T 575
ALV

**TESIS DE DOCTORADO**

**Título**

# ESTUDIO DE LOS FACTORES QUE AFECTAN LAS TASAS DE EVOLUCIÓN NUCLEOTÍDICA CON ESPECIAL ÉNFASIS EN LAS POSICIONES SINÓNIMAS

**AUTOR:**

Fernando Alvarez

## ORIENTADORES:

Prof. Giorgio Bernardi

Prof. Ruben Budelli


07432

ÍNDICE

**AGRADECIMIENTOS**

Agradezco a Oliver Clay, Antonio Miranda, Hector Musto y a Carlos Robello por las largas y fructíferas discusiones que hemos sostenido;

a la Profesora Scvortzoff (Katia) por haber contribuido en forma decisiva al desarrollo de la Genética Evolutiva en el país;

a CSIC y UNESCO/Twas por el apoyo económico;

a Monica Zambra y Luis Belcredi por haberme facilitado la infraestructura necesaria para la edición e impresión de esta tesis;

a Rafael Saa y a Jorge Elissalde puesto que gracias a su fantástico trabajo fue posible la existencia de genetica.edu.uy;

a Leo y Sandra por haberme soportado a diario durante varios meses;

finalmente, deseo agradecer a mis orientadores los Profesores Giorgio Bernardi y Ruben Budelli por su gran apoyo brindado desde todo punto de vista.

Este trabajo está dedicado a Andrea, Pedro y Gabriela.

**RESUMEN**

Los factores que gobiernan el cambio nucleotídico a escala evolutiva lejos de encontrarse dilucidados constituyen uno de los puntos que ha concitado mayor controversia en el área de la genética evolutiva. Esta controversia se ha centrado en los últimos años en un tipo particular de sustituciones nucleotídicas: las sustituciones sinónimas, es decir aquellas que no alteran la naturaleza codificante de los genes. El grupo biológico donde la polémica ha cobrado mayor relieve lo constituyen los vertebrados, y dentro de estos los mamíferos en particular. Dos visiones alternativas han sido propuestas. Una de ellas (neutraslista-mutacionista) sostiene que las sustituciones sinónimas son el resultado de procesos estocásticos donde variantes funcionalmente equivalentes (selectivamente neutras) se reemplazan unas a otras a lo largo del proceso evolutivo. En este proceso la mutación jugaría un rol preponderante. La visión alternativa propone en cambio, que tanto la composición de bases en las posiciones sinónimas como la velocidad con las que estas evolucionan tienen relación con el funcionamiento del aparato de expresión de los genes y que por tanto estarían gobernadas por la selección natural.

En el presente estudio hemos analizado las tasas de cambio nucleotídico en tres grupos biológicos que abarcan buena parte de la escala evolutiva de los eucariotas: los trypanosomátidos, las gramíneas y los mamíferos. Nuestros resultados nos permiten afirmar que en estos tres grupos biológicos las tasas de cambio aminoacídico y sinónimo así como la composición de bases de las posiciones sinónimas se encuentran correlacionados. Los genes que evolucionan más lentamente a nivel de aminoácidos también lo hacen a menor velocidad en sus posiciones sinónimas y presentan una determinada composición de bases en las mismas. En uno de los grupos (los trypansomátidos) fue posible además identificar que la composición de bases sinónima de los genes conservados coincide con la que se encuentra en genes que se expresan a niveles altos, lo que indicaría que los aminoácidos conservados estarían codificados por codones traduccionalmente óptimos. Este estudio se realizó además a nivel intragénico. Los resultados obtenidos en este nivel son compatibles con los que se observan de los valores promedio de los genes. Esto es, el patrón espacial (intragen) de cambio sinónimo está interrelacionado con el de cambio no-sinónimo y con la composición de bases sinónima. Las regiones de los genes conservadas a nivel de aminoácidos también tienden a estar conservadas a nivel sinónimo y presentan una composición de bases distinta (en general riqueza en G+C) a la de las zonas no conservadas.

Por otro lado se investigó en que medida el patrón de mutaciones determina el contenido en GC sinónimo de los genes de mamíferos. Dicho patrón mutacional fue inferido a partir de las sustituciones en seudogenes así como de la información de secuencia de mutaciones deletéreas que producen desórdenes genéticos en el hombre. Las estimaciones provenientes de ambas fuentes coinciden en la estimación del patrón mutacional y ponen en evidencia que ni el alto contenido en GC sinónimo de los genes de mamíferos ($GC_3$) así como tampoco la variación composicional de estos genomas (isocoros) pueda ser atribuible a una variación concomitante en el espectro mutacional.

Estos resultados tomados en conjunto apoyan la segunda hipótesis, y en particular descartan la incidencia de los sesgos en las mutaciones como moldeadores determinates de la arquitectura de los genomas de mamíferos.

Otros dos resultados merecen destaque. Por un lado se encontró, en todos los grupos analizados, correlación entre la tasa (velocidad) de cambio aminoacídico y el contenido de bases de la segunda posición de los codones ($GC_2$). Sin embargo dicha relación no pudo ser asociada a ninguna causa biológica clara. En segundo lugar, los análisis de las tasas de cambio en genes de trypanosomátidos muestran que los trypanosoma africanos han estado sujetos a una aceleración de por lo menos 400% en sus tasas de cambio sinónimo, siendo la aparición de las proteínas hipervariables (VSGs) de superficie la razón probable de la mencionada aceleración.

# I INTRODUCCIÓN Y ANTECEDENTES

Una de las primeras predicciones de la teoría neutralista de la evolución molecular (Kimura 1968, 1983; King & Jukes 1969) era que los cambios sinónimos (entre codones que codifican el mismo aminoácido) deberían estar completamente exentos de selección natural. Esta presunción estaba basada en el simple hecho que dichos cambios no alteran la naturaleza codificante de los genes, al no implicar modificaciones de ningún tipo en la estructura primaria de las proteínas por ellos codificadas. Debido a esta naturaleza supuestamente inocua de las mutaciones sinónimas, era de esperar que las diferencias de índole sinónima se acumularan a una tasa muy alta en el proceso de diferenciación entre genes y especies. Esta afirmación de la teoría neutral representa un caso particular de uno de sus postulados básicos: la tasa de cambio a nivel molecular es inversamente proporcional al grado de restricción funcional.

## I.1 USO DE CODONES SINÓNIMOS

### I.1.1 Uso de codones y disponibilidad de ARNts

A medida que se acumularon datos de secuencias nucleotídicas, resultó evidente que los distintos codones sinónimos aparecen en los genes con frecuencias que claramente se apartan de una distribución al azar. Este fenómeno conocido como "uso de codones sinónimos" se observa tanto en organismos procariotas como eucariotas (Grantahm et al 1980). Por otro lado, estos mismos estudios evidenciaron que los distintos genes de una especie dada (o de especies emparentadas) tienden a presentar el mismo patrón en la preferencia de codones sinónimos. Sobre la base de estos resultados Grantham et al (1980) formularon lo que se conoce como hipótesis del genoma ("genome hypothesis") que básicamente propone que la estrategia codificante (en lo que se refiere a las preferencias de codones sinónimos) es especie específica.

Por otro lado, Ikemura (1981,1982) presentó evidencia que indica que en *Escherichia coli* y *Saccharomyces cerevisiae* el uso de codones está, en gran medida, determinado por la abundancia relativa de los ARN de transferencia (ARNt) que reconocen a los correspondientes codones. Esto es, los codones más frecuentes (codones "óptimos") son reconocidos por los ARNt más abundantes. Esta correlación entre abundancia de codones y abundancia de ARNt es especialmente evidente en los genes que codifican proteínas que están presentes en altas concentraciones. Esto llevó inmediatamente a postular que en estos microorganismos el uso de codones estaría determinado por la selección para incrementar la eficiencia traduccional. Este incremento en la eficiencia de la síntesis proteica sería debido a que el número medio de interacciones entre el ARNt y el sitio-A del ribosoma se reduce considerablemente en un sistema que posea sesgo en el uso de codones y sesgo en las poblaciones de ARNt en relación a un sistema en el cual todos los codones y ARNt son equiprobables. Como resultado de este decremento en el número de interacciones, se reduce el tiempo medio de espera (del aminaoacil-ARNt correcto) por aminoácido, lo que a su vez resulta en una reducción neta del número de ribosomas necesarios para producir una determinada cantidad de proteína por unidad de tiempo.

El uso sesgado de codones acompañado con sesgos en las poblaciones de ARNt también ha sido implicado en el incremento de la fidelidad traduccional. Por un lado se ha demostrado experimentalmente en *E. coli* que la sustitución de un codón mayor (es decir reconocido por un ARNt abundante) por un codón menor, produce un incremento de casi 10 veces en la tasa de errores traduccionales en el aminoácido donde se realizó la sustitución (Precup & Parker, 1987). Por otro lado, Akashi (1994) analizando genes de *Drosophila melanoganster* y *D. simulans* mostró que los aminoácidos funcionalmente importantes (en general pertenecientes a motivos conservados o dominios proteicos cuya función es conocida) tienden a estar codificados por codones óptimos en una frecuencia significativamente más alta que los aminoácidos no conservados.

Por último, el uso de codones sinónimos ha sido involucrado también en el plegamiento de las proteínas durante la síntesis proteica (Purvis et al, 1987), en la estabilidad del ARN mensajero (ARNm) así como en la formación de estructuras secundarias del ARNm. Como se mencionó anteriormente, la existencia de sesgos en las poblaciones de ARNt como de las frecuencias de los codones sinónimos trae aparejado que el ribosoma tenga tiempos de espera distintos según se trate de un codón mayor o menor. Como resultado, la velocidad de avance de los ribosomas no es homogénea a lo largo del ARNm, sino que por el contrario existirían zonas de "tránsito rápido" así como pausas traduccionales. Resulta obvio que la existencia de pausas traduccionales implica que los ARNm en proceso de traducción posean zonas de "aglomeración" de ribosomas así como regiones en los cuales los ribosomas estarían prácticamente ausentes, generando por tanto zonas con distinta sensibilidad a las ribonucleasas endógenas ya sea directamente (protección del ARN) o afectando el plegamiento del ARNm (Gross et al, 1990; Zama, 1990). Debe tenerse en cuenta además, que la variación en la velocidad de avance de los ribosomas lleva implícita la variación en el ritmo de crecimiento del polipéptido naciente, lo que afectaría los patrones de plegamiento del mismo (estructura terciaria de la proteína) como ha sido puesto en evidencia en la proteína de cobertura de fago MS2 (Guisez et al, 1993) y en las colicinas de *E. coli* A, E1,E2, en las proteínas TEM, 1-B-lactamasa y OmpA (Varenne et al, 1984) a través de la demostración de la existencia de intermediarios de la traducción abortivos que coinciden con zonas de agrupamiento de codones menores.

De lo anteriormente expuesto surge que la distribución espacial (intragén) de codones mayores y menores debería depender de la distribución espacial de los aminoácidos importantes (incremento de la fidelidad traduccional), de la estructura tridimensional de la proteína (pausas de la traducción), así como de estructura secundaria del ARNm (degradabilidad del ARNm).

### I.1.2 Uso de codones y sesgos de origen mutacional

Las mutaciones puntuales (cambio de una base por otra) no ocurren generalmente en forma equiprobable. Por el contrario determinadas bases tienden a mutar con mayor probabilidad que otras, así como determinados tipos de cambios (en general las transiciones) ocurren con mayor frecuencia. A este sesgo en la direccionalidad de las mutaciones se le llama sesgo mutacional o presión mutacional (Sueoka, 1989,1992). La base bioquímica del mismo ha sido asociada con sesgos en la introducción de errores durante la replicación (por parte de las ADN polimerasas) y con mecanismos de "missmatch repair" que tienden a corregir preferentemente en un sentido (Sueoka, 1988, 1992). Como resultado de esta presión mutacional y en ausencia de selección natural, o con niveles bajos de selección, la composición de bases del ADN cambia, generalmente volviéndose rica en GC (presión mutacional GC) o rica en AT. Por esta razón, se asume que las secuencias nucleotídicas que no están bajo fuertes presiones selectivas (intrones, regiones flanqueantes, seudogenes), presentan una composición de bases que refleja la dirección del sesgo mutacional. Más precisamente, la dirección y magnitud del sesgo mutacional suelen inferirse tomando como base la composición de bases de dichas secuencias (Sharp & Devine 1989; Normura et al, 1987, Sharp, 1990; Moriyama & Hartl, 1993).

Se ha postulado que la presión mutacional podría tener efecto en la determinación de la composición de bases en las posiciones sinónimas (en general en la 3ra posición del codón) puesto que las restricciones selectivas son bajas en dichas posiciones. Por tanto, la presión mutacional podría en buena medida ser responsable de la existencia del uso de codones no azaroso (Sueoka, 1988, 1992). Estudios del uso de codones en varios organismos han demostrado la importancia del sesgo mutacional en la determinación de la preferencia de codones sinónimos. Sin embargo, el peso de la presión mutacional sería más importante en aquellos genes en los cuales la selección por incrementar la eficiencia traduccional es poco importante, es decir genes de baja expresión, mientras que en los genes expresados a altos niveles, la determinación de las preferencias de codones estaría fuertemente influida por la disponibilidad de ARNt (Alvarez et al, 1994; Sharp & Devine, 1989; Shields & Sharp, 1987). Por tanto cada gen poseería un uso de codones que reflejaría un punto de equilibrio entre el sesgo mutacional y la selección natural. La ubicación de este punto de equilibrio para cada gen dependería del nivel de expresión del mismo. Esto es lo que se conoce como la "Selection-Mutation-Drift Theory" (Bulmer, 1991). Es interesante resaltar que cuando la selección y el sesgo mutacional coinciden en la dirección, la diversidad intergén (intragenómica) en el uso de codones es baja, como ocurre en *Leishmania*, *Crithidia* (Alvarez et al 1994) y probablemente también en *Plasmodium* (Musto et al, 1997); mientras que en especias que presentan selección y sesgo mutacional empujando en direcciones opuestas, la diversidad intergén en la preferencia de codones es muy grande, como se observa en *Dyctiostelium* (Sharp & Devine, 1989) y *Trypanosoma* (Alvarez et al, 1994).

### I.1.3 Variación en el uso de codones debida a la compartimentalización genómica

Los genomas de los vertebrados de sangre caliente no son homogéneos en su composición de bases. Por el contrario, los genomas de aves y mamíferos presentan diversidad composicional que consiste en segmentos de ADN largos (mayores de 300 kb) que son composicionalmente homogéneos (Bernardi et al, 1985). Estos segmentos, que han sido llamados isocoros, no sólo se encuentran en vertebrados de sangre caliente sino también en plantas (Matassi et al, 1989; Salinas et al, 1988). Organizaciones genómicas de tipo isocorial han sido descritas en platelmintos (Musto et al, 1995) e incluso en la levadura *Saccharomyces cerevisiae*.

Además, la composición de bases (contenido G+C), especialmente en la tercera posición de los codones ($GC_3$), varía enormemente entre los genes de mamíferos. Teniendo en cuenta que el contenido en G+C en las posiciones sinónimas está altamente correlacionado con el contenido en G+C de los intrones y secuencias flanqueantes, esta variación entre genes estaría relacionada con la ubicación genómica de los mismos (Bernardi & Bernardi, 1985; D'Onofrio et al, 1991). Por tanto, el uso de codones en mamíferos no sería sino el resultado de la compartimentalización subyacente (Bernardi, 1989).

Dos hipótesis alternativas han sido postuladas para explicar el origen y preservación de los isocoros. La hipótesis seleccionista, sostiene que la organización isocorial tiene significado funcional y por tanto ha surgido como resultado de la selección natural (Bernardi, 1989). La otra hipótesis (mutacionista), sostiene que la diversificación composicional del genoma es el resultado de sesgos mutacionales distintos en distintas regiones del genoma. Puesto que durante el ciclo celular los isocoros ricos en G+C se replican temprano mientras que los ricos en A+T lo hacen tarde, una variación en la composición de las poblaciones de nucleótidos libres (citoplásmicos) a ser incorporados durante la replicación causaría un cambio en la dirección de la introducción de errores replicacionales y por tanto del patrón de mutaciones (Wolfe, Sharp & Li 1989). Por su parte Sueoka (1992), propone que el cambio en la dirección del sesgo mutacional entre isocoros sería debido a la estructura cromatínica laxa que presentarían los isocoros pesados (fundamentalmente H3) en las células de la línea germinal, lo que se debe a la alta concentración de genes "housekeeping" en estos isocoros. La discusión sobre el origen y mantenimiento de los isocoros ha sido uno de los temas mayor controversia en la discusión seleccionismo-neutralismo. Uno de los objetivos de esta tesis es aportar a esta discusión a través del análisis directo de los espectros mutacionales, como se describirá en detalle más adelante.

## I.2 SUSTITUCIONES SINÓNIMAS

Las sustituciones sinónimas (aquellas que no cambian el significado del codón en cuestión) han sido sujeto de estudio intenso en los últimos 20 años. Su importancia radica en el hecho que fueron los candidatos más firmes a estar sujetos a evolución neutral de acuerdo a las predicciones de la teoría neutralista. Se llegó incluso al extremo de postular la existencia de un reloj molecular universal basado en que las sustituciones sinónimas deberían tener la misma tasa de acumulación

en todos los genes de una especie dada (Ochman & Wilson, 1987), desconociendo el hecho que incluso bajo condiciones de estricto neutralismo, la tasa de sustitución varía si hay variación en la tasa de mutación.

Sin embargo la posible existencia de selección en las posiciones sinónimas basada en la evidencia aportada por el uso de codones no azaroso y su vinculación con la disponibilidad de ARNts isoaceptores, llevó a postular que la evolución sinónima debería estar sujeta a presión selectiva.

Posteriormente Sharp & Li (1987) mostraron que en enterobacterias (*E. coli, Salmonella typhimurium*) la tasa de sustituciones sinónimas es inversamente proporcional al grado de sesgo en el uso de codones, esto es, genes con mayor sesgo en sus preferencias de codones (y por tanto expresados a altos niveles y con mayor dependencia de la población de ARNts), evolucionan (a nivel sinónimo) significativamente más lentamente que aquellos genes con menor sesgo (y por lo tanto expresados más débilmente y con menor dependencia de los ARNts). Es interesante resaltar que si bien estos resultados indicarían la existencia de selección negativa para mantener un uso de codones dado, basada en la optimización de la eficiencia traduccional, resultados posteriores indican lo contrario. En efecto, Eyre-Walker & Bulmer (1995) estudiando también genes de enterobacterias demostraron que la tasa de sustitución sinónima no varía entre grupos de codones sujetos a diferentes niveles de presiones selectivas. Por ejemplo en *E. coli*, el aminoácido lisina prácticamente no varía su preferencia de sinónimos entre genes de alta y baja expresión, mientras que el cambio es muy notorio en para el aminoácido fenilalanina. En contraposición a lo que sería de esperar de acuerdo a los resultados de Sharp y Li mencionados más arriba, tanto Lys como Phe presentan el mismo rango de declinación de las tasas de sustituciones sinónimas entre genes de baja y alta expresión. Es de hacer notar que, como resultado de estos mismos estudios se llega a la conclusión de que las mutaciones sinónimas están definitivamente bajo restricción selectiva, puesto que el número medio de sustituciones sinónimas (por sitio sinónimo) entre *E. coli* y *S. typhimurium* es sólo 0.69 en relación a las 15 (aprox.) que se esperarían si las posiciones sinónimas fueran absolutamente neutras (Eyre-Walker & Bulmer 1995). Por otra parte Eyre-Walker y Bulmer (1993) han puesto de manifiesto que los genes de bacterias entéricas presentan niveles reducidos de sustituciones sinónimas en los primeros 50 codones sin que esto vaya acompañado por un incremento en el sesgo direccional del uso de codones. Por el contrario al comienzo de los genes de *E. coli* y *S. typhimurium* el CAI (Indice de Adaptación de Codones, que es una medida del sesgo direccional de codones) es significativamente mas bajo que en las restantes partes de los genes (Eyre-Walker & Bulmer, 1993). Estos resultados indican que si bien es claro que en genes bacterianos las mutaciones sinónimas no están totalmente exentas de selección, las causas funcionales que subyacen no están claras.

### I.2.1 Sustituciones sinónimas en mamíferos

Si bien es generalmente aceptado que en bacterias (y otros microorganismos) las sustituciones sinónimas no están totalmente exentas de presiones selectivas, la situación en vertebrados y particularmente en mamíferos es mucho más controversial. Por un lado se argumenta que los coeficientes de selección asociados a las posiciones sinónimas son excesivamente pequeños para que la selección tenga efecto en taxa que poseen tamaños poblacionales efectivos ($N_e$) comparativamente chicos en relación a los tamaños poblacionales de los organismos unicelulares (Bulmer et al, 1991). Los coeficientes de selección (S) para mutaciones sinónimas han sido estimados en aproximadamente $7.3 \times 10^{-9} \pm 4.7 \times 10^{-9}$ (Hartl et al, 1994), mientras que los tamaños poblacionales efectivos en mamíferos son menores de $10^6$.Por ejemplo Nei & Graur (1984) han estimado el $N_e$ de la especie humana en $10^4$. Puesto que para que la selección tenga efecto el producto Ne*S debe ser mayor 1 (Kimura, 1968) y en los mamíferos este producto (para las sustituciones sinónimas) nunca podría llegar a serlo (si las estimaciones de S fueran correctas), entonces las restricciones funcionales que puedan estar asociadas a las posiciones sinónimas serían demasiado débiles para poder producir algún efecto observable tanto en la determinación del uso de codones como en las tasas de sustituciones sinónimas.

Sin embargo resultados del análisis de genes de mamíferos aportan diversas evidencias en favor de que la selección tendría algún efecto en la determinación del uso de codones como en las velocidades de evolución sinónima. Li et al (1981) han reportado que si bien la tasa de cambio sinónimo en genes de mamíferos es mucho más alta que la tasa no-sinónima, esta última es bastante menor a la tasa de cambio encontrada en seudogenes, indicando que algún tipo de selección opera en contra de la fijación de mutaciones sinónimas. Además, el rango de variación de la velocidad sinónima entre genes ha sido estimado entre 5 (Li & Grauer, 1991) a 20 veces (Bernardi et al, 1993; Wolfe & Sharp, 1993). Dicho rango de variación es muy superior al que podría esperarse si sólo dependiera de la varianza mutacional. Recientemente se ha presentado evidencia adicional que sugiere que en los genes de mamíferos los cambios sinónimos no se encuentran totalmente libres de restricciones funcionales. Mouchiroud et al (1995) encontraron que las tasas sinónimas son gen-específicas puesto que procesos de divergencia independiente (humano-bovino; rata-ratón) producen distancias que están muy fuertemente correlacionadas. En congruencia con estos resultados, Cacciò et al (1995) reportaron que la tasa de divergencia sinónima en duetos (grupos de codones de doble degeneramiento) está correlacionada con la de los grupos de codones de cuádruple degeneramiento.

### I.2.2 Correlación entre sustituciones sinónimas y no-sinónimas

Varios autores han reportado independientemente que en genes de mamíferos las tasas de sustituciones sinónimas y no-sinónimas no son independientes. Por el contrario, existe una correlación relativamente alta (r entre 0.5 y 0.6) y muy significativa entre ellas (Ficth, 1980; Graur, 1985; Li et al, 1985; Mouchiroud et al, 1995; Otha & Ina, 1995; Wolfe & Sharp, 1993). Esta

correlación ha sido atribuida a distintas causas. Por un lado Wolfe & Sharp (1993) sostienen que la misma podría ser explicada como el resultado de mutaciones en dobletes. Esto es, dos bases consecutivas que mutan simultáneamente como resultado de un único evento. Sin duda, si estas mutaciones fueran frecuentes podrían explicar en parte la correlación en cuestión, puesto que al cambiar dos bases consecutivas, en aproximadamente 1/2 de casos se produciría una mutación sinónima y una no-sinónima. Sin embargo, Mouchiroud et al. (1995) han demostrado que si se eliminan de los alineamientos las sustituciones consecutivas en las posiciones del codón 2-3 y 3-1 (es decir sustituciones sinónimas y no-sinónimas consecutivas que podrían haber sido originadas a partir de una única mutación en doblete) la correlación baja pero se mantiene altamente significativa. Como explicación alternativa, los mismos autores propusieron que la correlación en cuestión debería ser la consecuencia de restricciones funcionales comunes a los cambios sinónimos y no-sinónimos.

## II OBJETIVOS E HIPÓTESIS DE TRABAJO

El objetivo de este estudio consiste en aportar ideas, resultados y nuevas aproximaciones que contribuyan a desentrañar los factores subyacentes al uso de codones y las tasas de cambio sinónimo. Se ha puesto especial énfasis en dos aspectos:

1- Estudio de las tasas de las sustituciones sinónimas en diversos sistemas biológicos. Uno de los aspectos de mayor relevancia en este trabajo lo constituye el estudio del papel de la selección natural en el incremento la fidelidad traduccional así como la relación entre esto y la covariación entre las tasas de cambio sinónimo y no-sinónimo.

2- Papel de los sesgos mutacionales en la determinación de la composición de bases sinónimas y uso de codones en mamíferos, así como en el rol de esta fuerza en el origen y mantenimiento de los isocoros.

## III MATERIAL DE ESTUDIO Y METODOLOGÍA

### III.1 ESTUDIO DE LAS SUSTITUCIONES SINÓNIMAS

#### III.1.1 Grupos biológicos analizados

El estudio de las tasas de sustituciones sinónimas se realizó en tres grupos taxonómicos distintos: mamíferos, gramíneas y en los trypanosomátidos. La elección del primer grupo está basada en que el grupo presenta fundamental importancia para la dilucidación en el mismo del posible papel de la selección. Como fue mencionado más arriba, el papel de la selección natural en la determinación del uso de codones y en la velocidad de evolución sinónima es un tema aún en disputa para los mamíferos. Por su parte las gramíneas resultan también de interés por ser un grupo que al igual que los mamíferos ha sufrido una transición composicional consistente en un incremento muy notorio en su contenido en G+C, particularmente en la tercera posición de los codones

Por último los trypanosomátidos han sido elegidos debido principalmente a dos razones. En primer lugar son microorganismos, y por tanto presentan tamaños poblacionales presumiblemente grandes. Como consecuencia, fuerzas selectivas asociadas a coeficientes de selección muy pequeños (como se presume es el caso de los cambios sinónimos) pueden ejercer una influencia muy significativa y por lo tanto son factibles de ser detectados con relativa facilidad. En segundo lugar los trypanosomátidos representan uno de los grupos de eucariotas más primitivos. Este hecho presenta gran importancia puesto que en esta tesis se analizan organismos que abarcan casi todos los grandes grupos de eucariotas (metazoarios, plantas y protozoarios primitivos). Resulta claro entonces que si los grupos acá analizados presentasen mecanismos similares gobernando la evolución en las posiciones sinónimas, los mismos podrían ser generalizados a todos los sistemas eucarióticos.

#### III.1.2 Genes utilizados y alineamientos de secuencias

Se analizaron genes ortólogos (homólogos cuyas relaciones filogenéticas coinciden con la de las especies), los cuales se alinearon para determinar los codones homólogos. El alineamiento

de genes homólogos se realizó usando el programa de alineamiento múltiple Clustalw (Higgins et al., 1992). La estrategia utilizada para obtener alineamientos de alta calidad consiste en alinear primero las secuencias traducidas (es decir a nivel aminoacídico) y luego realizar una "retrotraducción" (backtranslation) basándose en las secuencias de ADN que son conocidas (Alvarez et al, 1996). Esta modalidad para obtener alineamientos disminuye considerablemente el número de posiciones nucleotídicas cuya homología sea discutible debido al hecho que limita la introducción de "indeles" (inserciones-deleciones) a tamaños múltiplos de tres. Dicha limitación presenta la ventaja de impedir indeles que causen corrimientos de marco de lectura, los cuales obviamente son altamente improbables desde el punto de vista biológico.

En el análisis de genes de mamíferos se utilizaron grupos de genes ortólogos que estuvieran presentes al menos en cuatro órdenes diferentes. La lista completa de genes de mamíferos usados se encuentra en la Tabla 1.1. Los criterios elegidos para seleccionar dichos genes están descritos en la sección "Material and Methods" de la primera publicación que se anexa.

En el caso de genes de Gramíneas se utilizaron tres bases de datos. La primera de ellas consiste en 40 genes que son homólogos entre maíz (*Zea mais*) y arroz (*Oriza sativa*), la segunda en 47 genes homólogos entre maíz y cebada (*Hordeum vulgare*) o trigo (*Triticum aestivum*). El tercer grupo de datos involucra 32 genes homólogos entre arroz y cebada o trigo. Es necesario aclarar que cebada y trigo fueron considerados como un mismo taxón. Esto tiene su justificación en el hecho que para cualquier gen en particular, la distancia entre trigo y cebada es mucho menor que la distancia entre cebada y arroz (o maíz) o aquella entre trigo y arroz (o maíz). Además, los estudios de filogenias moleculares indican que trigo y cebada son muy cercanos y forma un grupo monofilético cuando se compara con maíz o arroz (Duval & Morton, 1996). Información adicional relativa a estas bases de datos, así como los criterios utilizados para seleccionar los homólogos están descritos en la sección material y métodos de la tercera publicación que se anexa.

El último grupo taxonómico analizado fue el de los trypanosomátidos. En este caso se utilizaron dos bases de datos. La primera consiste en 42 alineamientos (listados en la Tabla 3.1) conteniendo genes nucleares, homólogos entre *Leishmania*, *Trypanosoma cruzi* y entre *T. cruzi* y *T. brucei*. Dicha base de datos fue básicamente utilizada para testar la homogeneidad de las tasas de sustituciones sinónimas y no-sinónimas entre *T. cruzi* y *T. brucei* utilizando a *Leishmania* como el grupo externo ("outgroup"). La segunda base de datos contiene 19 genes codificantes de la metaloproteinasa de membrana (GP63). Es necesario aclarar que fue esta segunda base de datos y no la primera, la utilizada para analizar la covariación intragénica entre sustituciones sinónimas y no-sinónimas. La razón para la no-utilización del primer grupo de genes homólogos de trypanosomátidos en el análisis de la covariación intragénica es que las distancias sinónimas entre *T. cruzi* y *T. brucei* o entre trypanosomas y *Leishmania* son excesivamente grandes, e incluso en algunos casos lo es tanto que lleva a que el método utilizado para estimar las distancias sinónimas sea inaplicable. A partir de esto resulta claro que un análisis de la covariación intragénica entre sustituciones sinónimas y no-sinónimas es inviable en este grupo de genes pues este análisis

implica el cálculo de distancias en fragmentos de genes lo cual está sujeto a errores estocásticos mucho mayores que el cálculo para genes completos.

### III.1.3 Métodos y análisis

#### III.1.3.1 Estimación de las distancias sinónimas y no-sinónimas

Estimar distancias sinónimas (y no-sinónimas) entre dos genes homólogos presenta varias dificultades. Dicha estimación implica en primer lugar hacer un conteo del número de cambios sinónimos y no-sinónimos. En segundo lugar hay que contar el número de sitios sinónimos y no-sinónimos en cada gen. Por último hay que usar alguna aproximación que permita corregir para sustituciones múltiples, paralelas y hacia atrás. Los dos primeros pasos en la estimación de las distancias son complejos debido a las siguientes causas. Cuando comparamos dos codones homólogos que difieren en un sólo nucleótido la decisión de sí se trata de un cambio sinónimo o no-sinónimo es trivial, puesto que no hay ambigüedad, la sustitución implica o no una alteración en la naturaleza codificante del codón en cuestión. Sin embargo cuando los codones homólogos difieren en dos o en tres nucleótidos la estimación es considerablemente más complicada y además está sujeta a una fuente de error mayor pues depende del modelo evolutivo que utilicemos. Veamos los siguientes ejemplos: CTA (Leu) <=>CCT (Pro). Existen dos caminos alternativos (mutuamente excluyentes) para pasar de un codón a otro. 1) CTA (Leu)<=> CTC (Leu) <=> CCT (Pro) y 2) CTA (Leu)<=> CCA (Pro) <=> CCT (Pro). En este ejemplo la situación es relativamente sencilla pues en ambos caminos hay una sustitución sinónima y una no-sinónima obligadas. Pero el cálculo se torna mucho más complejo cuando los codones difieren en los tres nucleótidos Analicemos el cambio TTT (Phe)<=>GGG (Gly). En este caso hay seis caminos mutuamente excluyentes para pasar de un codón al otro:

| | Núm. de cambios | |
|---|---|---|
| | sinónimos | no-sinónimos |
| 1- (321) TTT (Phe)<=> TTG (Leu) <=>TGG (Trp) <=>GGG (Gly) | 0 | 3 |
| 2- (312) TTT (Phe)<=> TTG (Leu) <=>GTG (Val) <=>GGG (Gly) | 0 | 3 |
| 3- (123) TTT (Phe)<=> GTT (Val) <=>GGT (Gly) <=>GGG (Gly) | 1 | 2 |
| 4- (132) TTT (Phe)<=> GTT (Val) <=>GTG (Val) <=>GGG (Gly) | 1 | 2 |
| 5- (231) TTT (Phe)<=> TGT (Cys) <=>TGG (Trp) <=>GGG (Gly) | 0 | 3 |
| 6- (213) TTT (Phe)<=> TGT (Cys) <=>GGT (Gly) <=>GGG (Gly) | 1 | 2 |
| Total | 3 | 15 |

Podemos considerar a todos los caminos como igualmente probables, en este caso tendremos 3/18*3=0.5 sustituciones sinónimas y 15/18*3=2.5 sustituciones no-sinónimas. Alternativamente podemos considerar que aquellos caminos que incluyan cambios aminoacídicos drásticos (por ej. cambiar un aminoácido básico por otro ácido) son más improbables que aquellos que incluyan caminos que contengan cambios entre aminoácidos bioquímicamente similares o sustituciones sinónimas. Es posible usar cualquier índice de similitud química entre aminoácidos (por ej.

Gramtham, 1974) o medidas empíricas de intercambiabilidad evolutiva entre aminoácidos (por ej. Dayhoff et al, 1978) para darle peso diferencial a los diferentes caminos.

En cuanto al cálculo del número de posiciones sinónimas y no-sinónimas su determinación depende del codón en cuestión y del modelo de sustituciones nucleotídicas que asumamos. Las terceras posiciones de los codones de cuádruple degeneramiento son 100% sinónimas puesto que cualquier cambio en ellas nunca involucra un cambio aminoacídico. Las segundas posiciones de todos los codones así como la mayoría de las primeras posiciones y las terceras posiciones de los codones para triptofano y metionina son 100% no-sinónimas puesto que cualquier cambio en ellas implica un cambio en el aminoácido codificado. Por su parte las terceras posiciones de los codones de doble degeneramiento (duetos), la primera posición de los codones CTG,CTA, TTG y TTA para leucina y las primeras posiciones de los codones CGA, CGG, AGG y AGA para arginina son parcialmente sinónimas y parcialmente no-sinónimas, puesto que el cambio puede ser sinónimo o no-sinónimo dependiendo de que se trate de una transición (purina por purina o pirimidina por pirimidina) o una transversión (purina por pirimidina). En el caso particular de las terceras posiciones de los duetos las transiciones son siempre sinónimas mientras que las transversiones son no-sinónimas. En este tipo de posiciones la proporción con la cual las consideramos sinónimas o no-sinónimas depende del modelo de sustituciones que asumamos. Si consideramos que todos los cambios nucleotídicos son igualmente probables (modelo Jukes-Cantor, 1969) entonces las posiciones con doble degeneramiento serán 1/3 sinónimas y 2/3 no-sinónimas puesto que sólo uno de los tres cambios posibles mantiene la naturaleza codificante del codón, mientras que los otros dos cambios dan lugar a cambio de aminoácido. Si por el contrario usamos un modelo en el cual las transiciones son más probables que las transversiones, entonces la proporción en la cual la posición de doble degeneramiento será considera sinónima dependerá de la relación transiciones/transversiones que usemos.

