Evolución del Uso de Codones Sinónimos

Héctor Musto
hmusto@gmail.com

Laboratorio de Genómica Evolutiva

Código Genético

Características generales del código

a) Esencialmente universal (aunque hay excepciones: unos 30 "códigos").

b) No ambiguo: dado un codón, sabemos el aminoácido codificado.

c) Degenerado, o sea hay codones sinónimos:

2 aas son codificados por un solo codón, 1 aa lo es por tres, 5 aas por cuatro codones, 9 aas por dos, y 3 por seis.

Algunas excepciones (mitocondria de vertebrados)

Differences from the standard code

DNA codons	RNA codons	This code (2)	Standard code (1)
AGA	AGA	STOP = Ter (*)	Arg (R)
AGG	AGG	STOP = Ter (*)	Arg (R)
ATA	AUA	Met (M)	Ile (I)
TGA	UGA	Trp (W)	STOP = Ter (*)

Alternative initiation codons

- *Bos*: AUA
- *Homo*: AUA, AUU
- *Mus*: AUA, AUU, AUC
- *Coturnix*, *Gallus*: also GUG[8]

Algunas excepciones (equinodermos, pocos platelmintos)

DNA codons	RNA codons	This code (9)	Standard code (1)
AAA	AAA	Asn (N)	Lys (K)
AGA	AGA	Ser (S)	Arg (R)
AGG	AGG	Ser (S)	Arg (R)
TGA	UGA	Trp (W)	STOP = Ter (*)

Systematic range

- Asterozoa (starfishes)
- Echinozoa (sea urchins)
- Rhabditophora among the Platyhelminthes

Algunas excepciones (algunos hongos y levaduras)

DNA codons	RNA codons	This code (12)	Standard code (1)
CTG	CUG	Ser (S)	Leu (L)

Alternative initiation codons
•CAG may be used in *Candida albicans*.

Systematic range
· Endomycetales (yeasts): *Candida albicans*, *Candida cylindracea*, *Candida melibiosica*, *Candida parapsilosis*, and *Candida rugosa*

Consecuencias de la "degeneración": UCS

Cuando se elucida el código (faltaban 11 años para aprender a secuenciar el ADN) se pensó que si tomamos un gen suficientemente largo, o mejor aún, TODOS los genes de una especie, cada aa iba a ser codificado aproximadamente por el número de codones sinónimos que lo definen. Por ejemplo, dado un número grande de Phe, la mitad sería codificado por UUU y la otra mitad por UUC, o si consideramos todas las Val GUU, GUC, GUA y GUG estarían representados un 25% cada uno.

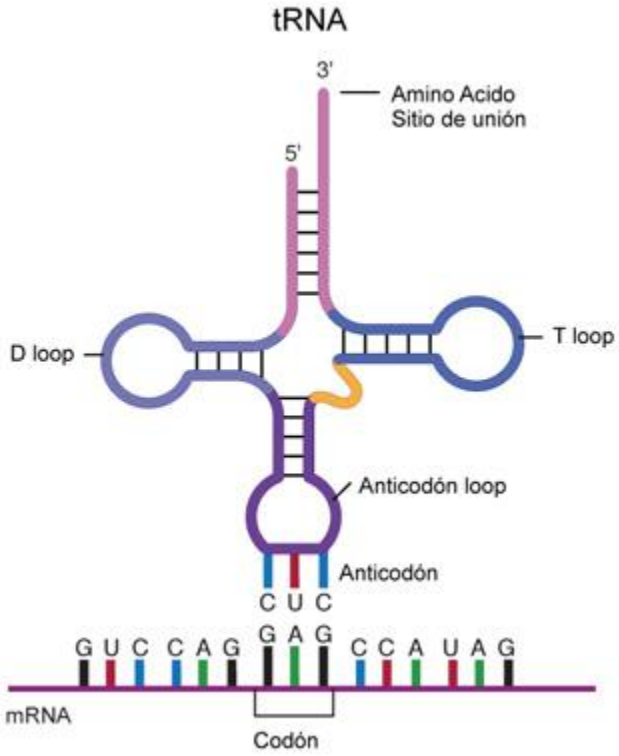
Más aún, Kimura plantea que el paradigma del neutralismo a nivel molecular eran, precisamente las terceras posiciones de los codones (son las que pueden cambiar más sin cambiar el aa, mirar y analizar la estructura del código).

Dos tipos de moléculas cruciales

1) Aminoacil t-RNA sintetasas (son las moléculas que hacen el pasaje de la información del lenguaje de los ácidos nucleicos al de las proteínas).

2) t-RNAs (se cargan con el aminoácido correspondiente).

Estructura secundaria del t-RNA



Hipótesis del "tambaleo" (wooble)

El código es:

-degenerado
-no ambiguo
-casi universal
-tripletes, no solapado y sin comas

If these bases are in first, or wobble, position of anticodon

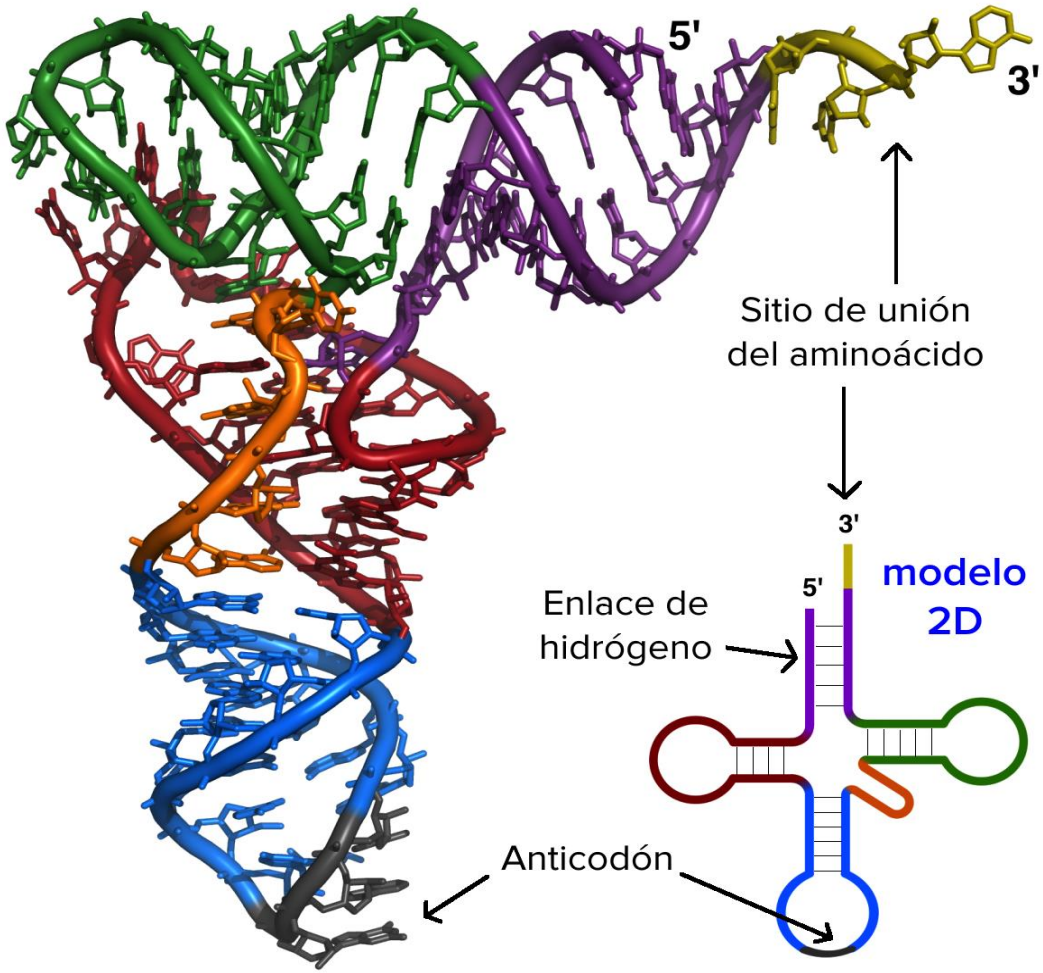
C	A	G	U	I	
G	U	C U	A G	C A U	then the tRNA may recognize codons in mRNA having these bases in third position

If these bases are in third, or wobble, position of codon of an mRNA

C	A	G	U	
G I	U I	C U	A G I	then the codon may be recognized by a tRNA having these bases in first position of anticodon

Papel de la AA-tRNA sintetasa

La enzima primero se une a un ATP y al aa correspondiente (o su precursor) para formar un aminoacil-adenilato, liberando una molécula de PPi. Luego se une la molécula apropiada de ARNt, y el aminoácido se transfiere desde el complejo aminoacil-AMP a un grupo OH (ya sea 2' o 3') del último nucleótido (A76, en el extremo 3') del ARNt.


