UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

"Retropseudogenes de las chaperoninas mitocondriales en el genoma de *Mus musculus* como modelo para estudiar la expresión genética en el pasado"

T E S I S

QUE PARA OBTENER EL TITULO DE:

B I Ó L O G O

P R E S E N T A:

LUIS DAVID ALCARAZ PERAZA

DIRECTOR DE TESIS: DR. VÍCTOR MANUEL VALDÉS LÓPEZ

ENERO 2005

Autorizo a la Dirección General de Bibliotecas de la UNAM a difundir en formato electrónico e impreso el contenido de mi trabajo recepcional.
NOMBRE: Luis David Alcaraz Peraza
FECHA: 14/II/05
FIRMA:

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Retropseudogenes de las chaperoninas mitocondriales en el genoma de Mus musculus como modelo para estudiar la expresión genética en el pasado."

realizado por Luis David Alcaraz Peraza

con número de cuenta 09950791-8 , quien cubrió los créditos de la carrera de: Biología

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dr. Víctor Manuel Valdés López
Propietario	Dr. Luis Felipe Jímenez García
Propietario	Biol. Luis José Delaye Arredondo
Suplente	M. en I.B.B. Claudia Andrea Segal Kischinevsky
Suplente	Biol. Alfonso José Vilchis Peluyera

Consejo Departamental de Biología

M. en C. Juan Manuel Rodríguez Chávez

FACULTAD DE CIENCIAS
UNIDAD DE ENSEÑANZA DE BIOLOGÍA

A mis padres,

Por el amor, la confianza y el apoyo incondicional en toda época. Este trabajo es un mínimo detalle, para todo lo agradecido que estoy y lo mucho que los quiero.

A mis hermanas y hermanos,

Por toda lo que me encanta y me divierte convivir con ustedes, además de que siempre, todo el tiempo los tengo presentes y aprendo mucho de cada uno, con cariño.

A mi familia,

En especial a mi tío Ricardo, por todo el cariño y apoyo brindados a lo largo de esta vida.

A la familia Arévalo López,

No tengo palabras para agradecer a cada uno de ustedes todo la ayuda y amistad a lo largo del tiempo.

A mis amigas y amigos,

De manera general y no discriminatoria, agradezco la convivencia, la plática, la crítica, la felicidad, en fin por cada momento compartido es un placer contar con ustedes.

Agradecimientos

Al Dr. Víctor Valdés López,

Por aceptarme en su laboratorio, ser mi maestro y tutor que se interesa en formar un espíritu crítico, sin descuidar la calidad humana, sin su apoyo hubiera sido imposible la realización de este trabajo. Muchas gracias.

A todos los integrantes del Laboratorio de Biología Molecular de la Facultad, por ser unos excelentes compañeros y amigos.

A mis sinodales, por el tiempo dedicado a la crítica y revisión del trabajo, que siempre son y serán de gran ayuda.

Al Dr. Luis Eguiarte y la Dra. Valeria Souza por todo el tiempo y la formación durante 2 años en su taller.

A mis profesores de la Facultad,
En la mayoría de los casos, además de una rigurosa formación académica, había posibilidades de establecer una amistad que se agradece sinceramente.

