elección de codones para cada aminoácido en los genomios nucleares de algunos vertebrados homeotermos (4). En este artículo nos limitaremos al genomio humano, centrándonos en los grupos de codones sinónimos compuestos por cuatro codones: esos grupos son los cuartetos

TABLA 1.

TTT	Fenilalanina	TCT	Serina	TAT	Tirosina	TGT	Cisteína
TTC	Fenilalanina	TCC	Serina	TAC	Tirosina	TGC	Cisteína
TTA	Leucina	TCA	Serina	TAA	Fin	TGA	Fin
TTG	Leucina	TCG	Serina	TAG	Fin	TGG	Triptófano
CTT	Leucina	CCT	Prolina	CAT	Histidina	CGT	Arginina
CTC	Leucina	CCC	Prolina	CAC	Histidina	CGC	Arginina
CTA	Leucina	CCA	Prolina	CAA	Glutamina	CGA	Arginina
CTG	Leucina	CCG	Prolina	CAG	Glutamina	CGG	Arginina
ATT	Isoleucina	ACT	Treonina	AAT	Asparagina	AGT	Serina
ATC	Isoleucina	ACC	Treonina	AAC	Asparagina	AGC	Serina
ATA	Isoleucina	ACA	Treonina	AAA	Lisina	AGA	Arginina
ATG	Metionina	ACG	Treonina	AAG	Lisina	AGG	Arginina
GTT	Valina	GCT	Alanina	GAT	Ac. aspártico	GGT	Glicocola
GTC	Valina	GCC	Alanina	GAC	Ac. aspártico	GGC	Glicocola
GTA	Valina	GCA	Alanina	GAA	Ac. glutámico	GGA	Glicocola
GTG	Valina	GCG	Alanina	GAG	Ac glutámico	GGG	Glicocola

La clave genética es redundante debido a que la mayoría de los aminoácidos están codificados por más de un triplete:

Tres sextetos (Serina, Arginina y Leucina), cinco cuartetos (Prolina, Valina, Glicocola, Alalina y Treonina), nueve duetos (Fenilalanina, Acido aspártico, Acido grutámico, Tirosina, Cisteína, Ilistidina, Glutamina, Asparagina y Lisina) y un trío (Isoleucina). Sólo dos aminoácidos están codificados por un triplete (Metionina y Triptófano).

correspondientes a los aminoácidos valina, prolina, treonina, alanina y glicocola, y los cuartetos de leucina, arginina y serina.

Para expresar la naturaleza de la relación entre utilización de cada codón (dentro de su grupo sinónimo) y el contenido de G+C procedimos de la siguiente manera. En primer lugar calculamos en cada gen la frecuencia de uso de cada codón mediante el índice RSCU (5), este índice se obtiene dividiendo la frecuencia observada por la frecuencia esperada si la utilización de codones sinónimos fuera uniforme. Esta tarea se llevó a cabo sobre la compilación de uso de codones que nos facilitó el grupo de Ikemura (6). Como medida de precaución, a la hora de estudiar el comportamiento dentro de cada grupo sinónimo sólo incluimos aquellos genes que tenían al menos cinco codones del grupo en cuestión.

En segundo lugar, construimos un diagrama para cada codón en el que cada gen esta representado por un punto cuyas coordenadas son el valor RSCU del codón bajo análisis en ese gen (eje Y) y el contenido de G+C de la secuencia codificadora (eje X). Posteriormente calculamos el coeficiente de correlación entre esas dos variables y ajustamos una línea de regresión. Un ejemplo de tales diagramas se muestra en la Figura 1 donde se representa la utilización (RSCU) del codón CGC de arginina en 277 genes humanos en función de sus contenido de G+C.

En la Tabla 2 se dan los resultados de los análisis de correlación y regresión que relacionan la frecuencia de utilización de los codones acabados en A, T, C y G en cada cuarteto con los cambios en el contenido de G+C de los genes. Como era de esperar, la frecuencia de utilización de codones acabados en C o G aumenta conforme lo hace el contenido de G+C de los genes (correlaciones positivas), y naturalmente lo contrario ocurre con los codones acabados en A o T (correlaciones negativas).