5'
3'
Sitio de unión
del aminoácido
3'
5'
modelo
2D
Enlace de
hidrógeno
Anticodón

Reacciones:

aminoácido + ATP → aminoacil-AMP + PPi

aminoacil-AMP + ARNt → aminoacil-ARNt + AMP

Reacción global:

aminoácido + ATP + ARNt → aminoacil-ARNt + AMP + PPi

Consecuencias del tambaleo

En teoría:

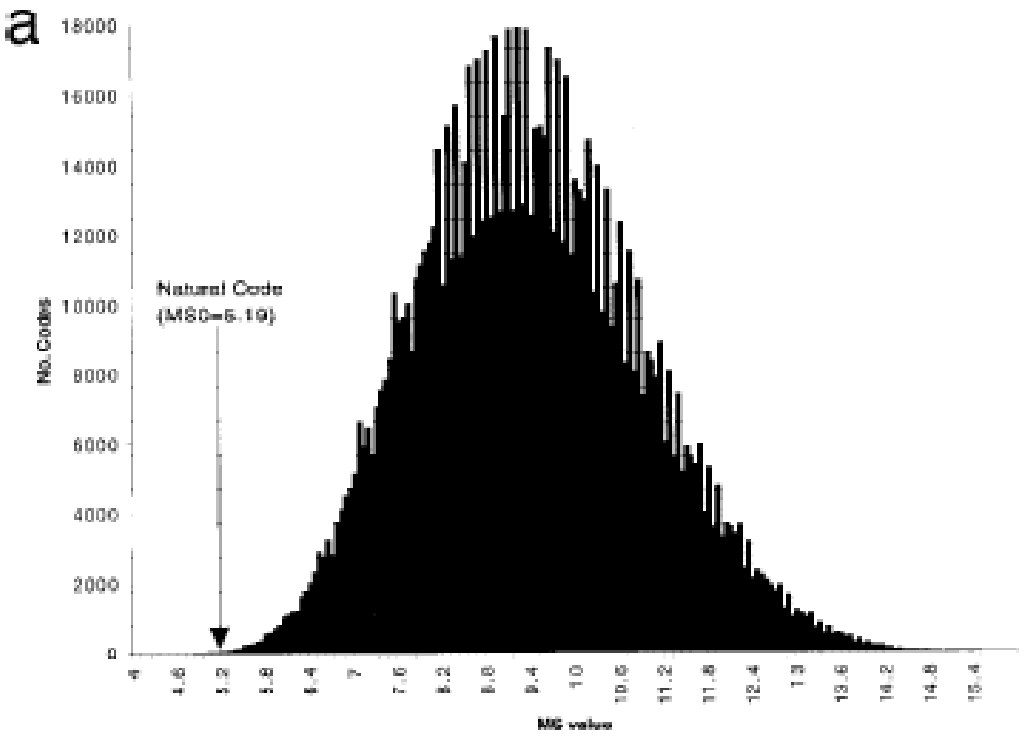
ARNm 5'---UUU/C---3'

ARNt 3'---AAG---5' o

ARNt 3'---AAA---5'

pero la A en el ARNt (habitualmente como Inosina) reconoce C, A o U en el codón, por lo que se podría incorporar Leu, mientras que la G reconoce C o U, por lo que no hay error. ¡Por eso no existen los 61 ARNt posibles! Minimizar los errores posibles.

¡El código es 1 en un millón!



Consecuencias de la evolución del código

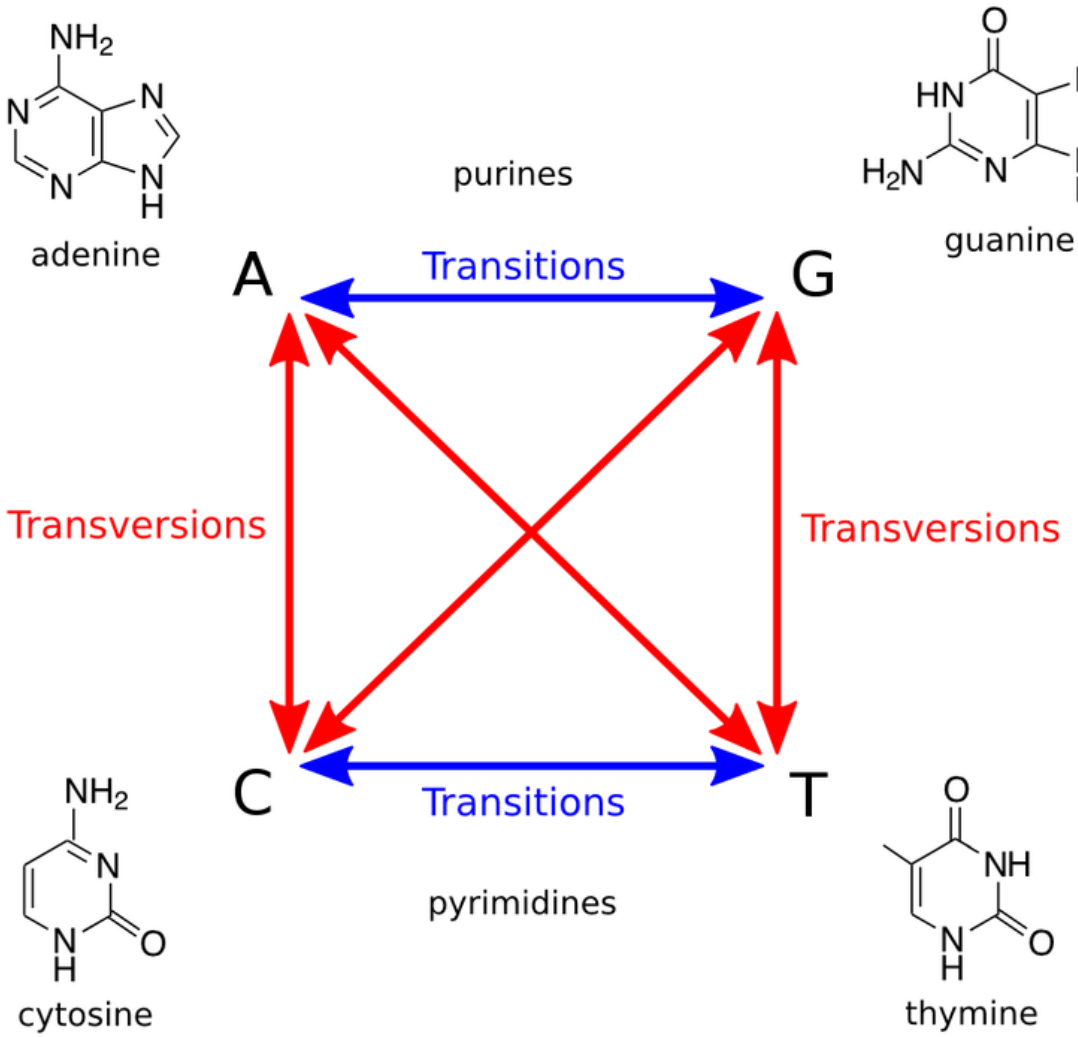
El código "universal" tiene como característica fundamental que minimiza el efecto de las mutaciones en la secuencia de las proteínas:

a) terceras posiciones en "cuartetos".

b) los sinónimos de "a dos" se diferencian por transiciones (ver fig. próxima).

c) los aas con características físico-químicas similares están agrupados.