Por último tenemos que considerar el modelo para corregir las sustituciones múltiples el cual obviamente está relacionado con el modelo de cambio nucleotídico que asumimos para contabilizar el número de posiciones sinónimas y no-sinónimas. El modelo más sencillo, y también el más ampliamente utilizado, es el desarrollado por Jukes & Cantor (1969) el cual asume que todos los tipos de cambios nucleotídicos son equiprobables. En dicho modelo la distancia se estima utilizando

$$D = -(\frac{1}{2}) \log[(1-2P-Q)\sqrt{(1-2Q)}]$$

la siguiente expresión:

$$D = -(3/4)^* \log[1-(4/3)^*S]$$

en la cual **S** simboliza la similitud observada, que para el caso de las posiciones sinónimas es:

**S**=(1-Núm. cambios sinónimos)/Núm. de sitios sinónimos.

Otro modelo ampliamente usado es el de dos parámetros de Kimura (1980). Este modelo diferencia entre transiciones y transversiones. La distancia se estima a partir de la siguiente expresión:

donde **P** simboliza la frecuencia observada de transiciones y **Q** la de transversiones. Se han desarrollado muchos otros modelos para estimar distancias nucleotídicas a partir del número observado de cambios. Dichos modelos se diferencian de los anteriores en que incorporan mayor cantidad de parámetros en la estimación. Aunque esto puede significar un avance pues la estimación está basada en un modelo más realista, en la práctica estos métodos resultan menos eficientes que los anteriores pues sus estimaciones presentan mayor varianza debido al gran número de parámetros a estimar (ver Zharkikh, 1994).

Los distintos métodos desarrollados para estimar distancias sinónimas y no-sinónimas se diferencian en cómo combinan los tres aspectos descriptos arriba. El método utilizado en la mayoría de los análisis en esta tesis, el cual a su vez es el más sencillo, fue desarrollado por Nei & Gojobori (1986). Este método no le da peso diferencial a los diferentes caminos para pasar de un codón a otro en el caso de los codones que se diferencien en más de un nucleótido. Además el método de Nei & Gojobori asume el modelo de Jukes & Cantor para calcular el número de sitios sinónimos y no-sinónimos así como para corregir para sustituciones múltiples. A pesar de su simplicidad, se ha demostrado que este método es muy confiable en la mayoría de las situaciones biológicas. Otro método usado en algunos análisis presentados en esta tesis es el desarrollado por Li et al (1985) con las correcciones anexadas por Li (1993) para estimar el número de sitios. Este método clasifica a las posiciones nucleotídicas en sitios de degeneramiento cero ($L_0$), sitios de doble degeneraniento ($L_2$) y sitios de cuádruple degeneramiento ($L_4$). Esta clasificación coincide (en términos generales) con la descripta arriba para la clasificación de posiciones sinónimas y no-sinónimas. Posteriormente se contabilizan el número de transiciones ($p_i$) y transversiones ($q_i$) que ocurren en cada tipo de sitio. Todas las sustituciones (tanto transiciones como transversiones) en los sitios $L_0$ son no-sinónimas mientras que todas las sustituciones en los sitios $L_4$ son sinónimas. Por su parte las transiciones en los sitios $L_2$ son sinónimas mientras que las transversiones en estos sitios son no-sinónimas. El propósito de esta clasificación es permitir estimar separadamente la tasa de transiciones y transversiones en cada tipo de sitio lo cual luego nos permitirá aplicar el método de dos parámetros de Kimura para estimar las tasas de sustituciones sinónimas y no-sinónimas.

Usando las fórmulas de Kimura (1980) podemos calcular la tasa de transiciones ($A_i$) y transversiones ($B_i$) para cada tipo de sitio:

$$A_i = (1/2)\log(a_i) - (1/4)\log(b_i) \quad \text{y} \quad B_i = (1/2)\log(b_i);$$

donde $a_i = 1/(1-2*P_i-Q_i)$ y $b_i = 1/(1-2*Q_i)$, siendo $P_i = p_i/L_i$ y $Q_i = q_i/L_i$.

La tasa total de sustituciones para cada tipo de sitio ($K_i$) estará dada por:

$$K_i = A_i + B_i$$

Posteriormente se calcula la tasa sinónima ($K_s$) y no-sinónima $K_a$ considerando los sitios $L_2$ como un tercio sinónimos y dos tercios no-sinónimos, es decir usando el modelo Jukes-Cantor (Li et al, 1985).

$$K_S = 3*(L_2*A_2 + L_4*K_4)/(L_2 + 3*L_4)^2,$$

$$K_a = 3*(L_2*B_2 + L_0*K_0)/(2*L_2 + 3*L_0)^2$$

Sin embargo esta forma de ponderar los sitios tiende a subestimar en $L_2$ el número de sitios sinónimos y sobrestimar el número de sitios no-sinónimos (puesto que normalmente las transiciones son más frecuentes que las transversiones) y como consecuencia se sobrestima $K_S$ y subestima $K_a$. Debido a esto Li (1993) y Pamilo & Bianchi (1993) han propuesto ponderar el número de sitios sinónimos y no-sinónimos proporcionalmente a la relación observada entre transiciones y transversiones. De esta manera se obtienen las siguientes fórmulas:

$$K_S = B_4+(L_2*A_2+L_4*A_4)/(L_2+L_4),$$

$$K_a = A_0+(L_0*B_0+L_2*B_2)/(L_2+L_0)$$

Las varianzas para $K_S$ y $K_a$ están dadas por

$$V(K_S)=V(B_4)+[L_2^{2*}\ V(A_2)+\ L_4^2\ *V(A_4)]/(L_2+L_4)^2-b_4*Q_4*[2*a_4*P_4-c_4*(1-Q_4)]/(L_2+L_4)$$

$$V(K_a)=V(A_0)+[L_0^{2*}\ V(A_0)+\ L_2^2\ *V(A_2)]/(L_0+L_2)^2-b_0*Q_0*[2*a_0*P_0-c_0*(1-Q_0)]/(L_2+L_0)$$

donde $c_i=(a_i-b_i)/2$.

**III.1.3.2 Comparación de las tasas de sustituciones entre linajes.**

En ausencia de información precisa sobre tiempos de divergencia el problema de testar la homogeneidad en las tasas de sustituciones nucleotídicas en distintos linajes evolutivos puede resolverse mediante del uso de una tercera especie de referencia que haya divergido antes que las dos especies que estamos considerando (grupo externo, ver figura **a** más abajo). Esta aproximación, conocido como test de las tasas relativas ("relative-rate test"; Sarich & Wilson,1973), presenta la ventaja de no requerir ningún tipo información de los tiempos de divergencia entre las especies que se están analizando.

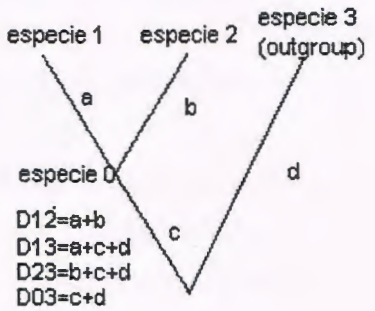



**Figura a**

En los análisis presentados en esta tesis se ha utilizado la aproximación basada en las varianzas de

$$Z = \frac{(D_{13} - D_{23})}{\sqrt{(\mathrm{var}(D_{13} - D_{23}))}}$$

las distancias (Wu y Li, 1985). En este test se calculan las distancias entre las especies 1 y 2 ($D_{12}$), así como las distancias entre cada una de estas especies y el grupo externo ($D_{13}$ y $D_{23}$). La hipótesis nula ($H_0$) es que: $D_{13}=D_{23}$, o lo que es lo mismo $D_{13}-D_{23}=0$. Para determinar si dicha diferencia es significativamente diferente de cero se usa la distribución normal. De esta forma tenemos que:

Por ser $D_{13}$-$D_{23}$ la diferencia entre dos variables entonces su varianza estará dada por:

$$Var(D_{13}-D_{23})=Var(D_{13})+ Var(D_{23})-2*Cov\ (D_{13},D_{23})$$

Wu y Li (1985) han utilizado la varianza de la distancia entre el outgroup con el ancestro común de las especies 1 y 2 ($D_{03}$) como medida de la covarianza, es decir se asume que $Cov(D_{13},D_{23})=Var(D_{03})$. Para poder calcular $Var(D_{03})$ resulta imprescindible estimar el número de transiciones ($P_{03}$) y transversiones ($Q_{03}$) en cada uno de los tipos de sitios ($L_0$, $L_2$ y $L_4$) que ocurrieron entre el grupo externo y el ancestro común a las especies 1 y 2. Las fórmulas aproximadas para $P_{03}$ y $Q_{03}$ son (Kimura, 1980):

$$Q_{03} =(1/2) * (1-e^{-2*B_{03}}),$$

$$P_{03} = (1/2)(1-Q_{03}-e^{-2*A_{03}-B_{03}}),$$

donde $B_{03}=(B_{13}+B_{23}-B_{12})/2$, $A_{03}=(A_{13}+A_{23}-A_{12})/2$.

Por último se debe tener en cuenta que si uno de los linajes que conduce a las especies que estamos comparando sufrió una aceleración o enlentecimiento de su tasa evolutiva (con relación al otro linaje) dicho cambio de velocidad necesariamente tiene que haber ocurrido luego de la separación de las especies. Resulta evidente entonces que si deseamos tener una estimación de la magnitud de dicho cambio debemos comparar las distancias entre cada una de las especies con su ancestro común. Para estimar estas distancias ($D_{10}$ y $D_{20}$) usamos las siguientes ecuaciones (Sarich & Wilson, 1973):

$$D_{10}=(D_{13}+D_{12}-D_{23})/2,$$

$$D_{20}=(D_{23}+D_{12}-D_{13})/2,$$

La relación $R=D_{10}/D_{20}$ nos da una idea aproximada de la magnitud del cambio en la tasa evolutiva.

### III.1.3.3 Estudio de la covariación intragénica entre sustituciones sinónimas y no-sinónimas y la composición de bases

El estudio de la relación intragénica entre las tasas evolutivas y la composición de bases se realizó mediante el uso de ventanas móviles. Concretamente, para cada alineamiento se obtienen perfiles de cada una de las propiedades que se desean analizar (distancias sinónimas y no-sinónimas, composición de bases en cada una de las posiciones del codón, etc.). Los mencionados perfiles son obtenidos de la siguiente forma: en un fragmento del alineamiento de tamaño preestablecido (es decir el tamaño de la ventana) se realizan las mediciones que nos interesan, a continuación se repite el mismo procedimiento en otro fragmento de igual tamaño el cual puede o no estar solapado con el fragmento que le precede. Este proceso se repite hasta llegar al final del gen. Se obtiene entonces un perfil de distancias para cada par de especies que formen parte del alineamiento. En los casos en los que el alineamiento contenga genes de más de dos especies, se usa la media aritmética de los perfiles de distancia individuales (de cada par de especies). Para las otras medidas (composición de bases) obtenemos un perfil para cada especie que forme parte del alineamiento siendo la media aritmética (entre especies) la medida que se usa en los subsiguientes análisis. En relación a los análisis usando ventanas móviles es importante tener en cuenta que el

tamaño de ventana que se elija no puede ser ni demasiado grande ni demasiado pequeña; éste aspecto es crucial para detectar algún tipo de señal. Si la ventana es demasiado pequeña nos enfrentamos al problema de que los errores de muestreo (de cada ventana) son excesivamente grandes. Esta complicación es particularmente delicada en el caso de los perfiles de distancias sinónimas cuyas estimaciones están en general asociadas a varianzas muestrales grandes. Puede ocurrir incluso que la distancia sinónima sea incalculable en algunas ventanas individuales porque el método para corregir sustituciones múltiples es inaplicable. Si por el contrario el tamaño de la ventana es muy grande reducimos considerablemente el "ruido" muestral de cada ventana pero incrementamos el error de muestreo del alineamiento. Esto se debe a que al incrementar el tamaño de la ventana tendremos menor número de ventanas para cada alineamiento. Por esta razón los genes menores a 150 codones de largo no han sido incluidos en el análisis de la variación intragénica. Además, el tamaño de ventana utilizado es hasta cierto punto proporcional al tamaño del gen. Para aquellos genes de entre 150 y 200 codones de largo el tamaño de la ventana es de 20 codones. Para genes entre 201 y 300 codones de largo el tamaño de la ventana que se usa es de 25 codones mientras que para aquellos genes con mas de 300 codones de largo el tamaño de la ventana es de 30 codones.

Una vez obtenidos los perfiles intragénicos se mide el grado de relacionamiento entre los mismos mediante el coeficiente de correlación de Pearson. Es importante resaltar que para el cálculo de las correlaciones intragénicas se usaron únicamente ventanas no solapantes con el objetivo de mantener la independencia de los puntos muestrales. Por su parte las ventanas solapantes fueron utilizadas con fines gráficos pues las mismas dan lugar a perfiles de variación más suaves.

Por último debe tenerse en cuenta que para cada grupo de datos se calculan varios coeficientes de correlación. Como resultado se espera que varios de estos coeficientes sean estadísticamente significativos sólo por azar. Por ejemplo en el caso de la base de datos de mamíferos donde fueron analizados 48 alineamientos se espera obtener 2.4 (48* 0.05) coeficientes significativos a niveles iguales o inferiores al 5%, entre éstos 1.92 [48*(0.05-0.01)] significativos sólo al 5% pero no a niveles inferiores de significación, 0.43 [48*(0.01-0.001)] coeficientes significativos al 1% pero no a niveles inferiores y 0.048 (48*0.001) coeficientes significativos al 0.1%. Por esta razón resulta necesario utilizar alguna aproximación que nos permita calcular la probabilidad de obtener por azar un conjunto particular de coeficientes de correlación estadísticamente significativos en cada base de datos. En otras palabras, debemos utilizar algún método que nos permita estimar la significación estadística conjunta en cada grupo de datos (considerando por separado cada medida que estemos analizando). Esto puede realizarse usando la distribución multinomial y teniendo en cuenta en forma discriminada la probabilidad para cada nivel de significación. La probabilidad exacta de obtener por azar nuestro conjunto particular de coeficientes y todos aquellos conjuntos más sesgados (es decir con menor probabilidad de surgir por azar) se obtiene mediante la

integración (sumatoria) de todos aquellos términos de la expansión multinomial cuyos valores de probabilidad sean menores o iguales a aquel que corresponde al de la Tabla que estamos testando. Cada término específico de esta expansión multinomial está dada por:

$$P_{ijk}=(n!/i!j!k!(n-i-j-)! *(\alpha_1-\alpha_2)^i *(\alpha_2-\alpha_3)^j *\alpha_3^k *(1-\alpha_1)^{(n-i-j-k)}$$

Esta expresión nos da la probabilidad para una Tabla de tamaño ***n*** que contenga ***i*** genes con coeficientes de correlación significativos al nivel de $\alpha_1$ (0.05>P>0.01), ***j*** genes con coeficientes significativos a nivel de $\alpha_2$ (0.01>P>0.001), ***k*** genes con coeficientes significativos a nivel de $\alpha_3$ (P<0.001) y por lo tanto ***n-i-j-k*** genes exhibiendo coeficientes de correlación no significativos (P>0.05).

## III.2 DETERMINACIÓN DE SESGOS MUTACIONALES EN MAMÍFEROS

En este estudio se estimaron la dirección y magnitud del sesgo mutacional (o los sesgos, si es que existen diferentes tipos de sesgos en distintas regiones del genoma) desde dos fuentes de información distintas e independientes:

1- de los espectros mutacionales reales .

2- del análisis del patrón de sustituciones en seudogenes

### III.2.1 Análisis de los espectros mutacionales en genes humanos

Este análisis es hoy posible gracias al enorme crecimiento en el volumen de datos sobre mutaciones que afectan a varios genes involucrados en desordenes genéticos humanos; ya que existen unas 30 bases de datos sobre mutaciones, las cuales contienen los datos de secuencia de varias mutaciones para cada gen. La mayoría (más del 90%) de las mutaciones descriptas corresponden a mutaciones puntuales, es decir el cambio de una base por otra, las cuales son precisamente las que nos interesan en este estudio.

El presente estudio se ha realizado en 9 bases de datos correspondientes a 7 genes distintos. Las mismas han sido clasificadas en 4 categorías de acuerdo al contenido en $GC_3$ del gen funcional al cual la base de datos pertenece:

1- genes pobres en $GC_3$(<0.45)

a- factor anti-hemofílico IX (factor B de coagulación o Christmas Factor) ($GC_3$=0.338)

b- factor anti-hemofílico VIII (factor A de coagulación) ($GC_3$=0.388)

2- genes con niveles medios de $GC_3$ (0.45<$GC_3$<0.60)

Fenilalanín hidroxilasa ($GC_3$=0.519)

3- genes ricos en $GC_3$ (0.6<=$GC_3$<0.75)

a- Receptor de andrógeno ($GC_3$=0.64)

b- P53 ($GC_3$=0.61)

4- genes muy ricos en GC3 (GC3>0.75).

a- Glucosa-6-phosphato dehidrogenasa ($GC_3$=0.84)

b- L1CAM, molécula de adhesión de neuronas ($GC_3$=0.77)

Para poder estimar el sesgo mutacional y la composición de bases que resultaría de ese sesgo, resulta necesario tener una medida de la probabilidad (P*ij*) de cambiar la base *i* por *j* (tomando i y j los valores 1,2,3,4, los cuales corresponden a las bases T, C, A y G). Esta probabilidad P*ij* puede ser estimada a partir de la matriz empírica O*ij*, es decir la matriz que surge del conteo directo de cada tipo de mutación. Cabe resaltar que O*ij*, no es un estimador directo de Pij, sino de Eij, es decir el número esperado de mutaciones (esperanza matemática). A partir de **E** (o de su estimador empírico **O**) es posible calcular **P**, puesto que E*ij*=P*ij**N*i**X, donde N*i* es el número de bases de tipo i en el gen en cuestión y X es el número de total de mutaciones que componen la base de datos. De esta forma calculamos las entradas que están fuera de la diagonal de **P**, mientras que los elementos de la diagonal de **P** (es decir la probabilidades que la base *i* no mute) están dados por:

$$1-\sum_{i} P_{ij}$$

De esta forma obtendremos una matriz **P** para cada gen.

Es importante resaltar varios aspectos de **P**:

-es una matriz estocástica, puesto que no presenta ninguna entrada negativa y sus filas suman 1.

- es una matriz regular, puesto que es posible pasar de cualquier estado (base) a cualquier otro, ya sea en uno o más pasos.

- por ser P estocástica y regular entonces presenta un vector de equilibrio, el cual además de representar un equilibrio estable es completamente independiente de las condiciones iniciales.

Precisamente, es este vector (o la versión normalizada del mismo) la composición de bases que uno esperaría en el equilibrio si el proceso de cambio estuviera gobernado por la matriz **P**. En otras palabras, si la composición de bases de los genes fuera el resultado de sesgos mutacionales, entonces el vector de equilibrio debería ser idéntico a la composición de bases.

Por lo antes expuesto, se desprende que utilizando los datos de espectro mutacional disponibles en las bases de datos no solamente es posible tener una estimación precisa del sesgo mutacional, sino además testar si la composición de bases de los genes corresponde con lo que esperaría de acuerdo a este sesgo mutacional. La existencia de desviaciones significativas entre la composición de bases real (la composición sinónima) y lo que se esperaría de acuerdo a al proceso mutacional, indicaría la existencia de "sesgos fijacionales", es decir selección natural.

### III.2.2 Análisis del patrón de sustituciones nucleotídicas en seudogenes de mamíferos

Puesto que los seudogenes están completamente libres de restricciones funcionales su patrón de sustituciones debe ser un reflejo del patrón mutacional subyacente. No obstante, si el patrón mutacional varía entre distintas regiones del genoma (es el punto que precisamente estamos analizando) es claro que no todos los seudogenes son útiles para estimar el patrón mutacional al cual están sujetos los genes funcionales. Este se debe a que la mayoría de los seudogenes pertenecen a la categoría de seudogenes procesados los cuales se retrotranscriben e insertan en

cualquier región del genoma usando un intermediario de ARN. En general la localización de este tipo de seudogén no está relacionada con la localización de su contraparte funcional. Es decir, el patrón mutacional que podríamos inferir de estos seudogenes procesados correspondería con el de la región cromosómica donde el seudogén se aloja pero no con la de su contraparte funcional. Por lo que el tipo de seudogenes valiosos para estimar un sesgo mutacional que pueda contrastarse con un contenido en GC sinónimo, son aquellos que se localizan en la vecindad inmediata de algún gen funcional. En general estos seudogenes surgen mediante el mecanismo de crossing-over desigual y se hallan localizados en la vecindad de su homólogo funcional.

Otra complicación con el análisis de seudogenes que limita aún más el número de seudogenes analizables ocurre en aquellos casos en los cuales la inactivación del gen ocurrió mucho tiempo después de la duplicación. En estos casos ocurre que muchas de las sustituciones que se observan en lo que hoy es un seudogén ocurrieron cuando el gen era todavía funcional y por lo tanto no estaba exento de restricciones funcionales. Si usáramos este tipo de seudogenes es claro que muchas sustituciones serían incorrectamente consideradas como libres de selección. Para detectar este tipo de seudogenes debemos computar la frecuencia de sustituciones sinónimas y no-sinónimas que ocurrieron en el linaje que conduce al seudogén. Si todas o la mayoría de las sustituciones tuvieron lugar cuando el gen ya era un seudogén (es decir después de la inactivación) entonces las frecuencias de sustituciones sinónimas y no-sinónimas deberían ser aproximadamente las mismas. Por el contrario si muchas sustituciones fueron incorporadas antes de la inactivación, es de esperar que las sustituciones sinónimas sean más frecuentes que las no-sinónimas. De hecho esperaríamos que la razón de sustituciones sinónimas sobre sustituciones no-sinónimas fuera cercana a la observada en la copia funcional del gen cuestión.

Por último, para que el seudogén pueda ser analizado es necesario que las copias funcionales del mismo hayan sido secuenciadas en varias especies de forma tal de poder discernir cuales sustituciones ocurrieron en el seudogén y cuáles en las copias funcionales.
Una búsqueda en la versión 107.0 (Junio 1998) del GenBank arroja 3350 entradas correspondientes a suedogenes de mamíferos de los cuales 2420 entradas corresponden a seudogenes humanos. Luego de eliminar los seudogenes procesados (que son la amplia mayoría) y aquellos que no cumplen con las otras dos condiciones exigidas se rescatan menos de 30 seudogenes. De este grupo hemos analizado sólo aquellos cuya copia funcional presenta alto contenido en $GC_3$.

La metodología usada para estimar el patrón de sustituciones de los seudogenes así, como el contenido en GC que uno esperaría si el mismo fuera selectivamente neutro y dependiera únicamente del sesgo mutacional, es igual al ya descripto en la sección anterior correspondiente al análisis de las bases de datos mutacionales. La única diferencia radica en cómo obtenemos la matriz de sustituciones $O_{ij}$. Obtener dicha matriz es equivalente a determinar el número de veces en que cada tipo de base fue cambiada en alguna de las restantes (en el linaje que conduce al seudogén), así como la frecuencia en la que cada tipo de base se mantuvo incambiada (elementos en la diagonal de la matriz $O_{ij}$). Para este propósito usamos como referencia la versión funcional del

gen sobre el cual necesitamos datos de secuencia en por lo menos 2 especies de mamíferos pertenecientes a órdenes distintos. Podemos asumir con un buen margen de confianza que una sustitución ocurrió en el seudogén si esa misma posición nucleotídica se mantiene incambiada en las dos copias funcionales del gen que estamos usando como referencia. Por el contrario, aquellas posiciones en las cuales las copias funcionales del gen difieren son excluidas del conteo de sustituciones pues en estos casos la sustitución ocurrió muy probablemente en una de las copias funcionales.

**III.2.3 Tratamiento del dinucleótido CpG**

Resultados provenientes de varias fuentes independientes coinciden en dar soporte a la hipótesis de que el dueto CG (una citosina seguida de una guanina) es hipermutable en los genomas de vertebrados como consecuencia de la metilación de la citosina. La citosona metilada (5mC) es inestable presentando una alta tendencia a la deaminación espontánea transformándose en timina (Bird, 1980). La deaminación de CpG da como resultado (luego de la replicación de ADN) a los dinucleótidos CpA (en una hebra) y TpG (en la otra hebra). En términos de mutaciones de bases individuales vamos a tener cambios del tipo C->T o G->A, dependiendo de en cual de las dos hebras del ADN haya ocurrido mutación. Es evidente que la hipermutabilidad de CpG contribuye verdaderamente al espectro mutacional sesgándolo en el sentido GC->AT (sesgo AT), el problema es que resulta extremadamente difícil tener una estimación de la contribución a largo plazo de este tipo de mutaciones. El inconveniente radica en que las mutaciones que involucran a CpG son muy frecuentes (en algunos casos llegan a frecuencias de hasta el 50% de todas las mutaciones) lo que provoca que las entradas de la matriz Oij correspondientes C->T y G->A van a contener valores muy altos. Pero estos valores no se corresponden con la contribución real a largo plazo de la hipermutabilidad CpG puesto que los duetos CpG son mantenidos en el gen pues participan de codones que codifican aminoácidos sujetos a restricción funcional que en general están conservados evolutivamente (Krawezak & Cooper, 1990). Si fuera posible liberar de restricciones funcionales al gen, veríamos que casi todos los duetos CpG serían rápidamente eliminados, luego de lo cual la contribución de la hipermutabilidad de CpG al espectro mutacional desaparecía o seria casi inexistente.

Debido a los problemas recién señalados, las matrices O*ij* se obtienen de dos formas alternativas: contabilizando a los dinucleótidos CpG y a las mutaciones que en ellos ocurren y excluyéndolos. La primera estas formas claramente sobrestima la ocurrencia de mutaciones G->A, C->T, la segunda la subestima. Esto último se debe a que a pesar de que no podemos obtener una estimación insesgada de la contribución a largo plazo de la hipermutabilidad de CpG, es real que las mutaciones en este dinucleótido contribuyen verdaderamente al espectro de mutaciones y por supuesto a la composición G+C en el largo plazo.

FACULTAD DE CIENCIAS

# IV RESULTADOS Y DISCUSIÓN

## IV.1 TASAS DE SUSTITUCIONES SINÓNIMAS EN MAMÍFEROS

### IV.1.1 Correlaciones intragénicas entre sustituciones sinónimas y no-sinónimas

Los lista de genes usados en este estudio así como su contenido en GC3, largo y la correlación intragénica entre las tasas de sustituciones sinónimas y no-sinónimas se presentan en la Tabla 1.1. El primer punto a remarcar de esta tabla es que la amplia mayoría de genes presentan coeficientes de correlación positivos entre los perfiles de sustituciones sinónimas y no-sinónimas. Sólo en una minoría de genes los coeficientes resultaron ser negativos pero en ninguno de éstos el coeficiente es estadísticamente significativo. Entre aquellos genes que presentan coeficientes positivos la correlación es tan alta en algunos casos que incluso es evidente mediante inspección visual de los perfiles (figura 1.1).

Como puede observarse en la Tabla 1.1, 16 genes, lo que representa aproximadamente un tercio de la muestra, presentan coeficientes de correlación estadísticamente significativos. En cuatro casos los coeficientes son significativos al nivel de 0.1%, en otros cuatro casos significativos al 1% mientras que ocho genes lo fueron al nivel de significación del 5%. La significación estadística global de esta Tabla (es decir la probabilidad de obtener por azar una tabla con este número o un número mayor de coeficientes de correlación estadísticamente significativos) es $6.97 \times 10^{-13}$. Es importante hacer notar que la probabilidad real es muchísimo menor pues todos los coeficientes de correlación significativos son positivos. En efecto, si el signo de los coeficientes de correlación se tiene en cuenta en el cálculo de la probabilidad conjunta de la Tabla, el valor que se obtiene es $2.06 \times 10^{-17}$. En conclusión, incluso si sólo un tercio de los genes analizados presentan correlaciones estadísticamente significativas, la probabilidad de que dicha proporción pueda haber surgido a azar es extremadamente baja indicando que la distribución espacial intragénica de las sustituciones sinónimas está indudablemente relacionada con la de las sustituciones no-sinónimas. Por otra parte, debe tenerse en cuenta que la ausencia de coeficientes estadísticamente significativos en muchos genes puede deberse al pequeño tamaño de los mismos. En este sentido debe resaltarse que la proporción de genes con coeficientes significativos es mayor en genes largos que en genes pequeños (Tabla 1.2).

### IV.1.2 Correlaciones intragénicas entre las tasas de sustituciones sinónimas y la composición de bases

La relación entre la variación intragénica en $GC_3$ (contenido G+C en la tercera posición de los codones excluyendo Met y Trp del cálculo) y la variación en la tasa de sustituciones fueron también investigadas mediante el uso de ventanas móviles. Para cada alineamiento se correlacionó el perfil de la tasa de sustituciones con el perfil promedio de $GC_3$ así como con el de $G_3$ y $C_3$ tomados separadamente. En relación a los resultados de estos análisis (presentados en la Tabla 1.1) hay varios puntos a destacar. En primer lugar, la mayoría de los genes ricos en $GC_3$ la correlación entre los perfiles de $GC_3$ y la tasa de sustituciones sinónimas es de signo negativo,

siendo en muchos de estos los coeficientes de correlación estadísticamente significativos. Además, la proporción de genes con correlaciones negativas baja sensiblemente en genes con menor contenido en $GC_3$ llegando incluso a predominar los coeficientes positivos en el extremo de la distribución más pobre en $GC_3$. Este cambio en la magnitud y signo de los coeficientes de correlación intragénicos a medida que nos movemos desde genes ricos en $GC_3$ hacia genes pobres puede observarse gráficamente en la Fig. 1.2a. En otras palabras, en los genes ricos en GC3, la distribución espacial de las sustituciones sinónimas está inversamente relacionada con la riqueza en $GC_3$ ya que las zonas del gen que presentan mayor conservación a nivel sinónimo son al mismo tiempo las zonas más ricas en $GC_3$. En cambio, en los genes pobres en $GC_3$ esta relación se invierte de forma tal que las zonas más conservadas a nivel sinónimo son las más pobres en $GC_3$. En la Fig. 1.3 podemos observar a modo de ejemplo los perfiles de un gen rico en $GC_3$. y los de un gen pobre en $GC_3$.

Cuando el análisis se realiza para $G_3$ y $C_3$ tomados separadamente podemos ver que el comportamiento de $C_3$ presenta la misma tendencia general que la de $GC_3$ (Tabla 1.1 y Fig. 1.2b-d) mientras que el comportamiento de $G_3$ es bastante más independiente (figura 1.2c,e). Además, el número de genes que presentan coeficientes de correlación significativos entre los perfiles de sustituciones sinónimas y $G_3$ no es mayor que lo que cabría esperar por azar (P=0.22) mientras que si lo es en el caso de las correlaciones entre los perfiles de $C_3$ y sustituciones sinónimas ($P=2.4 \times 10^{-3}$). Este resultado indicaría que la relación descripta arriba entre la variación intragénica de $GC_3$ y las sustituciones sinónimas depende principalmente de los codones terminados en C.

La relación entre el contenido en $GC_3$ y las sustituciones sinónimas que se ha descripto puede ser interpretada de dos formas diferentes no mutuamente excluyentes. Una posible interpretación es que la correlación es previsible. Esto se debería a que en los genes ricos en $GC_3$ esperamos que la mayoría de las sustituciones sinónimas sean desde C o G hacia T o A pues en estos genes C y G son las bases más abundantes en la tercera posición de los codones. Debe aclararse que este tipo de sustituciones sinónimas predominaría si las transversiones en los codones de cuádruple degeneramiento no fueran inusualmente frecuentes, en cuyo caso las transversiones C<=>G también serían muy frecuentes. Como resultado de esta posible direccionalidad composicional de las sustituciones sinónimas, es de esperar que en los genes ricos en $GC_3$ aquellos segmentos del gen que hayan sufrido mayor divergencia sinónima sean al mismo tiempo los segmentos en los cuales el contenido en $GC_3$ haya sufrido la mayor reducción. Es decir, se esperaría una correlación intragénica negativa. Por su parte en los genes pobres en $GC_3$, es de esperar que la mayoría de las sustituciones sean desde T o A hacia C o G, pues T y A son las bases más frecuentes en las posiciones sinónimas de estos genes. Esto daría como resultado que en estos genes con bajo contenido en $GC_3$, la divergencia sinónima sería acompañada por un incremento en $GC_3$, o lo que es equivalente en genes pobres en GC3 se esperaría correlación intragénica positiva entre las sustituciones sinónimas y el contenido en $GC_3$. En otras palabras, de acuerdo a esta primera interpretación, la relación entre el contenido total en $GC_3$ con la correlación intragénica

entre $GC_3$ y las sustituciones sinónimas no sería otra cosa que un resultado previsible del proceso de divergencia, puesto que es la propia divergencia quien contribuye a crear el patrón de zonas pobres y ricas en $GC_3$. La segunda interpretación posible es que en los genes ricos en $GC_3$, los segmentos más ricos en $GC_3$ tienden a evolucionar más lentamente desde el punto de vista sinónimo que los segmentos más pobres en $GC_3$, reflejando quizá algún tipo de restricción funcional que tiende a mantener la riqueza en $GC_3$ en los mencionados segmentos, mientras que lo opuesto ocurriría en los genes pobres en $GC_3$. Esta interpretación lleva implícita la existencia de selección negativa que tendería a mantener el contenido en GC de las posiciones sinónimas de los genes. En los genes ricos en $GC_3$, conservaría la riqueza en $GC_3$, mientras que en los genes pobres en $GC_3$ conservaría su riqueza en AT.

Con el objetivo de testar estas dos hipótesis se comparó el perfil de sustituciones sinónimas en un par de especies versus el perfil de contenido en $GC_3$ de una tercera especie en todos aquellos genes que en la Tabla 3 presentaron coeficientes de correlación significativos entre los perfiles de distancia sinónima y $GC_3$. Puesto que todas las especies pertenecen a órdenes de mamíferos diferentes y asumiendo que los órdenes de mamíferos divergieron en forma radial ("star phylogeny"), es claro que no podría esperarse correlación si sólo la primera hipótesis fuera correcta. Esto se debe simplemente a que de acuerdo a la primera hipótesis la correlación existiría debido a que el proceso de divergencia contribuye a la creación de zonas más o menos ricas en $GC_3$, sin embargo en este caso dicha contribución no es posible pues la divergencia entre las dos especies de las cuales se obtuvo el perfil de divergencia sinónima no afecta en ninguna forma a lo que ocurra en la tercera especie.

Los resultados presentados en la Tabla 1.3 muestran que todos aquellos genes que presentaron correlaciones significativas entre el perfil de divergencia sinónima medio con el perfil de $GC_3$ medio también presentan el mismo tipo correlación en la comparación antes descripta. Además, en la mayoría de los genes el coeficiente de correlación fue estadísticamente significativo; la única excepción está dada por el gen codificante para glucosa Glut3, cuyo coeficiente de correlación está en el límite de significación ($P=0.06$). Podemos concluir entonces que si bien no es posible descartar a los mecanismos propuestos en la primera hipótesis como posibles factores causantes de la correlación cuando consideramos los perfiles promedio, es claro que los mismos no pueden explicar los resultados que se observan en la Tabla 1.3. Esto lleva a la conclusión de que los mecanismos propuestos en la segunda hipótesis, es decir los mecanismos de selección, juegan un rol importante en la creación de la correlación intragénica entre las sustituciones sinónimas y el contenido en $GC_3$.