La intensidad de esa relación se puede cuantificar por las magnitud de las pendientes de las líneas de regresión ajustadas, los codones con mayores regresiones se prefieren conforme aumeta el G+C mientras que los codones con pendientes menores se eluden. Atendiendo a las pendientes puede apreciarse un patrón que caracteriza a los ocho cuartetos en tres grupos:

Grupo 1 compuesto por serina, prolina, treonina y alanina. En estos cuatro cuartetos al aumentar el G+C se prima el uso de los codones

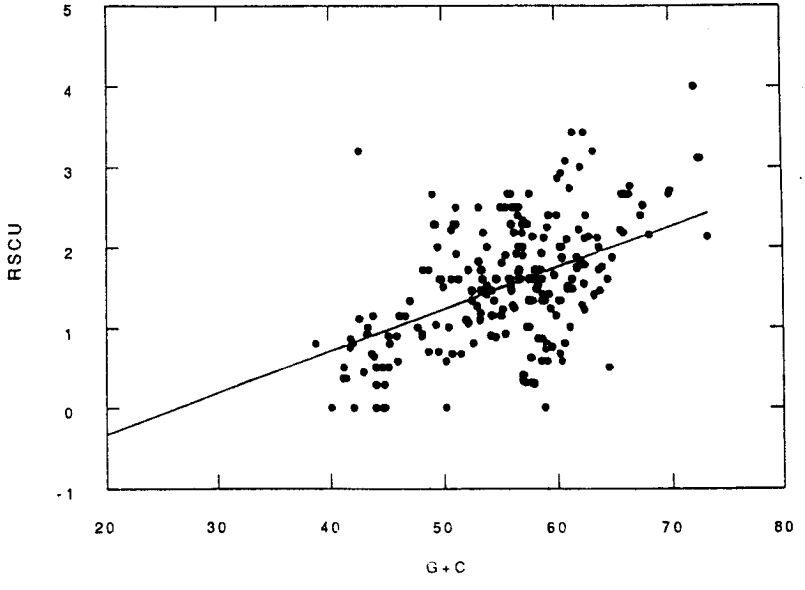
TABLA 2

Base en la tercera posicion del triplete

	N	A		T		C		G	
		R	s	R	s	R	s	R	s
Serina(TC-)	374	-.534	-4.04	-.445	-4.04	.599	5.96	365	2.12
Treonina(AC-)	371	-.536	-4.00	-.547	-4.20	.578	5.72	.397	2.48
Prolina(CC-)	370	-.450	-4.00	-.488	-4.16	.466	4.84	.570	3.32
Alanina(GC-)	395	-.526	-3.68	-.588	-4.32	.606	5.60	.453	2.40
Arginina(CG-)	277	-.532	-4.68	-.425	-3.52	.511	5.76	.249	2.44
Glicocola(GG-)	381	-.499	-4.52	-.406	-2.32	.506	4.64	.326	2.20
Leucina(CT-)	396	-.557	-2.08	-.631	-3.80	.076	0.44	.631	5.44
Valina(GT-)	379	-.543	-3.12	-.621	-4.28	.209	1.52	.570	5.84

Cuartetos. Valores de los coeficientes de correlación (R) y de las pendientes (s) cuando se representan los RSCU de los tripletes que terminan en A, T, C y G frente al contenido de G+C de la secuencia. N = número de genes. La dos primeras bases del triplete se dan entre paréntesis.

FIGURA 1. Valores RSCU del codón de Arginina CGC frente a los valores de G+C %, en una muestra de 277 genes humanos. El coeficiente de correlación es de 0'511, la pendiente vale 5'76.



acabados en C sobre el uso de los terminados en G, y la caída en la utilización de los codones acabados en A o T tiene la misma intensidad.

Grupo 2 compuesto por arginina y glicocola. En estos dos aminoácidos al aumentar el G+C existe también un sesgo en el uso preferencial de codones con C en tercera posición sobre codones con G, pero la atenuación en la frecuencia de uso de los codones acabados en A es mayor que la de los acabados en T.

Grupo 3 compuesto por leucina y valina. En estos dos cuartetos el aumento en G+C se traduce en una preferencia muy fuerte de los codones acabados en G frente a los acabados en C, el descenso en la frecuencia de uso de los codones acabados en T es más pronunciada que la de los acabados en A.

Resultados muy parecidos obtuvimos al analizar los genomios de ratón, de rata, de vaca y de pollo. En todos ellos el agrupamiento cualitativo es el mismo y las diferencias afectan solamente a la magnitud de las pendientes, más tarde volveremos sobre este particular.