Resumen

Los retropseudogenes (ψ) son secuencias producidas por un evento de retrotranscripción a partir de un molde de mRNA maduro y una reincorporación aleatoria en el genoma. Los ψ, usualmente, dejan de ser funcionales desde el momento de su reincorporación al genoma ya que las secuencias insertadas carecen de promotores y la selección natural libera su presión sobre dichas secuencias, con lo que comienzan a acumular mutaciones, inserciones y deleciones, por ende los ψ siguen un proceso de evolución neutral por lo que se consideran "fósiles moleculares". Si datamos, mediante el reloj molecular, a cada uno de dichas secuencias estamos en posibilidad de cuantificar la frecuencia de la generación de ψ y evaluar si existen tendencias temporales de generación con lo que se puede determinar el nivel de expresión genética en el pasado, tratando de correlacionar dichos eventos con presiones de selección que obligaran al organismo a elevar la tasa de transcripción y de esta forma aumentar la probabilidad de generar un ψ. Utilizamos las chaperoninas mitocondriales hsp10 y hsp60, debido a la sensibilidad a cambios ambientales y estrés para la regulación de su transcripción, además de que en nuestro grupo de trabajo contamos con datos del humano para estos ψ. La búsqueda inicial se hizo en la base de datos del Ensembl, encontrando 27 ψhsp10 con edades de 7 a 67 millones de años y 40 ψhsp60 con una antigüedad de entre 4 a 68 millones de años. Adicionalmente caracterizamos un retrogen, mediante la combinación de análisis bioinformático y la búsqueda de Etiquetas de Expresión de Secuencias (ESTs), correspondiente al Factor Temprano de Embarazo (EPF) producto de una retrotranscripción de la hsp10 que adquirió una nueva función, en el humano se desconoce el locus del EPF y se sugiere que es el mismo que el de la hsp10, por lo que el ratón tiene una ruta alternativa en la regulación de la expresión del EPF, ubicando el locus del EPF en un contexto genómico independiente.

Contenido

Índice de Figuras

Índice de Tablas

1. Introducción

¿Qué es pseudogen y retropseudogen?

Un pseudogen es una secuencia presente en el genoma de una población dada y se caracteriza por tener una alta similitud con uno o más genes parálogos (Mighell, et al. 2000). El consenso para definir a los pseudogenes es que son copias no funcionales, desactivadas y remanentes de un gen funcional. De manera general, existen dos sucesos principales en la generación de pseudogenes en los genomas: la duplicación génica y la retrotranscripción.

Los pseudogenes generados por duplicación génica, son conocidos como pseudogenes no procesados o de forma común, pseudogenes. Como la duplicación génica genera copias completas del DNA genómico se puede identificar la presencia de intrones del gen original.

Por el otro lado, los genes producidos por un evento de retrotranscripción son conocidos como retropseudogenes o pseudogenes procesados. La retrotranscripción se da a partir de un molde de mRNA maduro. Su posterior retrotranscripción a cDNA y reincorporación al genoma es aleatoria. Ya que los retropseudogenes son generados a partir del mRNA ya maduro, una característica es que no existen intrones del gen original.

El término pseudogen surge a partir de una investigación del genoma de *Xenopus laevis* en el año de 1977 (Jacq et al. 1977 en Mighell et al. 2000) Solo se puede aplicar el término, *sensu stricto*, a secuencias no codificantes relacionadas a una secuencia funcional. Existen varias formas de denotar formalmente a los pseudogenes, las más comunes son: anteponer la letra Ψ al nombre del gen, por ejemplo Ψhsp60; o bien un sufijo con la letra 'P', por ejemplo, hsp10P. No existe consenso sobre la simbología específica para representar un retropseudogen.

Características y génesis de un pseudogen no procesado

Un pseudogen no procesado, se genera cuando existe una duplicación génica de un gen funcional.

La duplicación génica es considerada como un proceso, posiblemente aleatorio, catalizado por las enzimas que se encargan de los procesos de recombinación. Los eucariontes, sin embargo, poseen un sistema enzimático encargado de unir las dos puntas de un DNA roto, por lo que eventos como las duplicaciones así como las inversiones, deleciones y traslocaciones pueden surgir como consecuencia de la reunión de fragmentos cromosómicos que han sido rotos en más de un punto. Cuando las secuencias de DNA duplicado se encuentran unidas cabeza con cola, se dice que se encuentran repetidas en tándem. Una vez que una repetición en tándem aparece, se puede extender rápidamente en series largas de repeticiones en tándem, debido a los eventos de recombinación desigual entre dos cromosomas homólogos. La duplicación del DNA seguida de una recombinación no homologa, o desigual, promueve la amplificación del DNA (Alberts et al. 2004). Los pseudogenes tienen una alta probabilidad de quedar adyacentes a sus parálogos funcionales, sin embargo, también pueden ser insertados en diferentes cromosomas (Mighell et al. 2000)

Después de que se ha generado el segmento duplicado, existe la posibilidad de un incremento en la producción del polipéptido, pero la diferenciación funcional entre las secuencias puede tomar dos caminos: En el primero, no existe el cambio funcional y sencillamente se da la duplicación de la producción del polipéptido. La otra opción, es que la función de la secuencia original se mantenga en el nuevo DNA, pero que exista cierto grado de diferenciación entre las secuencias por la acumulación de mutaciones, de tal forma que se generen variaciones de la secuencia de la proteína (Griffiths et al. 2000).