## Figura 1.1

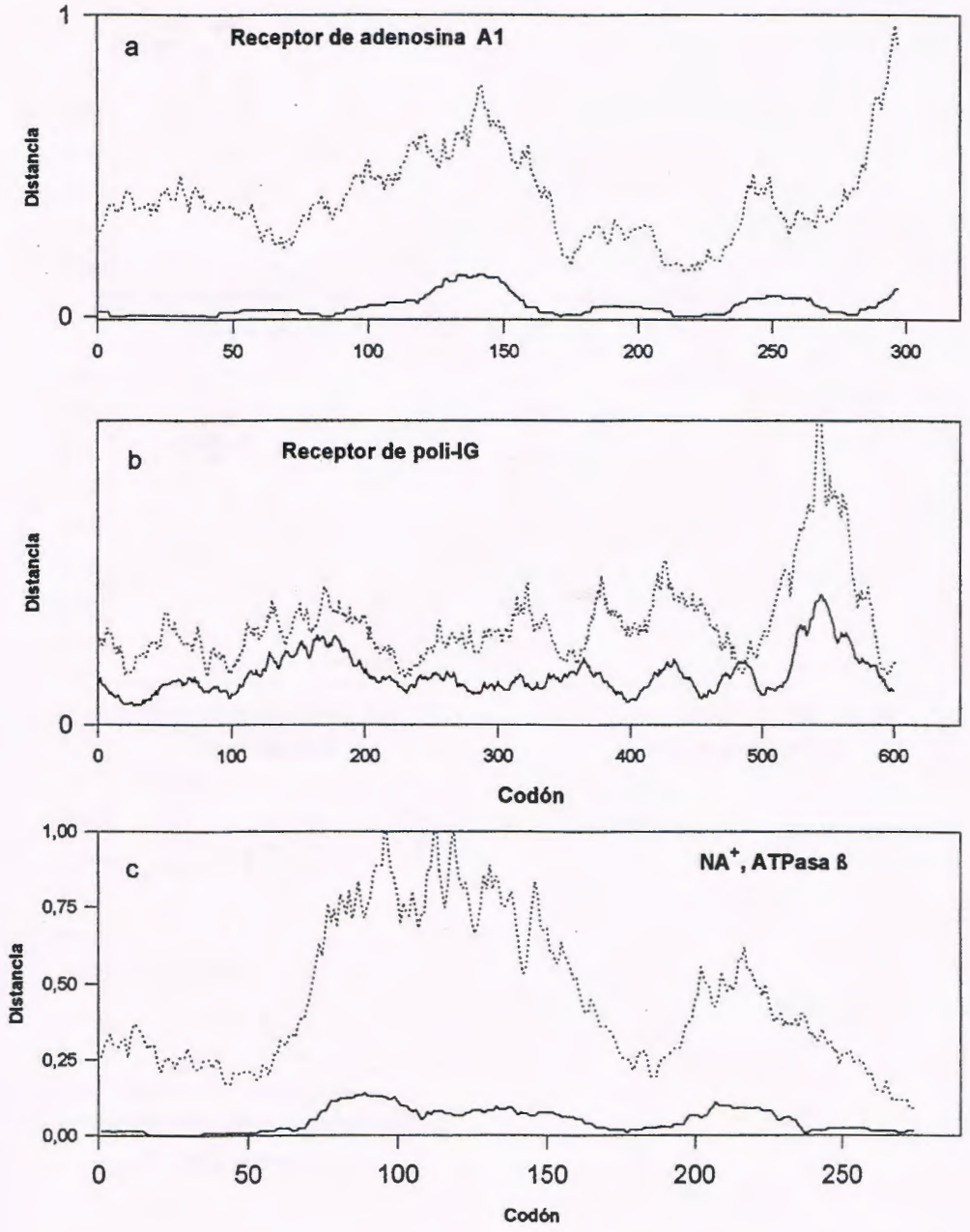



**Figura 1.1.** Perfiles de distancia sinónima (línea punteada) y distancia no sinónima (línea continua) en tres genes donde la correlación entre las variables recién mencionadas es evidente a la inspección visual. Los perfiles están construidos sobre la base de ventanas solapantes (tamaño de ventana usado es de 30 codones).

## Figura 1.2

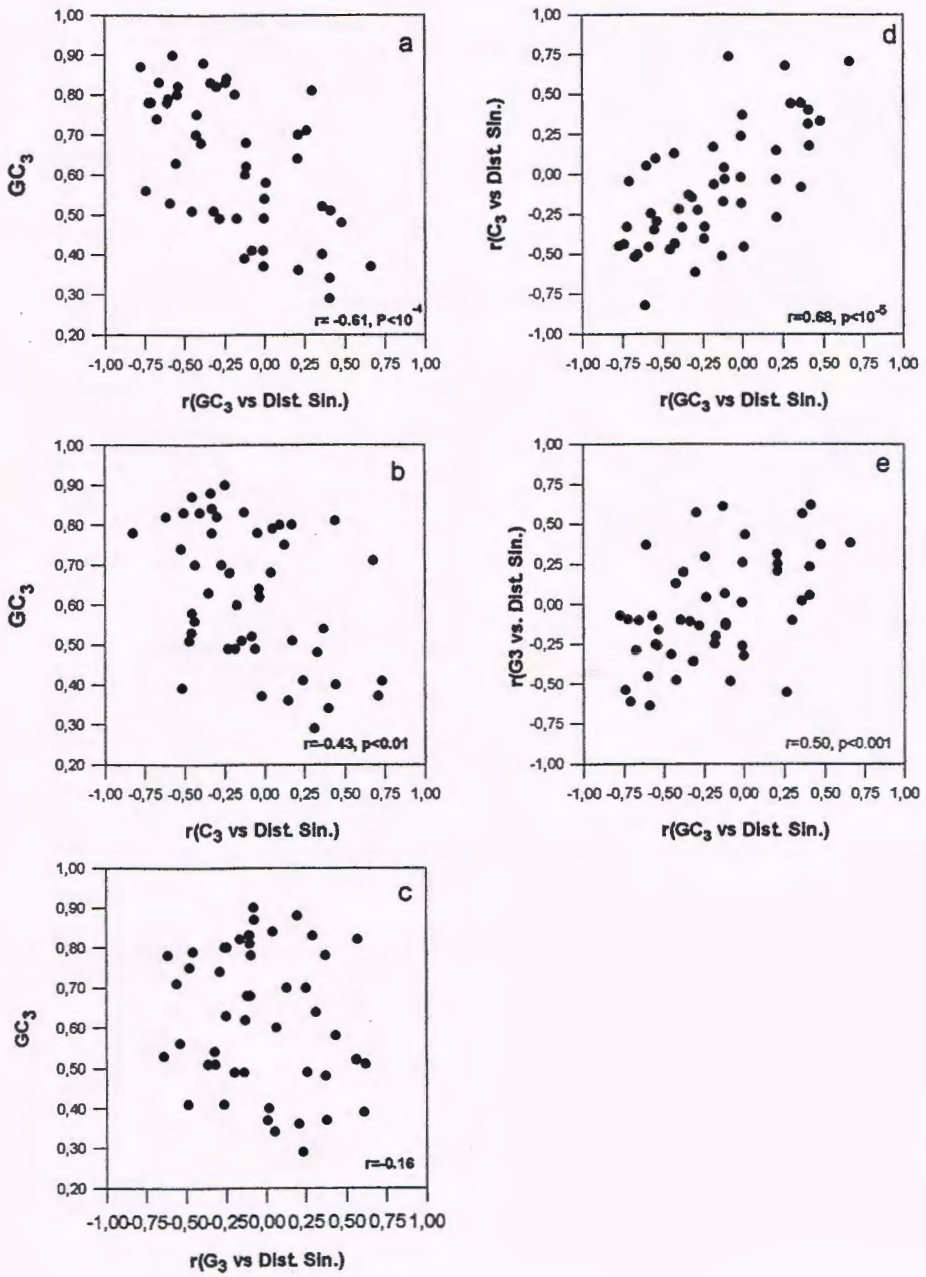



**Figura 1.2.** Gráficos donde se muestra la relación existente entre el contenido en $GC_3$ (total promediado entre los genes de todas las especies) y la correlación intragénica entre: (**a**) $GC_3$ y distancia sinónima, (**b**) $C_3$ y distancia sinónima y (**c**) $G_3$ y distancia sinónima. Las figuras **d** y **e** muestran la relación entre correlación intragénica distancia sinónima con $C_3$ y $G_3$ versus la correlación intragénica $GC_3$-distancia sinónima.

**Figura 1.3**

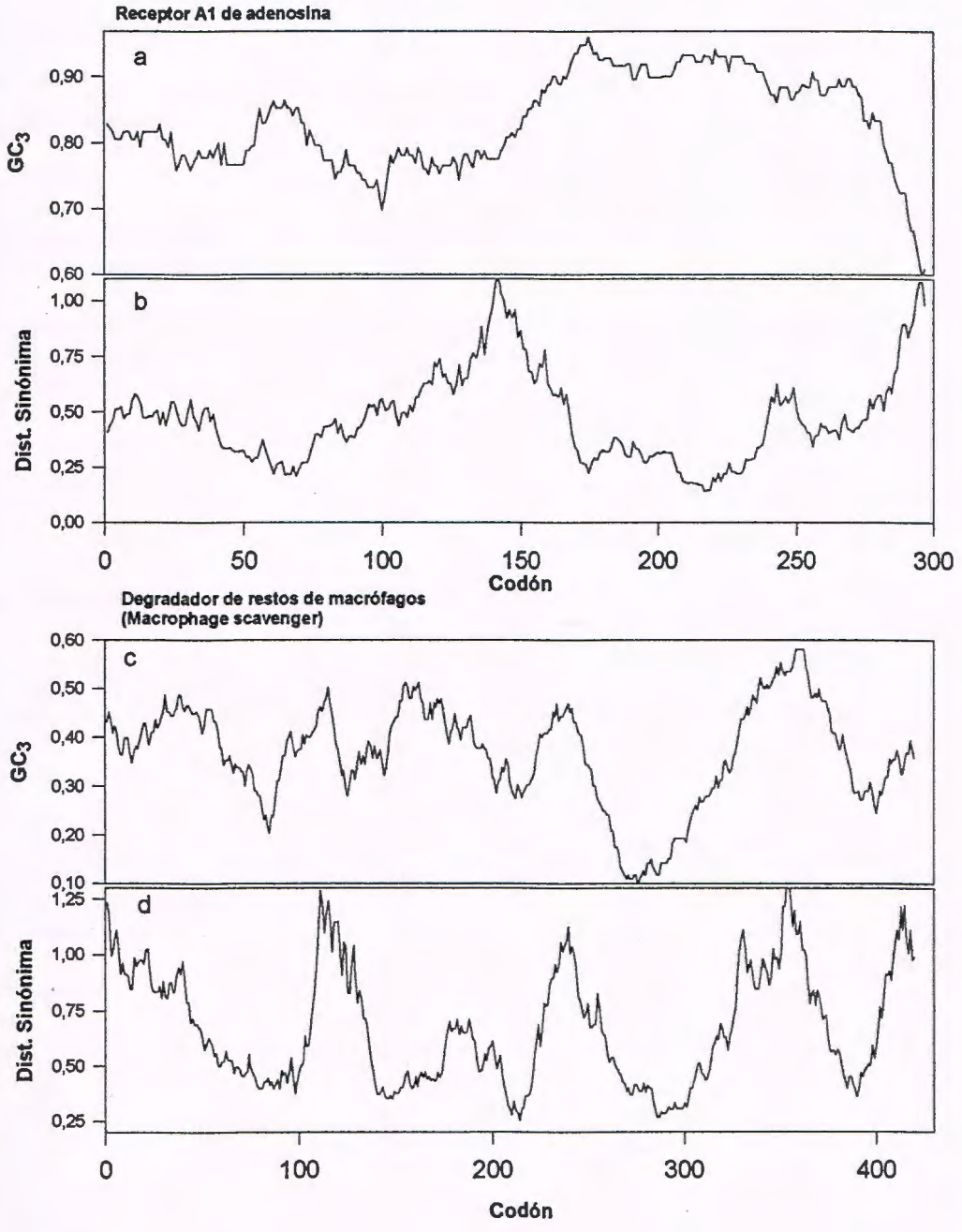



**Figura 1.3.** Perfiles de $GC_3$ (**a** y **c**) y distancia sinónima (**b** y **d**) en un gén rico en $GC_3$, el receptor de Adenosina (**a** y **b**) y un gén pobre en $GC_3$, el degradador de macrófagos de restos de macrófagos (macrophage scavenger) (**c** y **d**). Los perfiles están construidos sobre la base de ventanas solapantes (tamaño de ventana usado es de 30 codones). Los coeficientes de correlación para estos genes estan dados en la tabla 1.1.

**Tabla 1.1.** Genes de mamíferos analizados, sus largos, contenido en GC3 y correlaciones intrágenicas.

| Gén | Largo | $GC_3$ | r(DS-DNS) | r(DS-$GC_3$) | r(DS-$C_3$) | r(DS-$G_3$) |
|---|---|---|---|---|---|---|
| Creatin kinase B | 381 | 0.90 | 0.29 | -0.57 * | -0.25 | -0.07 |
| APO-E | 305 | 0.88 | 0.27 | -0.38 | -0.33 | 0.20 |
| A1-adenosine receptor | 326 | 0.87 | 0.81 ** | -0.77 ** | -0.45 | -0.07 |
| Apo A1 | 257 | 0.84 | -0.27 | -0.24 | -0.33 | 0.04 |
| Na-H exchanger | 810 | 0.83 | 0.64 *** | -0.67 *** | -0.50 ** | -0.10 |
| Serine Pyruvate aa. | 392 | 0.83 | 0.38 | -0.24 | -0.40 | 0.29 |
| CD8 | 227 | 0.83 | 0.04 | -0.34 | -0.13 | -0.11 |
| GMP-phospho | 852 | 0.82 | 0.02 | -0.54 ** | -0.30 | -0.16 |
| Dipetidase | 411 | 0.82 | 0.39 | -0.30 | -0.61 * | 0.57 |
| Prostaglandin E receptor | 335 | 0.81 | 0.15 | 0.30 | 0.44 | -0.10 |
| Glutatione peroxidase | 197 | 0.80 | -0.33 | -0.54 | 0.10 | -0.26 |
| Retinol | 199 | 0.80 | 0.00 | -0.18 | 0.17 | -0.25 |
| Creatin kinase M | 382 | 0.79 | 0.65 ** | -0.60 ** | 0.05 | -0.46* |
| H, K ATPase beta | 290 | 0.78 | -0.17 | -0.72 | -0.33 | -0.10 |
| TNF-alpha | 231 | 0.78 | 0.69 | -0.61 | -0.82 * | 0.37 |
| Glut III | 492 | 0.78 | -0.01 | -0.71 ** | -0.05 | -0.61* |
| Growth hormone | 214 | 0.75 | 0.71 * | -0.42 | 0.13 | -0.48 |
| Prolyll-4 Hydrolase | 502 | 0.74 | -0.35 | -0.67 ** | -0.52 * | -0.29 |
| TNF- beta | 193 | 0.71 | -0.33 | 0.27 | 0.68 | -0.55 |
| Polymeric Ig Receptor | 740 | 0.70 | 0.66 *** | -0.42 * | -0.43 * | 0.13 |
| Erythropoeitin | 185 | 0.70 | -0.30 | 0.21 | -0.27 | 0.25 |
| Na glucosa transporter | 602 | 0.68 | 0.68 *** | -0.40 | -0.22 | -0.10 |
| CD4 | 444 | 0.68 | 0.59 * | -0.11 | 0.04 | -0.12 |
| D-amino oxidase | 346 | 0.64 | 0.17 | 0.21 | -0.04 | 0.31 |
| TRNA ligase | 470 | 0.63 | 0.35 | -0.55 * | -0.35 | -0.25 |
| Endothelin | 201 | 0.62 | 0.79 * | -0.11 | -0.03 | -0.13 |
| Inter1B | 259 | 0.60 | 0.68 * | -0.12 | -0.17 | 0.07 |
| Inter6 | 203 | 0.58 | 0.46 | 0.00 | -0.45 | 0.43 |
| IntR2 | 265 | 0.56 | 0.78 * | -0.74 * | -0.44 | -0.54 |
| Phagocytic glyc. | 353 | 0.54 | 0.26 | -0.00 | 0.37 | -0.32 |
| Urate oxidase | 299 | 0.53 | 0.36 | -0.59 | -0.46 | -0.64* |
| Prolacting Receptor | 554 | 0.52 | 0.64 ** | 0.36 | -0.08 | 0.56* |
| Na, K ATPase | 303 | 0.51 | 0.83 ** | -0.32 | -0.15 | -0.36 |
| CD3 | 187 | 0.51 | 0.78 * | -0.45 | -0.47 | -0.32 |
| Tissue factor | 282 | 0.51 | 0.69 * | 0.41 | 0.18 | 0.62 |
| Selecting | 412 | 0.49 | 0.12 | -0.01 | -0.18 | 0.26 |
| Inter1A | 263 | 0.49 | 0.40 | -0.18 | -0.06 | -0.20 |
| Flavin cont. | 532 | 0.49 | -0.11 | -0.28 | -0.23 | -0.14 |
| LINK protein | 354 | 0.48 | -0.48 | 0.48 | 0.33 | 0.37 |
| Osteopontin | 253 | 0.41 | 0.22 | -0.08 | 0.73 * | -0.49 |
| Pancreatic | 460 | 0.41 | -0.01 | -0.01 | 0.24 | -0.26 |
| Serum albumin | 607 | 0.40 | 0.68 *** | 0.36 | 0.44 | 0.02 |
| Stem cell factor | 273 | 0.39 | 0.47 | -0.13 | -0.51 | 0.61 |
| HSP108 | 802 | 0.37 | 0.47 * | -0.01 | -0.02 | 0.01 |
| Macrophage Scavenger | 449 | 0.37 | -0.43 | 0.66 * | 0.70 * | 0.38 |
| APO-H | 345 | 0.36 | -0.17 | 0.21 | 0.15 | 0.21 |
| Calpastain | 593 | 0.34 | -0.10 | 0.41 | 0.40 | 0.06 |
| Ca ATP-M | 1176 | 0.29 | 0.33 * | 0.41 ** | 0.31 * | 0.23 |
| **Probabilidad** | | | **$2.06 \times 10^{-17}$** | **$2.6 \times 10^{-8}$** | **0.0024** | **0.21** |

**Nota.** (de la tabla 1.1) El largo de los genes está dado en codones.Los valores de $GC_3$ son promediados entre los genes homólogos. **r(DS-DNS)**, **r(DS-GC$_3$)**,**r(DS-C$_3$)**, **r(DS-G$_3$)** son los coeficientes de correlación intragénica entre las variables que se encuentran dentro de los paréntesis, siendo DS distancia sinónima y DNS distancia no sinónima. Los asteriscos a la derecha de las columnas indican la significancia estadística de los coeficientes de correlación. donde *= P<0.05, **=P<0.01, ***=P<0.001 y ****=P<$1x10^{-4}$ . Los valores de probabilidad en la parte inferior de cada columna expresan la significancia estadística conjunta (ver Sección III.1.3.3) de obtener esa combinación de coeficientes de correlación.

**Tabla 1.2.** Subgrupos de genes de tamaño progresivamente creciente y la proporción de genes que exhiben coeficientes de correlación estadísticamente significativos

| Límite inferior de tamaño de gén en el subgrupo | Número de genes | Número de genes con coef. significativos | Percentaje |
|---|---|---|---|
| 180 codones | 48 | 16 | 33% |
| 300 codones | 30 | 11 | 37% |
| 400 codones | 18 | 8 | 44% |
| 500 codones | 11 | 7 | 64% |

**Tabla 1.3.** Correlaciones entre el perfil de divergencia sinónima en un par de especies versus el perfil de **GC$_3$** de una tercera especie.

| Gene | Perfil de divergencia sinónima entre: | perfil de **CG$_3$** obtenido de: | coef. de correlación |
|---|---|---|---|
| Creatin kinase B | Humano-Conejo | Perro | -0.68 * |
| A1-adenosine receptor | Humano-Rata | Bovino | -0.62 * |
| Na-H exchanger | Cerdo-Conejo | Humano | -0.47 * |
| GMP-phospho | Humano-Perro | Bovino | -0.57 ** |
| Creatin kinase M | Humano-Conejo | Perro | -0.73 *** |
| H, K ATPase beta | Humano-Cerdo | Perro | -0.68 * |
| Glut III | Rata-Bovino | Humano | -0.47 |
| Prolyll-4 Hydrolase | Humano-Rata | Bovino | -0.56 * |
| Polymeric Ig Receptor | Bovino-Conejo | Rata | -0.56 ** |
| TRNA ligase | Bovino-Conejo | Humano | -0.78 *** |
| IntR2 | Humano-cat | Ratón | -0.76 * |
| Scavenger | Humano-Bovino | Conejo | 0.62 * |
| Ca ATP-M | Humano-Ovino | Conejo | 0.31 * |

**Nota-.** Los asteriscos a la derecha de la última columna expresan la signifcancia estadística de los coeficientes de correlación siendo la simbología igual a la usada en la tabla 1.1

## IV.2 TASAS DE CAMBIO NUCLEOTÍDICO EN GRAMÍNEAS

En esta sección analizaremos en genes de gramíneas las relaciones entre las tasas de sustituciones sinónimas y no-sinónimas así como las de éstas con la composición de bases de los genes. Los primeros resultados en ser presentados corresponden a los valores globales de los genes sin considerar la variación intragénica. Cabe resaltar que esta parte del estudio no fue llevada a cabo en el caso de los mamíferos pues existen varias publicaciones previas donde dichos resultados son descriptos exhaustivamente (ver Introducción, secciones I.2.1 y I.2.2).

### IV.2.1 Relación entre la velocidad de evolución nucleotídica y la composición de bases

Las Tablas 2.1a,2.1b y 2.1c contienen la información de cada uno de los grupos de genes homólogos gramíneas analizados. Cada Tabla contiene además de la lista de genes usados, información concerniente al contenido en G+C tanto en la segunda como tercera posición del codón así como también las distancias sinónimas y no-sinónimas entre cada par de genes homólogos. Es de resaltar la existencia de varios comportamientos que son comunes a las tres bases de datos. En primer lugar el contenido en G+C de la segunda posición de los codones ($GC_2$) está correlacionado positiva y significativamente con la distancia no-sinónima. Los coeficientes de correlación así como los niveles de significación se muestran en las Fig. 2.1a, 2.2a y 2.3a para cada uno de los grupos de datos analizados. Es interesante mencionar que cuando las bases se consideran individualmente, $T_2$ siempre exhibe correlaciones fuertemente negativas, pero $C_2$ y $G_2$ varían entre un grupo de datos y otro (Tabla 2.2). Para la base de datos que incluye a maíz/trigo-cebada, tanto $C_2$ como $G_2$ muestran coeficientes de correlación significativos con la distancia no-sinónima, mientras que solamente $C_2$ presenta correlación estadísticamente significativa en el grupo de genes que incluye arroz/trigo-cebada y $G_2$ en el grupo de genes homólogos entre maíz/arroz (ver Tabla 2.2).
Otros dos comportamientos que son comunes a los tres grupos de genes involucran a $GC_3$, el cual se encuentra correlacionado negativamente con la distancia sinónima (Figs. 2.1b, 2.2b y 2.3b) mientras que presenta correlación positiva con la distancia no-sinónima (Figs. 2.1c, 2.2c y 2.3c). Sin embargo como veremos más adelante, la correlación positiva entre $GC_3$ y la distancia no-sinónima no es el resultado de una relación continua entre estas dos variables sino que la misma está dada por la existencia de dos subpoblaciones de genes en el genoma de las gramíneas.

En lo que se refiere a la correlación entre las distancias sinónimas y no-sinónimas, los resultados obtenidos en esta investigación muestran que contrariamente a lo que ya ha sido descripto en mamíferos (ver Introducción, Sección I.2.2), las gramíneas no presentan correlación en ninguno de los tres grupos de datos estudiados. Sin embargo esta ausencia de correlación lineal no necesariamente implica ausencia de otro tipo de relación. Como puede apreciarse al graficar las distancias sinónimas versus las distancias no-sinónimas (Figs. 2.1d, 2.2d y 2.3d), existen dos subpoblaciones de genes en los genomas de estas plantas, la población muy rica en $GC_3$ ($GC_3$>=80%) y la población rica en $GC_3$ ($GC_3$<80%). La primera de estas subpoblaciones se caracteriza, además de por su alto contenido en $GC_3$, por poseer genes que presentan bajas tasas

de cambio sinónimo y altas tasas de sustituciones no-sinónimas. Por su parte la subpoblación caracterizada por su menor contenido en $GC_3$, contiene genes con altas tasas de cambio sinónimo y bajas tasas de cambio no-sinónimo. La diferenciación entre ambas supoblaciones es bastante manifiesta en los tres grupo de genes homólogos analizados.

**IV.2.2 Dos poblaciones de genes en gramíneas**

Con el fin de estudiar el problema de las subpoblaciones de genes en mayor profundidad, se repitieron los mismos análisis para cada una de las subpoblaciones de genes tomadas por separado. Mientras que algunas de las correlaciones presentaron el mismo comportamiento que las poblaciones completas, otras mostraron una clara diferenciación (Tabla 2.3). Las correlaciones entre $GC_2$ y la distancia no-sinónima son positivas y estadísticamente significativas en la subpoblaciones con mayor riqueza en $GC_3$, pero los coeficientes de correlación varían de un grupo de genes a otro en la subpoblaciones de genes con menor riqueza en $GC_3$. En lo que se refiere a la relación entre las distancias sinónimas y $GC_3$, los resultados presentados en la Tabla 2.3 muestran que ambas subpoblaciones exhiben coeficientes de correlación de signo negativo, pero sólo la subpoblación rica en $GC_3$ presenta valores altos y significativos en los tres grupos de genes. Contrastando con lo que sucede cuando los grupos de genes no se dividen en subpoblaciones, los coeficientes de correlación entre las distancias no-sinónimas y el contenido en $GC_3$ varían desde valores negativos y relativamente altos, hasta ausencia de correlación en las subpoblaciones ricas en $GC_3$. Por su parte, en las subpoblaciones con valores de $GC_3$ por debajo de 80%, los coeficientes de correlación también varían desde valores negativos a positivos. En este caso, resulta interesante contrastar para cada grupo de genes el comportamiento entre ambas subpoblaciones con el comportamiento de las poblaciones completas. En el primer grupo de genes (maíz/arroz), existe una correlación francamente positiva (y estadísticamente significativa) entre $GC_3$ y las distancias no-sinónimas (Tabla 2.2), mientras que cuando se separa en subpoblaciones, los coeficientes de correlación son cercanos a cero (ausencia de correlación) en ambas subpoblaciones. En el segundo grupo de genes (maíz/trigo-cebada) el coeficiente de correlación también es de signo positivo aunque no es estadísticamente significativo. Al separar este grupo de genes en dos subpoblaciones vemos que los coeficientes de correlación son negativos en ambas subpoblaciones. En el último grupo de genes, que incluye genes homólogos de arroz/trigo-cebada, vemos que para la población completa el coeficiente de correlación entre $GC_3$ y distancia sinónima también es positivo (y estadísticamente significativo), pero es negativo (y estadísticamente significativo) en la subpoblación de mayor riqueza en $GC_3$ y positivo (y en el límite de la significación estadística) en la subpoblación que contiene genes con menor riqueza en $GC_3$. Estos resultados muestran que la correlación positiva entre $GC_3$ y distancia no-sinónima descripta en la Tabla 2.2, no es el resultado de una relación continua a lo largo de toda la distribución, sino de la existencia de dos subpoblaciones de genes que difieren en sus valores promedio de distancia no-sinónima y por supuesto en su contenido en $GC_3$ (ya que esta es la variable usada para separar las subpoblaciones).

Otro punto de relevancia en relación a la existencia de dos subpoblaciones de genes lo constituye la correlación entre las distancias sinónimas y no-sinónimas. Como se recordará, a diferencia de lo que había sido descripto en otros grupos taxonómicos, ninguno de los grupos de genes presentó correlación (Tabla 2.2, Figs 2.1d, 2.2d y 2.3d). Sin embargo, la subpoblación que contiene genes con mayor riqueza en $GC_3$ exhibe correlaciones positivas en los tres grupo de genes, siendo los coeficientes de correlación estadísticamente significativos en los grupo de genes homólogos entre maiz-trigo/cebada y arroz-trigo/cebada.

Para finalizar con esta sección, es importante señalar que Carels et al (comunicación personal) han obtenido resultados que dan sustento adicional a la existencia de dos poblaciones de genes en el genoma de las gramíneas. Estos autores han encontrado que, además de la obvia diferencia en contenido en $GC_3$, los genes de gramíneas presentan diferencias en el número y tamaño de los intrones así como en la variablidad interna en el contenido de bases.

**IV.2.3 Correlaciones intragénicas**

Los resultados de este análisis, los cuales se presentan en las Tablas 2.4a, 2.4b y 2.4c, muestran que el comportamiento que se observa a nivel intragénico es compatible con el recién descripto para los valores promedio de los genes. En primer lugar, la correlación intragénica entre $GC_2$ y la distancia no-sinónima es positiva en la mayoría de los genes, siendo estadísticamente significativa en aproximadamente un tercio de estos. Por su parte, los coeficientes de correlación intragénica entre $GC_3$ y distancia sinónima son negativos en la mayoría de los genes, aunque estos varían desde valores negativos altos (en los genes muy ricos en $GC_3$) a valores cercanos a cero o suavemente positivos en los genes de menor contenido en $GC_3$ . De la misma forma, las correlaciones intragénicas entre $GC_3$ y la distancia no-sinónima, varían desde valores fuertemente negativos en los genes de gran riqueza en $GC_3$ , a valores positivos en los genes de menor riqueza en $GC_3$ . La proporción de genes que exhibe correlaciones intragénicas estadísticamente significativas es más alta de lo que sería esperado por azar en los tres grupo de genes y para los tres tipos de correlaciones (ver la parte inferior de cada columna en las Tablas 2.4a,2.4b y 2.4c)

Resulta de interés resaltar que los coeficientes de correlación intragénicos entre $GC_3$ y distancias no-sinónimas son dependientes del contendido en $GC_3$ global (medio) de cada uno de los genes, de forma tal que los genes muy ricos en $GC_3$ presentan coeficientes negativos mientras que los genes con valores menores en su contenido en $GC_3$ presentan coeficientes de signo positivo. De hecho, el contenido global en $GC_3$ de los genes se correlaciona negativamente con el coeficiente de correlación intragénico entre $GC_3$ y distancia no-sinónima (Figs 2.4a, 2.4d y 2.4g). Algo similar ocurre con el comportamiento de los coeficientes de correlación intragénicos entre la divergencia sinónima y el contenido en $GC_3$ . Sin embargo en este caso los coeficientes de correlación van desde valores negativos altos, en los genes muy ricos en $GC_3$ , hasta valores cercanos a cero o levemente positivos en los genes con menor contenido en $GC_3$ (Figs 2.4b, 2.4e y 2.4h). Este resultado significa en los genes muy ricos en $GC_3$ los fragmentos génicos que presentan mayor

riqueza en $GC_3$ son más conservados, tanto a nivel aminoacídico como a nivel sinónimo; en tanto los genes con menor riqueza en $GC_3$ , los fragmentos génicos de mayor contenido en $GC_3$ tienden a evolucionar más rápidamente. El aspecto importante de este resultado es que los patrones espaciales intragénicos de divergencia sinónima y no-sinónima se relacionan de una forma similar con la variación intragénica de $GC_3$ . Este paralelismo entre la divergencia sinónima y aminoacídica queda claramente en evidencia en las Figs 2.4c, 2.4f y 2.4i, las cuales muestran que en los tres grupos de genes homólogos de gramíneas, los coeficientes de correlación intragénicos entre distancia no-sinónima y variación interna de $GC_3$ están fuertemente correlacionados con los coeficientes intragénicos entre la distancia no-sinónima y variación de $GC_3$.

En lo que se refiere a las correlaciones intragénicas entre sustituciones sinónimas y no-sinónimas, varían desde valores positivos a negativos, pero sólo los coeficientes de signo positivo llegan a valores altos y estadísticamente significativos. Resulta evidente además que incluso si la proporción de genes que exhiben correlaciones significativas es mayor a lo que se esperaría por azar, estos representan sólo una modesta proporción de la muestra. De hecho sólo el grupo de genes que contiene genes homólogos entre arroz/trigo-cebada presenta una proporción de genes con correlaciones estadíscamente significativas comparable al observado en genes de mamíferos. Esta escasez de genes con coeficientes de correlación significativos es bastante llamativa pues como se describió más arriba, los procesos de divergencia sinónima y no-sinónima están seguramente relacionados en los genes de gramíneas dado que presentan un comportamiento paralelo en relación a la composición de bases sinónimas. Es probable que la ausencia de correlaciones significativas en muchos genes se deba a su pequeño número de codones. Esta afirmación tiene su apoyo en el hecho que muchos genes pequeños presentaron coeficientes de correlación en el límite de significación.

## Figura 2.1

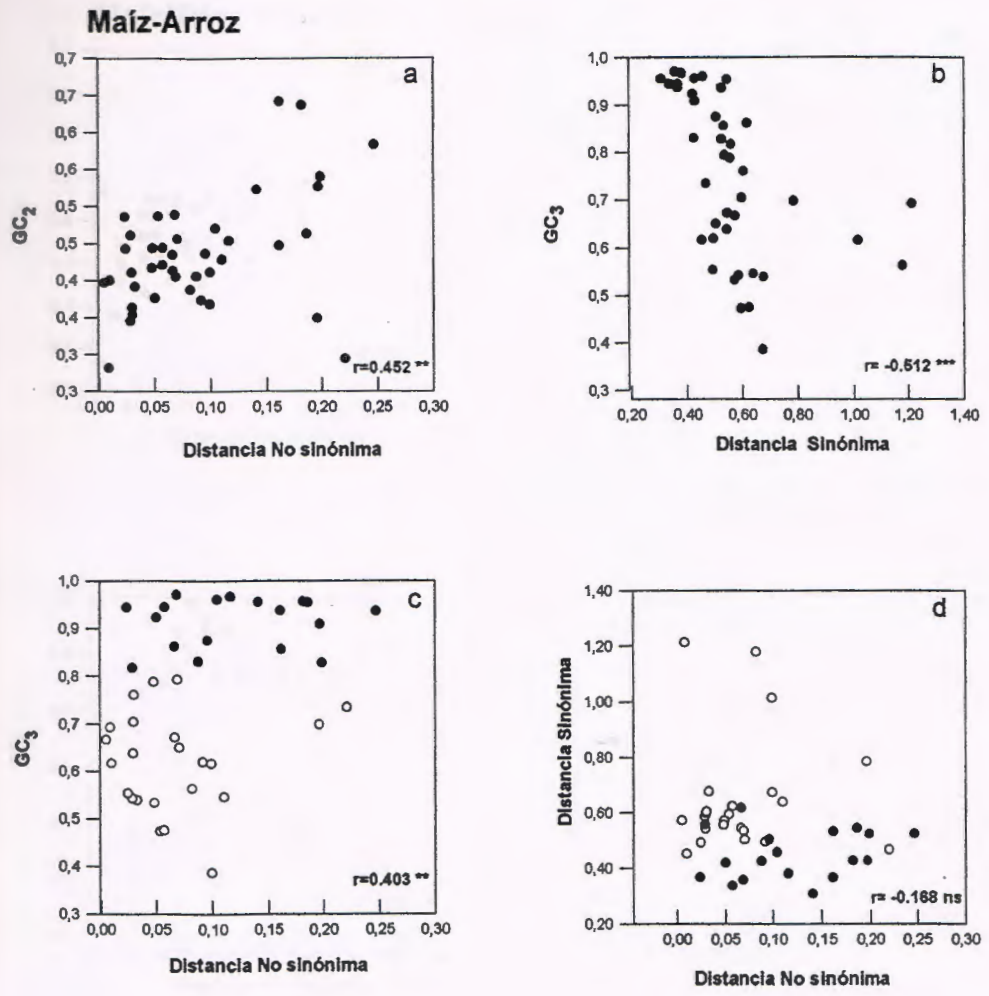



**Figura 2.1** (**a** y **c** ) Distancias no sinónimas entre genes homólogos de Maíz/Arroz graficados versus el contenido $GC_2$ y $GC_3$. ( b) Distancias sinónimas graficadas versus el contenido $GC_3$. (**d**) Distancias no sinónimas versus distancias sinónimas. Los coeficientes de correlacción y su significancia estadística se indican en cada caso. Los círculos negros representan genes con contenidos de $GC_3$ iguales o superiores al 80%, los círculos vacíos representan genes con valores de GC3 menores al 80%.