Un resultado interesante es que los cuartetos incluidos en cada grupo tienen en común la base que ocupa la posición II del codón: los del grupo 1 llevan C, los del grupo 2 llevan G y los del grupo 3 llevan T. Podría decirse que el modo de elección de la base que puede ocupar la posición III está bajo la influencia de la base que ocupa la posición II.

Frecuencias de dinucleótidos y contenido de G+C.

Desde hace treinta años (7) se sabe que la frecuencia de los pares de nucleótidos adyacentes -dinucleótidos o dobletes- en las secuencias de ADN no es uniforme. En el genomio humano algunos dobletes son muy abundantes como por ejemplo TG o CT, mientras que otros como CG o TA son muy escasos (8). Las razones del alejamiento de la uniformidad todavía no están claras, y entre ellas podrían encontrase algunas relacionadas con requerimientos estructurales de la molécula de ADN (9), aunque quizá es más probable que las diferencias entre las frecuencias de los dobletes sean el reflejo de sesgos mutacionales que pueden tener lugar durante los procesos de replicación y/o reparación del ADN. Efectivamente, se sabe que los patrones de mutación de una base

están influidos por la base más próxima (10), y es clásica la explicación del déficit de CG en el genomio nuclear de los vertebrados como una consecuencia de las mutaciones promovidas por la metilación de la citosina, esas mutaciones generalmente transforman por transición los dobletes CG en TG y CA (11).

El fenómeno de los dinucleótidos debe jugar un papel muy importante a la hora de explicar el patrón de elección de la tercera base en los cuartetos. Como se ha señalado anteriormente, las segundas posiciones de los codones son las más estables en una secuencia codificadora, también se ha dicho que no parece que existan diferencias selectivas importantes entre codones sinónimos. Con estas dos premisas, podemos suponer que las distribuciones de las bases en las terceras posiciones deben ser el reflejo de los eventos mutacionales. Por tanto, las diferencias en las frecuencias de las bases situadas en la posición III que existen entre los cuartetos se pueden achacar a la influencia de las bases que ocupan la posición II, esta puede ser la explicación más sencilla de los tres grupos de cuartetos descritos según la base que ocupa la posición II.

En resumen, se confirma que la utilización no uniforme de codones sinónimos en los genes humanos (y por extensión en los vertebrados homeotermos) es un reflejo de la composición de la región genómica donde estén localizados (2). Por lo que respecta a los cuartetos, se podría decir que la riqueza de G+C de un gen se acomoda poniendo en terceras posiciones de los codones C cuando hay C o G en II, y G cuando hay T en II.

Un aspecto evolutivo interesante es que la relación entre el uso de codones sinónimos y el contenido de G+C es bastante parecido entre las cinco especies de vertebrados homeotermos que analizamos. En los años 70 Richard Grantham (12) ya había señalado que las especies próximas tienen usos de codones parecidos, cuando comparamos las pendientes de nuestras regresiones encontramos también relación con la distancia taxonómica, de modo que las dos especies más próximas (ratón y rata) tienen los valores más parecidos, y las diferencias entre mamíferos son siempre más pequeñas que las que muestra cualquiera de ellos con el pollo.

Dinucleótidos en II-III y en intrones.

En una segunda aproximación hemos estudiado las frecuencias de dinucleótidos en posiciones II-III y en intrones de genes humanos

(13). Las variaciones de las frecuencias de dobletes también sigue las variaciones del contenido de G+C de los segmentos genómicos. El fenómeno habia sido estudiado en secuencias codificadoras humanas (14,15), entre las conclusiones más importantes de estos trabajos destaca que el castigo de los CG se atenúa cuando el contenido de G+C aumenta, mientras que el de TA se intensifica.

En la literatura no existía un análisis comparativo entre intrones y exones de los cambios en la frecuencia de dobletes en función de los contenidos de G+C. La idea aceptada era que las frecuencias de dobletes debían estar correlacionadas en secuencias codificadoras y no codificadoras y que los dobletes complementarios tenían la misma frecuencia (16,17,18). Sin embargo, nuestro trabajo ha puesto de manifiesto la existencia de diferencias importantes que afectan a algunos dobletes.