Se ha propuesto que, si además de la duplicación se da un evento de traslocación de la secuencia, se pueden llegar a adquirir nuevas funciones que aumenten la adecuación del organismo, ya que en una nueva ubicación genómica tiene la oportunidad de evolucionar independientemente a la secuencia en cuestión (Alberts et al. 1994). Si las variaciones de dichas proteínas siguen siendo funcionales y selectivamente positivas, se generan familias génicas, por ejemplo las globinas humanas.

En ocasiones, pueden llegarse a adquirir funciones cualitativas totalmente distintas de las del gen de origen. Un ejemplo de esto es el gen de la lisozima, gen compartido entre aves, mamíferos y algunos otros eucariontes. Este gen ha sido duplicado en mamíferos, siendo una secuencia que produce una proteína no enzimática, la α-lactoalbúmina. El gen de la α-lactoalbúmina muestra la misma estructura de intrones-exones que el gen de la lisozima, arreglo que sugiere un evento de duplicación múltiple en el origen de la lisozima y de la α-lactoalbúmina (Griffiths et al. 2000).

La generación de las familias génicas y su mecanismo de formación, así como la adquisición de nuevas funciones cualitativamente diferentes a las originales, enfatizan la importancia de la duplicación génica como una forma efectiva de generar nuevas funciones en los genes y su papel en la evolución molecular

Los pseudogenes representan genes que perdieron su capacidad funcional y que se mantienen en el genoma aunque no brinden ventajas selectivas al organismo en términos evolutivos; si bien hay que enfatizar que están en proceso de acumular mutaciones al grado de perder totalmente la identidad con la secuencia funcional de la cual provienen. Esta interpretación se hace considerando que la secuencia duplicada era funcional inmediatamente después de la duplicación. Posteriormente podría ser inhabilitada por mutaciones en codones que codifican aminoácidos funcionalmente irremplazables, inserciones y deleciones que generen cambios en los marcos de

lectura o mutaciones en regiones regulatorias y en sitios de procesamiento. Si la duplicación es incompleta, la secuencia sería un pseudogen *a priori* (Mighell et al. 2000; Harrison y Gerstein 2002).

Características y génesis de un retropseudogen.

Los retropseudogenes, originados por retrotranscripción y retrotransposición se caracterizan típicamente por la ausencia de promotores en el extremo 5', carencia de intrones, la presencia de repeticiones directas flanqueando la inserción conocidas como Secuencias con Molde Dirigido, TSDs (por sus siglas en inglés), y finalmente en algunos casos una secuencia de PoliA en el extremo 3' del retropseudogen. Los retropseudogenes son parecidos a los retrotransposones reinsertados en el genoma como una secuencia de doble cadena de DNA, generada a partir de una secuencia de RNA de cadena simple (ver Figura 1).

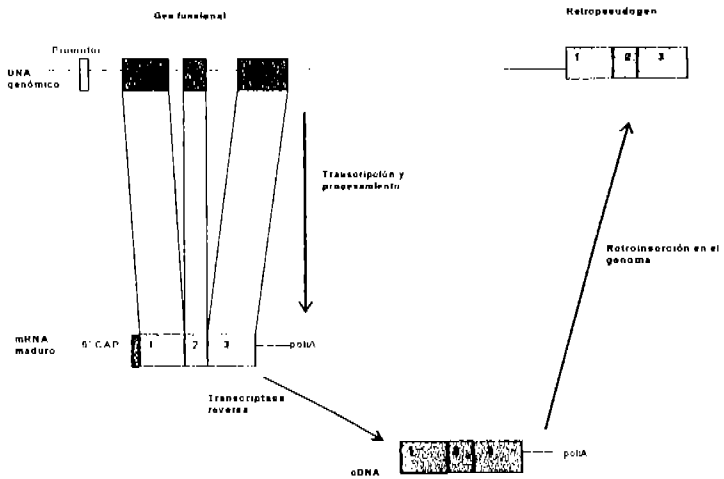


Figura 1. Mecanismo de formación de retropseudogenes. Los retropseudogenes son la consecuencia de la actividad accidental de la transcriptasa reversa que utiliza un mRNA como molde y genera una copia de DNA que se inserta en el genoma. Carecen de promotor e intrones y pueden contener restos de la cola de poliA. (Fuente: Valdés-López et al. 2004).