## Figura 2.2

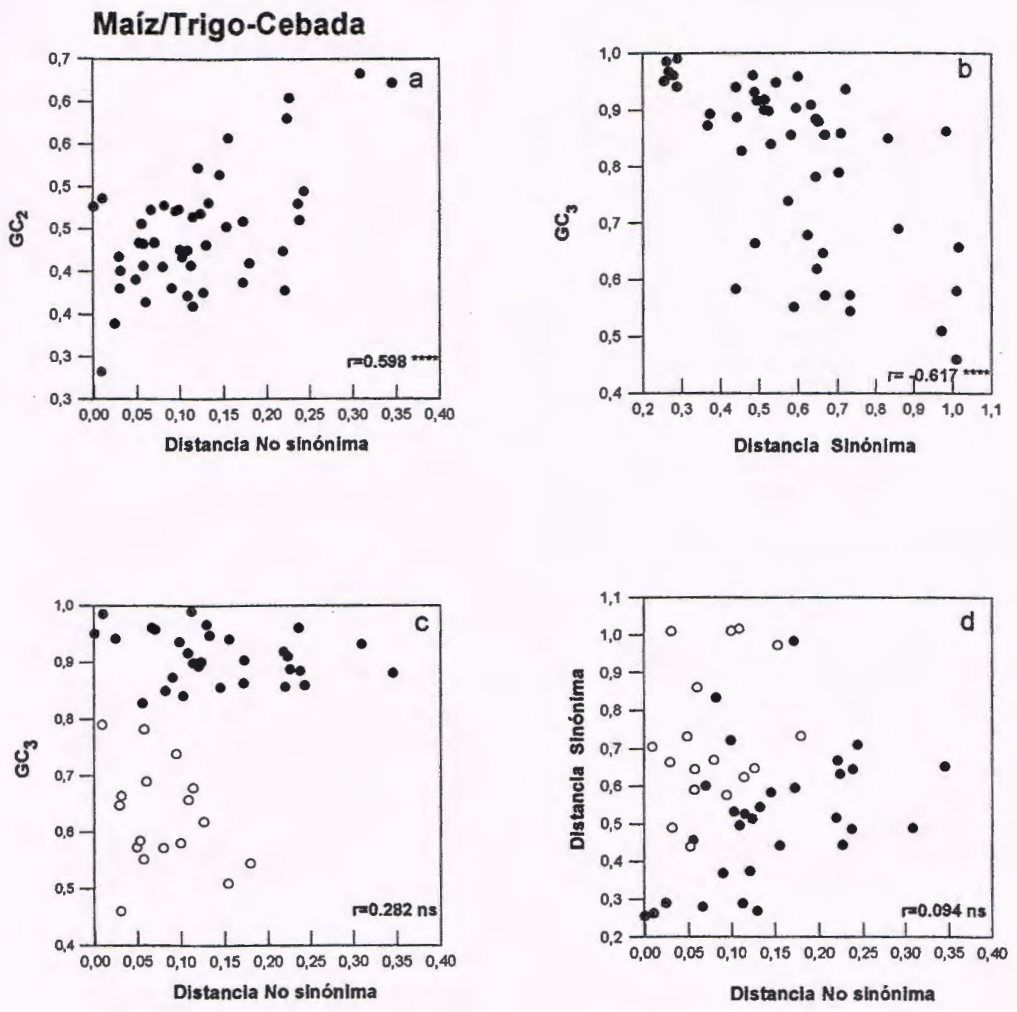



**Figura 2.2** (**a** y **c** ) Distancias no sinónimas entre genes homólogos de Maíz/Trigo-cebada graficados versus el contenido $GC_2$ y $GC_3$. ( **b**) Distancias sinónimas graficadas versus el contenido $GC_3$. (**d**) Distancias no sinónimas versus distancias sinónimas. Los coeficientes de correlacción y su significancia estadística se indican en cada caso. Los círculos negros representan genes con contenidos de $GC_3$ iguales o superiores al 80%, los círculos vacíos representan genes con valores de $GC_3$ menores al 80%.

**Figura 2.3**

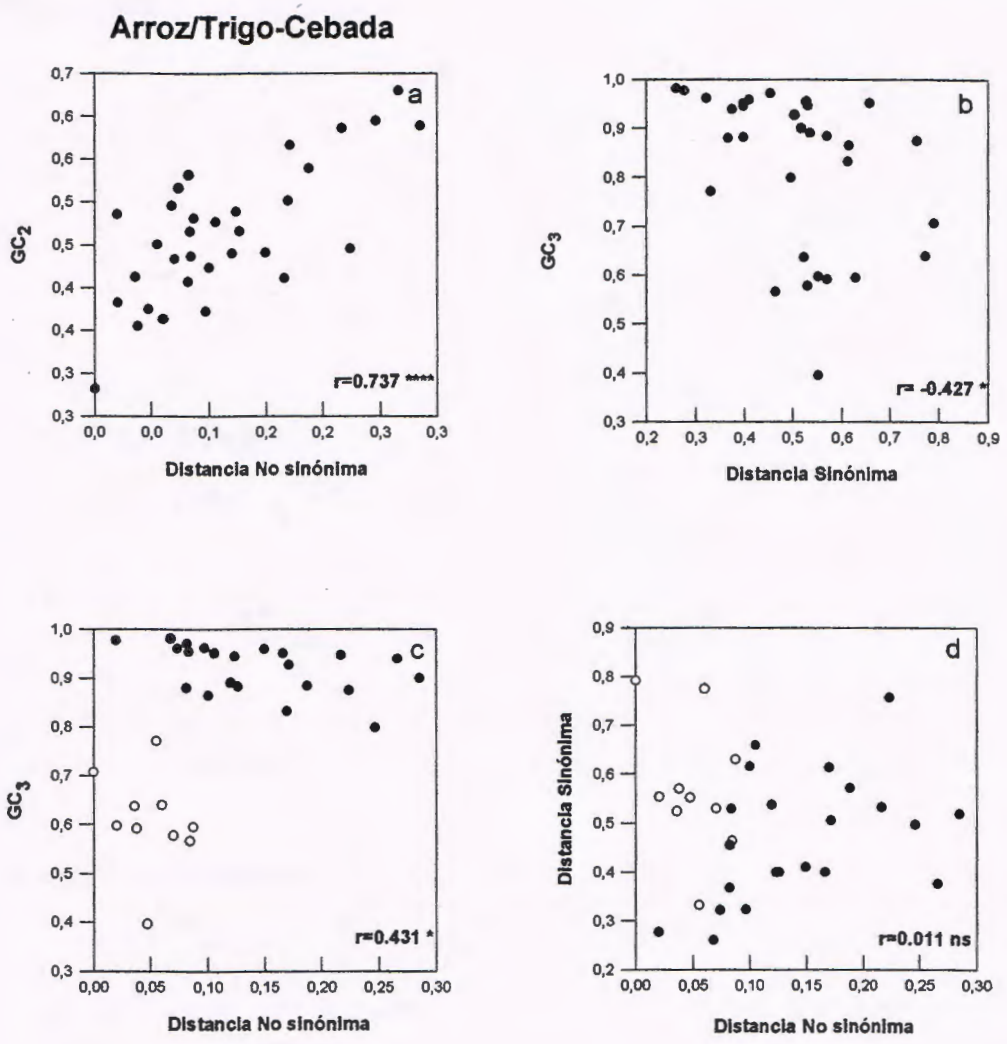
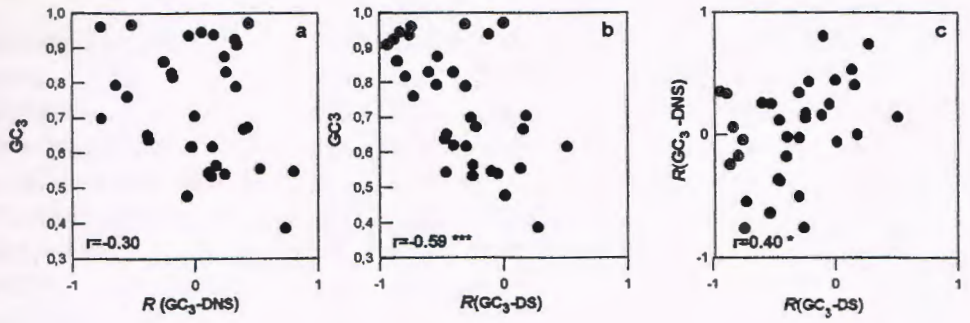


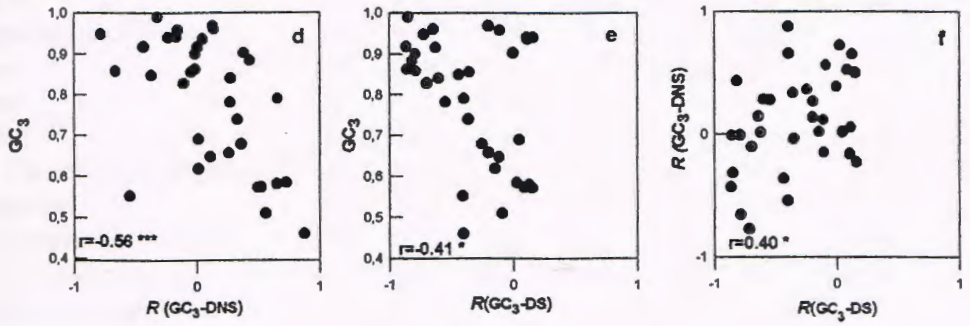

**Figura 2.3** (**a** y **c** ) Distancias no sinónimas entre genes homólogos de Arroz/Trigo-cebada graficados versus el contenido $GC_2$ y $GC_3$. ( **b**) Distancias sinónimas graficadas versus el contenido $GC_3$. (**d**) Distancias no sinónimas versus distancias sinónimas. Los coeficientes de correlacción y su significancia estadística se indcan en cada caso. Los círculos negros representan genes con contenidos de $GC_3$ iguales o superiores al 80%, los círculos vacíos representan genes con valores de $GC_3$ menores al 80%.

## Figura 2.4

**Maíz-Arroz**




**Maíz/Trigo-Cebada**




**Arroz/Trigo-Cebada**

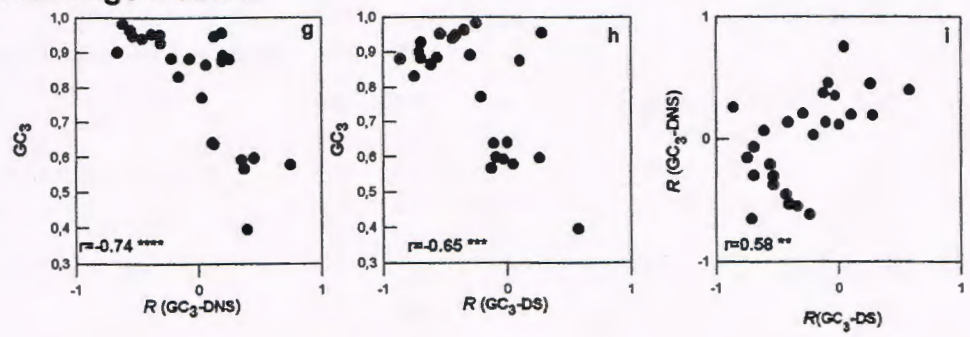



**Figura. 3.4.** (**a, b, d, e , g y h** ) Relación entre los coeficientes de correlación intragénicos y el contenido $GC_3$ (global) de los genes.(**c, f, e i)** Gráfico entre los coeficientes de correlación intragénicos de distancia sinónima y $GC_3$ versus los coeficientes de correlación intragénicos entre distancia no sinónima y $GC_3$. R($GC_3$-DNS) y R($GC_3$-DS) tienen el mismo significado que en la tabla 1.1. Los asteriscos indican la significación estadística de los coeficientes de correlación.

**Tabla 2.1a**. Genes homólogos entre Maíz y arroz

| | Gén codificante de: | Largo | $GC_3$ | $GC_2$ | Dist Sin. | Dist. No Sin. | Correlaciones intragénicas r($GC_2$-DNS) | | r($GC_3$-DNS) | | r($GC_3$-DS) | | r(DNS-DS) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Chlorophyll a/b binding protein | 265 | 0.97 | 0.49 | 0.36 | 0.07 | 0.62 | * | 0.45 | | 0.00 | | 0.54 | |
| 2 | Cyclophilin (CyP). | 172 | 0.97 | 0.45 | 0.38 | 0.12 | -0.06 | | -0.51 | | -0.31 | | -0.13 | |
| 3 | Lactate dehydrogenase. | 352 | 0.96 | 0.47 | 0.46 | 0.10 | 0.47 | | -0.76 | ** | -0.74 | ** | 0.62 | * |
| 4 | Phospholipid transfer protein | 117 | 0.96 | 0.64 | 0.43 | 0.18 | ---- | | ---- | | ---- | | ---- | |
| 5 | Embryogenic abscisic acid-inducible gen. | 91 | 0.96 | 0.52 | 0.31 | 0.14 | ---- | | ---- | | ---- | | ---- | |
| 6 | Ferredoxin (Fd) isoprotein pFD5. | 134 | 0.95 | 0.46 | 0.54 | 0.19 | ---- | | ---- | | ---- | | ---- | |
| 7 | Pyruvate decarboxylase. | 602 | 0.94 | 0.44 | 0.34 | 0.06 | 0.43 | | 0.06 | | -0.83 | **** | -0.16 | |
| 8 | Histone H3 (H3C4), . | 136 | 0.94 | 0.49 | 0.37 | 0.02 | ---- | | ---- | | ---- | | ---- | |
| 9 | Rubisco, small subunit | 169 | 0.94 | 0.45 | 0.37 | 0.16 | -0.12 | | 0.16 | | -0.12 | | -0.43 | |
| 10 | Acidic class I chitinase. | 316 | 0.94 | 0.58 | 0.52 | 0.25 | -0.02 | | -0.04 | | -0.75 | * | 0.26 | |
| 11 | Heat shock protein 17.2. | 150 | 0.92 | 0.38 | 0.42 | 0.05 | 0.72 | | 0.33 | | -0.88 | * | -0.33 | |
| 12 | Dehydrin (dhn3). | 155 | 0.91 | 0.53 | 0.43 | 0.20 | 0.62 | | 0.35 | | -0.94 | *** | -0.26 | |
| 13 | Amyloplast-specific transit protein | 603 | 0.87 | 0.44 | 0.50 | 0.10 | 0.59 | ** | 0.25 | | -0.53 | * | -0.06 | |
| 14 | Alpha-amylase. | 437 | 0.86 | 0.43 | 0.62 | 0.07 | -0.28 | | -0.25 | | -0.86 | *** | 0.25 | |
| 15 | Metallothionein. | 74 | 0.86 | 0.64 | 0.53 | 0.16 | ---- | | ---- | | ---- | | ---- | |
| 16 | Knotted-1 (Kn-1) | 350 | 0.83 | 0.41 | 0.42 | 0.09 | 0.68 | * | 0.26 | | -0.60 | * | -0.08 | |
| 17 | Viviparous-1. transcriptional activator | 679 | 0.83 | 0.54 | 0.52 | 0.20 | 0.51 | * | -0.18 | | -0.40 | | 0.64 | * |
| 18 | Glutamine synthetase, . | 356 | 0.82 | 0.46 | 0.56 | 0.03 | 0.08 | | -0.18 | | -0.79 | * | 0.27 | |
| 19 | Alcohol dehydrogenase 2. (Adh2-N) | 369 | 0.79 | 0.41 | 0.53 | 0.07 | 0.33 | | -0.64 | * | -0.53 | | 0.32 | |
| 20 | Fructose bisphosphate aldolase | 355 | 0.79 | 0.42 | 0.55 | 0.05 | 0.12 | | 0.34 | | -0.30 | | -0.30 | |
| 21 | Proliferating cell nuclear antigen. | 263 | 0.76 | 0.35 | 0.60 | 0.03 | 0.16 | | -0.55 | | -0.73 | * | 0.24 | |
| 22 | Cystatin I. | 102 | 0.73 | 0.29 | 0.47 | 0.22 | ---- | | ---- | | ---- | | ---- | |
| 23 | Sus1. | 816 | 0.70 | 0.36 | 0.60 | 0.03 | -0.08 | | 0.00 | | 0.18 | | 0.21 | |
| 24 | ZAG1 (homeotic gene) . | 232 | 0.70 | 0.35 | 0.78 | 0.20 | -0.06 | | -0.76 | * | -0.26 | | 0.29 | |
| 25 | Calmodulin. | 149 | 0.69 | 0.28 | 1.21 | 0.01 | ---- | | ---- | | ---- | | ---- | |
| 26 | Manganese superoxide dismutase (SOD-3) | 231 | 0.67 | 0.41 | 0.54 | 0.07 | 0.81 | * | 0.43 | | -0.22 | | -0.10 | |
| 27 | ß-6 tubulin. | 446 | 0.67 | 0.40 | 0.57 | 0.00 | -0.63 | * | 0.40 | | 0.16 | | 0.16 | |
| 28 | QM protein. | 218 | 0.65 | 0.46 | 0.50 | 0.07 | -0.04 | | -0.38 | | -0.46 | | 0.87 | * |
| 29 | Alcohol dehydrogenase 1(Adh1-1F). | 376 | 0.64 | 0.41 | 0.54 | 0.03 | 0.24 | | -0.37 | | -0.47 | | -0.14 | |
| 30 | Beta-amylase. | 488 | 0.62 | 0.37 | 0.49 | 0.09 | -0.22 | | -0.02 | | -0.40 | | -0.21 | |
| 31 | Alpha-tubulin. | 450 | 0.62 | 0.40 | 0.45 | 0.01 | 0.09 | | -0.03 | | -0.30 | | 0.06 | |
| 32 | Regulatory protein GF14-12. | 247 | 0.62 | 0.37 | 1.01 | 0.10 | 0.55 | | 0.14 | | 0.51 | | 0.16 | |
| 33 | Enolase. | 446 | 0.56 | 0.39 | 1.18 | 0.08 | -0.02 | | 0.18 | | -0.25 | | -0.04 | |
| 34 | ATP synthase beta subunit. | 550 | 0.55 | 0.44 | 0.49 | 0.02 | 0.72 | ** | 0.53 | * | 0.14 | | 0.18 | |
| 35 | NADP-dependent malic enzyme (Me1). | 631 | 0.55 | 0.43 | 0.64 | 0.11 | 0.61 | ** | 0.81 | **** | 0.10 | | -0.01 | |
| 36 | Cdc2 kinase. | 294 | 0.54 | 0.35 | 0.59 | 0.03 | 0.14 | | 0.12 | | -0.46 | | -0.18 | |
| 37 | Catalase-1 isoenzyme | 492 | 0.54 | 0.39 | 0.68 | 0.03 | 0.01 | | 0.25 | | -0.05 | | 0.70 | * |
| 38 | Adenine nucleotide translocator | 381 | 0.53 | 0.44 | 0.57 | 0.05 | 0.29 | | 0.14 | | -0.25 | | -0.03 | |
| 39 | Triosephosphate isomerase, . | 253 | 0.48 | 0.42 | 0.62 | 0.06 | 0.15 | | -0.06 | | 0.01 | | 0.62 | * |
| 40 | Starch branching enzyme II. | 798 | 0.39 | 0.41 | 0.67 | 0.10 | 0.59 | * | 0.74 | ** | 0.28 | | 0.29 | |
| | **Probabilidad** | | | | | | $3.4x10^{-5}$ | | $8.2x10^{-4}$ | | $1.5x10^{-8}$ | | 0,027 | |

**Tabla 2.1b**. Genes homólogos entre Maíz/trigo-cebada

| | Gén codificante de: | Largo | $GC_3$ | $GC_2$ | Dist Sin. | Dist. No sin. | Correlaciones intragénicas r($GC_2$-DNS) | r($GC_3$-DNS) | r($GC_3$-DS) | r(DS-DNS) |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Heat shock protein 18kDa | 157 | 0.99 | 0.41 | 0.29 | 0.11 | 0.58 | -0.32 | -0.85 * | 0.60 |
| 2 | Hstone H3 . | 136 | 0.99 | 0.49 | 0.26 | 0.01 | ---- | ---- | ---- | ---- |
| 3 | Flavanone 3-beta-hydroxylase (fht) | 369 | 0.97 | 0.43 | 0.27 | 0.13 | -0.45 | 0.13 | -0.20 | -0.45 |
| 4 | Ferredoxin I (Fd) isoprotein , pFD1'. | 143 | 0.96 | 0.48 | 0.48 | 0.24 | ---- | ---- | ---- | ---- |
| 5 | Transmembrane protein. | 287 | 0.96 | 0.47 | 0.28 | 0.07 | 0.65 * | 0.14 | -0.64 * | -0.27 |
| 6 | Chalcone synthase. | 398 | 0.96 | 0.43 | 0.60 | 0.07 | -0.06 | -0.15 | -0.11 | -0.05 |
| 7 | Histone H4 | 103 | 0.95 | 0.48 | 0.25 | 0.00 | ---- | ---- | ---- | ---- |
| 8 | Lactate dehydrogenase. | 353 | 0.95 | 0.48 | 0.54 | 0.13 | 0.52 * | -0.77 ** | -0.72 ** | 0.88 *** |
| 9 | Ubiquitin fusion protein (UBF9) | 155 | 0.94 | 0.34 | 0.29 | 0.02 | 0.15 | -0.16 | 0.11 | 0.05 |
| 10 | UDPglucose flavonoid glycosyl-transf | 452 | 0.94 | 0.56 | 0.44 | 0.16 | -0.12 | -0.23 | 0.16 | 0.00 |
| 11 | Chlorophyll a/b binding proetin (CAB-m7 gene) | 261 | 0.94 | 0.47 | 0.72 | 0.10 | 0.78 ** | 0.05 | 0.11 | 0.24 |
| 12 | MFS18 . | 121 | 0.93 | 0.63 | 0.49 | 0.31 | ---- | ---- | ---- | ---- |
| 13 | Alpha-amylase | 425 | 0.92 | 0.42 | 0.51 | 0.22 | -0.36 | -0.43 | -0.86 *** | 0.72 ** |
| 14 | Amyloplast-specific transit protein | 600 | 0.92 | 0.43 | 0.50 | 0.11 | 0.71 **** | 0.01 | -0.62 ** | 0.38 |
| 15 | Acidic class I chitinase | 313 | 0.91 | 0.58 | 0.63 | 0.22 | 0.05 | -0.48 | -0.89 *** | 0.19 |
| 16 | Rubisco small subunit | 169 | 0.90 | 0.46 | 0.59 | 0.17 | 0.13 | 0.39 | 0.00 | -0.70 |
| 17 | Heat shock protein 26 (HSP26) | 231 | 0.90 | 0.47 | 0.51 | 0.12 | 0.67 * | -0.01 | -0.79 * | -0.04 |
| 18 | Histone H2A | 138 | 0.90 | 0.46 | 0.53 | 0.11 | ---- | ---- | ---- | ---- |
| 19 | Embryogenic abscisic acid-inducible gene. | 91 | 0.89 | 0.52 | 0.37 | 0.12 | ---- | ---- | ---- | ---- |
| 20 | Metallothionein. | 72 | 0.89 | 0.60 | 0.44 | 0.23 | ---- | ---- | ---- | ---- |
| 21 | Chlorophyll a/b binding protein (Cab-1 gene) | 261 | 0.88 | 0.46 | 0.64 | 0.24 | 0.81 * | 0.44 | -0.82 * | -0.37 |
| 22 | Phospholipid transfer protein | 115 | 0.88 | 0.62 | 0.65 | 0.35 | ---- | ---- | ---- | ---- |
| 23 | H2B histone (gH2B4). | 121 | 0.87 | 0.38 | 0.37 | 0.09 | ---- | ---- | ---- | ---- |
| 24 | Heat shock protein 17.2. | 150 | 0.86 | 0.39 | 0.98 | 0.17 | 0.77 * | -0.01 | -0.86 * | 0.10 |
| 25 | Pathogenesis-related protein. | 164 | 0.86 | 0.49 | 0.71 | 0.24 | -0.57 | -0.66 | -0.79 * | 0.90 ** |
| 26 | Dehydrin (dhn3). | 146 | 0.86 | 0.51 | 0.58 | 0.15 | ---- | ---- | ---- | ---- |
| 27 | Lipase (LIP) | 248 | 0.86 | 0.38 | 0.67 | 0.22 | 0.19 | -0.04 | -0.36 | -0.50 |
| 28 | Chlorophyll a/b binding protein (Cab-m9 gene) | 264 | 0.85 | 0.48 | 0.83 | 0.08 | 0.51 | -0.36 | -0.44 | -0.33 |
| 29 | Knotted-1 (Kn-1) gene. | 349 | 0.84 | 0.42 | 0.53 | 0.10 | 0.69 * | 0.28 | -0.60 * | 0.06 |
| 30 | Glutamine synthetase | 356 | 0.83 | 0.46 | 0.46 | 0.06 | 0.12 | -0.10 | -0.70 * | 0.28 |
| 31 | Calmodulin. | 149 | 0.79 | 0.28 | 0.70 | 0.01 | -0.81 * | 0.66 | -0.40 | -0.84 * |
| 32 | Adh2-N alcohol dehydrogenase 2. | 379 | 0.78 | 0.41 | 0.64 | 0.06 | -0.51 | 0.28 | -0.55 | 0.23 |
| 33 | Cysteine proteinase, clone CCP2 | 359 | 0.74 | 0.47 | 0.57 | 0.09 | -0.18 | 0.34 | -0.37 | -0.17 |
| 34 | Ssus1 gene | 816 | 0.69 | 0.36 | 0.86 | 0.06 | -0.04 | 0.02 | 0.05 | 0.43 * |
| 35 | Protein disulfide isomerase (pdi) | 508 | 0.68 | 0.36 | 0.62 | 0.11 | 0.52 | 0.37 | -0.26 | 0.29 |
| 36 | GapC2 gene. | 337 | 0.66 | 0.40 | 0.49 | 0.03 | -0.16 | 0.29 | 0.01 | 0.33 |
| 37 | Regulatory protein GF14-12 | 248 | 0.66 | 0.37 | 1.02 | 0.11 | 0.57 | 0.27 | -0.20 | -0.14 |
| 38 | Alcohol dehydrogenase (Adh1-S) | 379 | 0.65 | 0.42 | 0.66 | 0.03 | -0.16 | 0.12 | -0.12 | 0.33 |
| 39 | ß-amylase. | 488 | 0.62 | 0.38 | 0.65 | 0.13 | -0.45 | 0.02 | -0.15 | -0.08 |
| 40 | Mit. ATP synthase ß subunit. | 552 | 0.58 | 0.43 | 0.44 | 0.05 | 0.61 * | 0.73 *** | 0.03 | 0.20 |
| 41 | ADP-glucose pyrophosphorylase. | 517 | 0.58 | 0.43 | 1.01 | 0.10 | 0.26 | 0.65 ** | 0.13 | 0.32 |
| 42 | Catalase-1 isoenzyme . | 492 | 0.57 | 0.39 | 0.73 | 0.05 | 0.49 | 0.53 | 0.09 | -0.02 |
| 43 | Porin. | 275 | 0.57 | 0.41 | 0.67 | 0.08 | 0.29 | 0.50 | 0.15 | 0.00 |
| 44 | Cysteine synthase. | 324 | 0.55 | 0.43 | 0.59 | 0.06 | 0.16 | -0.54 | -0.41 | 0.56 |
| 45 | Acyl carrier protein. | 121 | 0.54 | 0.41 | 0.73 | 0.18 | ---- | ---- | ---- | ---- |
| 46 | Dihydrodipicolinate synthase | 376 | 0.51 | 0.45 | 0.97 | 0.15 | 0.81 ** | 0.56 | -0.09 | -0.08 |
| 47 | TATA-binding protein | 200 | 0.46 | 0.38 | 1.01 | 0.03 | 0.06 | 0.88 ** | -0.40 | -0.50 |
| | **Probabilidad** | | | | | | **$5.0x10^{-6}$** | **$3.3x10^{-4}$** | **$6.5x10^{-7}$** | **$2,3x10^{-3}$** |

**Tabla 2.1c,** genes homólogos entre Arroz/trigo-cebada

| | Gén codificante de: | Largo | $GC_3$ | $GC_2$ | Dist. Sin. | Dist. No Sin. | Correlaciones intragénicas r($GC_2$-DNS) | | r($GC_3$-DNS) | | r($GC_3$-DS) | | r(DNS-DS) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gamma-Tip. | 250 | 0.98 | 0.50 | 0.26 | 0.07 | -0.07 | | -0.61 | | -0.25 | | -0.12 | |
| 2 | Histone H3 | 136 | 0.98 | 0.49 | 0.28 | 0.02 | ---- | | ---- | | ---- | | ---- | |
| 3 | Chloroplast transit peptide | 142 | 0.97 | 0.53 | 0.46 | 0.08 | ---- | | ---- | | ---- | | ---- | |
| 4 | Heat shock protein 16.9 kD. | 150 | 0.96 | 0.37 | 0.32 | 0.10 | 0.52 | | -0.55 | | -0.34 | | 0.72 | |
| 5 | Emp1 gene. | 93 | 0.96 | 0.52 | 0.32 | 0.07 | ---- | | ---- | | ---- | | ---- | |
| 6 | Ferredoxin. | 136 | 0.96 | 0.44 | 0.41 | 0.15 | ---- | | ---- | | ---- | | ---- | |
| 7 | Lactate dehydrogenase. | 352 | 0.95 | 0.47 | 0.53 | 0.08 | 0.02 | | 0.19 | | 0.28 | | -0.23 | |
| 8 | type I light-harvesting chlorophyll a/b | 261 | 0.95 | 0.48 | 0.66 | 0.11 | 0.88 | *** | -0.38 | | -0.53 | | 0.48 | |
| 9 | Lipoxygenase L-2. | 855 | 0.95 | 0.41 | 0.40 | 0.17 | 0.34 | | -0.31 | | -0.53 | ** | 0.38 | * |
| 10 | Endochitinase (Cht-2 gene). | 332 | 0.95 | 0.59 | 0.53 | 0.22 | -0.25 | | -0.53 | | -0.41 | | 0.84 | ** |
| 11 | ß-0-glucanase. | 334 | 0.95 | 0.49 | 0.40 | 0.12 | 0.40 | | 0.13 | | -0.42 | | 0.02 | |
| 12 | Lectin. | 212 | 0.94 | 0.63 | 0.38 | 0.27 | -0.28 | | -0.45 | | -0.44 | | 0.88 | *** |
| 13 | Endochitinase (Cht-1 gene). | 320 | 0.93 | 0.57 | 0.50 | 0.17 | 0.48 | | -0.30 | | -0.69 | * | 0.11 | |
| 14 | Thaumatin-like protein. | 169 | 0.90 | 0.59 | 0.52 | 0.29 | 0.26 | | -0.65 | | -0.71 | * | 0.59 | |
| 15 | Alpha-amylase (amy2A). | 435 | 0.89 | 0.44 | 0.54 | 0.12 | -0.08 | | 0.20 | | -0.30 | | 0.30 | |
| 16 | Water-stress inducible protein | 152 | 0.88 | 0.54 | 0.57 | 0.19 | 0.49 | | -0.21 | | -0.56 | | 0.86 | *** |
| 17 | Chloroplast carbonic anhydrase . | 262 | 0.88 | 0.47 | 0.40 | 0.13 | 0.78 | ** | -0.07 | | -0.69 | * | 0.68 | * |
| 18 | Homeobox protein (OSH1 gene). | 355 | 0.88 | 0.41 | 0.37 | 0.08 | 0.59 | * | 0.26 | | -0.86 | *** | 0.05 | |
| 19 | type II light-harvesting chlorophyll a/b | 258 | 0.87 | 0.45 | 0.76 | 0.22 | 0.59 | | 0.19 | | 0.11 | | 0.12 | |
| 20 | Glycogen (starch) synthetase. | 602 | 0.86 | 0.42 | 0.62 | 0.10 | 0.65 | | 0.06 | | -0.61 | ** | 0.34 | |
| 21 | Peroxidase. | 312 | 0.83 | 0.50 | 0.61 | 0.17 | 0.09 | | -0.16 | | -0.75 | * | -0.11 | |
| 22 | Metallothionein-like protein | 74 | 0.80 | 0.59 | 0.50 | 0.25 | ---- | | ---- | | ---- | | ---- | |
| 23 | Cytosolic glutamine synthetase | 356 | 0.77 | 0.45 | 0.33 | 0.06 | 0.21 | | 0.03 | | -0.21 | | 0.47 | |
| 24 | Calmodulin . | 149 | 0.71 | 0.28 | 0.79 | 0.00 | ---- | | ---- | | ---- | | ---- | |
| 25 | Sucrose-UDP glucosyltransferase | 816 | 0.64 | 0.36 | 0.77 | 0.06 | 0.02 | | 0.11 | | 0.00 | | 0.28 | |
| 26 | Alcohol dehydrogenase 1. | 376 | 0.64 | 0.41 | 0.52 | 0.04 | -0.02 | | 0.13 | | -0.11 | | -0.14 | |
| 27 | Kinase C inhibitor homologue | 260 | 0.60 | 0.38 | 0.55 | 0.02 | -0.36 | | 0.46 | | -0.09 | | 0.22 | |
| 28 | Chloroplastic glutamine synthetase | 428 | 0.60 | 0.48 | 0.63 | 0.09 | 0.61 | ** | 0.45 | | 0.26 | | 0.84 | *** |
| 29 | Sucrose synthase. | 807 | 0.59 | 0.36 | 0.57 | 0.04 | 0.09 | | 0.35 | | -0.03 | | -0.04 | |
| 30 | Mitochondrial F1-ATPase. | 550 | 0.58 | 0.43 | 0.53 | 0.07 | 0.59 | ** | 0.75 | *** | 0.05 | | 0.19 | |
| 31 | Aspartic protease. | 506 | 0.57 | 0.44 | 0.46 | 0.08 | -0.03 | | 0.38 | | -0.13 | | -0.01 | |
| 32 | ADP-glucose pyrophosphorylase | 472 | 0.40 | 0.38 | 0.55 | 0.05 | 0.30 | | 0.40 | | 0.58 | * | 0.65 | ** |
| | **Probabilidad** | | | | | | $1.03 \times 10^{-4}$ | | | | $1.18 \times 10^{-5}$ | | $8.7 \times 10^{-8}$ | |

**Nota.** El largo de los genes está dado en codones. $GC_2$ y $GC_3$ son valores promediados entre los dos genes homólogos. **r($GC_2$-DNS)**, **r($GC_3$-DNS)**, **r($GC_3$-DS)** y **r(DS-DNS)** son los coeficientes de correlación intragénica entre las variables que se encuentran dentro de los paréntesis, siendo DS distancia sinónima y DNS distancia no sinónima. Los asteriscos a la derecha de las columnas indican la significancia estadística de los coeficientes de correlación. Los valores de probabilidad la parte inferior de cada columna expresan la significancia estadística conjunta de obtener esa combinación de coeficientes de correlación. Las celdas vacias (línea quebrada) indican que el gén no fue incluído en el análisis de ventanas debido a su pequeño tamaño.

| Tabla 2.2 | Correlación SD-$GC_3$ | Correlación NSD-GC2 | Correlación NSD-GC3 | Correlación NSD-C2 | Correlación NSD-G2 | Correlación NSD-T2 |
|---|---|---|---|---|---|---|
| Maiz/Arroz | -0.512*** | 0.452** | 0.405** | 0.303 | 0.402** | -0.568*** |
| Maiz/Trigo-Cebada | -0.617**** | 0.598**** | 0.282 | 0.541*** | 0.333* | -0.428** |
| Rice/Trigo-Cebada | -0.427* | 0.737**** | 0.431* | 0.273 | 0.662**** | -0.530** |

**Nota.-*** significativo al 5%, * * significativo al 1% ,*** significativo 0.1% y **** significativo al 0.01%

| Tabla 2.3 | Núm. de genes | Correlación DNS-$GC_2$ | Correlación DNS-$GC_3$ | Correlación DS-GC3 | Correlación DS-DNS |
|---|---|---|---|---|---|
| **Población muy rica en GC3** | | | | | |
| Maiz/Arroz | 18 | 0,64 ** | 0,1 | -0,61 ** | 0,18 |
| Maiz/Trigo-Cebada | 30 | 0,53 ** | -0,25 | -0,58 *** | 0,44 * |
| Arroz/Trigo-Cebada | 22 | 0,59 ** | -0,47 * | -0,53 ** | 0,42 * |
| **Población menos rica en GC3** | | | | | |
| Maiz/Arroz | 22 | -0,26 | 0,08 | -0,09 | 0,03 |
| Maiz/Trigo-Cebada | 17 | 0,32 | -0,29 | -0,35 | 0,26 |
| Arroz/Trigo-Cebada | 10 | 0,82 ** | 0,43 | -0,02 | -0,33 |

**Nota.-** *,**,***, igual que en la Tabla 2.2

## IV.3 TASAS DE SUSTITUCIONES SINÓNIMAS EN TRYPANOSOMÁTIDOS

De manera similar a lo realizado en plantas, empezaremos presentando los resultados correspondientes a los valores globales de los genes sin considerar la variación intragénica, la cual será tratada más adelante. También se presentarán y discutirán resultados acerca de la aceleración en las tasas de cambios nucleotídicos a la cual estuvo sometido (y probablemente continua estando) el linaje que conduce a los trypanosomas africanos (Salivaria).