En nuestro trabajo utilizamos 1208 secuencias codificadoras completas de más de 600 nucleótidos y 651 intrones de más de 400 nucleótidos. La variable comparada fue la 'preferencia' -DP- de cada doblete en posiciones II-III y en intrones, dentro de intervalos de variación del 4% del contenido de G+C. La preferencia de un doblete se puede medir por un cociente cuyo numerador es la diferencia entre su frecuencia observada y esperada y cuyo denominador es la frecuencia esperada. La asignación a los intervalos de contenido de G+C se hizo para las secuencias codificadoras por el G+C en terceras posiciones y para los intrones por el G+C en terceras posiciones de las secuencias codificadoras a las que pertenecen.

Los resultados de la comparación se pueden presentar en forma de diagramas en los que en el eje horizontal se mide el contenido de G+C de las terceras posiciones y en el eje vertical se mide la preferencia para el doblete en cuestión. A lo largo del eje horizontal corrimos una ventana de 4% de amplitud de variacion de G+C en pasos de 2%. En cada paso representamos dos puntos (uno para la medida en II-III, y otro para la medida en intrones) cuyas coordenadas verticales respectivas fueron los promedios en ese intervalo de DP en II-III y en intrones, y la horizontal el valor de G+C. La figura 2 es un ejemplo con dos de tales diagramas: los correspondientes a los dobletes CC y GG.

Como puede verse, no existen diferencias entre las posiciones II-III y los intrones en sus preferencias por el doblete CC, y tampoco hay variaciones en la preferencia en todo el rango de variación del

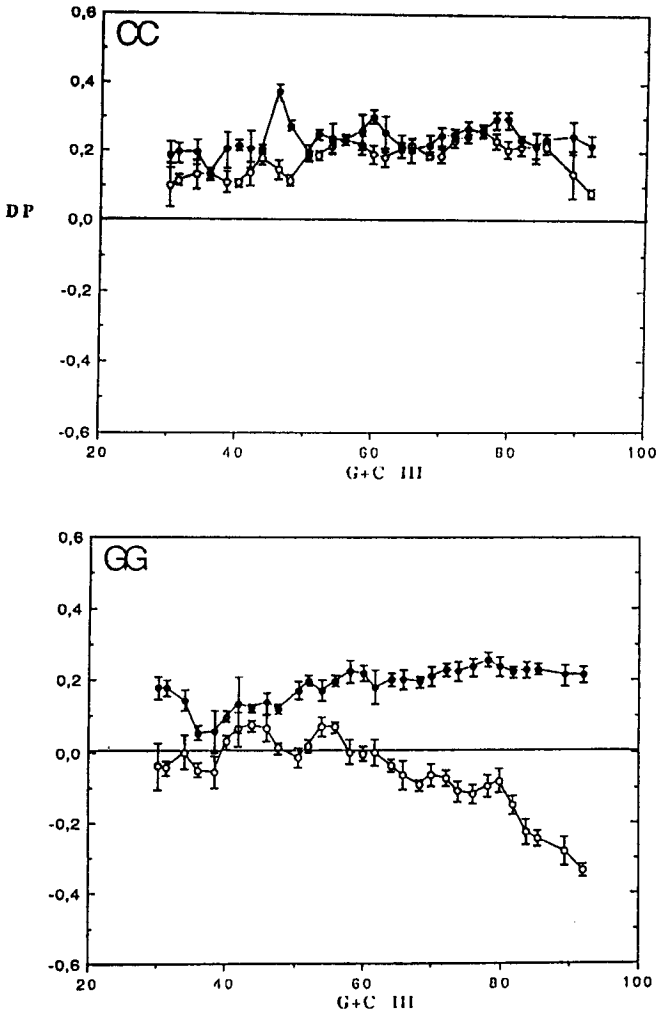


FIGURA 2. Gráfica que representa las variaciones del índice DP de los dinucleótidos complementarios GG y CC, frente al contenido en G+C medido en las terceras posiciones de los codones. (o) Posiciones II-III de los codones, (•) intrones. Las barras verticales indican el error típico de la media dentro de cada intervalo.

contenido de G+C. Por el contrario, al contemplar las gráficas correspondientes al doblete GG, se puede observar que las diferencias entre II-III e intrones se van incrementando conforme aumenta el contenido de G+C. La preferencia de GG en intrones es casi constante, mientras que en II-III no existe preferencia y para contenidos de G+C superiores al 50% existe un castigo creciente. Conviene resaltar que aunque la frecuencia absoluta del doblete GG aumenta con el contenido de G+C, es su preferencia (en relación con los valores esperados) la que disminuye. Los ejemplos de CC y GG nos dicen también que los comportamientos de los dobletes complementarios en II-III pueden ser diferentes.