Generalmente los retropseudogenes no son funcionales desde el momento de su reincorporación al genoma (Ophir y Graur, 1997). Esto se debe a que las secuencias insertadas carecen de promotores. Por esta razón no se transcriben y por lo tanto no tienen presiones de selección sobre la secuencia, de tal manera que pueden empezar a acumular inserciones, deleciones y sustituciones que dan origen a corrimientos del marco de lectura y la aparición de codones de término prematuros (Strichman-Almashanu et al. 2003). Sin embargo, existen reportes de casos donde la secuencia retroinsertada queda en un ambiente genómico en el que existe un promotor y la secuencia puede llegar a ser expresada y seguir siendo funcional (Rogalla et al. 2000; Birger et al. 2001; Strichman-Almashanu et al. 2003; Kornev et al.; en Harrison et al. 2001). Si la secuencia es expresada se le denomina retrogen y algunos autores los denominan copias retropuestas (Almashanu, 2003).

Un ejemplo de un retrogen es el gen del Factor de Embarazo Temprano (EPF) en el genoma del ratón, reportado por Summers *et al.* (2001). Se trata de una secuencia sin intrones, que tiene tres diferencias a nivel de nucleótidos con respecto a la de la hsp10. Se sabe que el EPF es esencial para la iniciación y el mantenimiento del embarazo (Athanasas et al. 1989), pero su función no está limitada a la gestación. El EPF es secretado por células normales, transformadas y neoplásicas durante el crecimiento y la división celular; también se sabe que es requerido para el crecimiento celular. Aparte de estas funciones regulatorias en el crecimiento celular el EPF tiene una actividad inmuno-moduladora (Fletcher *et al.* 2001 en Barberán-Soler, 2002).

Se han descrito características generales, que debe cumplir un gen para generar retropseudogenes que logren fijarse en el genoma del hospedero (Devor et al. 2003; Gonçalves et al. 2000)

a) Deben de proceder de genes con un alto nivel de expresión. En realidad este no es un requisito como tal. Sin embargo, dado que la probabilidad de que un mRNA específico sea usado como sustrato por la transcriptasa

reversa depende de su abundancia, genes con un bajo nivel de transcripción tiene menos probabilidades de ser retrotranscritos.

b) Para que se fije en el genoma del hospedero y para que el retropseudogen sea heredado, debe aparecer en la línea germinal. Por supuesto, también se da la generación de retropseudogenes a nivel somático, pero que en este caso no serían heredados
c) Preferencialmente la retrotransposición se da en mRNAs cortos, considerando que el tamaño promedio de un mRNA es de 1000 pb aproximadamente.
d) Los retropseudogenes se generan a partir de genes con bajo contenido de GC (Barberán-Soler, 2002; Gonçalves et al. 2000).

Así, quizá los hechos más trascendentes para la generación de retropseudogenes dependen de las probabilidades independientes de retrotranscripción del mRNA maduro, su transferencia al núcleo y la final reincorporación al cromosoma.

Características Generales de las Proteínas de Choque Térmico de 60 y 10 kDa

El plegamiento de las proteínas ha sido visto como el proceso resultante de las propiedades inherentes de la estructura primaria de los polipéptidos (Anfinsen, 1973). En algunos casos sin embargo, se requiere la participación de otras proteínas para lograr un plegamiento correcto y el subsecuente ensamblaje de los oligómeros (Hemmingsem et al. 1988). Estás moléculas auxiliares son conocidas como chaperonas moleculares o asistentes de plegamiento. Una subfamilia de estas son las chaperoninas: las heat shock proteins: la hsp60 y hsp10.