Transiciones, transversiones



Origen del código

Tema en el que se continúa trabajando.

*** Accidente congelado (Crick, 1971).**

*** Aas físicoquímicamente similares tienen codones similares (Woese, 1967).**

*** Hipótesis de coevolución del código genético: la síntesis bioquímica de los aas acompañó la asignación de codones (Wong, 1975).**

El código continuaría evolucionando

El código continuaría evolucionando:

a) el hecho de que haya variantes implica que cambia, pero ver problema para fijar "nuevos" códigos.

Hay más de 20 aas en las proteínas "naturales": el 21 es la selenocisteína (1986) y está presente en los tres dominios; y el 22 es la pirrolisina que aparece solo en algunas arqueas.

Selenocisteína (Sec)

Se codifica en el ARNm mediante un triplete UGA. Este codón codifica habitualmente la finalización de la traducción, pero en conjunción con una región del ARNm denominada SecIS (secuencia de inserción de la selenocisteína) pasa a codificar la incorporación de este aa a la cadena polipeptídica. La secuencia SecIS se localiza en la región 3' no traducida (3'UTR) en arqueas y eucariotas, e inmediatamente después del codón UGA en bacterias.

Pirrolisina (Pyl)

La pirrolisina es un aminoácido natural, derivado de la lisina, codificado en el genoma de algunas arqueas metanógenas y en pocas bacterias. Se encuentra codificada en el ARNm por el codón UAG. Fue descubierta en el año 2002.

¿Qué se entiende por "uso de codones sinónimos" (UCS)?

Estrictamente, es el conteo de codones (sea para un grupo de genes o para toda una especie).

Pero se toma como *el desvío* respecto al uso determinado por la cantidad de sinónimos que codifican a cada aa.

Puede observarse tanto a nivel de genes individuales como a nivel genómico.

Primer hipótesis

Volume 8 Number 1 1980 Nucleic Acids Research

Codon catalog usage and the genome hypothesis

R.Grantham, C.Gautier, M.Gouy, R.Mercier and A.Pavé

Equipe Evolution Moléculaire, Laboratoire de Biométrie, Université Lyon I, 69622 Villeurbanne Cedex, France

ANALISIS DE CORRESPONDENCIA (COA)

El propósito de la estadística ha sido definido como “resumir, simplificar y eventualmente explicar” (Greenacre 1984).

El uso de codones es, por naturaleza, multivariado, y es por lo tanto necesario analizarlo usando técnicas estadísticas multivariadas.

El COA es uno entre varios análisis multivariados.

Los análisis multivariados se usan para simplificar matrices rectangulares en las cuales (para nuestro propósito) las columnas representan alguna medida de uso de codones y las filas representan genes individuales.

Dicho en forma extremadamente simplificada:

Cada gen puede ser representado en un espacio multidimensional por un vector.

El espacio multidimensional es Euclideano, siendo cada eje ortogonal.

Las distancias son definidas en filas y/o columnas, y estas distancias son aproximadas por distancias Euclideanas en una representación de pocas dimensiones de la tabla.

El objetivo del COA (como de cualquier otro análisis multivariado) es identificar el "subespacio" de pocas dimensiones que mejor representa los puntos (genes).

Es una herramienta poderosa para "aislar" tendencias, pero su mayor debilidad consiste en que no brinda claves para la interpretación de estas tendencias.

La interpretación corre por cuenta del investigador.

Hipótesis "genómica" de Grantham et al., 1980

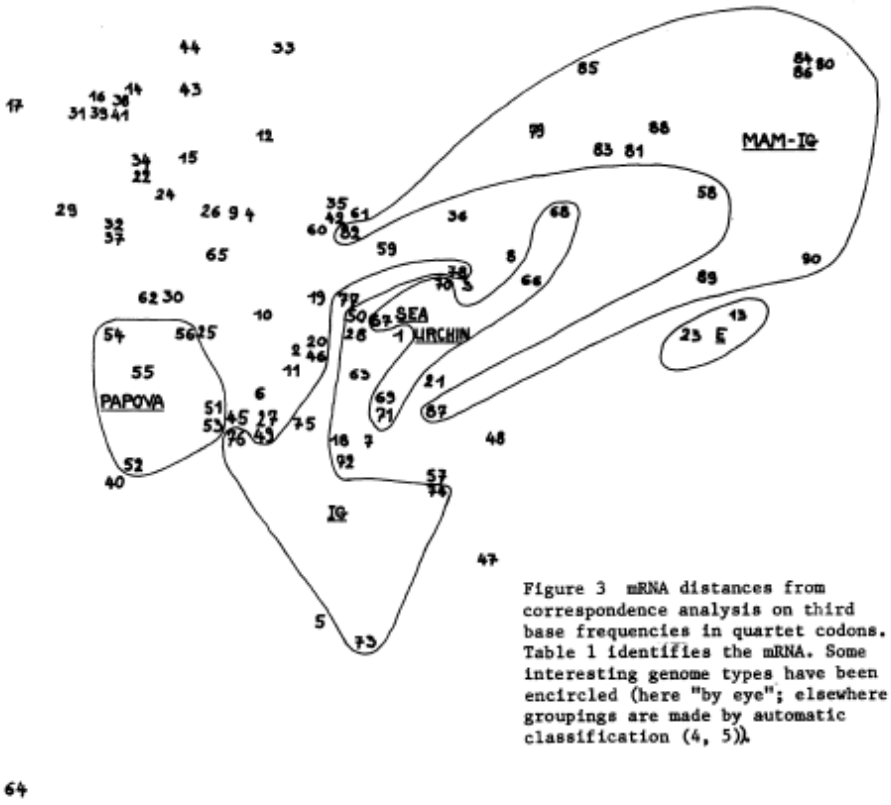
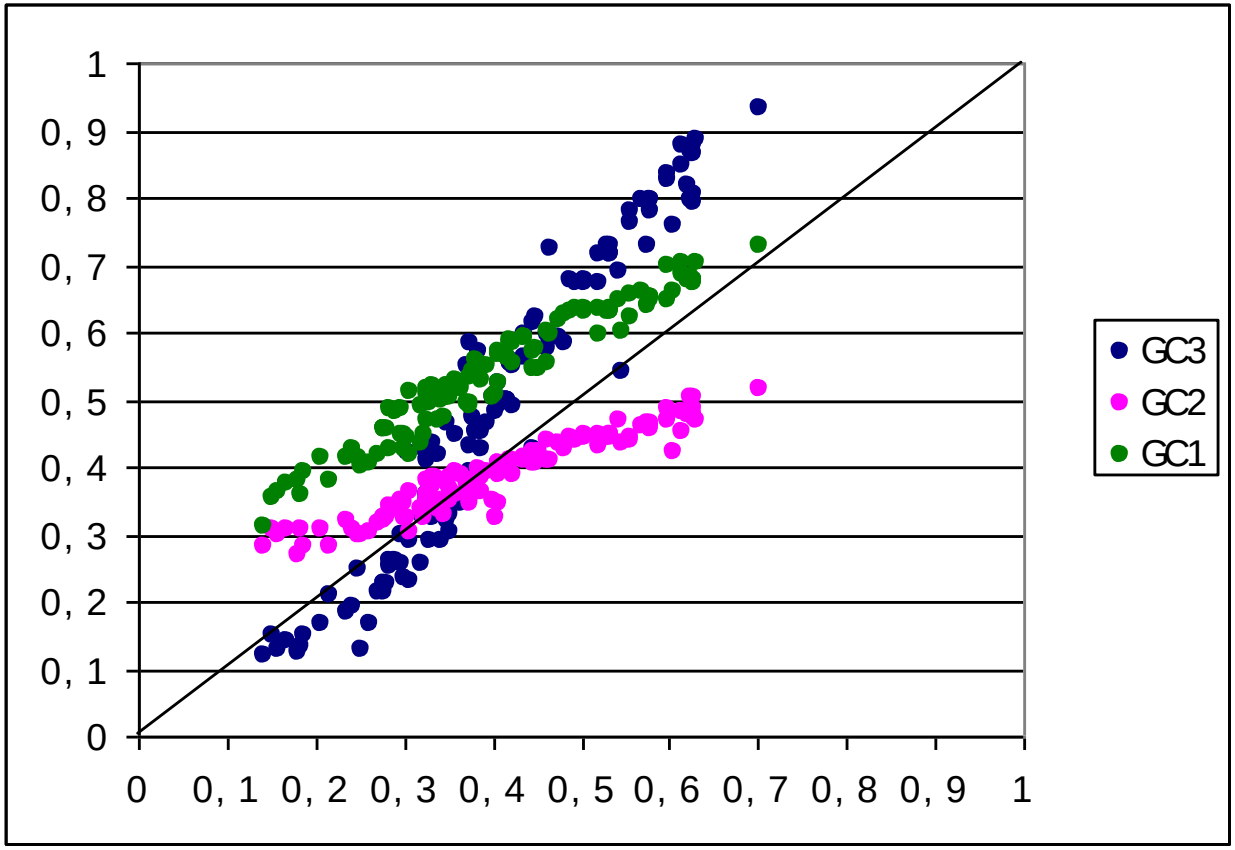