### IV.3.1 Relación entre las tasas evolutivas y la composición de bases.

La Figura 3.1 resume los resultados de los análisis realizados con el grupo genes homólogos de trypanosomatidos que se presentan en la Tabla 3.1. A partir de los resultados presentados en las Figuras 3.1a-h podemos ver que el comportamiento de las sustituciones sinónimas y no-sinóminas es, en varios aspectos, similar a lo descrito en el capitulo anterior para plantas y también, de acuerdo a publicaciones previas, a lo que ocurre en los mamíferos. En primer lugar, la tasa de sustituciones no-sinónimas se encuentra positivamente correlacionado con el contenido en G+C en la segunda posición de los codones ($GC_2$), siendo los coeficientes de correlación estadísticamente significativos en ambos grupos de genes homólogos pero particularmente alto en el grupo de genes que son homólogos entre *Trypanosoma cruzi* y *Leishmania* (figuras 3.1a y 3.1e). Como discutiremos más adelante, de acuerdo a los resultados que se presentan en esta tesis y a otros obtenidos por quien suscribe en colaboración con otros colegas, la correlación entre las tasas no-sinónimas y la composición de bases en la segunda posición de los codones parece ser (por razones que se desconocen) una constante en todos los grandes grupos biológicos tanto eucariotas como procariotas. Por otro lado, y contrastando lo que ocurre en las gramíneas, la tasa de sustituciones no-sinónimas no presenta correlación positiva con el contenido en $GC_3$, por el contrario, la correlación es de signo negativo siendo estadísticamente significativo en ambos grupos de genes homólogos.

En lo que tiene relación a la tasa de cambio sinónimo, podemos resaltar tres grandes puntos. En primer lugar, y de forma similar a lo que ocurre en genes de gramíneas, la distancia sinónima presenta correlación negativa con el contenido en $GC_3$. Sin embargo en el caso de los trypanosomátidos, los coeficientes de correlación presentan valores extremadamente altos (Figs. 3.1b y 3.1f). La mencionada correlación negativa implica que los genes ricos en $GC_3$ se encuentran muchísimo más conservados a nivel de las posiciones sinónimas de lo que lo están los genes pobres en $GC_3$. Esta observación es del todo compatible con los resultados sobre la evolución del uso de codones sinónimos que hemos obtenido anteriormente (Alvarez et al, 1994). En este sentido cabe resaltar que en los trypanosomas, los genes ricos en $GC_3$ son en su mayoría genes de alta expresión, siendo sus preferencias en el uso de codones altamente conservadas entre *Leishmania, Crithidias fasciculata, T. cruzi* y en menor medida *T. brucei*. Por el contrario, los genes pobres en $GC_3$, que en general corresponden a genes expresados a niveles bajos, son mucho más divergentes en sus preferencias de codones sinónimos. Esta conservación en la preferencia de codones en los

genes de alta expresión indicaría que el conjunto de codones traduccionalmente óptimos estaría conservado en los trypanosomátidos. Los resultados sobre uso de codones sinónimos tomados en conjunto con la correlación negativa entre la velocidad de cambio sinónimo y el contenido en $GC_3$ presentado en esta tesis, indican de forma contundente la existencia de selección negativa seguramente determinada por la presión para incrementar en la eficiencia traduccional. Esta fuerza selectiva no sólo tendería a mantener las preferencias de codones sinónimos en los genes de alta expresión, sino, y lo que es mucho más sugestivo, a impedir las sustituciones nucleotídicas en las posiciones sinónimas de los genes que se expresan a altos niveles.

En segundo lugar, las distancias sinónimas y no-sinónimas se encuentran altamente correlacionadas entre sí (Figs 3.1d y 3.1h), siendo los coeficientes de correlación no sólo altamente significativos desde el punto de vista estadístico, sino que los mismos son sorprendentemente más altos a aquellos reportados por varios autores en genes de mamíferos (ver Introducción Sección I.2.2). Si asociamos esta correlación entre las tasas de cambio sinónimo y no-sinónimo con el hecho de que ambas tasas evolutivas están a su vez correlacionadas con el contenido en $GC_3$ (que en los trypanosomátidos es equivalente a sesgo direccional en el uso de codones, pues todos los codones que parecen ser traduccionalmente óptimos terminan en C o G), llegamos a la conclusión que la selección para conservar un determinado uso de codones en los trypanosomátidos estaría determinada no sólo por la eficiencia traduccional sino que también por el grado de conservación aminoacídica de los genes. Postergaremos por el momento la discusión acerca de las implicancias esta cuestión, que sin duda, es uno de los puntos centrales en esta tesis.

El último punto que surge de la Figs. 3.1b y 3.1 (en relación a las distancias sinóminas), es que las mismas son grandes en todos los genes. Sin embargo, el aspecto más llamativo es que las distancias sinónimas entre *T. cruzi* y *T. brucei* son promedialmente mayores que aquellas entre *T. cruzi* y *Leishmania*, siendo esto válido incluso cuando comparamos el mismo par de genes homólogos (por ejemplo, actina, $\alpha$y $\beta$ tubulina etc.). Este resultado indicaría un incremento en la tasa de cambio sinónima en *T. brucei* (y probablemente las restantes especies de la sección Salivaria), dado que *Trypanosoma* es con toda seguridad un género monofilético (Alvarez et al, 1996, 1998) por lo que sería esperable que las distancias entre *T. brucei* y *T. cruzi* fueran menores que aquellas entre *T.cruzi* y *Leishmania*. Este tema será el sujeto de estudio de la siguiente sección.

**IV.3.2 Aceleración de las tasas sinónimas y no-sinónimas en genes de Salivaria**

Debido a la ausencia de registro fósil, resulta imposible comparar las tasas absolutas de evolución nucleotídica en los trypanosomátidos. Sin embargo, incluso en ausencia de datos sobre los tiempos de divergencia, es posible comparar las tasas de cambio mediante el uso del test de las tasas relativas (ver Sección III.1.3.2). Este test no nos da una medida del número de cambios nucleotídicos por unidad de tiempo (digamos por cada millón de años), pero si nos permite determinar que linaje evoluciona más rápido que otro. Mediante el test de las tasas relativas comparamos las tasas de cambio entre *T. cruzi* y *T. brucei*. Es importante aclarar que aparte de la

gliceraldehido 3-fosfato deshidrogenasa, no hay datos de secuencia en ningún grupo taxonómico que pueda ser apropiadamente usado como grupo externo de los trypanosomátidos. No obstante, si asumimos que el género *Trypanosoma* es monofilético, podemos usar como grupo externo otras especies de trypanosomátidos que no pertenezcan a *Trypanosoma*. Concretamente podemos usar genes de *Leishmania* o *Crithidia*. La asunción de monofilia para el género *Trypanosoma* posee un fuerte soporte a la luz de los resultados que hemos presentado anteriormente (Alvarez et al, 1996) así como de los resultados que se muestran en el manuscrito número 3 que se anexa.

La Tabla 3.2 muestra los resultados del test de las tasas relativas realizado para 19 genes nucleares. El primer aspecto de importancia que surge de esta Tabla es que para todos los genes analizados, las distancias no-sinóninas entre *T. cruzi* y *T. brucei* ($D_{12}$) son mucho más pequeñas que las aquellas entre cada especie de *Trypanosoma* y *Leishmania* o *Crithidia* ($D_{13}$ y $D_{23}$). Evidentemente, este resultado da soporte adicional a la asunción de que *Trypanosoma* es un género monofilético. En segundo lugar, varios genes muestran tasas mayores de cambio no-sinónimo en *T. brucei* en relación a *T. cruzi*, siendo la diferencia estadísticamente significativa en 5 de estos genes. No obstante, *T. brucei* no parece evolucionar consistentemente más rápido que *T. cruzi* a nivel de sustituciones sinónimas puesto que 9 de 19 genes analizados (47%) acumularon mayor número de sustituciones no-sinónimas en el linaje que lleva a *T.cruzi*. Cuando se consideran todos los genes juntos (como si fueran un único gen) la diferencia se hace bastante pequeña aunque estadísticamente significativa, mostrando que *T. brucei* evoluciona aproximadamente un 12% más rápido que *T. cruzi* en lo que se refiere a sustituciones no-sinónimas.

En la sección b de la Tabla 3.2 podemos observar las comparaciones llevadas a cabo para las sustituciones sinónimas. Como puede verse, a excepción del gen para la topoisomerasa II, todos los genes restantes evolucionan más rápido a nivel de las sustituciones sinónimas en *T. brucei* que en *T. cruzi*. Es de destacar que en 6 de los 19 genes analizados fue imposible realizar el test de las tasas relativas. Esto se debe al hecho de que las distancias sinónimas entre *T. brucei* y *Leishmania* son de tal magnitud que el método utilizado para corregir las sustituciones múltiples (Li, 1993) resulta inaplicable. En la mayoría de los genes para los cuales fue posible realizar el test, las diferencias en las tasas de sustituciones sinónimas son estadísticamente significativas. Este resultado nos lleva a la conclusión de que las tasa de sustituciones sinónimas son definitivamente mayores en *T. brucei* en relación a *T. cruzi*. Si todos los genes se consideran en conjunto vemos que *T. brucei* presenta tasas de evolución sinónima que son aproximadamente un 70% mayores a las que se observan en *T. cruzi*. Debe tenerse en cuenta sin embargo, que esta estimación de la magnitud del incremento en la tasa sinónima es muy imprecisa. La causa de esta impresión reside en el hecho de que los errores estándar de las estimaciones de distancia son muy grandes (particularmente entre *T. brucei-Leishmania*). Como resultado, el test de las tasas relativas no nos permite obtener una estimación confiable de cuanto más rápido evolucionan las posiciones sinónimas en *T. brucei*. Con el objetivo de estimar la magnitud de este incremento en forma más exacta y confiable se calcularon los cocientes entre distancia sinónima sobre distancia no-sinónima

en procesos de divergencia independientes que involucren a especies más estrechamente emparentadas. Debe tenerse en cuenta que el mencionado cociente se espera que varíe fundamentalmente como resultado del incremento o disminución de la tasa sinónima, puesto que los factores que tienden a aumentar las tasas de cambio nucleotídico (tiempo de generación del organismo, tasa de mutación, etc.) producen un efecto muy pequeño en las tasas de cambio no-sinónimo (Ohta, 1993, 1995). *T. brucei* pertenece a la sección Salivaria (también llamados trypanosomas africanos), la cual integran los trypanosomas trasmitidos en las piezas bucales de insectos dípteros y que se caracterizan entre otras cosas por poseer antígenos de superficie hipervariables (VSGs). Teniendo en cuenta que este grupo de trypanosomas en sin lugar a duda monofilético, podemos afirmar que las procesos de divergencia entre dos especies de Salivaria como *T. brucei* y *T. vivax* o entre *T. brucei* y *T. congolesense* son completamente independientes de los procesos de divergencia entre dos trypasomatidos inferiores como *Leishmania* y *Crithidia fasciculata* o entre dos trypanosomas de la sección Stercoraria como *T. cruzi* y *T. rangeli.* La Tabla 3.3 muestra los resultados de estos cocientes en 4 genes nucleares. Como uno puede apreciar, el cociente es aproximadamente 4 veces mayor (en todos los genes analizados) en Salivaria que en los restantes pares de especies. Además estas últimas presentan cocientes bastante similares entre sí. Este incremento del 400% en los cocientes experimentado por los Salivaria, sólo puede ser atribuible cambios de igual magnitud en la velocidad de evolución sinónima (lo más probable que se deba a una aceleración en Salivaria mas que a una disminución en los restantes trypanosomátidos). Podemos descartar que la razón del incremento del cociente sea una disminución de la tasa no-sinónima en los Salivaria, pues como fue puesto en evidencia a través del análisis de tasas no-sinónimas (Tabla 3.2a), *T. brucei* presenta tasas de cambio no-sinónimo levemente mayores (o mínimamente iguales) a *T. cruzi.*

**IV.3.3 Correlaciones intragénicas**.

Las distancias sinónimas entre las especies de trypanosomátidos incluidas en los análisis de las secciones precedentes son excesivamente grandes para permitir un análisis confiable de las correlaciones intragénicas entre estas distancias y otras variables. A causa de esta complicación dicho análisis fue omitido en estos genes. En su lugar se ha realizado un análisis equivalente en los genes codificantes para la glicoproteína de membrana conocida como GP63, la que además de ser la proteína de membrana más abundante presenta actividad metaloproteinasa (Button et al, 1989). Esta proteína es específica de los trypanosomátidos inferiores puesto que copias de sus genes (Inverso et al ; 1993) o actividad enzimática han sido descriptas en todas las especies analizadas de *Leishmania, Crithidia, Endotrypanum, Herpetomonas, Leptomonas*, *Phytomonas* y *Blastocrithdia* (Etges, 1992; Schneider & Glaser, 1993; Medina-Acosta et al, 1994), pero nunca ha podido ser detectada en especies del género *Trypanosoma* (McConville & Schneider, 1993; Medina-Acosta et al. 1994).

La figura 3.2a muestra los perfiles de distancias sinónimas y no-sinónimas en un alineamiento que contiene 19 secuencias de la GP63 pertenecientes a varias especies de *Leishmania*. Como claramente pone en evidencia esta figura ambos perfiles de distancia están estrechamente relacionados en forma tal que aquellas zonas del gen que presentan poca diferenciación a nivel de aminoácidos también están conservadas a nivel sinónimo, mientras que las zonas del gen con mayor divergencia aminoacídica también son más divergentes desde el punto de vista sinónimo. La figura 3.2b muestra los perfiles de $GC_3$ y CAI (Indice de Adaptación de Codones, Sharp & LI, 1987) para el mismo gen. En esta figura pueden observarse claramente dos aspectos importantes. En primer lugar los perfiles de $GC_3$ y CAI son muy similares lo que pone de manifiesto que el valor $GC_3$ es un excelente indicador de la frecuencia de codones mayores en trypanosomátidos. En segundo lugar podemos apreciar que la variación de el CAI (y de $GC_3$ por supuesto) es inversa a la de ambos tipos de sustituciones ($r=-0.76$, $P<10^{-4}$). Esto es, en las regiones del gen que presentan bajos valores de CAI muestran a su vez altas tasas de ambos tipos de sustituciones mientras que aquellas regiones con altos valores de CAI presentan bajas tasas de sustituciones nucleotídicas. En otras palabras, en los genes que codifican la proteína GP63, aquellas regiones del gen que presentan alta frecuencia de los codones considerados traduccionalmente óptimos presentan bajas tasas de cambio tanto sinónimo como no-sinónimo.

**IV.3.4 Conversión génica y correlaciones intragénicas**

La conversión génica es un proceso de transferencia uni-direccional de material genético en el que dos genes o segmentos de genes interactúan de forma tal que uno ellos (el donador) "convierte" al otro gen (receptor) en su propio tipo. Es decir, el gen o segmento de gen receptor se vuelve idéntico al donador. Este proceso tiene lugar entre genes que presenten homología de secuencias, ya sea alelos de un mismo locus (Nicolas & Rossignol, 1983), o miembros de familias multigénicas (Weiner & Denison, 1983).

Debido al hecho que los genes codificantes de la proteína GP63 pertenecen a una familia multigénica, es probable que la covariación entre ambos tipos de sustituciones que se observa en los mismos sea el resultado de la conversión génica en segmentos. Hay que tener en cuenta que la conversión génica es potencialmente capaz de generar el patrón de covariación debido a que luego de la ocurrencia de la misma los segmentos convertidos pasan a ser idénticos desde el punto de vista sinónimo y no-sinónimo. Como resultado, es posible que cuando comparamos miembros de una familia multigénica encontremos segmentos génicos con poca o ninguna diferenciación, tanto a nivel sinónimo como no-sinónimo (los segmentos convertidos relativamente reciente) y segmentos con mayor diferenciación.

Sin embargo es posible testar si el patrón de covariación que se observa pueda ser atribuible a la conversión génica o a otra causa. La conversión es al azar desde el punto de vista espacial. Esto es, cualquier zona de un gen tiene la misma chance de participar en el proceso de conversión. Por lo que si la conversión génica es realmente operativa en esta familia multigénica, es

muy improbable que encontremos el mismo patrón de covariación cuando comparamos dos pares de genes codificantes de GP63. Por el contrario si el patrón de covariación obedece a una causa funcional, como por ejemplo a la distribución espacial de los aminoácidos importantes desde el punto de vista funcional, es de esperar que el mismo patrón espacial de sustituciones sinónimas y no-sinónimas sea encontrado cuando comparamos procesos de divergencia filogeneticamente independientes.

Para poder comparar procesos de divergencia que sean independientes desde el punto de vista evolutivo, es necesario conocer a priori las relaciones filogenéticas de los genes que se desean analizar. Con este propósito se construyó el árbol filogenético que se muestra en la figura 3.3. A partir de la información aportada por este árbol podemos afirmar que el proceso de divergencia entre las secuencias señaladas en la figura 3.3 como 1 y 2 (AF039721, proveniente de *Leishmania major* y LIGP63 de *L. infantum*) , es independiente de la divergencia entre las secuencias señaladas como 3 y 4 (LMGP63C1de *L. mexicana* y LEIMSP52 de *L. donovani*) y de aquellas señaladas como 5 y 6 ( AF037165 de *L. panamensis* y LEIGO63K de *L. guyanensis*). Asimismo, la divergencia entre 3 y 4 es independiente de 5 y 6.

Las figuras 3.4a, 3.4b y 3.4c muestran los perfiles de variación sinónima y no-sinónima en estos tres pares de genes. Puede observarse que los tres pares de genes presentan perfiles de variación donde las tasas de sustituciones sinónimas y no-sinónimas se encuentra claramente correlacionadas. Pero lo que sin duda es más importante es que los tres pares de genes dan lugar a perfiles que son extraordinariamente similares entre sí ; de hecho el perfil de divergencia sinónima entre las secuencias 1 y 2 está correlacionado con el perfil de divergencia no-sinónima entre las secuencias 3 y 4 y con el de entre las secuencias 5 y 6 ($r=0.52$, $p<0.05$, y $r=0.51$, $p<0.05$, respectivamente). La divergencia sinónima entre 3 y 4 está correlacionada con el perfil de divergencia no-sinónimo entre 5 y 6 ($r=0.84$. $p<10^{-5}$ ), así como lo está el perfil sinónimo entre 5 y 6 con el perfil no-sinónimo de 3 y 4 ($r=0.71$, $p<<10^{-3}$). El perfil no-sinónimo entre las secuencias 1 y 2 presenta correlaciones positivas pero bajas con los perfiles sinónimos de las otras comparaciones lo cual puede atribuirse a la poca divergencia aminoacídica (y por ende poca varianza) entre las secuencias 1 y 2. Podemos afirmar entonces, que tres procesos independientes de divergencia dieron lugar a patrones espaciales de divergencia sinónima y no-sinónima claramente similares. En otras palabras, la figura 3.4 nos muestra convergencia de los patrones de divergencia.

Por último se desea recalcar el hecho de que los perfiles de $GC_3$ y CAI presentan correlación negativa con el de ambas tasas de sustituciones, lo cual resultaría completamente inexplicable si la covariación fuera causada a través de procesos de conversión génica recurrente.

## Figura 3.1

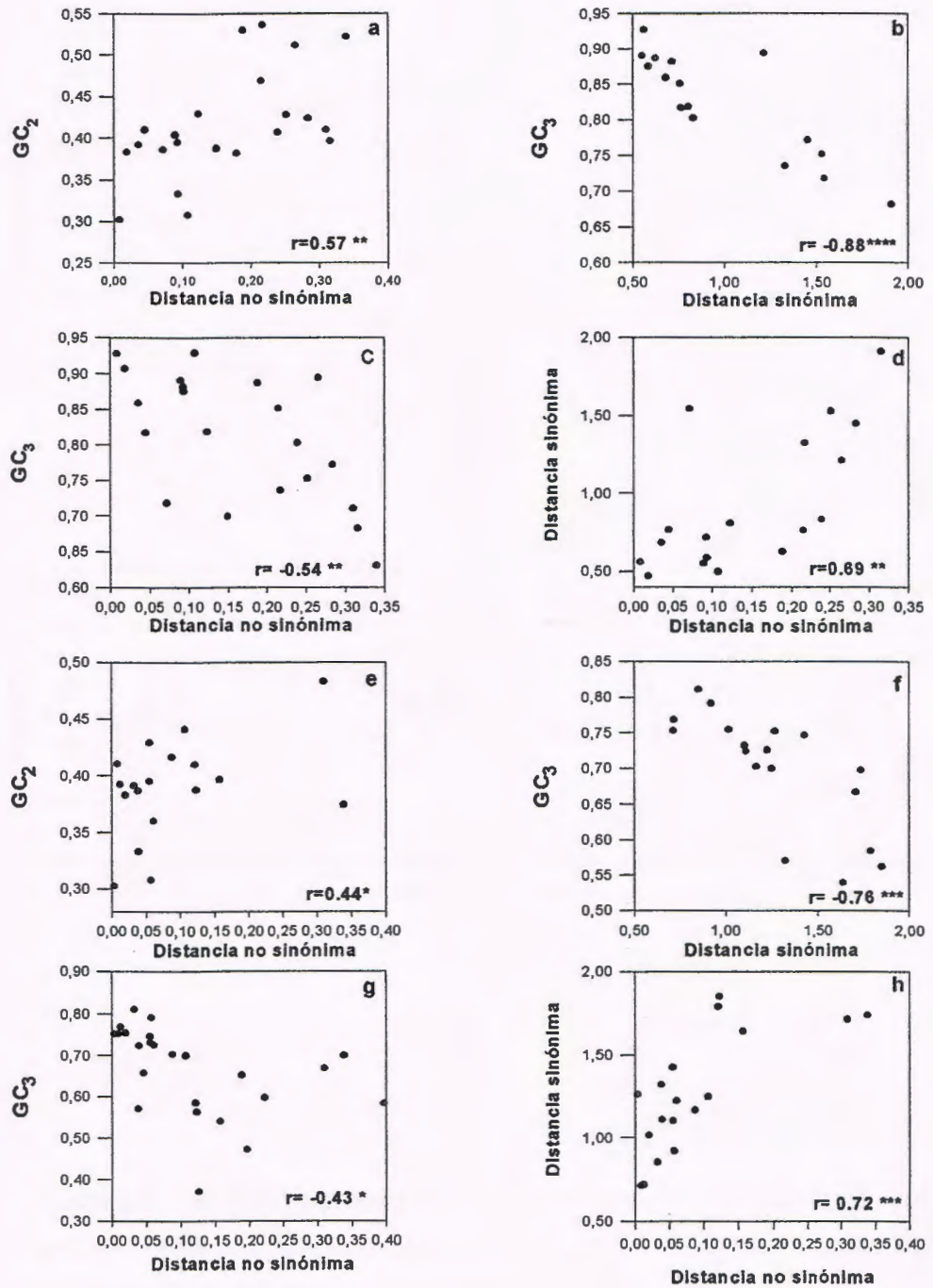



**Figura 3.1.** Relaciones entre la composición de bases y las distancias nucleotídicas en genes de trypanosomátidos. Las figuras 3.1a-d corresponden a genes que son homólogos entre *Leishmania* y *T. cruzi.* Las figuras 3.1e-h corresponden a genes que son homólogos entre *T. brucei* y *T. cruzi.*

**Figura 3.2**

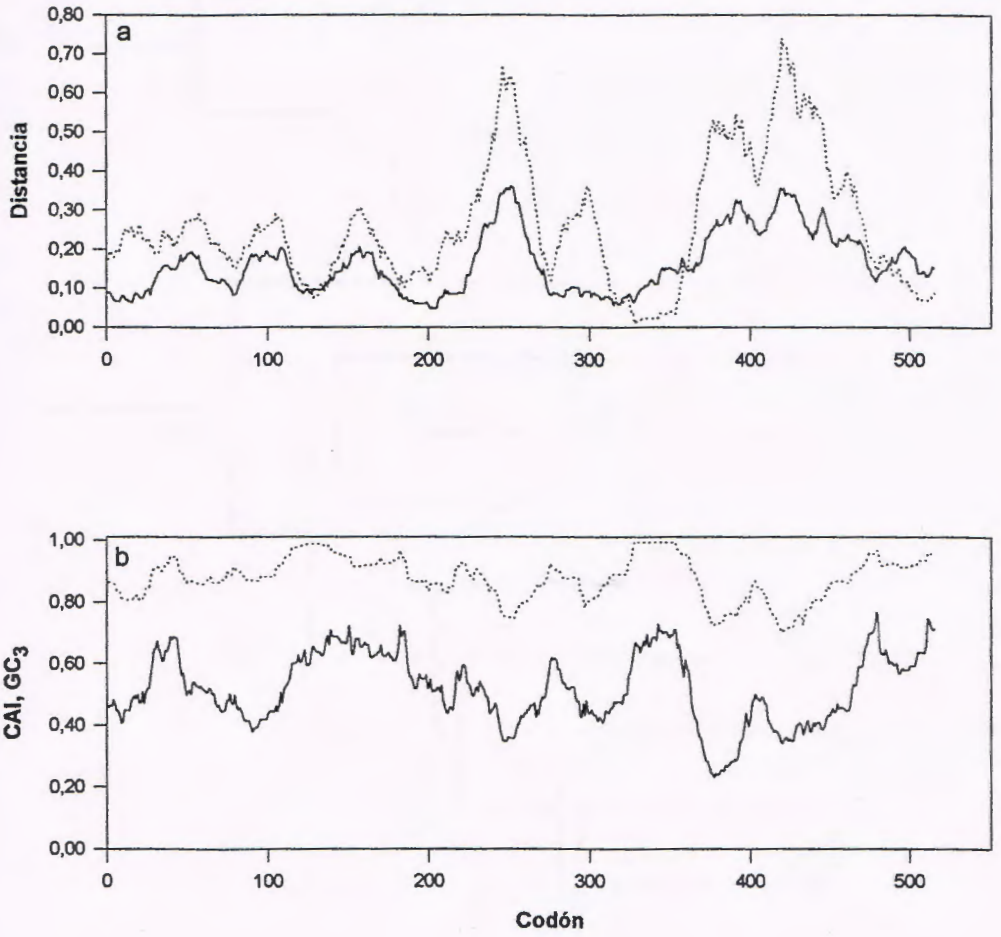



**Figura 3.2.** 3.2a. Perfiles promedio de distancia sinónima (línea punteada) y no sinónima (línea continua) en genes de GP63. 3.2b. Perfiles promedio de $GC_3$ (línea punteada) y Codon Adaptation index (línea continua).

## Figura 3.3

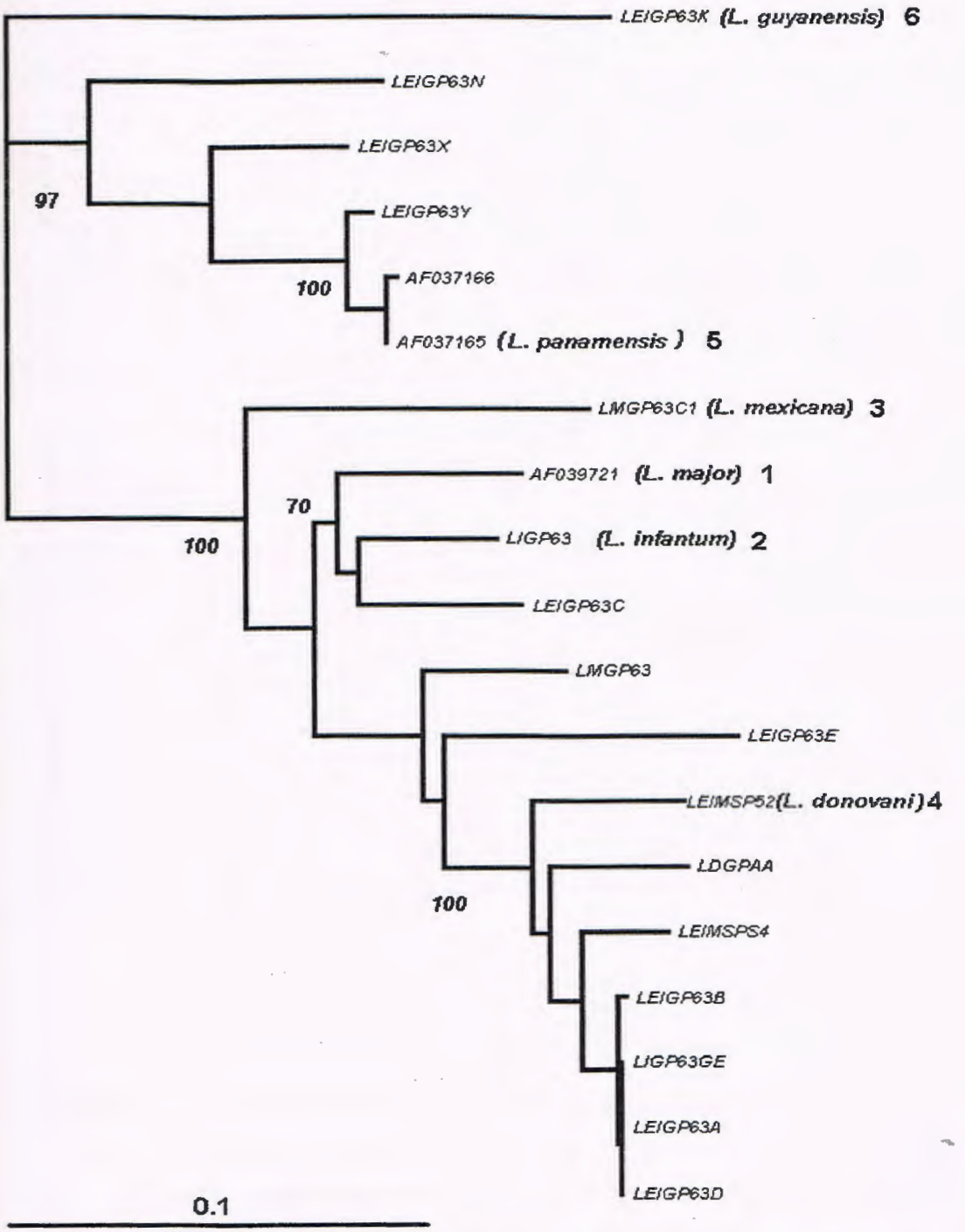



**Figura 3.3.** Árbol filogenético de 19 genes codificantes de la proteína GP63. El método de reconstrucción filogenética fue el neighbor-joining (Saitou & Nei, 1987) a partir de distancias estimadas usando la primera y segunda posición de los codones. La correción para sustituciones mútiples se llevó a cabo usando el modelo de 2 parámetros de Kimura (ver material y métodos sección ). Los valores de boostrap están dados para los nodos significativos (1000 seudoréplicas). El largo de las ramas del árbol es propocional al número de sustituciones estimadas para la rama. La barra ubicada en la parte inferior izquierda de la figura equibale una distancia de 0.1 sustituciones por sitio.

**Figura 3.4**

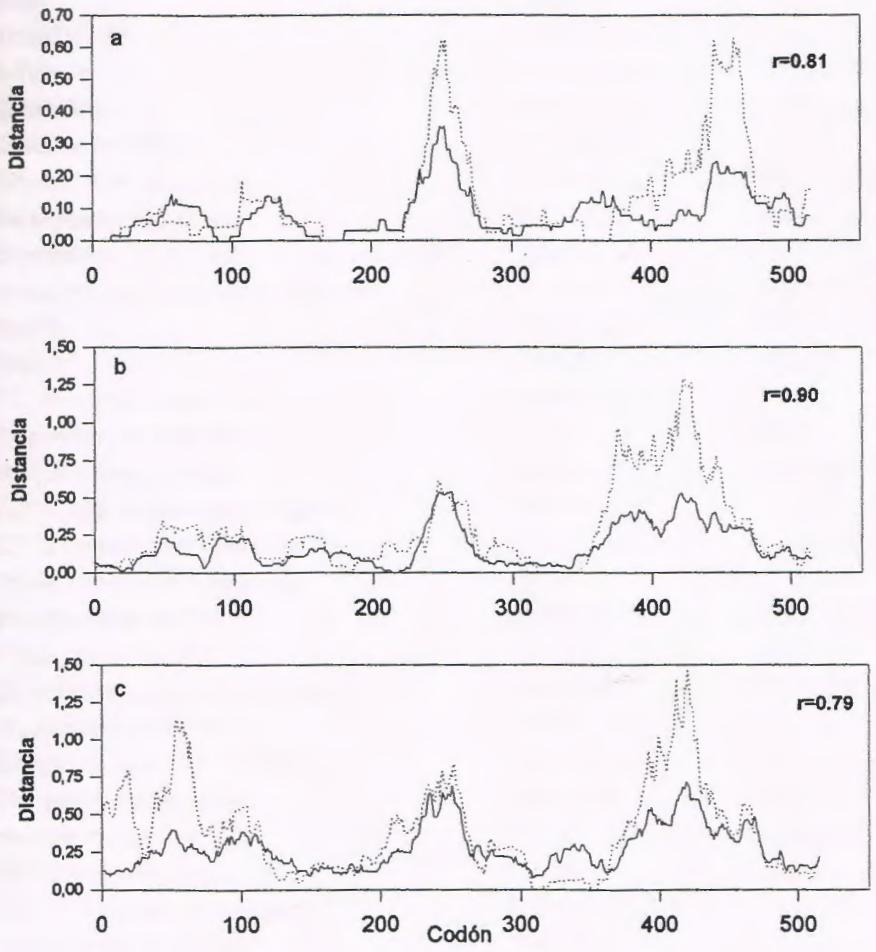



**Figura 3.4.** Perfiles de distancias sinónimas (línea punteada) y no sinónimas (línea continua) en tres pares de genes homólogos que divergen en forma independiente. 3.4a, 3.4b y 3.4c respectivamente muestran los perfiles de divergencias entre las secuencias señaladas en la figura 3.3 como 1y 2, 3y 4 y 5 y 6.