Las chaperoninas son requeridas para un crecimiento normal, demostrado por el hecho de que no existen mutantes de *E. coli* para los genes de las chaperoninas en un rango de temperatura de 20 a 43°C. Estos genes se

sobreexpresan por estrés, ayudando a estabilizar o proteger de la denaturalización a polipéptidos bajo condiciones de choque térmico (Prasad y Stewart, 1992). Las chaperoninas de tipo I se encuentran presentes en eubacterias, mitocondrias y cloropastos y requieren la acción concertada de dos polipéptidos, la hsp60 y la hsp10, que integran un gran oligómero de 14 subunidades de hsp60 y siete de hsp10. Las chaperoninas de tipo II se encuentran en el citoplasma de eucariontes y en arqueobacterias y el oligómero está constituido por un solo tipo de polipéptido homologo a la hsp60.

En condiciones de estrés fisiológico se da una desnaturalización de las proteínas dentro de la célula. Bajo estas condiciones, se piensa que las proteínas de choque térmico, actuando como asistentes de plegamiento, disminuyen el daño celular. Una gran cantidad de proteínas mitocondriales han sido observadas asociadas a la hsp60-10 a temperaturas elevadas (Martin et al. 1992, en Langer y Neupert, 1996). También se ha demostrado que el mecanismo de replegamiento mediado por la hsp60-10 es dependiente de ATP (Langer y Neupert, 1996).

La chaperonina de 10 kDa, hsp10, cpn10 ó groES en bacterias, es un oligómero en forma de anillo, de entre 6 a 8 subunidades idénticas, mientras que la hsp60, cpn60 ó groEL en bacterias, conforma una estructura que contiene 2 anillos apilados, cada anillo con 7 subunidades idénticas (Hemmingsem et al. 1988). Estas estructuras en forma de anillo se ensamblan en presencia de Mg^{2+} y ATP. La cavidad central del tetradecámero cilíndrico de la hsp60 da un ambiente aislado que favorece el plegamiento de las proteínas, mientras que la hsp10 se une a la hsp60 y sincronizan la liberación de la proteína plegada en una dependencia de Mg^{2+}–ATP (Schmidt et al. 1992; Prasad y Stewart, 1992). La unión de la hsp10 con la hsp60 inhibe la débil actividad de ATPasa de la hsp60.

3. Objetivo General.

Un motivo para estudiar a los retropseudogentes de las chaperoninas mitocondriales es que se trata de secuencias conservadas en prácticamente todo ser viviente (eubacterias, arqueobacterias y homólogos citosólicos en eucariontes), además de las características de expresión de las mismas.

En el 2002, Barberán-Soler, llevó a cabo su tesis de licenciatura en la cual describe los retropseudogenes de la hsp10 y la hsp60 en el genoma del humano. Además de la búsqueda, realizó la calibración de un reloj molecular para calcular la edad de cada retropseudogen.

En el presente trabajo se busca identificar los retropseudogenes de las chaperoninas mitocondriales del genoma del ratón, realizar la datación de cada retropseudogen y ver si existe algún tipo de patrón temporal distinguible. Hemos mencionado que la expresión de las chaperoninas se vincula a estrés ambiental, choque térmico, respuesta inmune, etc. (Langer y Neupert, 1996; Hemmingsem et al. 1988; Coates, 1996). Como hipótesis de trabajo proponemos que el patrón cuantitativo de expresión de los genes funcionales podría variar evolutivamente y estar vinculado a respuestas ambientales. En función de esta premisa, la probabilidad de que aparezca un retropseudogen estará en función del nivel particular de expresión del gen, y este podría variar en el tiempo. Dicho de otro modo, de igual manera que los niveles de expresión de un gen pueden variar en el desarrollo ontogenético de un organismo o presentar diferencias de expresión en diferentes tejidos del adulto, evolutivamente un organismo puede adaptarse a presiones del medio ambiente aumentando o disminuyendo los niveles de expresión de genes específicos. El estudio de los niveles de expresión en celulas y tejidos se han realizado usando principalmente hibridación de ácidos nucleicos *v. gr.* Experimentos de tipo "Northern" y actualmente usando los llamados microarreglos. En este sentido, los retropseudogenes podrían ser considerados como paleo-microarreglos, con la diferencia de que la hibridación

de secuencias se lleva a cabo de manera virtual en las bases de datos. Resultados de nuestro grupo de trabajo efectivamente sugieren que los retropseudogenes podrían ser una ventana molecular para certificar dichos eventos.