Figure 3 mRNA distances from correspondence analysis on third base frequencies in quartet codons. Table 1 identifies the mRNA. Some interesting genome types have been encircled (here "by eye"; elsewhere groupings are made by automatic classification (4, 5)).

Correlaciones composicionales entre las tres posiciones de los codones y el GC genómico (regiones intergénicas) en procariotas



Uso de codones en *H. sapiens*

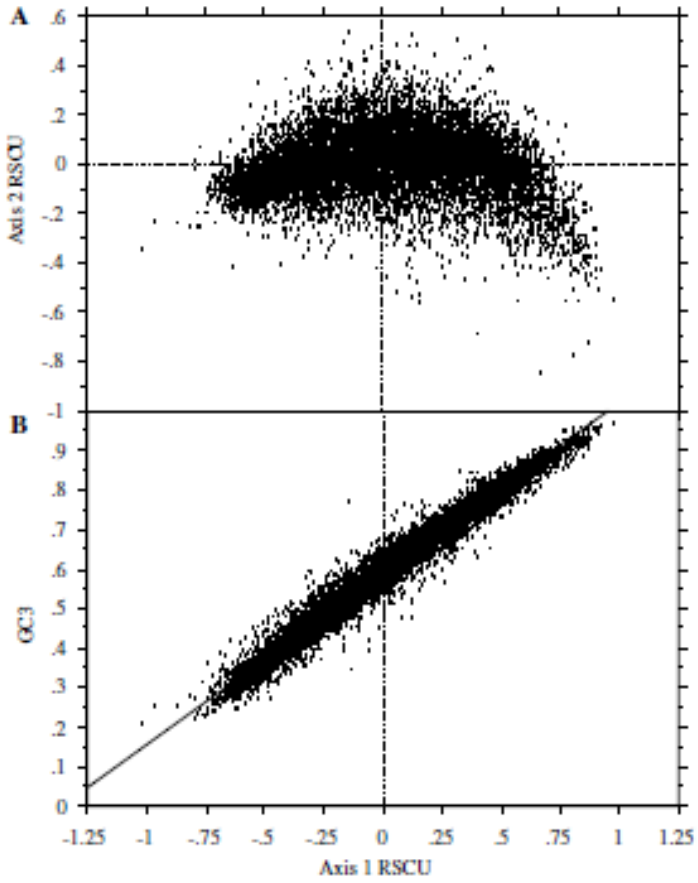


Fig. 3. (A) Distribution of the genes on the plane defined by the two main axes of the correspondence analysis. (B) Correlation between the position of each gene along the first axis and the GC3 content of the respective sequences. $R^2 = 0.97$; $P < 0.0001$.

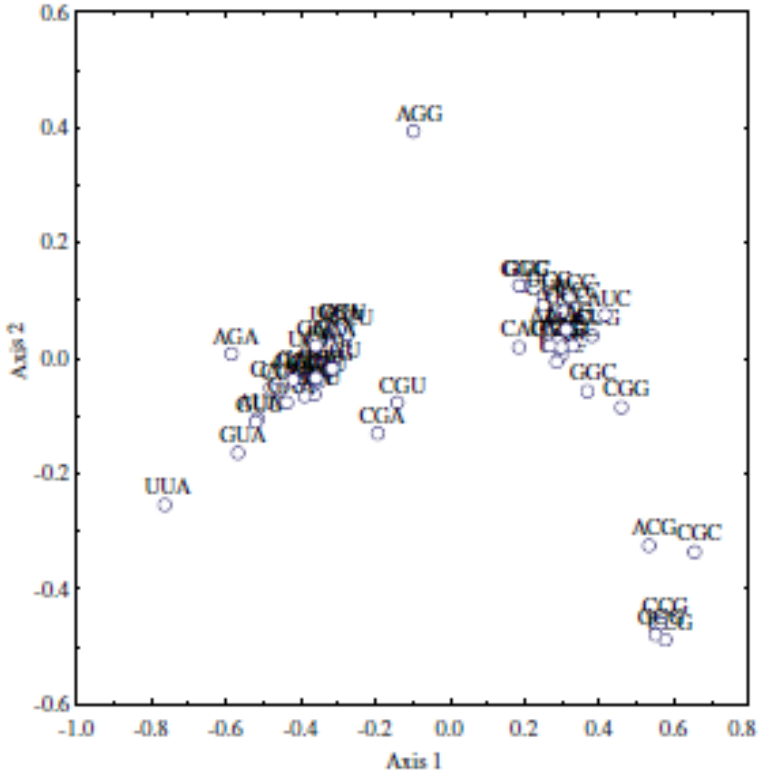
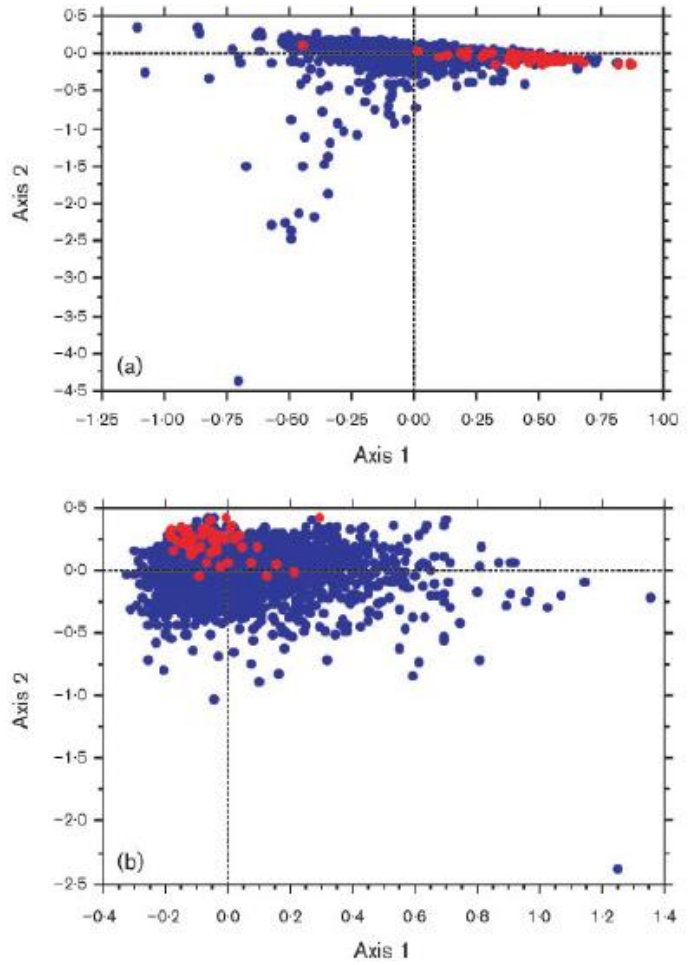


Fig. 4. Distribution of the codons from the human genes on the plane defined by the two main axes of the correspondence analysis.