Con la ayuda de un test estadístico no paramétrico comparamos en cada intervalo las distribuciones de DP en II-III y en intrones. Encontramos que los dobletes AA, AC, CC, CT y GA tienen el mismo comportamiento en II-III que en intrones. Los dobletes AT, AG, GT, TC, TT, GG, GC, CG y CA sólo difieren por encima de contenidos de G+C en III superiores al 55 %. Y finalmente, los dobletes TG y CA presentan diferencias a lo largo de todo el rango de variación del contenido de G+C.

Quizá el resultado más sorprendente de este trabajo fue que además de GG, los dobletes GC y TC tienen comportamientos divergentes en sus preferencias en el tramo de alto contenido de G+C: la preferencia de GG y TC disminuye en II-III y aumenta (GG) o es constante (TC) en intrones, y la de GC aumenta en II-III y disminuye en intrones.

Aunque todavía no disponemos de una explicación adecuada para el comportamiento divergente de los dobletes GG, GC y TC en II-III y en intrones, el fenómeno podría estar relacionado con la existencia de impedimentos o restricciones que afecten a los procesos postranscripcionales tales como la maduración y la traducción. En otras palabras, si los intrones y los exones comparten sitio físico en el cromosoma y con ello presiones mutacionales y requerimientos estructurales, y sólo después de la transcripción sus circunstancias varían, entonces las diferencias podrían asociarse a estos últimos eventos.
Otra posibilidad resulta de la existencia de variaciones en las tasas de mutación según el contenido de G+C (19) que deben tener efectos -por selección- diferentes sobre los intrones y los exones.

REFERENCIAS

1. Holmquist, G. P. (1989) *J. Mol Evol.* **28**, 469-486.

2. Bernardi, G. (1989) *Annu. Rev. Genet.* **23**, 637-661.

3. Aota, S., I. & Ikemura, T. (1986) *Nucleic Acids Res.* **14**, 6345-6355.

7. Josse, J. , Kaiser, A. D. & Kornberg, A. (1961) *J. Biol. Chem.* **236**, 864-875.

4. Marín, A., Bertranpetit, J., Oliver, J. L. & Medina, J. R. (1989) *Nucleic Acids Res.* **17**,6181-6189.

5. Sharp, P. M. & Li, W. H. (1986) *Nucleic Acids Res.* **14**, 7737-7749.

6. Aota, S., Gojobori, T., Ishibashi, F., Maruyama, T. & ikemura, T. (1988) *Nucleic Acids Res.* **16**, r315-r402.

8. Ohno, S. (1988) *Proc. Natl. Acad. Sci.* USA. **85**, 9630-9634.

9. Nussinov, R. (1984) *J. Mol. Evol.* **17**, 237-244.

10. Bulmer, M. (1986) *Mol. Biol. Evol.* **3**, 322-329.

11. Bird, A. P. (1980) *Nucleic Acids Res.* **8**, 1499-1504.

12. Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pave, A. (1980) *Nucleic Acids Res.* 8, r49-62.

13. Gutiérrez, G., Oliver, J. L. & Marín, A. (1993) *J. Mol. Evol.* **37**, (en prensa).

14. Hanai, R. & Wada, A. (1988) *J. Mol. Evol.* **27**, 321-325.

15. Wada, K-N., Watanabe, I., Tsuchiya, R. & Ikemura, T. (1991) En: Kimura, M. 6 Nakahata, T., (eds). Japan Sci. Soc. Press, Tokyo.

16. Nussinov, R. (1981) *J. Mol. Biol.* **149**, 125-131.

17. Nussinov, R. (1981) *J. Mol. Evol.* **17**, 237-244.

18. Hanai, R. & Wada, A. (1990) *J Mol. Evol.* **30**, 109-115.

19. Wolfe, K. H., Sharp, P. M. & Li, W. H. (1989) *Nature*. **337**, 283-285.