Objetivos particulares.

- Identificar las coordenadas cromosómicas de los genes *hsp60* y *hsp10* en el genoma del ratón.
- Confirmar la estructura genómica de los genes funcionales de las chaperoninas mitocondriales.
- Caracterizar e identificar a los retropseudogenes y pseudogenes de las chaperoninas mitocondriales en el genoma del ratón.
- Datar a los retropseudogenes.
- Analizar el potencial de codificación de los retropseudogenes.
- Buscar el EPF en el genoma del ratón.
- Analizar posibles causas de la generación y ubicación de los retropseudogenes en el genoma del ratón.
- Contrastar los resultados entre los retropseudogenes de las chaperoninas mitocondriales del humano y el ratón.

4. Metodología

Caracterización de los retropseudogenes

Para realizar la búsqueda inicial de retropseudogenes de *hsp10* y *hsp60* en el genoma del ratón se llevó a cabo en la versión 14.30.1 del genoma del ratón en el sitio de Ensembl (http://www.ensembl.org), mediante un BLASTn (Altschul *et al.*, 1990) genómico en la misma base de datos del Ensembl y en el NCBI (http://www.ncbi.nih.gov) con un valor E de 0.01, además de utilizar un filtro para secuencias simples.

Nuestras secuencias iniciales de búsqueda fueron el cDNA de la hsp10 y de la hsp60 con los números de acceso ENSMUST00000043474 y ENSMUST00000027123 respectivamente, de la base de datos del Ensembl, que se utilizaron como 'query' en el BLASTn para la búsqueda genómica de dichas secuencias.

Para buscar ψ utilizando el cDNA de los genes de origen, permite realizar una búsqueda centrada en retropseudogenes, ya que como se mencionó anteriormente, se trata de secuencias reincorporadas al genoma a partir de un mRNA maduro, por lo que el cDNA del gen funcional es la secuencia idónea para ejecutar estas búsquedas. Si utilizáramos el gen completo no tendríamos sensibilidad de búsqueda, al menos no para retropseudogenes. Para efectuar la búsqueda de pseudogenes no procesados se realizó el mismo procedimiento pero en lugar del cDNA se utilizó toda la secuencia genómica de la hsp60 y la hsp10 (incluidos los intrones).

Con los resultados del BLAST se genera el primer panorama general sobre la ubicación cromosómica de cada alineamiento resultante. El siguiente paso es revisar uno a uno los alineamientos resultantes y depurar los que formen parte de una misma secuencia pero que el BLAST divida en varios

independientes, dada la característica de que este programa busca el mejor alineamiento local por lo que una inserción puede dividir a una secuencia real en múltiples. Con el siguiente ejemplo se ilustra la manera de depurar los resultados del BLAST:

[A] [S] [C] Chr6	Chr6:39234945-39235144 +/-	423	2.5e-19
[A] [S] [C] Chr6	Chr6:39235141-39235225 +/-	235	2.5e-19

En la primera columna tenemos accesos directos a la alineación [A], a la obtención de la secuencia [S], y la vista de la región cromosómica en la que se encuentra [C]. La segunda, nos dice la ubicación en las coordenadas cromosómicas, la polaridad de la secuencia (+/-), el valor del alineamiento y el valor E del mismo.

Si se observa el ejemplo anterior, se puede notar que tenemos ubicaciones colindantes en nuestros resultados. Por ejemplo las secuencias 6.39234945-39235144 y 6.39235141-39235225 ambas con polaridad (+/-) se pueden juntar en un solo alineamiento si unimos los extremos de 6.39235144 y de 6.39235141, lo que indica que existe una inserción de 3 pb que no permite un alineamiento continuo. A continuación, ambas secuencias son empalmadas *in silico* y se observa la congruencia del empalme en un alineamiento pareado, de la secuencia resultante con el cDNA.