Conclusión 1 (sesgo mutacional)

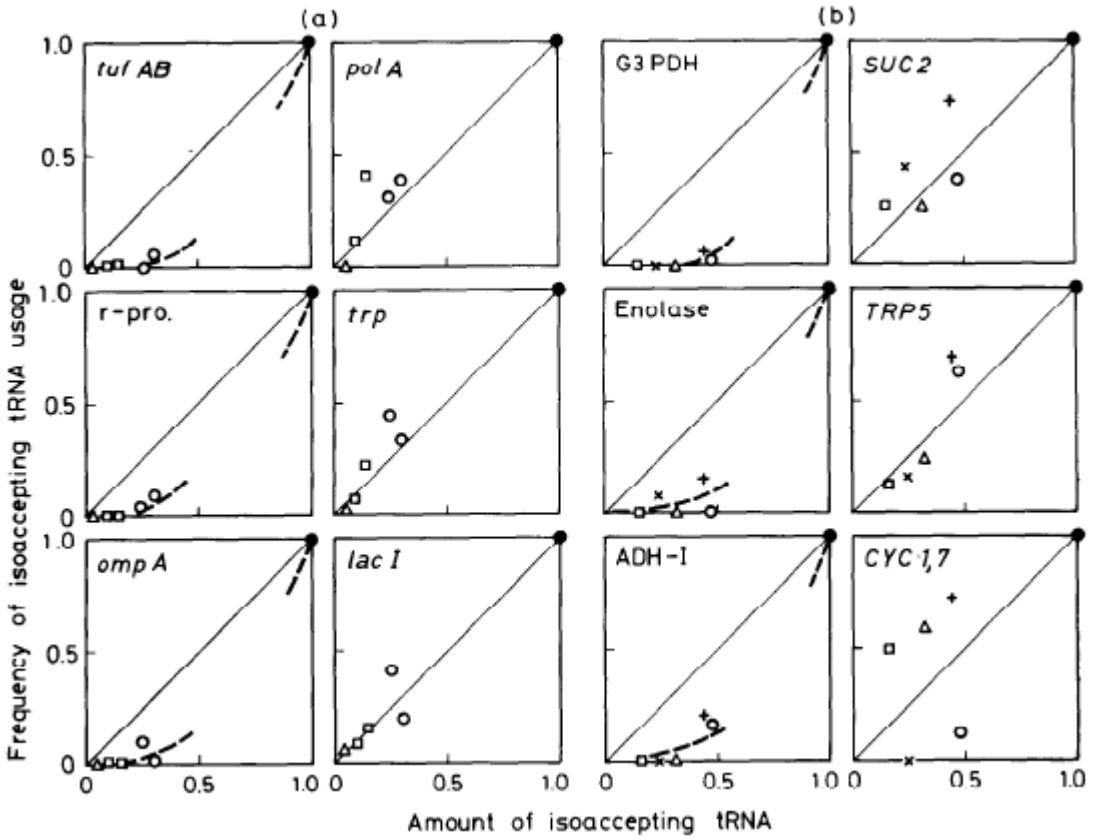
El principal factor que afecta el uso de codones sinónimos, tanto en procariotas como en eucariotas y virus, es el contenido en GC genómico de cada especie. O sea, especies ricas en G+C tenderán a usar codones terminados en estas bases mientras que lo contrario ocurre con especies cuyo en genoma es pobre en G+C. ***Concepto de sesgo mutacional*****.**

Además de la variación intergenómica en el UCS, se detectó también (mediante análisis multivariado) que existe una variabilidad intragenómica.



Entre un grupo de codones sinónimos reconocidos por varios ARNt, los reconocidos por el ARNt más abundante son usados más frecuentemente (adaptación del uso de codones al pool de ARNt). El sesgo es más extremo en los genes de más alta expresión (selección para velocidad, concepto de *codones mayores*) (ver figura próxima).

Ikemura, 1985 (cantidad de t-RNAs isoaceptores y la frecuencia de su uso en genes individuales)



Ejemplo: *Plasmodium falciparum*

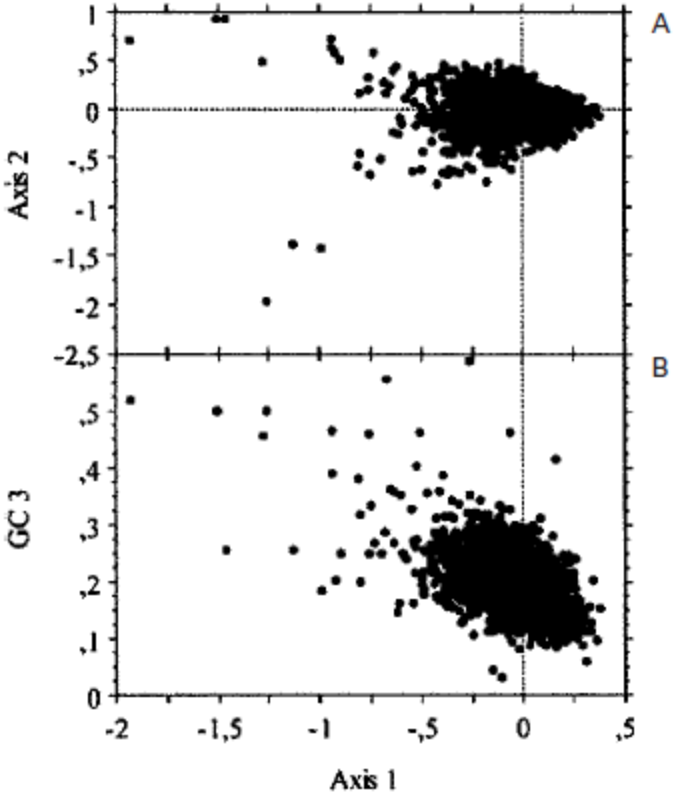


Fig. 1. The position of each gene along the first axis generated by the COA (calculated on RSCU values) is plotted against the second axis of the same analysis (A) and the respective GC3 (B).

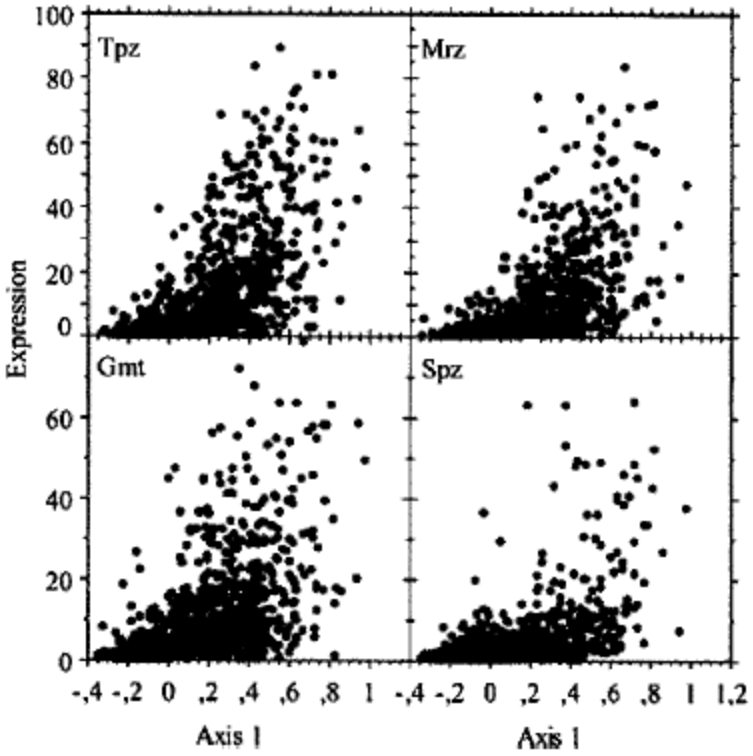


Fig. 2. The position of each gene along the first axis generated by the COA (calculated on codon usage numbers) is plotted against the expression levels of proteins for trophozoites (Tpz), merozoites (Mrz), gametocytes (Gmt) and sporozoites (Spz).