**Tabla 3.1.** Lista de genes homólogos en trypanosomátidos.

| | | *T. brucei* | *T. cruzi* | *Leishmania* |
|---|---|---|---|---|
| 1 | Actin | TRBACTA | TCU20234 | LEIACTIN |
| 2 | AlphaTubulin | TRBTUBBAB | TRBBEALTU | LDU09612 |
| 3 | ß-Tubulin | TRBTUBBAB | TRBBEALTU | LEITUBBAA |
| 4 | Calmodulin | TBCALUBG | TRBCALB2 | LTCAMA |
| 5 | Cysteine proteinase | TBCYSPRO | TRBCYSPRO | LMCPBMR |
| 6 | Dihydrofolate reductase | TBU20781 | TRBDHFRTS | LMDHFRTS |
| 7 | Paraflagellar rod protein | TRBFRAB | TRBPAR2A | LMU45884 |
| 8 | Glyceraldehyde 3-p dehydro. (glycosomal) | TBGAPDHB | TRBGAP | LMGAP |
| 9 | Hypoxanthine phosphoribosyl-transf. | TRBHGPRT | TRBHGPRTA | LEIHGPT |
| 10 | Hsp70 | TBHSP70A | TRBHSP70A | LDHSP70 |
| 11 | Hsp83 | TBHSP83 | TRBHSC | LEIHSP01 |
| 12 | P2- ribosomal protein (ß) | TBARP2PA | X75030 | LILIPA |
| 13 | Trypanothione reductase | TBTRG | TRBTTR | LDTRYREDA |
| 14 | Ubiquitin fusion protein | UBIQEP52/1 | TRBUBIA | LMUBIQUTN |
| 15 | CDC2- like (protein kinase type 1) | TBKIN1 | TCU74762 | LMCDC2 |
| 16 | CDC2 related (protein kinase 3) | TBCDC2RPK | TCU69958 | LMCRK3 |
| 17 | Triose phosphate isomerase | TBTIN | TCU53867 | LMTPIGE |
| 18 | Mitochondrial HSP70 | TRBMHSP70H | TRBMTP70 | 1170375 |
| 19 | Toposisomerase II | TBTOPII | TRBTOPII | 2944450 |
| 20 | Glycosomal protein 60 Kd (Peck) | TBGLP60 | TRBPEPCK | ---- |
| 21 | Mitochondrial HSP60 | TRBMTHSP | TRBHSP60A | ---- |
| 22 | BiP/grp78 gene (HSP70 related) | TRBBIPGRP | TRBGRP78 | ---- |
| 23 | Elongation Factor EF1 | TBU10662 | TRBEF1A | ---- |
| 24 | Hexose transporter | TBTHT | TCU05588 | ---- |
| 25 | Calcium binding protein | TBTBS17 | S43664 | ---- |
| 26 | GP72 , flagellar glycoprotein | TBU43717 | TRBGP72 | ---- |
| 27 | Ornithine decarboxylase | TRBORD | ---- | LEIOD |
| 28 | Esag10 | TBESAG10 | ---- | LEIIMNBA |
| 29 | Phosphoglycerate kinase (glycosomal) | GKG_B | ---- | LEIGPGK |
| 30 | Phosphoglycerate kinase (cytosolic) | GKB_C | ---- | LEICPGK |
| 31 | Glyceraldehyde 3 P dehydrogenase (cyt.) | TBGAP | ---- | LMGAPGENE |
| 32 | Pyruvate kinase | TBPYK1 | ---- | LMPYK |
| 33 | S-adenosylmethionine decarboxylase | TBU20092 | ---- | LDU20091 |
| 34 | Phospho-glucoisomerase | TBPGIG | ---- | LMPGI |
| 35 | EFH5 (calcium binding protein) | TBCALUGB | ---- | LTEFH5 |
| 36 | Ribosomal Protein -2(alpha type) | ---- | TCP2BMRNA | LILIPB |
| 37 | Histone H2A | ---- | TCHISH2A | LDH2A |
| 38 | Histone H2B | ---- | TRBRH2B | LEIHISH2B |
| 39 | Histone H3 | ---- | TRBHIH3A | LEIHISH3 |
| 40 | Histone H1 | ---- | TRBHIC2AA | U01031 |
| 41 | P0, ribosomal Protein | ---- | TRBRPB0X | LEIRIBPP |

**Nota.** Las líneas quebradas significan que la secuencia de el gen correspondiente no se halla disponible en la especie.

**Tabla 3.2.** Test de las tasas relativas en genes de *Trypanosoma cruzi* y *Trypanosoma brucei*

| **Tabla 3.2a** | | | **Distancias No Sinónimas** | | | | |
|---|---|---|---|---|---|---|---|
| **Gén:** | **$D_{13}$** | **$D_{23}$** | **$D_{13}$-$D_{23}$ st** | **$D_{12}$** | **$D_{10}$** | **$D_{20}$** | **D10/D20** |
| Actin | 0,0719 | 0,0689 | 0,0030 ± 0,0072 | 0,039 | 0,0209 | 0,0179 | 0,8572 |
| Alpha tubulin | 0,0451 | 0,0366 | 0,0085 ± 0,0031 ** | 0,009 | 0,0086 | 0,0001 | 0,009 |
| Beta tubulin | 0,0358 | 0,0377 | -0,0019 ± 0,0036 | 0,013 | 0,0054 | 0,0072 | 1,3495 |
| Cysteine proteinase | 0,5326 | 0,5543 | -0,0216 ± 0,0302 | 0,310 | 0,1443 | 0,1659 | 1,1499 |
| Dihydrofolate reductase | 0,2838 | 0,3224 | -0,0386 ± 0,0183 * | 0,223 | 0,0923 | 0,1308 | 1,4177 |
| Paraflagellar rod protein | 0,1075 | 0,1222 | -0,0147 ± 0,0073 * | 0,057 | 0,0214 | 0,0361 | 1,6885 |
| glycosomal GAP | 0,1228 | 0,1469 | -0,0241 ± 0,0099 * | 0,056 | 0,0157 | 0,0398 | 2,5314 |
| Hypoxanthine transferase. | 0,4345 | 0,6085 | -0,1740 ± 0,0535 ** | 0,423 | 0,1245 | 0,2984 | 2,3972 |
| Heat shock protein 70 | 0,0927 | 0,0828 | 0,0098 ± 0,0068 | 0,055 | 0,0325 | 0,0227 | 0,6977 |
| Heat shock protein 83 | 0,0934 | 0,0992 | -0,0058 ± 0,0086 | 0,039 | 0,0168 | 0,0226 | 1,345 |
| Triose phosphate isomerase | 0,2519 | 0,2332 | 0,0187 ± 0,0227 | 0,189 | 0,1039 | 0,0852 | 0,8198 |
| Calmodulin | 0,0083 | 0,0118 | -0,0035 ± 0,0035 | 0,004 | 0 | 0,0035 | NA |
| Ubiquitin fusion protein | 0,0185 | 0,0143 | 0,0042 ± 0,0083 | 0,02 | 0,0121 | 0,0079 | 0,6517 |
| CDC2- like (1) | 0,1489 | 0,2183 | -0,0693 ± 0,0163 **** | 0,123 | 0,0269 | 0,0962 | 3,5793 |
| CDC2- like (2) | 0,1786 | 0,1708 | 0,0077 ± 0,016 | 0,126 | 0,067 | 0,0593 | 0,8845 |
| Mitochondrial Hsp70 | 0,0894 | 0,0848 | 0,0046 ± 0,0098 | 0,046 | 0,0253 | 0,0207 | 0,8184 |
| P2- ribosomal protein | 0,3394 | 0,3139 | 0,0255 ± 0,0402 | 0,197 | 0,1112 | 0,0857 | 0,7707 |
| Trypanothione reductase | 0,3098 | 0,3277 | -0,0179 ± 0,0143 | 0,121 | 0,0517 | 0,0696 | 1,3458 |
| Toposisomerase II | 0,3156 | 0,2922 | 0,0233 ± 0,0102 * | 0,157 | 0,0902 | 0,0669 | 0,7413 |
| **Total** | **0,1845** | **0,1907** | **-0,0063 ± 0,003 *** | **0,109** | **0,0515** | **0,0578** | **1,1214** |
| **Tabla 3.2b** | | | **Distancias Sinónimas** | | | | |
| Actin | 1,5449 | 1,8325 | -0,2876 ± 0,4127 | 1,318 | 0,5154 | 0,8029 | 1,558 |
| Alpha tubulin | 0,7646 | 0,8385 | -0,0738 ± 0,0821 | 0,713 | 0,3197 | 0,3935 | 1,231 |
| Beta tubulin | 0,6815 | 0,8406 | -0,1591 ± 0,0925 | 0,719 | 0,2798 | 0,4389 | 1,569 |
| Cysteine proteinase | 1,1938 | 2,0138 | -0,8201 ± 0,3746 * | 1,712 | 0,4458 | 1,2658 | 2,839 |
| Dihydrofolate reductase | 1,4495 | 2,2363 | -0,7868 ± 0,4387 | 2,053 | 0,6332 | 1,4201 | 2,243 |
| Paraflagellar rod protein | 0,4967 | 1,0969 | -0,6003 ± 0,1200 ***** | 0,919 | 0,1592 | 0,7595 | 4,771 |
| glycosomal GAP | 0,8050 | 0,9732 | -0,1681 ± 0,1273 | 1,101 | 0,4666 | 0,6347 | 1,36 |
| Hypoxanthine transferase | 1,6870 | ---- | ---- ---- | 1,424 | ---- | ---- | ---- |
| Heat shock protein 70 | 0,7146 | 1,0794 | -0,3648 ± 0,1062 *** | 1,424 | 0,5297 | 0,8945 | 1,689 |
| Heat shock protein 83 | 0,5861 | 1,2088 | -0,6227 ± 0,1605 *** | 1,108 | 0,2424 | 0,8651 | 3,569 |
| Triose phosphate isomerase | 1,5266 | 2,2158 | ---- | ---- | | ---- | ---- |
| Calmodulin | 0,5595 | 1,4086 | -0,8491 ± 0,3255 ** | 1,261 | 0,2058 | 1,0550 | 5,126 |
| Ubiquitin fusion protein | 0,4693 | 1,1056 | -0,6363 ± 0,2111 ** | 1,016 | 0,1896 | 0,8259 | 4,356 |
| CDC2- like 1 | 2,0735 | ---- | ---- ---- | 1,845 | ---- | ---- | ---- |
| CDC2- like 2 | ---- | ---- | ---- ---- | 2,187 | ---- | ---- | ---- |
| Mitochondrial Hsp70 | 0,5506 | ---- | ---- ---- | 2,039 | ---- | ---- | ---- |
| P2- ribosomal protein | ---- | ---- | ---- ---- | ---- | ---- | ---- | ---- |
| Trypanothione reductase | ---- | ---- | ---- ---- | 1,790 | ---- | ---- | ---- |
| Toposisomerase II | 1,9091 | 1,5802 | 0,3289 ± 0,2019 | 1,641 | 0,9851 | 0,6562 | 0,666 |
| **Total** | **0.9551** | **1,2658** | **-0.3107 ± 0,0432 *****| **1,191** | **0.4400** | **0.7507** | **1,7061** |

**Nota-.** $D_{13}$: distancia entre *T. cruzi* y el outgroup; $D_{23}$: distancia entre *T. brucei* y el outgroup. $D_{12}$ distancia entre *T. cruzi* and *T. brucei*; $D_{10}$ y $D_{20}$ son las distancias entre las especies (T. *cruzi* y *T. brucei*) con su ancestro común. st= desvío standard. La línea quebrada significa que la comparasión no es posible debido a que el método para corregir sustituciones múltiples es inaplicable. Los astericos indican la significación estadística de $Z=(D_{23}-D_{13})/st$.

**Tabla 3.3.** Cocientes entre distancias sinónimas y no-sinónimas

| Especies comparadas | Distancia sinónima | Distancia no-sinónima | cociente (DS/DNS) |
|---|---|---|---|
| **Glyceraldehyde 3-P dehydrogenase (glycosomal)** | | | |
| *T. brucei - T vivax* | 1.1641 | 0.0399 | 29.2 |
| *L. mexicana- L. seymoury* | 0.4394 | 0.0639 | 6.8 |
| *L. mexicaca- C.fasciculata* | 0.4528 | 0.0625 | 7.2 |
| *T. cruzi- T. rangeli* | 0.5674 | 0.0660 | 8.6 |
| *T. cruzi- L. mexicaca* | 0.779 | 0.1193 | 6.52 |
| **Phosphoglycerate kinase** | | | |
| *T. brucei- T. congolense* | 2.899 | 0.1222 | 23.7 |
| *Leishmania- C. fasciculata* | 0.367 | 0.0661 | 5.5 |
| **Cystein proteinase** | | | |
| *T. brucei-T. congolense* | 1.641 | 0.2300 | 7.13 |
| *T. cruzi- T. rangeli* | 0.410 | 0.2077 | 1.97 |
| **Trypanothione reductase** | | | |
| *T.brucei-T.congolense* | 2.305 | 0.107 | 21.54 |
| *Leishmania-C. fasciculata* | 0.728 | 0.155 | 4.69 |

## IV.4 SESGOS MUTACIONALES EN GENES DE MAMÍFEROS

### IV.4.1 Consideraciones previas sobre los análisis de bases de datos mutacionales

Antes de pasar a la presentación de los análisis llevados a cabo es necesario plantear y discutir las premisas sobre las cuales los análisis están basados. Como ya fuera planteado en la presentación de objetivos, se intenta determinar el espectro (o espectros) mutacional de los genes de mamíferos para contrastar dicho patrón con la composición de bases sinónima de los genes. La idea conductora es que las mutaciones presentes en las bases de datos representan una muestra al azar del conjunto de mutaciones que surgen en las poblaciones, y por lo tanto el espectro de mutaciones que obtengamos a partir de ellas es un buen estimador del espectro mutacional real. ¿En qué se basa esa presunción? Las sustituciones nucleotídicas, es decir las diferencias de bases entre genes homólogos, son el resultado de dos eventos: la mutación y la fijación. La primera es una alteración en la secuencia de bases de una (y sólo una) molécula de ADN durante la formación de los gametos, muy probablemente durante la replicación. Cuando una nueva mutación es introducida en la población (mutación "de novo") esta presenta una frecuencia de 1/2N (siendo N el tamaño de la población), es decir sólo un individuo, el que recibió el gameto mutante porta la mutación en cuestión. Para que una mutación se "transforme" en una sustitución es necesario que la misma se imponga en la población y en la especie, lo que significa que llegue a ser la única variante o la más frecuente. A este proceso se le llama fijación. Las fuerzas evolutivas que dirigen la fijación son la selección natural direccional positiva (o "darwiniana") y la deriva genética. Sin embargo la amplísima mayoría de las mutaciones nuevas que surgen en las poblaciones naturales son deletéreas (perjudiciales) y los individuos portadores de las mismas tienen chance reducida (o en la mayoría de los casos imposibilidad) de pasar sus genes a la siguiente generación. En consecuencia, en la casi totalidad de los casos, las mutaciones deletéreas desaparecen de la población. A este proceso lo llamamos selección natural purificadora o selección negativa. En cambio, las mutaciones que llegan a fijarse, y por ende llegan a verse como sustituciones, o bien son beneficiosas o no producen ningún cambio sustancial en la función del gen del cual forman parte (mutaciones neutras). De lo anteriormente expuesto surge que el espectro de mutacional no puede ser inferido a partir del patrón de sustituciones puesto que los sesgos de fijación (selección natural) eclipsan el patrón mutacional subyacente, es decir vemos lo que la selección natural nos "permite" ver. Por el contrario, el patrón de mutaciones debe ser inferido analizando las mutaciones tal cual estas surgen en las poblaciones previamente a que hayan sido sometidas al efecto tamizador de la selección natural.

La ventaja de inferir el patrón mutacional a partir de la información contenida en las bases de datos de mutaciones, es que dicha información no adolece del defecto de representar una muestra sesgada por la selección natural. En realidad estas mutaciones todavía no han sido sometidas a la selección. Sin embargo hay ciertos problemas que es necesario tener presente pues la muestra de mutaciones que contienen las bases de datos no son una muestra completamente al azar del espectro total de mutaciones.

Podemos clasificar a estos problemas en tres grandes grupos. En primer lugar no es completamente cierto que las mutaciones que contienen las bases de datos no hayan sido sometidas aun a la selección natural puesto que las mismas son aisladas a partir de pacientes (o fetos abortivos) implicando por tanto que la etapa de selección gamética y de desarrollo embrionario ya ha sido superada. Hay que tener presente que algunas mutaciones, las llamadas letales dominantes, que en general involucran a genes fundamentales, son eliminadas durante esta etapa pues dan lugar a desarrollos abortivos generalmente no detectables. Sin embargo, esta etapa es sorteada (independientemente de la mutación que porten) por los genes codificantes de enzimas monoméricas pues las mutaciones en estos genes son recesivas en la casi totalidad de los casos. Otro tipo de genes que pasan esta etapa de selección natural son aquellos genes que no participan en la formación del fenotipo hasta la etapa de feto (etapa en la cual ya son detectables) o más tardíamente. Este primer problema nos introduce entonces un sesgo en el tipo de genes sobre los cuales hay bases de datos, pero no introduce sesgos en el tipo de mutaciones que contienen las bases de datos.

El segundo problema es el del sesgo de detectabilidad. Para que la información de secuencia de una mutación pueda llegar a una base de datos es necesario que la mutación llame la atención clínica pues se determina la secuencia en aquellos individuos que se sabe a priori son mutantes. Muy pocas de las mutaciones presentes en las bases de datos son del tipo sinónimo o no deletéreas. Esto significa que las mutaciones que usaremos para inferir el patrón mutacional constituyen un sub-conjunto de las mutaciones posibles: el de las mutaciones deletéreas.

Uno puede preguntarse si este tipo de filtro introduce algún sesgo consistente en nuestra muestra de forma tal que algunas mutaciones nucleotídicas aparezcan sobre-representadas y otras sub-representadas. Podemos afirmar que el sesgo mayor se encuentra en el tipo de mutaciones que ocurren en la tercera posición de los codones. ¿A que se debe este sesgo?. En esta posición del codón la amplísima mayoría de mutaciones detectables, por ser potencialmente deletéreas, son las transversiones en los duetos. Esto se debe a que sólo las transversiones producen cambios de aminoácidos en los codones de doble degeneramiento. Por su parte, las transiciones en la tercera posición de los codones son sinónimas, tanto en los codones de doble como cuádruple degeramiento. Los únicos cambios posibles de aminoácido causados por transiciones en la tercera posición del codón ocurren en los codones para metionina, isoleucina y triptofano. De lo anteriormente expuesto se deduce que toda la información referente a la tercera posición de los codones debe ser excluida en este análisis. En caso contrario estaríamos usando una muestra de mutaciones donde las transversiones se encuentren sobre-representadas.

Analicemos ahora en qué medida el sesgo de detectabilidad nos puede afectar la primera y segunda posición de los codones. En estas posiciones la amplísima mayoría de las mutaciones producen cambio de aminoácido. Concretamente de un total de 354 mutaciones individuales posibles (59*3*2), solamente 4 (1.13%) de ellas no implican cambio de aminoácido, C<->T en primera posición de los codones CTA<->TTA, CTG<-> TTG que codifican para leucina y C<->A en la

primera posición de los codones CGA<->AGA, CGG<-> AGG que codifican para arginina. En consecuencia casi el 99% de las mutaciones en las dos primeras posiciones del codón son potencialmente detectables por producir alteraciones en la naturaleza codificante del codón al que afectan. Podría ocurrir sin embargo que algunos tipos de cambios nucleotídicos estuvieran sobre- o sub-representados como consecuencia de que no todos los tipos de cambios aminoacídicos pueden producir el mismo efecto deletéreo. En efecto, los cambios entre aminoácidos bioquímicamente disimiles tienen mayor chance de estar representados en las bases de datos que aquellos cambios de tipo conservativos entre aminoácidos similares. Es preciso reconocer que este es el aspecto más crítico sobre el uso de las bases de datos para inferir los patrones mutacionales. Dada su importancia, el mismo ya ha sido objeto de estudio por otros autores. Los estudios realizados con anterioridad (Cooper & Krawezak, 1990; Krawezak & Cooper; 1996) en la base de datos de mutaciones del gen que codifica el factor IX de coagulación muestran que la frecuencia de determinado tipo de cambio aminoácido no se correlaciona con la distancia química entre los aminoácidos intercambiados. Además las frecuencias de cada tipo de cambio nucleotídico se correlaciona con el espectro de mutaciones predicho en base a los estudios "in vitro" sobre errores de incorporación de nucleótidos durante la replicación de las subunidades $\alpha$ y ß de la ADN polimerasa de vertebrados. Otro aspecto a tener en cuenta es que la mayoría de los cambios entre aminoacídicos bioquímicamente similares son atribuibles a transversiones en la tercera posición de los codones las cuales no son consideradas en este estudio. Esto indicaría que en nuestro estudio en particular, el problema de la dectetabilidad diferencial dependiente de la naturaleza química del cambio tenga incluso un efecto menor. Además, si la detectabilidad representara realmente un problema, este nos afectaría sólo parcialmente pues lo haría en forma simétrica. Expliquemos el punto. Cabe recordar que nuestro objetivo es estimar las entradas O*ij* de la matriz empírica **O**, entonces si una entrada O*ij* (por ejemplo G->A) estuviera afectada por la detectabilidad, también lo estaría la entrada O*ji* (A->G). Esto es debido que si un tipo particular de cambio fuera muy detectable por ser drástico, también lo sería el cambio inverso por ser igualmente drástico. Un ejemplo de cambio drástico que podría dar lugar a mutaciones con alta probabilidad de ser detectables sería: Glu->Lys, es decir ácido glutámico (GAR) hacia la lisina (AAR), el único aminoácido básico al que puede pasar cambiando sólo un nucleótido. Esto nos incrementaría la entrada de Oij correspondiente al cambio G->A. Pero si este cambio es muy detectable, también lo es el cambio inverso, desde lisina hacia ácido glutámico, lo que nos llevaría a incrementar la entrada A->G. Este aspecto de simetricidad, es de fundamental importancia para el tipo de estimación que se pretende en nuestro estudio, pues en caso de existir un sesgo debido a la detectabilidad, el mismo no nos afectaría la relación GC<->AT, que sin lugar a dudas es la estimación de mayor importancia, pues afecta directamente el contenido en G+C.

El último problema a considerar es el de la identidad por ascendencia. Algunas mutaciones están representadas más de una vez en las bases de datos. Las razones de esta repetición son, o bien que la misma variante aparezca más de una vez debido a eventos de mutación recurrentes y

por tanto independientes, o que la variante haya sido aislada en pacientes que compartan la mutación porque la mutación en cuestión surgió en un ancestro común a ambos pacientes. En el primer tipo de repetición debemos contar cada aparición como una mutación distinta, pues eso es lo que son. En el segundo tipo de situación, la mutación debe ser contabilizada una sola vez. Sin embargo esta distinción no siempre es posible, o bien porque no se sabe o porque la información no está disponible en la base de datos. En los resultados que se presentan en la siguiente sección se calculan las matrices O de dos maneras distintas: teniendo en cuenta la repetición (pero excluyéndola cuando se tiene certeza de que no son mutaciones independientes) y contado cada mutación una sola vez independientemente del número de veces que aparezca en la base de datos.

**IV.4.2 Análisis de los espectros mutacionales en genes humanos**

**IV.4.2.1 Bases de datos de genes pobres en $GC_3$**

**IV.4.2.1.a- Factor anti-hemofílico B (factor IX o Christmas Factor)**

Este gen se encuentra localizado en el cromosoma X, siendo sus mutaciones recesivas (pérdida de función) y de herencia ligada al sexo. Al ser recesiva y ligada al sexo, los varones hemicigotas y las mujeres homocigotas presentan la afección. El compilado de mutaciones de este gen constituye la base de datos de mutaciones de la línea germinal de mayor tamaño. Existe abundante y confiable información para cada mutación. Es particularmente resaltable que no existe prácticamente ningún tipo ambigüedad sobre cuáles mutaciones son idénticas por ascendencia y cuáles son independientes.

El contenido en $GC_3$ del gen es 0.336; y el número de mutaciones puntuales en la base de datos (en la primera y segunda posición del codón), incluyendo aquellas que ocurren en los sitios CpG es de 1243, siendo 604 las mutaciones que ocurren en estos sitios.

A continuación se muestran las matrices que contienen el conteo de la ocurrencia de cada tipo de mutación:

Matriz de mutaciones número 1

Se incluyen las múltiples apariciones de cada mutación y los sitios CpG.

| | a | T | C | A | G |
|---|---|---|---|---|---|
| desde | | | | | |
| T | | 210 | 114 | 40 | 28 |
| C | | 351 | 158 | 39 | 23 |
| A | | 24 | 21 | 303 | 45 |
| G | | 84 | 44 | 430 | 253 |

Matriz de mutaciones número 2
Se incluyen las múltiples apariciones de cada mutación y excluyen los sitios CpG.

| desde \ a | T | C | A | G |
|---|---|---|---|---|
| T | 210 | 114 | 40 | 28 |
| C | 50 | 146 | 31 | 20 |
| A | 24 | 20 | 303 | 45 |
| G | 67 | 42 | 158 | 234 |

Matriz de mutaciones número 3
Se excluyen las múltiples apariciones de cada mutación y se incluyen los sitios CpG.

| desde \ a | T | C | A | G |
|---|---|---|---|---|
| T | 210 | 46 | 29 | 25 |
| C | 41 | 158 | 30 | 20 |
| A | 17 | 16 | 303 | 31 |
| G | 50 | 31 | 93 | 253 |

Matriz de mutaciones número 4
Se excluyen las múltiples apariciones de cada mutación y de los sitios CpG.

| desde \ a | T | C | A | G |
|---|---|---|---|---|
| T | 210 | 46 | 29 | 25 |
| C | 30 | 146 | 25 | 18 |
| A | 17 | 16 | 303 | 31 |
| G | 46 | 28 | 83 | 234 |

A continuación se presenta, para cada una de las matrices de arriba, la composición de bases esperada en el equilibrio si esta fuera selectivamente neutra y por tanto gobernada por el patrón mutacional

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| matriz 1 | 0.2920 | 0.1582 | 0.4375 | 0.1124 | 0.2705 |
| matriz 2 | 0.1980 | 0.2250 | 0.4248 | 0.1522 | 0.3772 |
| matriz 3 | 0.2135 | 0.1732 | 0.4447 | 0.1686 | 0.3418 |
| matriz 4 | 0.2042 | 0.1852 | 0.4393 | 0.1713 | 0.3565 |

**IV.4.2.1.b- Factor anti-hemofílico VIII (factor A de coagulación)**

Esta afección genética (que probablemente es la mejor conocida de todas) es muy similar a la hemofilia B. Al igual que la hemofilia B, es recesiva y se transmite ligada al sexo siendo las variantes mutantes igualmente severas. La información contenida en la base de datos es sin embargo de menor calidad que en la hemofilia B. Esto se debe a que dada la alta frecuencia de este tipo de hemofilia en las poblaciones resulta muchas veces imposible determinar que repeticiones de una mutación dada son independientes. La base de datos contiene 537 mutaciones puntuales (en las posiciones 1 y 2 de los codones) de las cuales 304 se ubican en sitios CpG. El contenido en $GC_3$ del gen es: 0.3888

Matriz 1: Se incluyen las múltiples apariciones de cada mutación y los sitios CpG.

| desde | a T | C | A | G |
|---|---|---|---|---|
| T | 1096 | 20 | 4 | 8 |
| C | 208 | 1024 | 38 | 12 |
| A | 21 | 9 | 1573 | 52 |
| G | 31 | 9 | 125 | 1009 |

Matriz 2: Se incluyen las múltiples apariciones de cada mutación y excluyen los sitios CpG.

| desde | a T | C | A | G |
|---|---|---|---|---|
| T | 1096 | 20 | 4 | 8 |
| C | 27 | 974 | 12 | 7 |
| A | 21 | 9 | 1573 | 52 |
| G | 16 | 7 | 50 | 953 |

Matriz 3: Se excluyen las múltiples apariciones de cada mutación y se incluyen los sitios CpG.

| desde | a T | C | A | G |
|---|---|---|---|---|
| T | 1096 | 18 | 4 | 6 |
| C | 47 | 1024 | 13 | 11 |
| A | 9 | 8 | 1573 | 32 |
| G | 24 | 8 | 48 | 1009 |

Matriz 4 : Se excluyen las múltiples apariciones de cada mutación y de los sitios CpG.

| desde | a T | C | A | G |
|---|---|---|---|---|
| T | 1096 | 18 | 4 | 6 |
| C | 19 | 974 | 10 | 7 |
| A | 9 | 8 | 1573 | 32 |
| G | 16 | 6 | 32 | 953 |

Composición de bases esperada en el equilibrio si la secuencia fuera selectivamente neutra y por tanto gobernada por el patrón mutacional

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| matriz 1 | 0.5623 | 0.0608 | 0.2815 | 0.0954 | 0.1561 |
| matriz 2 | 0.3902 | 0.2070 | 0.2386 | 0.1642 | 0.3712 |
| matriz 3 | 0.4287 | 0.1433 | 0.2974 | 0.1306 | 0.2739 |
| matriz 4 | 0.3328 | 0.2166 | 0.2847 | 0.1659 | 0.3825 |

### IV.4.2.2 Bases de datos de genes con valores medios de $GC_3$

#### IV.4.2.2.a-Locus de la Fenilalanín-hidroxilasa

Las mutaciones en este gen pueden lar lugar a la afección metabólica conocida como fenil-cetonuria que se caracteriza entre otras cosas por retardo mental severo. El gen se hereda en forma autosómica y los alelos mutantes pueden dar lugar al desarrollo de la enfermedad (si los individuos no son tratados) cuando se encuentran en homocigosis, lo que significa que son recesivos. Dicho modo de herencia es compatible con que se trate mutaciones de pérdida de función en genes codificantes de enzimas. Uno de los inconvenientes de esta base de datos es que resulta muy difícil (casi imposible) distinguir entre repeticiones debidas a identidad por ascendencia de aquellas debidas a eventos mutacionales recurrentes (independencia). Por esta razón únicamente se consideran las matrices sin repetición. El contendido en CG3 del gen es: 0.519

Matriz 1: se incluyen los sitios CpG.

| | a | T | C | A | G |
|---|---|---|---|---|---|
| desde | | | | | |
| T | | 224 | 23 | 3 | 9 |
| C | | 36 | 202 | 13 | 8 |
| A | | 7 | 8 | 277 | 20 |
| G | | 9 | 11 | 26 | 199 |

Matriz 2: se excluyen las los sitios CpG.

| | a | T | C | A | G |
|---|---|---|---|---|---|
| desde | | | | | |
| T | | 224 | 23 | 3 | 9 |
| C | | 25 | 184 | 12 | 7 |
| A | | 7 | 8 | 277 | 20 |
| G | | 9 | 8 | 15 | 179 |

Composición de bases esperada en el equilibrio para una secuencia selectivamente neutra.

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Matriz 1 | 0.3182 | 0.1830 | 0.3180 | 0.1852 | 0.3682 |
| Matriz 2 | 0.2927 | 0.2015 | 0.2828 | 0.2229 | 0.4240 |

### IV.4.2.3 Bases de datos de genes con valores altos de $GC_3$

#### IV.4.2.3.a- Receptor de andrógeno

Las mutaciones en este gen producen alteraciones con distinto grado de severidad en la diferenciación sexual masculina, dando lugar (en los casos más extremos) a fenotipos genitales de apariencia femenina. También se han aislado varias mutaciones somáticas asociadas a desarrollos tumorales metastásicos de próstata.

La información sobre la historia familiar de cada mutación es bastante precisa, por los que en la base de datos, es posible distinguir entre aquellas mutaciones repetidas que están relacionadas por ancestro común y aquellas que corresponden a eventos de mutación recurrente. Es

importante aclarar que el gen codificante del receptor de andrógeno contiene 6 exones, el primero de los cuales es muy largo comprendiendo aproximadamente el 60% del tamaño total de la secuencia codificante. Esta región del gen fue excluida del análisis por las siguientes razones: sólo un 3% de las mutaciones afectan al exon 1 y en segundo lugar posee una composición aminoacídica muy sesgada, conteniendo varias repeticiones de poli-glicina, poli-prolina, poli-leucina, poli-alanina y poli-asparagina, cada una de estas repeticiones tiene unos 20 aminoácidos de largo. Por lo que, dado el tamaño del exón, la escasez de mutaciones dentro del mismo y su composición aminoacídica extremadamente sesgada, su inclusión generaría un gran desbalance de las matrices mutacionales (particularmente en las entradas de la diagonal) lo que conduciría a una estimación errónea del patrón mutacional.

El contenido en $GC_3$ del gen es 0.64; y el número de mutaciones puntuales (de la primera y segunda posición del codón) contenido en la base de datos, incluyendo aquellas que ocurren en los sitios CpG es de 216, de las cuales 6 fueron excluidas por pertenecer al primer exón. El número de mutaciones ubicadas en sitios CpG es 97.

Matriz 1: Se incluyen las múltiples apariciones de cada mutación y los sitios CpG.

| | a | T | C | A | G |
|---|---|---|---|---|---|
| desde | | | | | |
| T | | 193 | 17 | 4 | 8 |
| C | | 41 | 168 | 5 | 11 |
| A | | 2 | 3 | 233 | 19 |
| G | | 20 | 7 | 73 | 172 |

Matriz 2 : Se incluyen las múltiples apariciones de cada mutación y excluyen los sitios CpG.

| | a | T | C | A | G |
|---|---|---|---|---|---|
| desde | | | | | |
| T | | 193 | 17 | 4 | 8 |
| C | | 12 | 153 | 5 | 10 |
| A | | 2 | 3 | 233 | 19 |
| G | | 11 | 3 | 19 | 148 |

Matriz 3:Se excluyen las múltiples apariciones de cada mutación y se incluyen los sitios CpG.

| desde \ a | T | C | A | G |
|---|---|---|---|---|
| T | 193 | 14 | 4 | 7 |
| C | 19 | 168 | 5 | 6 |
| A | 2 | 3 | 233 | 14 |
| G | 14 | 5 | 28 | 172 |

Matriz 4:Se excluyen las múltiples apariciones de cada mutación y los sitios CpG.

| desde \ a | T | C | A | G |
|---|---|---|---|---|
| T | 193 | 14 | 4 | 7 |
| C | 11 | 153 | 5 | 5 |
| A | 2 | 3 | 233 | 14 |
| G | 10 | 2 | 15 | 148 |

Composición de bases esperada en el equilibrio para una secuencia selectivamente neutra.