Este tipo de criterios para depurar los resultados del BLAST ya ha sido utilizado y justificado por otros autores y por lo general el factor limitante es el tamaño de la inserción para considerar si se trata de una sola secuencia o dos. El criterio estándar es tomar en cuenta inserciones no mayores de 60 pb. Este razonamiento se basa en la demostración de que el 95% de los intrones en mamíferos son mayores a 60 pb (Lander et al. 2001, en Zhang et al. 2002; Zhang et al. 2004). Además, se puede revisar en el alineamiento si la inserción corresponde o no con la posición exacta de un intrón en el gen funcional. Adicionalmente, en nuestra depuración de resultados, se analizó la posición de

los "hits" por pares dentro de un mismo cromosoma, examinando a la secuencia adyacentes para verificar si se trata de la continuación de la secuencia anterior y no descartar la posibilidad de una inserción mayor entre ambos fragmentos. Con este método se pueden detectar inserciones mayores a 60 pb entre dos fragmentos.

Posterior a la localización y depuración de los alineamientos del BLAST, se obtuvieron las coordenadas cromosómicas de cada secuencia y la polaridad de la misma. Más adelante, con las coordenadas genómicas se bajaron dichas secuencias con un exceso de 200 pb en cada extremo, para tener mayor probabilidad de identificar repeticiones en los flancos de las secuencias y buscar posibles promotores y otras señales como colas de poliadenilación.

Cada secuencia obtenida de la base de datos, se almacenaron localmente y se procedió a convertir a todas las secuencias a la misma polaridad, 5' → 3' con el programa BioEdit (Hall, 1999). El objetivo de lo anterior es trabajar con todas las secuencias en una misma orientación, con lo cual alineamientos posteriores sean coherentes entre sí.

Con todas las secuencias en una misma polaridad se procede a hacer los alineamientos pareados de la secuencia genómica y el cDNA con el programa ClustalX v1.81 (Thompson et al. 1997). Este alineamiento permite obtener los porcentajes de identidad entre cada secuencia y el cDNA.

Los porcentajes de identidad obtenidos de los alineamientos pareados pueden considerar la longitud del alineamiento global o solo considerar el grado de conservación del hit (Identidad relativa). La otra forma de considerar el alineamiento, es ponderar la longitud del tamaño de ambas secuencias para calcular la identidad global entre ambas (Identidad Real). En este trabajo obtuvimos ambos valores ya que un resultado nos habla de la conservación de

la secuencia en general con respecto al cDNA y el otro nos dice el porcentaje del cDNA original que cubre la secuencia en cuestión.

Con cada una de las secuencias obtenidas alineada con su respectivo cDNA, se procedió a analizar cada secuencia para ver si existía evidencia de que dicha secuencia fuera un retropseudogen o un pseudogen y dentro de cada caso ver si existen duplicaciones de estos eventos o si efectivamente se trata de eventos independientes por cada copia con la que se cuente.

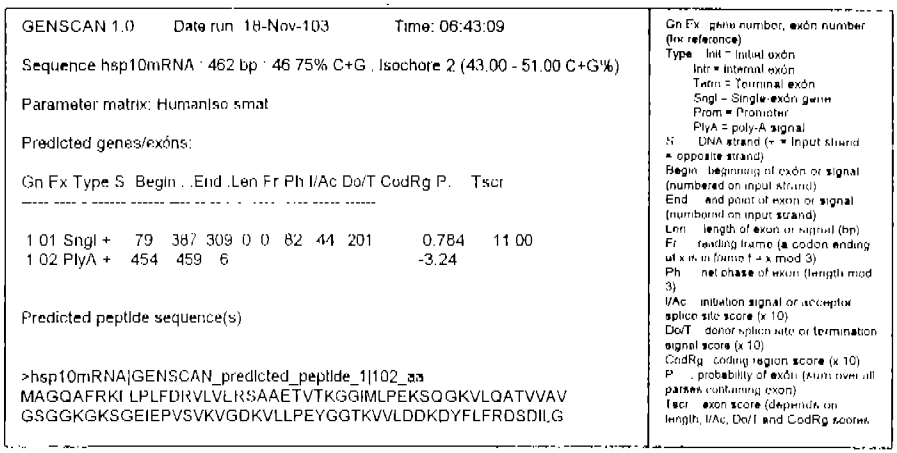
Para identificar si se trata de un retropseudogen se buscan e identifican algunos elementos exclusivos a eventos de retrotranscripción, como por ejemplo, ubicar los sitios de procesamiento unidos, o dicho de otra manera, encontrar una estructura de un gen ya procesado (solamente los exones) y observar la unión entre los mismos y buscar sitios de poliadeninación. La identificación de dichos elementos de retrotranscripción se realizó en el programa BioEdit, realizando un alineamiento pareado de la secuencia genómica con el cDNA ubicando a los exones del mismo. Los sitios de poliadeninación fueron analizados mediante el programa Genscan (Burge y Karlin, 1997).