Conclusión 2: selección natural a nivel de la traducción

***Mínimamente*, tres aspectos de la traducción podrían ser afectados por el UCS:**

a) la tasa de elongación (velocidad)

b) el costo del "proofreading"

c) la fidelidad en la traducción

a) la tasa de elongación

1) **Los genes de más alta expresión presentan un uso de codones más sesgado.**
2) **Muestran un incremento significativo de codones mayores.**
3) **Usualmente estos codones aparean perfectamente con el ARNt isoaceptor más abundante.**
4) **La tasa de sustituciones sinónimas es más baja en los genes de alta expresión, y viceversa (ver figura próxima).**
5) **Evidencias experimentales: En *E. coli*, la traducción de los codones mayores ocurre 3 a 6 veces más rapido que la de los menores, permitiendo un mejor uso de los ribosomas.**

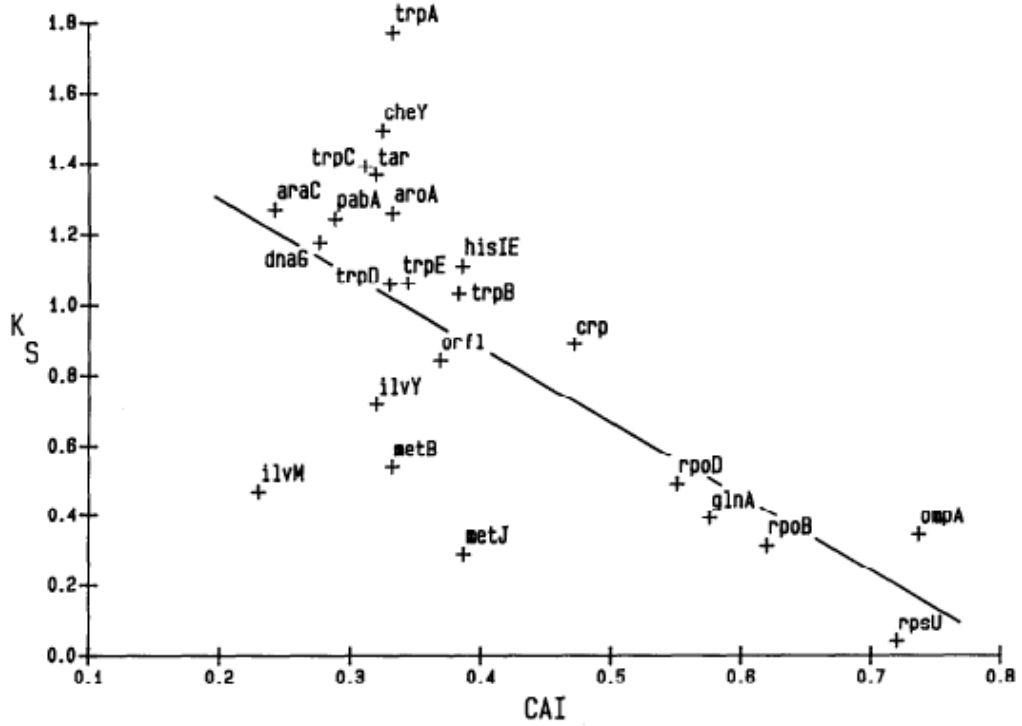


FIG. 1.—Relationship between synonymous-codon usage bias and K_S in *Escherichia coli* and *Salmonella typhimurium*. K_S = Estimated number of nucleotide substitutions per synonymous site; CAI = mean of the CAI values for the two species. For a sequence in which all 61 codons are equiprobable, CAI = ~0.17. The least-squares linear regression of K_S on CAI is drawn; the linear correlation coefficient is 0.68, $P < 0.01$.

b) el costo del “proofreading”

Los codones mayores pueden reducir el costo energético de rechazar un ARNt equivocado, lo que cuesta un GTP por reacción

c) la fidelidad en la traducción

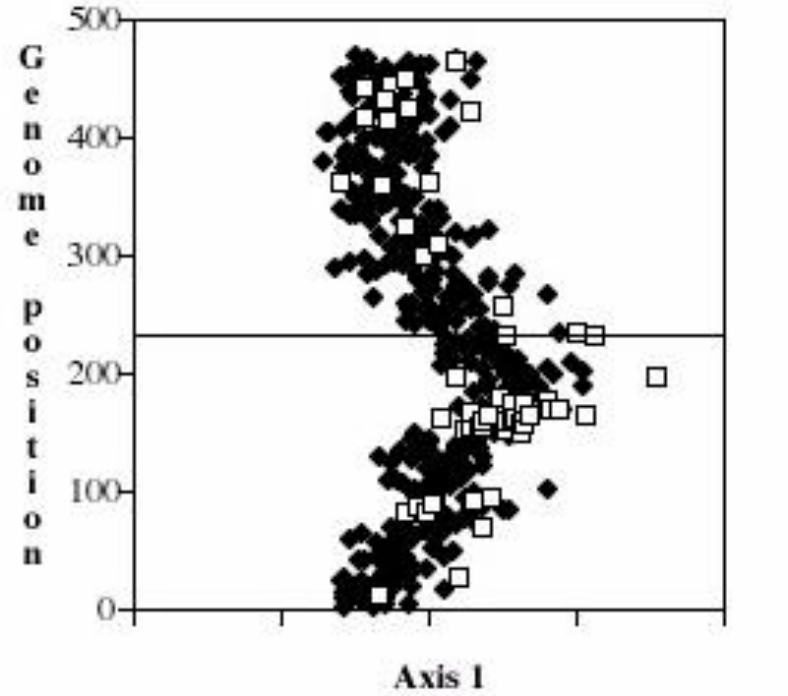
Las propiedades críticas de las proteínas dependen de aas particulares en posiciones específicas de sus estructuras. La traducción "fiel" reduce los costos de producir proteínas no funcionales por incorporaciones erróneas y/o errores en la procesividad ("frameshifting" y/o terminaciones prematuras).

En *E. coli*, la traducción de los codones mayores es 10x más "fiel" que la de los codones menores, permitiendo también un uso más eficiente de los ribosomas.

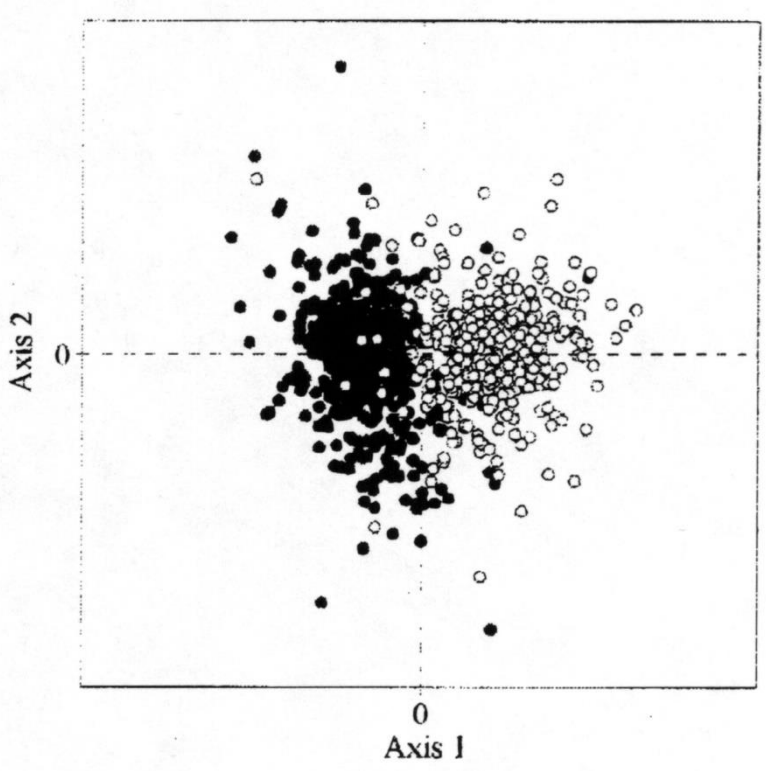
Además...