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Matriz 1 | 0.29 | 0.10 | 0.51 | 0.10 | 0.2058 |
| Matriz 2 | 0.225 | 0.17 | 0.37 | 0.23 | 0.4028 |
| Matriz 3 | 0.28 | 0.145 | 0.43 | 0.14 | 0.2858 |
| Matriz 4 | 0.23 | 0.19 | 0.36 | 0.22 | 0.4026 |

**IV.4.2.3.b-Gen codificante de la proteína P53**

La proteína P53 participa en la regulación del ciclo celular como regulador (inhibidor) del pasaje a la fase de duplicación de ADN (fase de desencadenamiento del ciclo). Esto significa que la proteína posee propiedades antiproliferativas. Las mutaciones en este gen (que en la gran mayoría de los casos surgen en células somáticas) pueden desinhibir la proliferación celular con alta probabilidad de conducir al desarrollo de varios tipos de tumores malignos (Finlay et al, 1989). Existen mutaciones de P53 nulas (pérdida de función), mutaciones dominantes negativas (la copia mutada del gen inactiva el producto normal de la otra copia) y mutaciones dominantes positivas (ganancia de función) (Milner & Medcalf, 1991). Además de las mutaciones que surgen somaticamente, se han aislado varias mutaciones heredables (las cuales se originan en células de la línea germinal) en pacientes con el síndrome de Li-Fraumeni. Los individuos portadores de esta afección desarrollan varios tipos de tumores malignos a edad muy temprana.

La base de datos de mutaciones somáticas de P53 usada en este estudio contiene más de 6000 mutaciones, mientras que la última versión (que no está disponible públicamente) de esta base de datos contiene más de 8000 mutaciones. Se analizan 4423 mutaciones puntuales (en la primera y segunda posición del codón), de las cuales 1776 ocurren en sitios CpG. La base de datos de mutaciones de la línea germinal contiene 600 mutaciones pertenecientes a 122 genealogías independientes. Un aspecto interesante de la base de datos somática de P53 es que contiene 204

mutaciones sinónimas de las cuales 193 son en la tercera posición del codón. Lo verdaderamente significativo de este conjunto de mutaciones es que las mismas fueron aisladas casualmente como mutaciones acompañantes de otras que sí son deletéreas. Es claro que este grupo de mutaciones no adolece de ningún tipo de sesgo de detección por representar una muestra completamente al azar. Esto nos brinda la oportunidad de comparar las estimaciones basadas en uno y otro grupo de mutaciones con el fin de testar el problema del sesgo de detección anteriormente discutido. El conteo de mutaciones considerando una sola vez cada tipo de mutación no se llevó a cabo para P53 por no ser necesario. Las mutaciones somáticas son independientes pues no se heredan, y en el caso de las mutaciones de la línea germinal sólo se consideraron las 122 mutaciones (de un total de 600) pertenecientes a genealogías que sabemos son independientes.

Matrices en las que se incluyó los sitios CpG en el conteo de mutaciones

Mutaciones somáticas (en posiciones 1 y 2):

| a / desde | T | C | A | G |
|---|---|---|---|---|
| T | 147 | 148 | 121 | 104 |
| C | 1033 | 234 | 107 | 136 |
| A | 139 | 73 | 215 | 435 |
| G | 657 | 226 | 1244 | 190 |

Mutaciones en la línea germinal (en posiciones 1 y 2):

| a / desde | T | C | A | G |
|---|---|---|---|---|
| T | 147 | 4 | 3 | 1 |
| C | 27 | 234 | 3 | 2 |
| A | 5 | 1 | 215 | 9 |
| G | 7 | 3 | 33 | 190 |

Mutaciones sinónimas en la tercera posición de los codones (somáticas)

| a / desde | T | C | A | G |
|---|---|---|---|---|
| T | 86 | 22 | 2 | 3 |
| C | 64 | 131 | 8 | 3 |
| A | 1 | 2 | 60 | 23 |
| G | 10 | 7 | 48 | 100 |

Matrices en las que se excluyó los sitios CpG en el conteo de mutaciones.

Mutaciones somáticas

| a / desde | T | C | A | G |
|---|---|---|---|---|
| T | 147 | 148 | 121 | 104 |
| C | 363 | 204 | 78 | 102 |
| A | 139 | 73 | 215 | 435 |
| G | 449 | 137 | 498 | 157 |

Mutaciones de la línea germinal

| a / desde | T | C | A | G |
|---|---|---|---|---|
| T | 147 | 4 | 3 | 1 |
| C | 6 | 204 | 2 | 1 |
| A | 5 | 1 | 215 | 9 |
| G | 5 | 1 | 6 | 157 |

Mutaciones sinónimas en la tercera posición del codón

| a / desde | T | C | A | G |
|---|---|---|---|---|
| T | 86 | 22 | 2 | 3 |
| C | 59 | 118 | 7 | 3 |
| A | 1 | 2 | 60 | 23 |
| G | 7 | 7 | 42 | 90 |

Composición de bases esperada en el equilibrio para una secuencia selectivamente neutra.

incluyendo los sitios CpG

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Somático | 0.4335 | 0.1227 | 0.3472 | 0.0965 | 0.2192 |
| Línea germ. | 0.4452 | 0.1106 | 0.3601 | 0.0842 | 0.1947 |
| Sinónimas | 0.3897 | 0.2082 | 0.2352 | 0.1669 | 0.3752 |

excluyendo los sitios CpG

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Somático | 0.3745 | 0.2202 | 0.2715 | 0.1339 | 0.3540 |
| Línea germ. | 0.3390 | 0.2583 | 0.2301 | 0.1727 | 0.4310 |
| Sinónimas | 0.3809 | 0.2038 | 0.2386 | 0.1767 | 0.3804 |

### IV.4.2.4 Bases de datos de genes con valores muy altos de $GC_3$

#### IV.4.2.4.a- Glucosa-6-fosfato deshidrogenasa

Las mutaciones en este gen causan anemia hemolítica aguda (conocida como favismo) cuando el individuo afectado por la mutación ingiere semillas o inhala polen de *Vicia faba*. Esta anemia, se hereda en forma recesiva y ligada al sexo por lo que los varones hemicigotas y las mujeres homocigotas presentan la afección.

La base de datos sobre mutaciones de este gen contiene 118 entradas correspondientes a mutaciones puntuales que ocurren en la primera y segunda posición del codón. De estas mutaciones 48 están ubicadas en sitios CpG. La información concerniente a la independencia de origen en las mutaciones repetidas es en general buena, sin embargo existen 3 grupos de mutaciones repetidas donde la independencia no es clara.

Matriz de mutaciones número1

Se incluyen las múltiples apariciones de cada mutación y los sitios CpG.

| | a T | C | A | G |
|---|---|---|---|---|
| desde | | | | |
| T | 235 | 11 | 0 | 0 |
| C | 26 | 225 | 1 | 3 |
| A | 4 | 1 | 313 | 13 |
| G | 14 | 9 | 36 | 257 |

Matriz de mutaciones número 2

Se incluyen las múltiples apariciones de cada mutación y excluyen los sitios CpG.

| | a T | C | A | G |
|---|---|---|---|---|
| desde | | | | |
| T | 235 | 11 | 0 | 0 |
| C | 13 | 192 | 1 | 2 |
| A | 4 | 1 | 313 | 13 |
| G | 9 | 5 | 11 | 183 |

Matriz de mutaciones número 3

Se excluyen las múltiples apariciones de cada mutación y se incluyen los sitios CpG.

| | a T | C | A | G |
|---|---|---|---|---|
| desde | | | | |
| T | 235 | 9 | 0 | 0 |
| C | 16 | 225 | 1 | 3 |
| A | 3 | 1 | 313 | 11 |
| G | 8 | 5 | 24 | 257 |

Matriz de mutaciones número 4
Se excluyen las múltiples apariciones de cada mutación y de los sitios CpG.

| | a T | C | A | G |
|---|---|---|---|---|
| desde | | | | |
| T | 235 | 9 | 0 | 0 |
| C | 7 | 192 | 1 | 2 |
| A | 3 | 1 | 313 | 11 |
| G | 6 | 1 | 9 | 183 |

Composición de bases esperada en el equilibrio para una secuencia selectivamente neutra.

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Matriz 1 | 0.64 | 0.24 | 0.092 | 0.03 | 0.266 |
| Matriz 2 | 0.54 | 0.325 | 0.08 | 0.05 | 0.37 |
| Matriz 3 | 0.55 | 0.256 | 0.13 | 0.056 | 0.31 |
| Matriz 4 | 0.435 | 0.341 | 0.133 | 0.094 | 0.43 |

**IV.4.2.4.b- L1CAM (molécula de adhesión neuronal)**

L1CAM es una proteína que tiene una función importante en el desarrollo del sistema nervioso. El gen que codifica a L1CAM se encuentra localizado en el cromosoma X, por lo que las alteraciones genéticas en esta proteína se transmiten ligadas al sexo. Las mutaciones en este gen conducen a varios tipos de síndromes de retardo mental, incluyendo la hidrocefalia ligada al X entre otros (Wong et al, 1995). Este gen posee un contenido de $GC_3$ de 0.77. El número de mutaciones analizadas es de 41, de las cuales 12 ocurren en sitios CpG. La base de datos contiene solamente una repetición (ubicada en un sitio CpG) de la que se tiene certeza pertenece a una genealogía independiente; por esta razón no se realizaron conteos de mutaciones considerando las apariciones sin repetición de cada mutación.

Matriz 1: se incluyen las mutaciones ubicadas en los sitios CpG.

| | a T | C | A | G |
|---|---|---|---|---|
| desde | | | | |
| T | 498 | 5 | 0 | 1 |
| C | 9 | 610 | 1 | 1 |
| A | 0 | 0 | 741 | 4 |
| G | 1 | 1 | 11 | 665 |

Matriz 2: se excluyen las mutaciones ubicadas en los sitios CpG.

| | a T | C | A | G |
|---|---|---|---|---|
| desde | | | | |
| T | 498 | 5 | 0 | 1 |
| C | 4 | 526 | 1 | 0 |
| A | 0 | 0 | 741 | 4 |
| G | 1 | 1 | 5 | 548 |

Composición de bases esperada en el equilibrio para una secuencia selectivamente neutra.

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Incluyendo sitios CpG | 0.13 | 0.09 | 0.59 | 0.185 | 0.275 |
| Excluyendo sitios CpG | 0.175 | 0.21 | 0.42 | 0.20 | 0.41 |

**IV.4.3 Patrón de sustituciones nucleotídicas en seudogenes de mamíferos**

Se estudian 3 grupos de seudogenes derivados de genes ricos o muy ricos en $GC_3$. El primer grupo comprende a genes del "cluster" de la globina β. El segundo grupo está integrado por un único seudogén derivado del receptor de la interleukina 8. Y el tercer grupo de seudogenes lo integran los del "cluster" de la globina α.

**IV.4.3.a- Seudogenes del "cluster" de la globina β**

El "cluster" de la globina β constituye un grupo de genes de estructura, organización y regulación compleja, incluyendo varios genes funcionales y seudogenes. La historia evolutiva de este grupo de genes (particularmente en los mamíferos) también es compleja, mostrando que las distintas variantes de la globina β se han generado por duplicación, mezcla de regiones codificantes de genes preexistentes mediante conversión génica en segmentos, y varios genes han sufrido inactivaciones y deleciones en distintos linajes evolutivos (ver Hardison & Miller, 1993). Este proceso de duplicaciones, deleciones e inactivaciones ha generado varios seudogenes en distintos grupos de mamíferos. Sin embargo nos es posible incluir a todos estos seudogenes en el presente análisis puesto que el "status" de seudogén de algunos de ellos (por ej. la globina δ humana) es dudoso, además en muchos de estos seudogenes la inactivación parece haber ocurrido bastante tiempo después de la duplicación. Usando un criterio conservador se han seleccionado únicamente aquellos seudogenes sobre los que existe consenso sobre su carácter de seudogén. Estos son la globina $\Psi\eta$ humana (locus HSBGLOP, ACC. X02133), las globinas $\Psi\beta^X$,y $\Psi\beta^Z$ de cabra (locus GOTHBBPS1 y CHBGL5, ACC. J00047y V00154) y la suedo-delta globina de conejo ($\Psi\delta$, locus RABBGLOB, ACC. M18818 X07786). Los genes funcionales usados para inferir la dirección de las sustituciones fueron los genes de la globina beta humanos, de cabra y conejo (loci HSBGLOP y RABBGLOB). El contenido en $GC_3$ de las copias funcionales de los genes varía entre 0.64 en conejo a 0.67 en cabra, por lo que podemos clasificar a estos genes como ricos en $GC_3$.

Los patrones de sustituciones inferidos para cada un de estos seudogenes son:

**seudo globina $\Psi\eta$ humana**

incluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 104 | 10 | 7 | 4 |
| C | 13 | 103 | 5 | 5 |
| A | 3 | 2 | 87 | 8 |
| G | 8 | 7 | 23 | 128 |

excluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 104 | 10 | 7 | 4 |
| C | 13 | 105 | 5 | 5 |
| A | 3 | 2 | 87 | 8 |
| G | 8 | 7 | 24 | 134 |

**globina Ψδ de conejo**

incluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 105 | 6 | 2 | 3 |
| C | 11 | 107 | 5 | 2 |
| A | 2 | 4 | 90 | 4 |
| G | 4 | 5 | 12 | 137 |

excluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 104 | 6 | 2 | 3 |
| C | 9 | 104 | 5 | 2 |
| A | 2 | 4 | 90 | 4 |
| G | 4 | 5 | 11 | 134 |

**globinas $\Psi\beta^X$,y $\Psi\beta^z$ de cabra**

incluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 208 | 4 | 3 | 0 |
| C | 23 | 204 | 5 | 4 |
| A | 3 | 3 | 165 | 6 |
| G | 5 | 7 | 36 | 270 |

excluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 206 | 4 | 3 | 0 |
| C | 20 | 194 | 3 | 4 |
| A | 3 | 3 | 163 | 6 |
| G | 5 | 7 | 34 | 262 |

Valores del contenido de bases esperado en el equilibrio si la secuencia fuera selectivamente neutra y gobernada por el patrón de sustituciones derivados de las matrices presentadas arriba:

incluyendo sitios CpG

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| **Ψη de humano** | 0.2405 | 0.1883 | 0.3888 | 0.1824 | 0.3706 |
| **Ψδ de conejo** | 0.3252 | 0.2221 | 0.2830 | 0.1697 | 0.3918 |
| **$\Psi\beta^X$, y $\Psi\beta^z$ de cabra** | 0.5526 | 0.1239 | 0.2554 | 0.0681 | 0.1920 |

excluyendo sitios CpG

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| **$\Psi\eta$ de humano** | 0.2370 | 0.1890 | 0.3889 | 0.1851 | 0.3741 |
| **$\Psi\delta$ de conejo** | 0.3037 | 0.2382 | 0.2842 | 0.1740 | 0.4122 |
| **$\Psi\beta^X$,y $\Psi\beta^z$ de cabra** | 0.5356 | 0.1324 | 0.2570 | 0.0750 | 0.2074 |

**IV.4.3.b- Seudogén humano derivado del receptor de la interleukina 8 (receptor de baja afinidad).**

Este seudogén es derivado de un gen funcional ubicado en la el cromosoma 2, banda **q35.** De acuerdo con los estudios de hibridación "in situ" el seudogén también se localiza en la región **2q35** (Morris et al, 1992). El gen funcional puede ser clasificado en el límite entre gen rico en $GC_3$ y muy rico en $GC_3$, pues posee valores de $GC_3$ de 0.71 en el hombre y de 0.81 en conejo. Los genes funcionales usados para inferir cuales sustituciones tuvieron lugar en el seudogén fueron: hombre (locus HUMIL8R), conejo (locus RABIL8RSB) y *Macaca mulatta* (locus MMMCIL8RB).

Matrices de sustituciones nucleotídicas

Incluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 277 | 9 | 2 | 3 |
| C | 20 | 321 | 8 | 4 |
| A | 4 | 4 | 207 | 7 |
| G | 9 | 2 | 23 | 241 |

Excluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 277 | 9 | 2 | 3 |
| C | 12 | 295 | 8 | 4 |
| A | 4 | 4 | 207 | 7 |
| G | 8 | 1 | 18 | 214 |

Valores del contenido de bases esperado en el equilibrio si la secuencia fuera selectivamente neutra y gobernada por el patrón de sustituciones derivados de las matrices presentadas arriba:

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Incluyendo sitios CpG | 0.4283 | 0.1993 | 0.2597 | 0.1127 | 0.3120 |
| Excluyendo sitios CpG | 0.3779 | 0.2226 | 0.2706 | 0.1289 | 0.3515 |

**IV.4.3.c Seudogenes del "cluster" del la globina $\alpha$.**

Similar a lo ya descripto para el "cluster" de la globina $\beta$, la historia del grupo de genes codificantes de las distintas variantes de la globina $\alpha$ es compleja, signada por duplicaciones,

eventos de conversión génica e inactivaciones (Hardison & Miller, 1993). En el "cluster" de la globina α encontramos varios seudogenes. Por ejemplo en humanos hay 5 seudogenes descriptos. Sin embargo no todos los seudogenes son apropiados para nuestro análisis, ya sea porque su divergencia es muy reciente, por lo que son virtualmente idénticos a las copias funcionales (excepto por codon de terminación que los convierte en no-funcionales) o porque son extremadamente divergentes. Se eligieron 3 grupos de seudogenes. El primer grupo está integrado únicamente por el seudogén humano de la globina α (Ψα1, locus HUMHBA4, ACC. J00153). Los genes funcionales usados para inferir la dirección de las sustituciones fueron las copias de globina α humana (locus HUMHBA4, ACC. J00153), de conejo (locus OCATGL1, ACC. X04751), de caballo (locus ECPZA2GL, ACC. X07053) y de cabra (locus GOTHBAI, ACC. J00043). El segundo grupo de seudogenes está integrado por las tres copias inactivas de la globina Z de conejo (ζ1,locus RABHBZ1; ζ2, locus RABHBZ2 y ζ3, locus RABHBZ3). Los tres seudogenes de globina Z de conejo se generaron muy tempranamente por duplicación a partir de un seudogén ancestral y han evolucionado en forma independiente desde entonces. Esto nos permite agrupar la información proveniente de estos tres seudogenes en una única matriz de sustituciones pues no son linajes filogeneticamente correlacionados. Las copias funcionales de la globina Z usadas para inferir la dirección de las sustituciones fueron: la humana (locus HUMHBA1, ACC. J00181), la de caballo (locus ECZGL1, ACC. X07051) y la de cabra (locus CHAGLZ1, ACC. X04726). El último grupo de seudogenes de globina tipo α usados, está compuesto por dos seudogenes de la globina Z, el humano (Ψα2, locus HSAGL2, ACC. ) y el de caballo (Ψζ, locus ECPZA2GL, ACC. X07053). A diferencia de los otros seudogenes mencionados arriba, estos dos se encuentran localizados dentro de islas CpG, por lo que su patrón de sustituciones puede ser diferente al encontrarse dentro de regiones que presentan una organización más laxa de la cromatina y no existir (o es muy baja) metilación de la citosina en los dinucleóticos CpG (Bird, 1986). Los genes funcionales usados para inferir la dirección de las sustituciones son los mismos a los usados con los seudogenes de globina Z de conejo. Un inconveniente en relación a estos dos seudogenes, es que su divergencia en relación a las copias funcionales es bastante alta. Esto implica que muchos sitios pueden haber sufrido más de una única sustitución, por lo que los resultados basados en estos seudogenes presenta un menor margen de confianza que aquellos basados en los de los seudogenes anteriores.

Por último es importante recalcar que todos los genes funcionales de globina tipo α exhiben valores de $GC_3$ muy altos los cuales varían entre 0.89 y 0.97.

Matrices de sustituciones
**seudogén Ψα1, humano**
Incluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 67 | 3 | 2 | 2 |
| C | 20 | 137 | 12 | 7 |
| A | 1 | 1 | 69 | 8 |
| G | 2 | 4 | 14 | 110 |

Excluyendo sitios CpG.

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 67 | 3 | 2 | 2 |
| C | 10 | 105 | 10 | 3 |
| A | 1 | 1 | 69 | 8 |
| T | 1 | 4 | 8 | 87 |

Valores del contenido de bases esperados en el equilibrio si la secuencia fuera selectivamente neutra y gobernada por el patrón de sustituciones derivados de las matrices presentadas arriba

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Incluyendo sitios CpG | 0.2353 | 0.0943 | 0.3696 | 0.3008 | 0.3951 |
| Excluyendo sitios CpG | 0.2035 | 0.1326 | 0.3367 | 0.3272 | 0.4598 |

**Seudogenes ζ1, ζ2 y ζ3 de conejo**

Incluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 205 | 7 | 12 | 6 |
| C | 24 | 456 | 21 | 33 |
| A | 9 | 6 | 214 | 11 |
| T | 21 | 21 | 40 | 362 |

Excluyendo sitios CpG

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 205 | 7 | 12 | 6 |
| C | 11 | 330 | 9 | 17 |
| A | 9 | 6 | 212 | 11 |
| T | 11 | 9 | 17 | 250 |

Valores del contenido de bases esperados en el equilibrio si la secuencia fuera selectivamente neutra y gobernada por el patrón de sustituciones derivados de las matrices presentadas arriba

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Incluyendo sitios CpG | 0.2782 | 0.1794 | 0.3550 | 0.1874 | 0.3668 |
| Excluyendo sitios CpG | 0.2470 | 0.2248 | 0.2968 | 0.2314 | 0.4562 |

**Seudogenes localizados en islas CpG: globinas Z de caballo y hombre**

incluyendo sitios CpG

caballo (ΨζΨ)

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 68 | 18 | 6 | 9 |
| C | 25 | 154 | 10 | 19 |
| A | 3 | 8 | 78 | 19 |
| G | 6 | 7 | 22 | 116 |

humano (Ψα2)

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 68 | 12 | 4 | 3 |
| C | 16 | 148 | 5 | 27 |
| A | 5 | 9 | 71 | 17 |
| G | 6 | 14 | 14 | 120 |

excluyendo sitios CpG
caballo (Ψζ)

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 68 | 18 | 6 | 9 |
| C | 15 | 112 | 6 | 14 |
| A | 3 | 8 | 76 | 19 |
| G | 6 | 4 | 10 | 79 |

humano (Ψα2)

| desde/ a | T | C | A | G |
|---|---|---|---|---|
| T | 67 | 12 | 4 | 3 |
| C | 14 | 111 | 4 | 17 |
| A | 5 | 9 | 70 | 17 |
| G | 4 | 8 | 9 | 80 |

Contenido de bases esperado en el equilibrio si la secuencia fuera selectivamente neutra y gobernada por el patrón de sustituciones derivados de las matrices presentadas arriba

excluyendo sitios CpG

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| Ψζ caballo | 0.1529 | 0.2407 | 0.2574 | 0.3490 | 0.5897 |
| Ψα2 humana | 0.2070 | 0.2985 | 0.1573 | 0.3373 | 0.6357 |

incluyendo sitios CpG

| | T | C | A | G | G+C |
|---|---|---|---|---|---|
| **Ψζ caballo** | 0.1664 | 0.2533 | 0.2053 | 0.3750 | 0.6283 |
| **Ψα2 humana** | 0.2217 | 0.2951 | 0.1552 | 0.3281 | 0.6232 |

**IV.4.4 Comparición entre los valores esperados de acuerdo al patrón de mutaciones y los valores reales de los genes.**

La Figura IV.3.1 nos muestra la relación entre los valores reales de los genes y aquellos inferidos partir del patrón mutaciones y sustituciones en los seudogenes. Los puntos en negro indican es para valores esperados sin corrección para el efecto del dinucleótido CpG, mientras que los círculos en blanco representan los valores estimados con corrección para el efecto de CpG. Esta corrección se llevó a cabo de la siguiente forma:
GC-corregido=GCe- [**a***(1/2*GCe)$^2$], donde GCe es el valor esperado de GC3 de acuerdo al patrón inferido de mutaciones, resulta claro entonces que (1/2*GC2)$^2$ es la proporción de dinucleótidos CpG que se espera por azar de acuerdo a la composición de bases, mientras **a** representa la proporción de dinucleótidos CpG que son eliminados por la metilación/hipermutación. Debe tenerse en cuenta que la reducción de los valores de G+C es exactamente igual a la reducción del dinucleótido CpG, ya que por cada CpG que se transforma en TpG o CpA se pierde exactamente una C o una G. Los valores de **a** utilizados fueron de 0.8 (se eliminan el 80% de los CpGs) para genes que se encuentran fuera de las islas CpG, y 0.2 (se elimina el 20% de los CpGs) para genes que están ubicados dentro de las islas CpG (fue el caso de todos los genes tipo globina α). Aunque

esta aproximación no es exacta, nos da valores relativamente cercanos a aquellos obtenidos mediante modelos más complejos (Alvarez y Vietez, sin publicar). La línea recta se usa como referencia para indicar en que lugar deberían estar ubicados los genes si sus valores esperados fueran iguales a los observados. La distancia vertical entre los puntos y la recta indica cuan por encima (o por debajo) de los valores esperados se encuentra el $GC_3$ real de los genes.

Como resulta evidente a partir de la figura 4.1, el rango de variación de los valores reales de CG3 es de mas del 60% mientras que la variación de los esperados cubre un rango muchísimo menor (del 35% al 45% para la gran mayoría de genes). Resulta evidente además que todos los genes ricos y muy ricos en $GC_3$ los que presentan valores esperados de contenido en G+C muy por debajo a los valores reales, esto es cierto incluso para aquellos valores estimados a partir de los seudogenes ubicados en las islas CpG que presentan valores reales de $GC_3$ 50% por encima de los valores esperados.

Figura 4.1

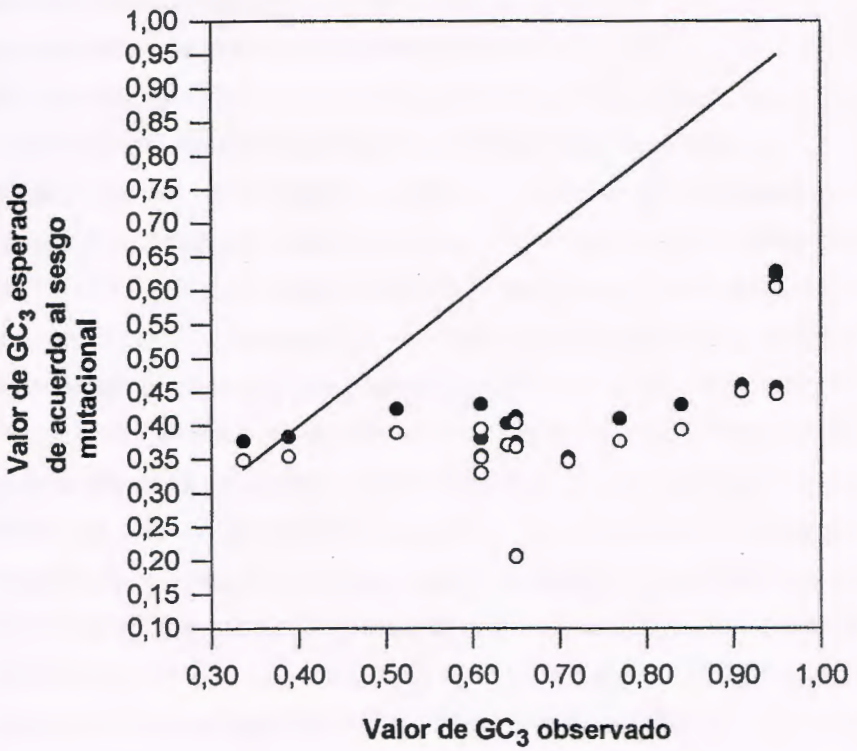



**Figura 4.1.** Relación entre los valores reales de $GC_3$ y aquellos esperados de acuerdo al sesgo mutacional. La línea indica la posición que deberían ocupar los puntos si el sesgo mutacional coincidiera con el contenido real en $GC_3$ .

## V DISCUSIÓN Y CONCLUSIONES

### V.1 Implicancias de la correlación entre $GC_2$ y la tasa de evolución no-sinónima.

Uno de los resultados más inesperados del análisis de sustituciones nucleotídicas fue la correlación entre la distancia no-sinónima y el contenido en $GC_2$ de los genes. Dicha correlación se observa en los tres grupos de genes homólogos de gramíneas, así como en los dos grupos de genes de los trypanosomátidos. Por otra parte la correlación fue encontrada previamente en genes de mamíferos (Alvarez, sin publicar) así como en un conjunto de 290 genes ortólogos entre *Mycobacterium tuberculosis* y *Mycobacterium leprae* (Miranda, Alvarez, Jabbari, Degrave & Bernardi, en preparación). Es decir la correlación en cuestión parece abarcar toda la escala evolutiva, desde bacterias a metazoarios superiores y plantas. La razón por la cual los genes con mayor riqueza en $GC_2$ evolucionan más rápidamente desde el punto de vista aminoacídico no es clara. Una posible causa es que se encuentre relacionado con la hidrofobicidad, pues la correlación siempre es fuerte con el contenido de T en la segunda posición del codón. En este sentido cabe aclarar que los codones hidrofóbicos poseen T como segunda base del codon, mientras que los hidrofílicos poseen A. Sin embargo, no parece existir una relación clara entre la hidropatía de las proteínas y la tasa de cambio aminoacídico, pues a excepción de los genes que *Mycobacerium*, no existe correlación entre la tasa no-sinónima y la hidrofobicidad promedio de las proteínas.

Otro posible nexo entre la tasa de cambio aminoacídico y el contenido en GC2 podría encontrarse en la estabilidad evolutiva de ciertos aminoácidos. Graur (1985) ha presentado evidencia indicando que la tasa de cambio aminoacídico de las proteínas tiene relación directa con la composición aminoacídica de las mismas. Esta relación se debería a que algunos aminoácidos serian más intercambiables evolutivamente debido a la relación entre la similitud bioquímica de los aminoácidos y la estructura del código genético. Por ejemplo, los derivados de la metionina realizando un sólo cambio de bases son: arginina (1 de los 9 cambios posibles llevan de Met a Arg), isoleucina (3 de 9 cambios), leucina (2 de 9 cambios), lisina (1 de 9 cambios), treonina (1 de 9 cambios) y valina (1 de 9 cambios). Por otro lado las distancias bioquímicas entre la metionina con cada uno de estos aminoácidos son (Gramtham, 1974): Arg=91, Ile=10, Leu=15, Lys=10,. Teniendo en cuenta estos dos elementos podemos ver que los aminoácidos más estables (más conservativos) serán aquellos que estén conectados por un mayor número de caminos mutacionales con aminoácidos disímiles y poco (o nada) conectados con aminoácidos similares, mientras que serán más inestables aquellos aminoácidos que estén conectados a través de varios cambios de una sóla base con aminoácidos similares. Dicha combinación arroja que los 5 aminoácidos más estables serian empezando desde el más estable (Graur, 1985): cisteina (TGY), triptofano (TGG), tirosina (TAY), glicina (GGN) y serina (TCN, AGY), mientras que los 5 más inestables serian (empezando desde el más inestable): metionina (ATG), glutamina (CAR), isoleucina [AT(Y/A)], leucina(CTN y TTY) e histidina (CAY). Este ordenamiento no sólo no coincide con lo que se esperaría sobre la base de la correlación $GC_2$-Distancia aminoacídica, sino que es lo inverso puesto que los 5

aminoácidos más inestables tienen A o T en la segunda posición de sus codones, mientras que los más estables (a excepción de tirosina) tienen G o C.
En base a lo anteriormente expuesto se desprende que la relación $GC_2$-distancia aminoacídica no presenta un correlato biológico trivial, sin embargo dada su posible "universalidad", es altamente probable que posea una base biológica. Será necesario estudiar esta relación desde una óptica nueva para arrojar algo de luz al problema. Sin duda el punto lo merece.

**V.2 Factores que afectan las tasas de cambio sinónimo en los trypanosomátidos**

El segundo punto que analizaremos concierne la aceleración de las tasas de cambio sinónimo que se observa en trypanosomas de la sección Salivaria, y en particular *T. brucei*. En esta tesis (sección IV.3.2) y en el manuscrito número 4 que se anexa, se presenta evidencia muy clara indicando que este grupo de trypanosomas ha por lo menos cuadruplicado sus tasas en cambio sinónimo en relación a los restantes trypanosomátidos. Esta aceleración también afecta, pero en mucho menor medida, la tasa de cambio aminoacídico. Varios factores biológicos podrían estar incidiendo en esta aceleración. A primera vista, este incremento parece estar relacionado con la duración de la generación organísmica (Hafner et al, 1994; Konhe, 1970; Wu & Li, 1985) puesto que de acuerdo a lo propuesto por Ohta (1973) en su "nearly-neutral theory of molecular evolution", la variación en los tiempos de generación afectaría la magnitud de la aceleración del cambio nucleotídico en forma inversa al grado de restricción funcional. Es decir, a menor grado de restricción, mayor de aceleración (o enlentecimiento) de la tasa evolutiva. Esta predicción ha recibido cierto soporte a partir de los resultados obtenidos de la comparación de las tasas de evolución nucleotídica entre distintos ordenes de mamíferos (Ohta, 1993, 1995). donde es posible apreciar que aquellos grupos de mamíferos con tiempos de generación más cortos (ej. múridos) presentan incrementos importantes en las tasas de evolución sinónima y en menor medida en la tasa de cambio aminoacídico. Sin embargo no es posible afirmar que dicha hipótesis puede explicar sencillamente lo que ocurre en los trypanosomátidos. En primer lugar hay que tener en cuenta que los tiempos de generación en el largo plazo (el tiempo de generación promedio del linaje evolutivo) no se conocen en este grupo de organismos. Es posible sin embargo que los trypanosomas africanos posean una duración de generación promedio más corta que *Trypanosoma cruzi,* puesto que los primeros son parásitos de la sangre mientras que *T. cruzi* es una parásito intracelular. Sin embargo resulta difícil concebir que los tiempos de generación *de T. brucei* sean más cortos que los de *Crithidia fasciculata* la que además de ser parásito exclusivamente en insectos, presenta fases de vida libre en su ciclo de vida.