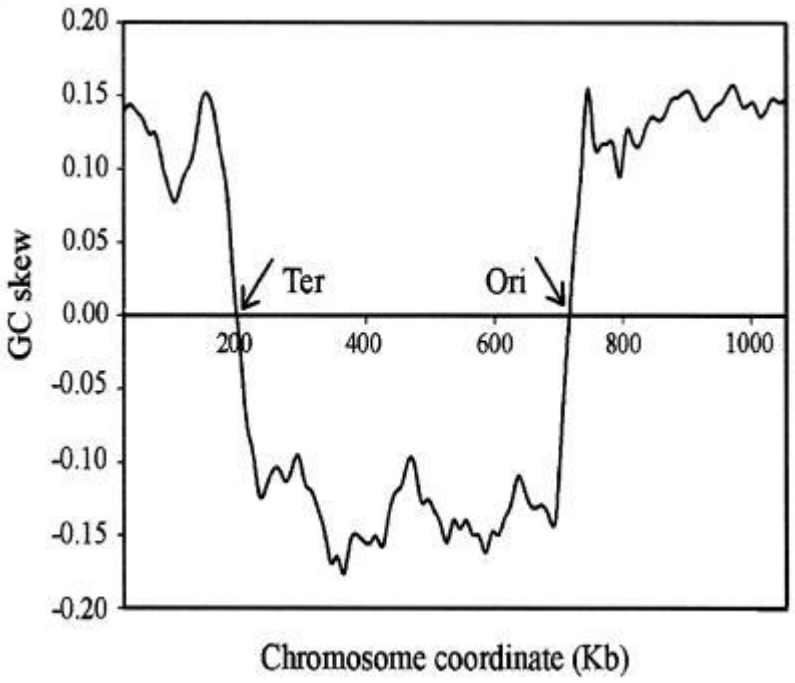
En los últimos años, la disponibilidad de genomas completos permitió detectar nuevos factores que influyen en el uso de codones sinónimos, por ejemplo:

1) La localización en el genoma de cada secuencia determina el GC3, y por lo tanto el uso de codones sinónimos en *Mycoplasma genitalium*.



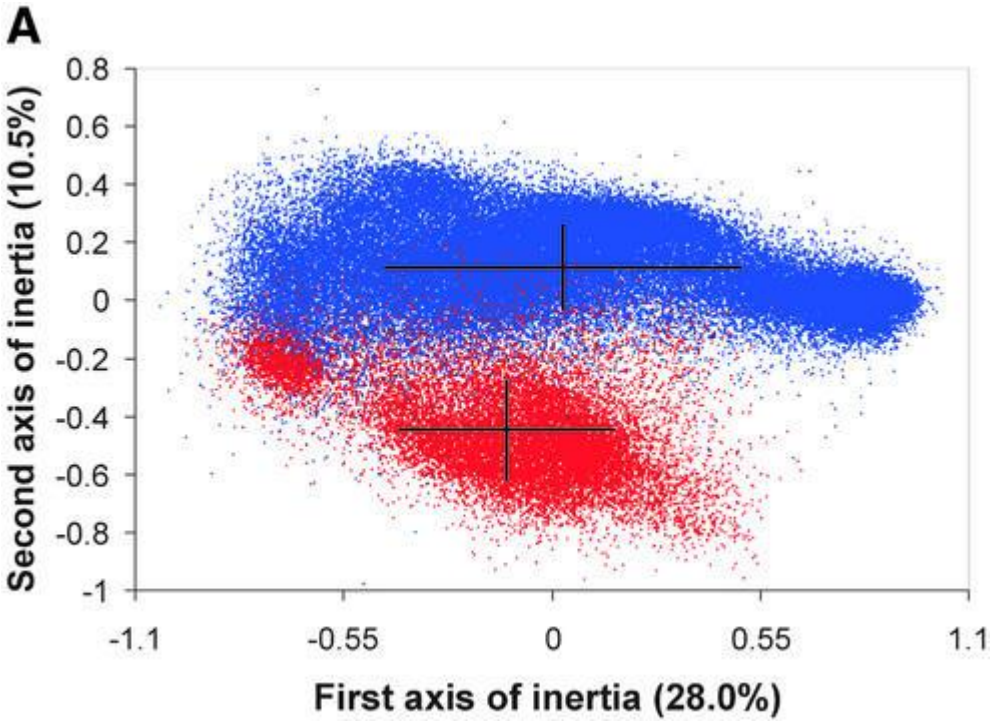
2) En especies como *Borrellia burgdorferi*, *Treponema pallidum* y *Chlamydia trachomatis*, la localización de los genes en la hebra "leading" o "lagging" de la replicación es el factor principal en el UCS (GC-skew).



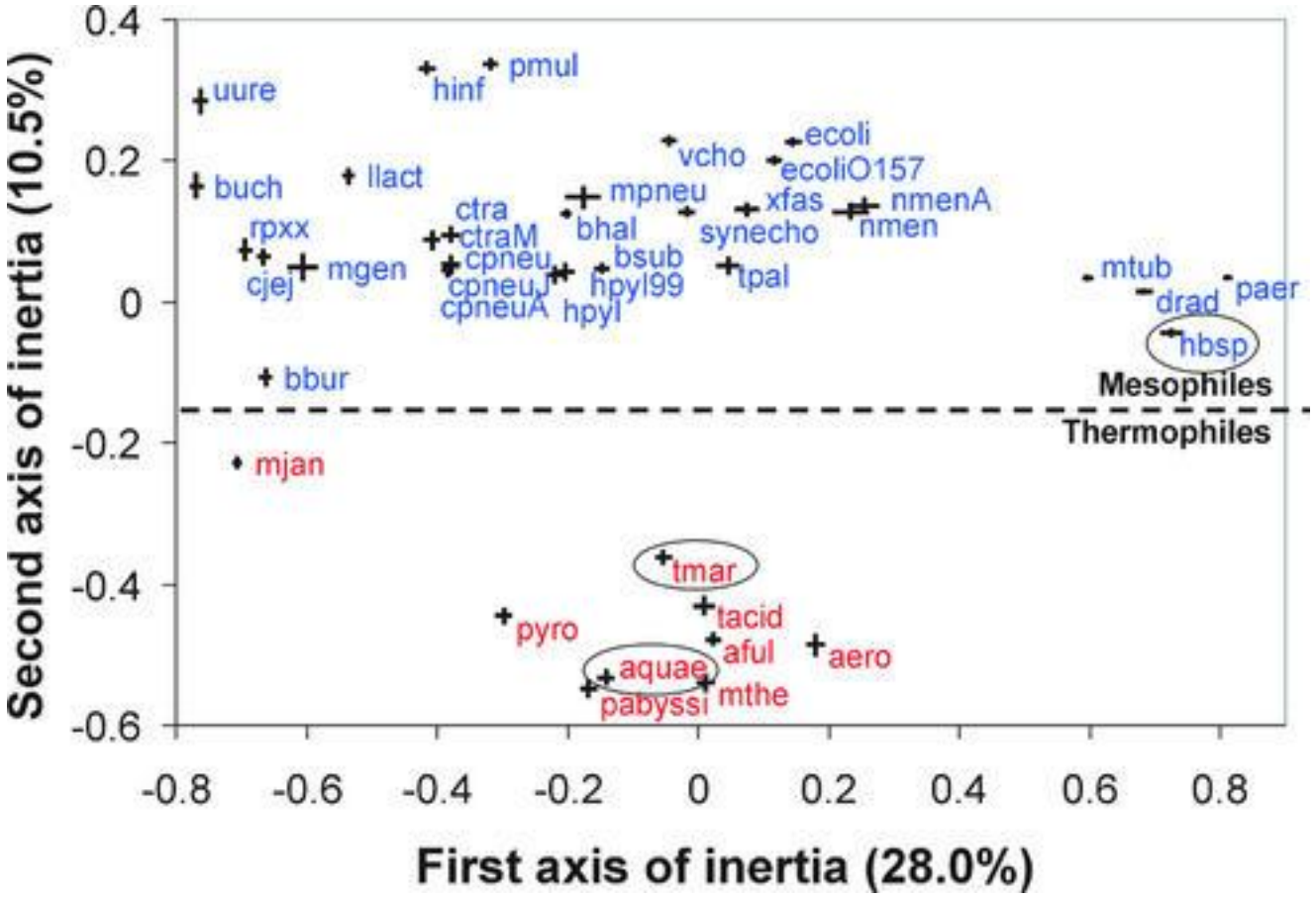


3) En varios procariotas, la hidropatía de cada proteína codificada, está entre los factores de más influye en el uso de codones. Habría, además, un UCS diferencial "intragen", en las proteínas de membrana, de forma que las zonas del gen que codifican las regiones intramembrana (hidrofóbicas) difieren en el UCS del resto del gen que codifica regiones hidrofílicas.

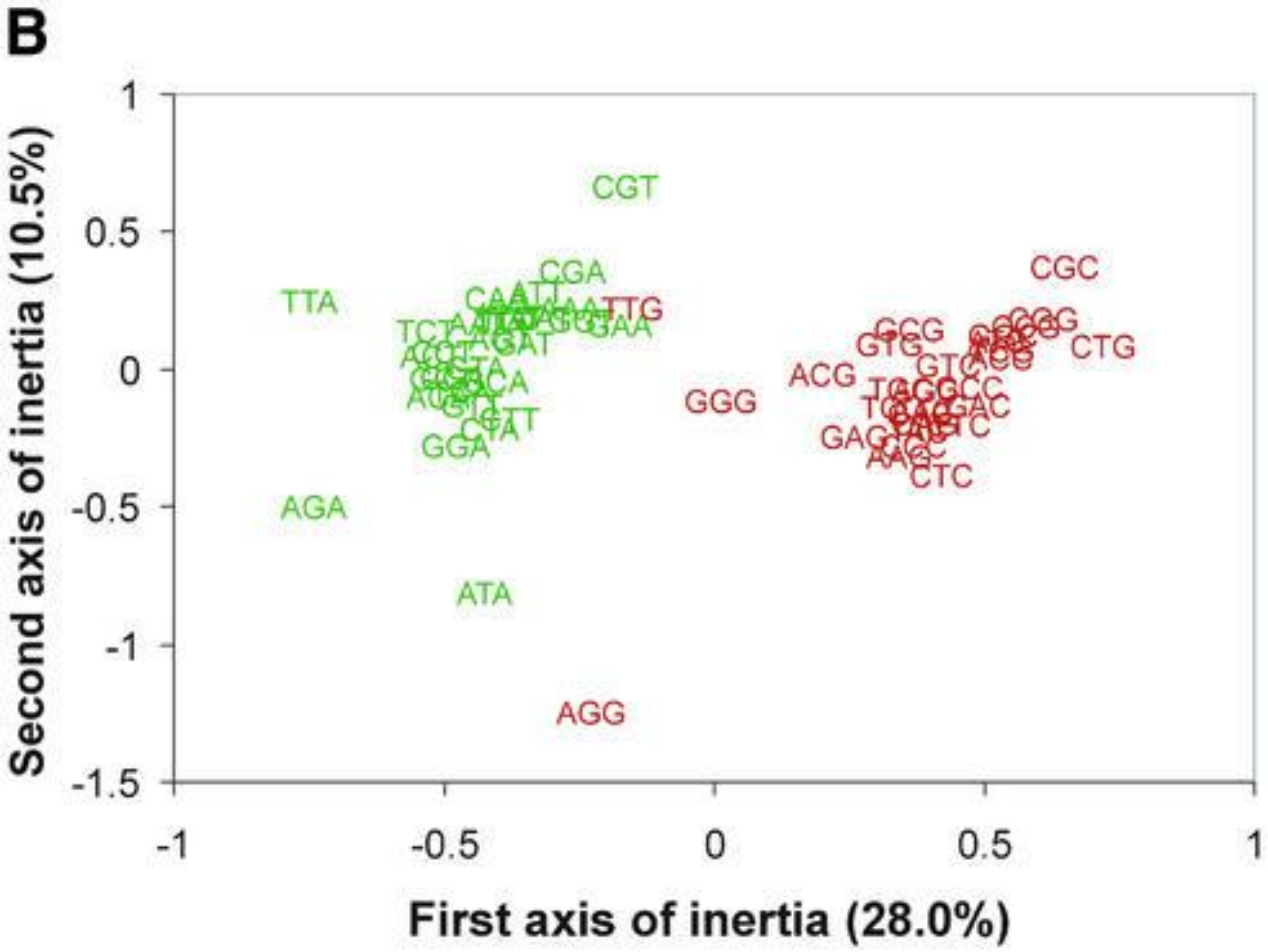
4) Los procariotas mesofílicos y termofílicos difieren en el uso de codones a nivel de Arg: los primeros prefieren CGN y los segundos AGR.



Distribución de las especies procariotas en el plano definido por eje 1 y 2



Distribución de los codones en procariotas temofílicos y mesofílicos (ver ejes)



Otros factores que influyen enel UCS

5) Frecuencia de dinucleótidos, en particular, se evita el uso de codones que en las posiciones 1-2, 2-3, y 3-1 contengan el dinucleótido CpG. A su vez, sobre todo en virus con genoma ARN, se evita el dinucleótido UpA.

6) Estructura secundaria del mARN (vida media, afinidad por el ribosoma).

7) En regiones cortas, el uso de codones que aceptan el mismo t-RNA serían más frecuentes que los que requieren otro isoaceptor.

8) Habría una adaptación a la frecuencia relativa de los isoaceptores a fin de evitar que los ribosomas se "atasquen" en el mARN, lo que lleva a cambios de marco y terminaciones prematuras.

9) Etc… Cabe esperar que se sigan detectando otros factores.

En resumen...

1) El uso de codones sinónimos (UCS) o sesgo en el uso de codones, es cuando todos los sinónimos no son usados con la frecuencia esperada por azar.

2) El principal factor determinante para ese sesgo, es el contenido en G+C, sea a nivel genómico global (procariotas) o a nivel de isocoros (mamíferos y aves).

3) Para entender el sesgo en el UCS hay que tener presente que a) existe un tambaleo entre la 3 posición del codón en el mARN y la 1 en el anticodón del tARN, b) los isoaceptores (tARNs con distintos anticodones pero que cargan el mismo aa) están a nivel intracelular a distinta concentración en distintas especies y muy probablemente tejidos.

4) A la composición genómica (distintos valores en G+C) se le superponen distintos factores que interactúan, a modo de ejemplo:

a) Selección natural a nivel traduccional (velocidad, fidelidad). Los codones más usados (sobre todo en los genes de más alta expresión) se los llama codones óptimos o mayores.

b) Frecuencia de dinucleótidos (evitar CpG y UpA).

c) Nivel de hidropatía.

d) Estructura secundaria de la proteína (el UCS puede variar entre hojas β, α hélices y estructuras desordenadas).

e) El UCS difiere (a nivel de la Arg) entre procariotas mesófilos y termófilos.

f) Todos estos factores actúan, tanto a nivel del genoma global como a nivel de cada gen individual, con distinta "fuerza".

g) Por lo tanto, el fenotipo final (uso de codones de todos los genes sumados de una especie) es el resultado de múltiples factores (tanto selectivos como neutros) que actúan de distinta forma, y hasta quizás en distinta dirección, en cada gen individual.