Un factor adicional que puede haber afectado en el incremento de la tasa de evolución sinónima de *T. brucei* es el aparente relajamiento de la presión selectiva sobre el uso de codones. Hemos demostrado anteriormente (Alvarez et. al 1994) que si bien *T. brucei* presenta las mismas preferencias de codones en sus genes de alta expresión que los restantes grupos de trypanosomátidos, la intensidad de dicha preferencia es mucho menor. Además, en algunos grupos

de sinónimos, el codón preferido en *T. brucei* no es el mismo que en los restantes trypanosomátidos. Esto indicaría que la preferencia de codones en *T. brucei* se encuentra en fase de cambio, y/o que las presiones selectivas sobre las posiciones sinónimas son menores por lo que el grado de sesgo en el uso de codones disminuye, con el concomitante incremento en la tasa de cambio nucleotídico. Ambas causas podrían estar estrechamente relacionadas por obedecer a la misma razón biológica: el surgimiento y evolución de las glicoproteínas hipervariables de superficie (VSGs). En efecto, los trypanosomas de la sección Salivaria se caracterizan por presentar una familia multigénica de proteínas de membrana que contiene, en el caso de *T. brucei*, más de 1000 genes altamente divergentes unos de otros. En un momento dado toda la población de trypanosomas (en una infección en particular) expresa una única copia de estos genes, aquella que se encuentra localizada en la región de cromatina telomérica activa (Pays et al, 1989). Las restantes copias de genes de VSG se encuentran "apagadas" y exentas de cualquier tipo de presión selectiva, incluida por supuesto aquella que pueda afectar al uso de codones sinónimos. Podriamos decir que las copias de VSGs inactivas se comportan como seudogenes, puesto que no sólo acumulan muchas mutaciones que dan lugar a reemplazo de aminoácidos, sino que incluso algunas de estas mutaciones dan lugar a codones de terminación. Cada un determinado número de generaciones, en un individuo en particular de la población infectante, se produce un rearreglo génico y una nueva variante de VSGs comienza a ser expresada. A partir de este individuo se genera una nueva población clonal de trypanosomas que expresan la nueva variante de VSG. Los trypanosomas que expresan la vieja variante son eliminados por sistema inmunitario. Esta es la estrategia que "usan" los trypanosomas africanos para evadir al sistema inmune. Sin embargo dicho proceso implica dos cosas decisivas: en primer lugar genera un "cuello de botella evolutivo", pues cada cierto número de generaciones el tamaño de la población pasa a ser igual a 1, el individuo del cual surge el nuevo clón. Este proceso de continua contracción al mínimo de la población trae aparejado una reducción muy notoria en el tamaño efectivo de la población (Ne), lo que a su vez resulta en un decremento de la eficacia de la selección natural (ver introducción sección I.2.1). En segundo lugar, los VSGs son las proteínas más abundantes en los trypanosomas africanos, representando más del 10% de la masa proteica (Turner 1982). La abundancia de estas proteínas y el hecho que nos es posible generar o mantener en las mismas ningún tipo de sesgo en el uso de codones para optimizar el proceso traduccional (pues sus genes están en promedio el 99.9% del tiempo inactivos y por lo tanto libres de presión selectiva), ejerce una presión enorme sobre el aparato traduccional y plantea una contradicción: traducir los genes que codifican el producto proteico (por lejos) más abundante sin que exista ninguna optimización a tales efectos. Frente a esta "disyuntiva" existen dos soluciones: eliminar los genes codificantes VSGs y su sistema de evasión al sistema inmune, camino que es muy caro pues "en esto se les va la vida" y que además evidentemente los trypanosomas no tomaron; o cambiar las preferencias en el uso de codones optimizando las poblaciones de ARNts a las frecuencias de bases de los VSGs. La segunda posibilidad también es muy costosa, pues los restantes genes de alta expresión (tubulinas, GAP etc.) presentan preferencias de codones muy

distintas a la de los VSGs. Estas preferencias de codones en los restantes genes de alta expresión son el fruto de su herencia evolutiva, pues las mismas son las preferencias compartidas por todos los trypanosomátidos y por lo tanto las ancestrales en la familia Trypanosomatidae (Alvarez et al, 1994). Este segundo camino parece ser el que han tomado los trypanosomas de la sección Salivaria, bajar la estringencia del sesgo en las poblaciones de ARNt (acercándose a las de los genes de VSG) y consecuentemente cambiar las preferencias en el uso de los restantes genes de alta expresión acompañando el proceso. Probablemente el efecto "cuello de botella" mencionado más arriba juegue un rol decisivo en hacer esto posible pues disminuye el tamaño efectivo de la población. En este sentido es importante resaltar que cambiar las preferencias en el uso de codones es equivalente (a nivel molecular) a pasar de un pico adaptativo a otro. La teoría de los valles y picos adaptativos de Sewall Writght ( "Writght's shifting balance theory", Writght, 1931, 1932 y 1977) concibe la posibilidad de dicho pasaje a través del relajamiento de las presiones selectivas. Es importante tener presente que no es posible pasar de un pico adaptativo a otro (que siempre implica pasar por un valle adaptativo) si la selección natural predomina, pues como Ronald Fisher demostró (en su "teorema fundamental de la selección natural") el valor adaptativo promedio de una población nunca disminuye, o lo que es lo equivalente, la selección siempre presiona en dirección a los picos de la superficie adaptativa. En poblaciones pequeñas sin embargo, el pasaje de un pico a otro sería viable debido a que la deriva genética y los errores de muestreo en general, juegan un rol determinante. Es resaltable además que, en el caso de los trypanosomas de la sección Salivaria, la situación sería incluso más dramática (en el largo plazo), pues implicaría bajar a un valle adaptativo para permanecer ahí como consecuencia de que cambiar el uso de codones en dirección a las preferencias de los genes codificantes de VSGs es equivalente a cambiar en dirección a la ausencia de preferencias.

Para concluir esta sección se desea recalcar el hecho que muy poco se sabe acerca de que factores pueden dar lugar a un cambio en las preferencias de codones. Shields (1991) ha sugerido que el cambio podría ser causado por alteraciones drásticas en los sesgos mutacionales. Sin embargo, es bastante difícil concebir como un cambio en el sesgo mutacional, que a lo sumo implica modificaciones en algunas (pocas) proteínas relacionadas con la duplicación del ADN, pueda ser mantenido y llegue a imponer una alteración radical de las preferencias de codones en todos los genes de alta expresión, cuando el mismo implica costos energéticos altísimos por las alteraciones que provoca en la dinámica del aparato traduccional. Es probable en cambio, que las modificaciones en las preferencias de codones sinónimos ocurran en situaciones tales como la recién descripta en trypanosomas, es decir, donde no sea posible "evitar" el cambio.

### V.3 Correlación entre las tasas de evolución nucleotídica y el contenido $GC_3$

Varios estudios previos han reportado la relación existente entre la tasa de cambio sinónima y no-sinónima en genes de mamíferos (ver introducción). A partir de los resultados presentados en esta tesis (y las publicaciones que acompañan) es posible afirmar que la relación entre ambas tasas

evolutivas no es exclusiva de los genomas de mamíferos sino que parecería abarcar a todos los eucariotas. Las implicaciones de este conjunto de resultados son evidentes: los factores que afectan las posiciones sinónimas parecen ser parcialmente coincidentes con aquellos que afectan la conservación y variabilidad aminoacídica. El posible nexo funcional entre posiciones sinónimas y no-sinónimas podría estar dado por la selección para incrementar la fidelidad traduccional. En efecto, como ha sido discutido en la "Introducción" (ver sección I.1), el uso de codones puede incidir en la tasa de errores traduccionales, por lo que podría haber selección para mantener un determinado uso de codones en aminoácidos conservados (que se presumen importantes desde el punto de vista funcional) con la consiguiente disminución de la tasa de cambio sinónimo. Con el fin de investigar esta posibilidad, el análisis de las correlaciones fue extendido al nivel intragénico.

¿Cómo puede influir la selección para disminuir la tasa de errores traduccionales en la variación intragénica de las tasas de cambio sinónimo y no-sinónimo así como del contenido de bases sinónimas? Se reconoce ampliamente que muchos genes presentan dominios que son conservados o divergentes en términos de sustituciones de aminoácidos. Se acepta además que esta distribución espacial de la conservación aminoacídica refleja el efecto de la selección negativa por mantener los aminoácidos funcionalmente importantes (ver Kimura, 1991). Por otro lado, es precisamente en estos codones que codifican aminoácidos importantes donde la selección por disminuir la tasa de errores traduccionales debería tener mayor efecto, puesto que errores en la traducción de los mismos tendrían como efecto que la proteína pierda funcionalidad (el equivalente a un mutante nulo). Es de esperar entonces que éstos sitios estén ocupados por codones traduccionalmente óptimos y que por tanto exista selección negativa para mantener los codones sinónimos en cuestión. En otras palabras, aquellos sitios que son conservados a nivel aminoacídico deberían exhibir también tasas bajas de sustituciones sinónimas. De esta forma, además de la correlación que existe cuando los genes son considerados como una unidad, es esperable encontrar una correlación intragénica entre las tasas de cambio sinónimo y no-sinónimo. Es decir, aquellos genes que presenten a nivel aminoacídico dominios conservados y divergentes, deberían presentar el mismo patrón de conservación y divergencia a nivel sinónimo, resultando entonces en covariación intragénica. Esto además debería estar acompañado de una variación intragénica acorde en el uso de codones, esto es, las zonas conservadas desde e punto de vista aminoacídico deberían presentar mayor frecuencia de codones mayores que las zonas divergentes.

Los resultados presentados dan soporte a esta predicción. En los tres grupos biológicos analizados, las tres variables en cuestión presentan el comportamiento que se esperaría si la selección para incrementar la fidelidad traduccional tuviera incidencia en la distribución intragénica de la conservación sinónima y el uso de codones. Esto es, la tasa de sustituciones sinónimas (a lo largo de los genes) se encuentra correlacionada con la tasa de cambio no-sinónimo y también con el contenido en $GC_3$. Podemos decir que en los genes de mamíferos (y hasta cierto punto también de gramíneas) existe una triple relación: la distribución de zonas conservadas a nivel a nivel sinónimo se relaciona con las de zonas conservadas a nivel aminoacídico y con contenido de bases en las

posiciones sinónimas. Si bien no es posible afirmar que exista selección para un determinado uso de codones en los mamíferos, si es posible afirmar, en base a los resultados del estudio de los espectros mutacionales (los cuales serán discutidos en la siguiente sección) y de covariación intragénica, que ni el contenido en $GC_3$ total de los genes ni su variación intragénica sea selectivamente neutro. Sin embargo, son los resultados provenientes de genes de trypanosomátidos quienes nos permiten ver esta triple relación con mayor claridad. Los estudios que involucran a la proteína GP63 de *Leishmania* son básicamente similares a los descritos para genes de mamíferos y gramíneas (correlación intragénica entre ambos tipos de sustituciones nucleotídicas y con el contenido en $GC_3$). Sin embargo estos se diferencian de aquellos no sólo en la intensidad de la correlación (que en este caso es muchísimo mayor), sino que en los trypanosomátidos existen resultados independientes (basados en los niveles de expresión de los genes) que positivamente permiten identificar que codones son probablemente traduccionalmente óptimos. Cabe recordar que este análisis nos muestra no sólo que las zonas del gen más conservadas a nivel aminoacídico son también más conservadas desde el punto de vista sinónimo, sino que además estas regiones doblemente conservadas presentan valores de CAI (y $GC_3$) mayores que las regiones no conservadas. En otras palabras, aquellas regiones del gen que presentan alta frecuencia de los codones considerados traducionalmente óptimos presentan también bajas tasas de cambio tanto sinónimo como no-sinónimo. Esto que quiere decir que los aminoácidos conservados (y probablemente importantes desde el punto de vista funcional) se encuentran preferentemente codificados por codones mayores y que al mismo tiempo existe menor reemplazo evolutivo de los mismos.

Un aspecto que merece discusión adicional concierne a porqué sólo una proporción de genes despliega correlaciones intragénicas significativas entre la tasa de cambio sinónima y no sinónima. De hecho en el caso de los mamíferos aproximadamente un tercio de los genes entran en esta categoría siendo en las gramíneas la proporción incluso menor. Dos aspectos pueden estar afectando la ausencia de correlaciones significativas. Uno de ellos es tamaño de los genes y en parte fue discutido en la sección IV.1.1 (Tabla 1.2). En este sentido debe tenerse en cuenta que el tamaño pequeño de algunos genes atenta en forma muy severa contra la posibilidad de considerar significativo un coeficiente de correlación. Por ejemplo en genes de 200 codones de largo usando una ventana de un tamaño igual a 20 codones, solamente aquellos genes con coeficientes de correlación mayores a 0.63 son considerados estadísticamente significativos (0.63 es el valor crítico de r al 5% para 8 grados de libertad). Si todos los genes tuvieran 200 codones de largo y la correlación paramétrica fuera por ejemplo de 0.5 (esto es la correlación verdadera en la población), solamente un 30.85% de los genes serían detectados como estadísticamente significativos, mientras que esta proporción subiría a aproximadamente el 50% sí el largo fuera de 500 codones (el intervalo de confianza para *r* se obtiene con la transformación de Fisher $Z= 0.5 * \log[(1+r)/(1-r)]$, $\sigma =1/\sqrt{(n-3)}$ ). De lo anterior se desprende que incluso si la correlación intragénica estuviera presente en todos los genes (algo que es altamente improbable como veremos en el siguiente punto) detectaríamos

significación estadística en menos del 50% de los genes. El segundo aspecto que influye negativamente en la posibilidad de detectar correlación (incluso en genes infinitamente largos) tiene que ver con la distribución espacial de los aminoácidos funcionalmente importantes. Si asumimos que esta determina el patrón de distribución espacial de las sustituciones no-sinónimas y que también/de influye en la distribución espacial de las sustituciones sinónimas, se desprende que el proceso de divergencia daría lugar a covariación intragénica (entre los dos tipos de sustituciones) cuando dicha distribución espacial adopta determinadas formas. La ausencia de regiones definidas (dominios conservados y dominios divergentes), ya sea porque el gen es homogéneamente conservado u homogéneamente variable influye negativamente en la posibilidad de detección de la relación entre los dos tipos de sustituciones.

**V.4 Sesgos mutacionales en los genomas de mamíferos.**

Como ya ha sido mencionado en la Introducción (Sección I.1.3 ), se ha postulado que los sesgos mutacionales podrían explicar la variabilidad intragenómica en el uso de codones de los mamíferos así como el origen y mantenimiento de los isocoros. Sin embargo la estimación de la dirección e intensidad del sesgo mutacional se ha basado en las composiciones de bases de secuencias supuestamente exentas de restricciones funcionales. Es decir, se asume que la composición de bases de secuencias tales como intrones, flanqueantes (en incluso terceras posiciones de los codones), refleja el estado de equilibrio para un determinado patrón mutacional. Es claro sin embargo que asumir neutralidad total en secuencias como estas implica entrar en un círculo vicioso (asumimos neutralidad para demostrar neutralidad). Por otra parte, existe abundante evidencia que indica que dichas secuencias presentan al menos cierto grado de restricción funcional. Por un lado las secuencias flanqueantes contienen innumerables señales regulatorias, así como señales para el inicio y terminación de la transcripción, y por su parte los intrones poseen varias secuencias que actúan como señales para el splicing además de poseer señales que participan en la regulación génica.

En los análisis presentados en esta tesis se ha estimado el espectro mutacional a partir de dos fuentes de información independientes: las mutaciones deletereas y las sustituciones en los seudogenes. La idea de estimar el patrón intrínseco de mutaciones a partir de la información aportada por las mutaciones detectables (deletéreas) es relativamente reciente debido a la casi total ausencia de datos de secuencia sobre mutaciones hasta no hace demasiado tiempo. Varios trabajos han reportado patrones de mutación para algunos de los genes estudiados en esta tesis. Concretamente los espectros mutacionales de los factores de coagulación VIII y IX , han sido descriptos por varios autores (Cooper & Krawezak 1990; Krawezak & Cooper 1996; Wacey et al, 1994), así como del gen que codifica a la proteína P53 (Hartmann et al, 1997). A diferencia del encare que se ha dado a este tipo de análisis en la presente tesis, en ninguno de estos estudios mencionados anteriormente se utiliza el patrón de mutaciones para inferir la composición de bases que resultaría a partir de dicho patrón mutacional. Por otro lado varios autores han utilizado el

patrón de sustituciones de los seudogenes con el objetivo de inferir el espectro mutacional subyacente así como la composición de bases que se esperaría en el equilibrio para una secuencia selectivamente neutra (Bulmer, 1986; Gojobori et al, 1982; Li et al, 1984; Cassane et al, 1997). Los resultados provenientes de estos análisis, aunque en términos generales coinciden con los resultados presentados acá, no son sin embargo adecuados para contrastar con el contenido en $GC_3$ de genes funcionales pues la mayoría de los seudogenes usados en estos trabajos son del tipo procesados.

El objetivo que guió los análisis que se presentan en esta tesis fue estimar el espectro de mutaciones con el fin de testar la hipótesis que sostiene que el origen y mantenimiento de los isocoros, así como la diversidad en el contenido en GC3 de los genes de mamíferos se debe a la variabilidad subyacente en el sesgo mutacional. En efecto, a partir de los espectros mutacionales es posible inferir la composición de bases de una secuencia nucleotídica teórica completamente libre de restricciones funcionales. Esto nos permite comparar la predicción mutacionista-neutralista con el contenido en $GC_3$ real de los genes de mamíferos. Si la composición de bases en las posiciones sinónimas ($GC_3$) fuera completamente neutra y determinada por el sesgo mutacional, se desprende que la misma debería coincidir con la composición de bases que se infiere a partir del espectro mutacional. Además, dada la enorme variación en el contenido en $GC_3$ de los genes de mamíferos (más un 60%), de acuerdo a la hipótesis mutacionista, es esperable que dicha variación se encuentre acompañada por una variación equivalente en el patrón de mutaciones. Los resultados presentados en la Sección IV.4 nos permiten afirmar que ni el contenido de bases sinónimo de los genes ricos o muy ricos en $GC_3$, ni variación en $GC_3$, pueden ser explicados por el sesgo mutacional. Para ilustrar el punto permítasenos comparar los extremos: el contenido en G+C (para una secuencia neutra) que se espera a partir del espectro de mutaciones de un gen pobre en $GC_3$, como lo es el que codifica al Factor IX de coagulación ($GC_3$=0.33), es 0.38 o 0.36 dependiendo de sí se tienen en cuenta o no las repeticiones de mutaciones. Por su parte, el contenido en G+C esperado a partir el espectro mutacional de un gen muy rico en $GC_3$, como la glucosa 6-P deshidrogenasa ($GC_3$=0.84), es 0.37 o 0.43 dependiendo de sí las mutaciones repetidas se incluyen o no. Tenemos entonces que dos genes cuyos valores reales de $GC_3$ difieren en un 51%, dan lugar estimaciones G+C neutro que a lo sumo difieren en un 7%, aunque para ambos genes la primera estimación (que incluye las repeticiones) es más confiable pues sus bases de estos genes se encuentran bastante depuradas en lo que tiene que ver con la identidad por ascendencia. El gen de la glucosa 6-P deshidrogenasa presenta un valor en $GC_3$ más de un 40% por encima de lo que se esperaría de acuerdo a su esquema de mutaciones. La situación no es muy distinta cuando observamos los resultados provenientes de los restantes genes. En todos los casos, independientemente de la forma en que se obtenga la matriz de mutaciones, los genes ricos o muy ricos en $GC_3$ (y por supuesto que también los genes pobres en $GC_3$), dan lugar a estimaciones de patrones mutacionales que indican que los mismos se encuentran bajo presión mutacional AT. A

esto se le debe añadir además el efecto de la hipermutabilidad del dinucleótido CpG, el cual es un factor que tiende a empobrecer aún más el contendido en G+C.

Por otro lado hemos inferido los patrones mutacionales a partir de las sustituciones en seudogenes. Hemos analizado un grupo de seudogenes que se encuentran en la vecindad inmediata de sus contrapartes funcionales de forma tal que sea posible inferir el espectro de mutaciones en las regiones donde los genes se encuentran localizados. Los resultados obtenidos a partir de estos seudogenes apuntan en la misma dirección que aquellos obtenidos en base a la frecuencia de mutaciones deletéreas: el contenido en $GC_3$ de los genes se encuentra muy por encima de lo que cabría esperar si este fuera selectivamente neutro y gobernando por el patrón de mutaciones. Sin embargo, en el caso de los seudogenes ubicados dentro de las islas CpG parecería existir un leve sesgo mutacional hacia GC; aunque esto no es posible afirmarlo con absoluta certeza debido al grado de divergencia que se observa en esas secuencias. No obstante, incluso asumiendo que los resultados provenientes de estos seudogenes ubicados en las islas CpG sean confiables, es posible observar que la intensidad de este posible sesgo mutacional GC es ampliamente insuficiente para explicar la gran riqueza en $GC_3$ de los genes funcionales. En efecto, al comparar la composición neutra a la cual darían lugar los patrones de sustituciones de estos seudogenes con el contenido en $GC_3$ real de sus homólogos funcionales, nuevamente se observa que la composición esperada en el equilibrio para una secuencia selectivamente neutra se encuentra más de un 30% por debajo del valor en GC3 de los genes funcionales.

Un aspecto que requiere ser discutido concierne la validez de los resultados y afirmaciones del párrafo anterior. Como fuera planteado anteriormente (Sección IV.4.1), la estimación del patrón intrínseco de mutaciones a partir de las mutaciones deletéreas presenta dos posibles fuentes de sesgo: el de la detectabilidad diferencial de las mutaciones y el problema de la identidad por ascendencia. En lo que se refiere al segundo punto, estamos en condiciones de descartar rápidamente que el mismo pueda haber tenido algún tipo de incidencia en sesgar los resultados pues las estimaciones basadas en las matrices que tienen en cuenta la repetición no difieren demasiado de aquellas que no la consideran. El problema de sesgo de detectabilidad en cambio requiere una discusión un poco más detallada. Existen tres elementos que sugieren que este problema no nos está introduciendo ningún sesgo serio. En primer lugar, las estimaciones basadas en todas las bases de datos de mutacionales que hemos analizado coinciden en el tipo de sesgo mutacional que se infiere de ellas, es decir sesgo AT. Parece altamente improbable (aunque no totalmente descartable) que un conjunto de genes con funciones tan diversas como el analizado acá coincidan por azar en presentarnos (en forma artificiosa) un patrón mutaciones nucleotídicas donde los cambios G/C-->A/T sean predominantes en relación a las mutaciones A/T-->G/C. El segundo elemento que sugiere en forma bastante clara la ausencia de sesgo de detección surge de los resultados obtenidos con las tres bases de datos del gen P53. Es particularmente destacable que el patrón de mutaciones inferido a partir de las sustituciones sinónimas sea muy similar a aquel obtenido a partir de las mutaciones de la primera y segunda posición de los codones. Teniendo en

cuenta esta similitud y considerando además que la muestra de sustituciones sinónimas claramente no adolece de ningún tipo de sesgo de detección por representar una muestra completamente al azar, podemos deducir que la muestra de mutaciones deletéreas de la proteína P53 tampoco adolece de tal sesgo y que por lo tanto es un muy buen estimador del patrón real de mutaciones. El último punto que apoya la idea de que las bases de datos mutacionales no arrojan estimaciones sesgadas está dado por la similitud entre el espectro mutacional inferido a partir de estas bases de datos y aquel obtenido a partir de las sustituciones en los seudogenes.

## REFERENCIAS


-Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. ***Genetics*** 136:927-935.

-Alvarez, F.; Cortinas, M; & H. Musto, H. (1996). The analysis of protein coding genes suggests monoply of *Trypanosoma*. ***Mol Phylogenet Evol.*** 5:333-343.

-Alvarez, F., Robello, C. & M. Vignali. (1994). Evolution of codon usage and base contents in kinetoplastid protozoans. ***Mol. Biol. Evol.*** 11:790-802.

-Bernardi, G. & G. Bernardi (1985). Codon usage and genome composition. ***J. Mol. Biol.*** 22:363-365.

-Bernardi, G; Olofsson, B.; Fillipski, J.; Zerial, M.; Salinas, J.; Cuny, G.; Meunier-Rotival, M & F. Rodier (1985). The mosaic genome of warm-blooded vertebrates. ***Science*** 228:953-958.

-Bernardi, G.; Mouchiroud, D. & C. Gautier (1993). Silent substitutions in mammalian genomes and their evolutionary implications. ***J. Mol. Biol.*** 37:583-589.

-Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. ***Nucleic Acids Res.*** 8:1499-1504.

-Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. ***Nature*** 321:209-213.

-Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. ***Genetics*** 129:897-907.

-Bulmer, M. (1986). Neighboring base effects on substitutions rates in pseudogenes. ***Mol. Biol. Evol.*** 3(4): 322-329.

-Bulmer, M., Wolfe, K.H. & P.M. Sharp (1991). Synonymous nucleotide substitutions rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. ***Proc. Natl. Acad. Sci. (USA)*** 88:5974-5978.

Casane, D., Boissinot, B.H., Chang, B.H.-J., Shimmin, L.C. & W.-H Li (1997). Mutation pattern variation among regions of primate genome. ***J. Mol. Evol.*** 45:216-226.

-Cooper, D.N. & M. Krawezak (1990). The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. ***Hum. Genet.*** 85:55-74.

-Dayhoff, M. O., R. M. Schwartz, & B. C. Orcutt. (1978). A model of evolutionary change in proteins. In ***Atlas of Protein Sequence and Structure***, Vol 5, Suppl. 3 (ed M. O. Dayhoff), National Biomedical Research Foundation, Washington D.C., pp. 345-352.

-D'Onofrio, G; Mouchiroud, D.; Aíssani, B; Gautier, C. & G. Bernardi (1991). Correlations between the compositional properties of huma genes, codon usage and aminoacid composition of proteins. ***J. Mol. Biol.*** 32:504-510.

-Duvall, M.R. & B. R. Morton (1996). Molecular phylogenetics of Poaceae: An expanded analysis of rbcl sequence data. ***Mol. Phyl. Evol***. 5: 352-358

-Eyre-Walker A, & Bulmer M (1995). Synonymous substitutions rates in enterobacteria. ***Genetics*** 140:1407-1412.

-Finlay C.A., Hinds P.W. & A.J. Levine (1989). The p53 proto-oncogene can act as a suppressor of transformation. ***Cell*** 57:1083-1093.

-Fitch, W. M. (1980). Estimating the total number of nucleotide substitutions since the common ancestor of a pair of genes: comparison of several methods and three beta hemoglobin messenger RNA's. ***J. Mol. Evol.*** 16:153-209.

-Gojobori, T., W.-H Li & D. Graur (1982). Patterns of nucleotide substitutions in pseudogenes and functional genes. J. Mol. Evol. 18:360-369.

-Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. ***Science*** **185**:862-864.

-Grantham, R.; Gautier, C; Gouy, M.; Mercier, R. & R. Pave. (1980). Codon catalog usage and the genome hypothesis. ***Nucleic Acids Res.*** 8:49-62.

-Gross, G., Mielke, C., Hollatz, I., Blocker H., & R. Frank (1990).RNA primary sequence or secondary structure in the translational initiation region controls expression of two variant interferon-beta genes in Escherichia coli. ***J . Biol. Chem .*** 265(29):17627-17636

-Guisez, Y.; Robbens, J.; Remaut, E. & W. Fiers (1993). Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon usage: a hypothesis. ***J. Theor. Biol.*** 162:243-252.

-Hafner, M.S., Sudman, P. D., Villablanca, F. X., Demastes, J. W., & S. A. Nadler (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. ***Science*** 265:1087-1090.

-Hardison, R. & W. Miller (1993). Use of long sequence alignments to study the evolution and regulation of mammalian globin gene cluster. ***Mol. Biol. Evol.*** 10(1):73-102.

-Hartl, D. L.; Moriyama, E. T. & S. A. Sawyer (1994). Selection intesity for codon bias. ***Genetics*** 138:227-234.

-Hartmann, A., Blasyk, H., Kovach, J. & S.S. Sommer (1997). The molecular epidemiology of P53 mutations in human breast cancer. ***Trends in Genetics***, 13(1):27-33.

-Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the ocurrence of the respective codons in proteins genes. ***J. Mol. Biol.*** 146:1-21.

-Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the ocurrence of the respective codons in proteins genes. ***J. Mol. Biol.*** 158:573-587.

-Inverso, J. A., E. Medina-Acosta, J. O'connor, D. G. Russell & G. A. Cross. (1993). *Crithidia fasciculata* contains a transcribed leishmanial surface proteínase (gp63) gene homologue. ***Mol. Biochem. Parasitol.*** 57:47-54.

-Jukes, T.H. & C. R. Cantor (1969). Evolution of protein molecules. En ***Mammalian Protein Metabolism***, editor H. N. Munro; New York Academic Press:21-132.

-Kimura, M. (1968). Evolutionary rate at the molecular level. ***Nature*** 217:624-626.

-Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.

-Kimura, M. (1983). "***The Neutral Theory of Molecular Evolution***", Cambridge Univ. Press, Cambridge,U.K.

-Kimura, M. (1991). Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. ***Proc. Natl. Acad. Sci. (USA)*** 88:5969-5973.

-King, J.L. & T.H. Jukes (1969). Non-Darwinian evolution. ***Science*** 164:788-798.

-Kohne, D.E. (1970). Evolution of higher-organism DNA. ***Quart. Rev. Biophys***. 33:327-375.

-Krazewak, M & D.N. Cooper (1996). Mutational processes in pathology and evolution. en "***Human Genome Evolution***", edit. M. Jackson, T. Strachan y G. Dover. BIOS Scientific Publishers, Oxford.

-Li, ,W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitutions. ***J. Mol. Evol.*** 36:96-99.

-Li, W.-H; Gojobori, T & M. Nei (1981). Pseudogenes as a paradigm of molecular evolution. ***Nature*** 292:237-239.

-Li, W.-H., & D. Graur (1991). "***Fundamentals of Molecular Evolution***," Sinauer, Sunderland, MA.

-Li, W.-H., Wu, C.-I & C.C .Luo (1984). Nonrandomness of point mutations reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. ***J. Mol. Evol.*** 21:58-71.

-Li, W.-H., Wu, C.-I & C.C. Luo (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide of codon changes. ***Mol. Biol. Evol.*** 2:150-174.

-Matassi, G.; Montero, L.,M.; Salinas, J.; & G. Bernardi (1989). The isochore organization and the compositional distribution of homologous coding genes in the nuclear genomes of plants. ***Nucleic Acids Res.*** 17:5273-5290.

-Milner J. & E.A.Medcalf (1991). Cotranslation of activated mutant p53 with wild type drives the wild-type p53 protein into the mutant conformation. ***Cell*** 65: 765-774.

-Moriyama, E. & D.L. Hartl (1993). Codon usage biases and base composition of nuclear genes in *Drosophila*. ***Genetics*** 134:847-858.

-Morris, S.W., N. Nelson, N., Valentine, M.B.,Shapiro, D.N., Look, A.T. Kozlosky, C.J., Beckmann, M.P. & D.P. Cerretti (1992). Assignment of the genes encoding human interleukin-8 receptor types 1 and 2 and an interleukin-8 receptor pseudogéne to chromosome 2q35. ***Genomics*** 14:685-691.

-Mouchiroud D, Gautier C & G. Bernardi (1995). Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of non-synonymous substitutions. ***J Mol Evol*** 40:107-113

-Musto, H., Cacciò, S., Rodriguez-Maseda, H. & G. Bernardi (1997). Compositional constraints in the extremely GC-poor genome of *Plasmodium falciparum*. ***Mem. Inst. Oswaldo Cruz***. 92(6):835-841

-Musto, H.; Rodriguez, H. & F. Alvarez (1995). Compositional correlations in the nuclear genes of the flatworm *Schistosoma mansoni*. ***J. Mol. Evol.***, 40:343-346.

-Nei, M. (1987). ***"Molecular Evolutionary Genetics".*** Columbia University Press, New York.

-Nei, M & T. Gojobori (1986). Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. ***Mol. Biol. Evol.*** 3:418-426.

-Nei, M & D. Grauer (1984). Extent of protein polymorphism and the neutral mutation theory. ***Evol. Biol.*** 17:73-118.

-Nicolas A., & J.L. Rossignol (1983). Gene conversion: point-mutation heterozygosities lower heteroduplex formation. ***EMBO J.*** 1983;2(12):2265-2270

-Normura, M.; Sor, F. Yamagishi, M.; & M. Lawson (1987). Heterogeneity of GC content within a single bacterial genome and its implications for evolution. ***Cold Spring Harbor Symp. Quant. Biol.*** 52:658-663.

-Ochman H., & A.C. Wilson (1987). Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. ***J. Mol . Evol*** .26(1-2):74-86

-Ohta, T. (1973).Slightly deleterious mutant substitutions in evolution. ***Nature***, 246:96-98.

-Ohta, T. (1993). An examination of the generation time effect on molecular evolution. ***Proc. Natl. Acad. Sci. (USA)*** 90: 10676-10680.

-Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly nuetral theory. ***J. Mol. Evol.*** 40:56-63.

-Otha, T & Y. Ina (1995). Variation in synonymous substitutions rates among mammalian genes and correlations between synonymous and nonsynosymous divergences. ***J. Mol. Evol.*** 41:717-720.

-Pamilo, P. & Bianchi, N.O. (1993). Evolution of the Zfx and Zfy genes: rates and interdependence between teh genes. ***Mol. Biol. Evol.*** (Fecha)

-Pays, E., Tebabi, P., Coquelet, H., Revelard, P., Salmon,D. & M. Steiner. (1989). The genes and transcripts of an antigen gene expression site from *T. brucei.* ***Cell*** 57:835-845.

-Precup, J., & J. Parker (1987). Missense misreading of asparagine codons as a function of codon identity and context. ***J. Biol. Chem***. 262:11351-11356.

-Purvis, I.J.; Bettany, A.; Chinnappan-Santiago, T.; Coggins, J.; Duncan, K.; Esason, R. & A.J. Brown (1987). The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*. A hypothesis. ***J. Molec. Biol.*** 193:423-417.

-Saitou, N., & M. Nei. (1987), The nieghbor-joining method: a new method for reconstructing phylogenetic trees. ***Mol. Biol. Evol.*** 4:406-425.

-Salinas, J.; Matassi, G.; Montero, L.M. & G. Bernardi (1988). Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. ***Nucleic Acids Res.*** 16:4269-4285.

-Sarich, V.M. & A.C. Wilson (1973). Generation time and genomic evolution in primates. ***Science*** 179(78):1144-1147

-Sharp, P.M. (1990). Processes of genome evolution reflected by base frequency differences among *Serratia marcenscens* genes. ***Molecular Microbiology*** 4(1):119-122.

-Sharp, P. M. & K.M. Devine (1989). Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do "prefer" optimal codons. ***Nucleic Acids Res.*** 17: 5029-5039.

-Sharp, P. M. & W-H. Li (1987). The rate of synonymous substitutions in enetrobacterial genes is inversely related to codon usage bias. ***Mol. Biol. Evol.*** 4:222-230.

-Shields, D.C. & P.M. Sharp (1987). Synonymous codon usage in *Bacilis subtilis* reflects both translational selection and mutational biases. ***Nucleic Acids Res.*** 15: 8023-8040.

-Ticher, A. & D. Graur (1989). Nucleic acid composition, codon usage, and the rate of synonymous substitutions in protein-coding genes. ***J. Mol. Evol.*** 37:441-456.

-Turner, M. J. (1982). Biochemistry of the variant surface glycoproteins of salivarian trypanosomes. ***Adv. Parasitol***. 21: 69-153.

-Varenne, S, Buc, J.; Lloubés, R. & Lazdunski (1984).Translation in a nonuniform process: effect of tRNA availability on the rate of elongation of nascent polypeptide chains. ***J. Molec. Biol.*** 80:549-576.

-Weiner AM & R.A. Denison (1983). Either gene amplification or gene conversion may maintain the homogeneity of the multigene family encoding human U1 small nuclear RNA. ***Cold Spring Harb. Symp Quant. Biol.*** 47(2):1141-1149

-Wacey, A.I., Krawezak, M., Kabbar, V.V. & D.N. Cooper (1994). Determinants of the factor IX mutational spectrum in haemophilia B: an analisys of missense mutations using a multi domain molecular model of activated protein. ***Human Genet.*** 94:594-608.

-Wolfe, K. H. & P.M. Sharp (1993). Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. ***J. Mol. Evol*** 37:441-456.

-Wolfe, K. H.; Sharp, P. M. & W.-H Li (1989). Mutation rates differ among regions of the mammalian genome. ***Nature*** 337:283-285.

-Wong et al. (1995). Mutations in the cell adhesion molecule L1 cause mental retardation. ***Trends Neurosci***. 18:168-172.

-Wright, S. (1931). Evolution in Mendelian populations. ***Genetics*** 16:97-159.

-Wright, S. (1932). The roles of mutation, inbreeding, crossbrreding, and selection in evolution. ***Proc. 6th Intl. Cong. Genet.*** 1:355-366 (citado por Nei, 1987).

-Wright, S. (1978). "***Evolution and the Genetics of Populations Vol. 4: Variability within and among populations***". University of Chicago Press, Chicago, Ill.

-Zama, M (1990). Codon usage and secondary structure of mRNA. ***Nucleic. Acids. Symp. Ser.*** 22:93-94

-Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. ***J. Mol. Evol*** 39:315-329.