Midiendo el potencial de transcripción

Otro elemento importante para la caracterización de retropseudogenes es la identificación de las deshabilitaciones que hacen que dichas secuencias pierdan su potencial de codificación. Para la búsqueda de estas características primero se realizó una inspección visual del alineamiento pareado con el cDNA, en BioEdit y se anotó si la secuencia se encuentra truncada hacia 3' o 5'.

Mediante Genscan es posible detectar el potencial de que una secuencia tenga la capacidad de transcribirse. Este es el programa que utiliza el Ensembl para la predicción de ORFs y además da estimados de la probabilidad de que una secuencia en cuestión sea codificante o no, independientemente de las

bases de datos de expresión (ESTs). Genscan busca potencial de transcripción en los 6 marcos de lectura posibles además de señales de promotores (ver Figura 2).



GENSCAN 1.0 Date run 18-Nov-103 Time: 06:43:09

Sequence hsp10mRNA : 462 bp : 46.75% C+G . Isochore 2 (43.00 - 51.00 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exóns:

Gn.Ex	Type	S	Begin	End	Len	Fr	Ph	I/Ac	Do/T	CodRg	P.	Tscr
1.01	Sngl	+	79	387	309	0	0	82	44	201	0.784	11.00
1.02	PlyA	+	454	459	6						-3.24	

Predicted peptide sequence(s)

>hsp10mRNA|GENSCAN_predicted_peptide_1|102_aa
MAGQAFRKI LPLFDRVLVLRSAAETVTKGGIMLPEKSQGKVLQATVVAV
GSGGKGKSGEIEPVSVKVGDKVLLPEYGGTKVVLDDKDYFLFRDSDILG

Gn.Ex : gene number, exón number (for reference)
Type : Init = Initial exón
Intr = Internal exón
Term = Terminal exón
Sngl = Single-exón gene
Prom = Promoter
PlyA = poly-A signal
S : DNA strand (+ = input strand; - = opposite strand)
Begin : beginning of exón or signal (numbered on input strand)
End : end point of exon or signal (numbered on input strand)
Len : length of exon or signal (bp)
Fr : reading frame (a codon ending at x is in frame f = x mod 3)
Ph : net phase of exon (length mod 3)
I/Ac : initation signal or acceptor splice site score (x 10)
Do/T : donor splice site or termination signal score (x 10)
CodRg : coding region score (x 10)
P : probability of exón (sum over all parses containing exon)
Tscr : exon score (depends on length, I/Ac, Do/T and CodRg scores)

Figura 2. Ejemplo de un resultado típico del programa GENSCAN. En la columna de la izquierda se puede observar la ventana de resultados y en la columna de la derecha se explican las abreviaciones utilizadas por el programa para denotar sus resultados.

Adicionalmente a Genscan, se llevó a cabo una búsqueda exhaustiva de ORFs en los 6 marcos de lectura para cada una de las secuencias, utilizando el programa Artemis v6.0 (Rutherford et al. 2000). Dicho algoritmo indica gráficamente la longitud de ORF, cada codón de inicio y término por cada marco de lectura, además de graficar el contenido de GC a lo largo de cada secuencia.

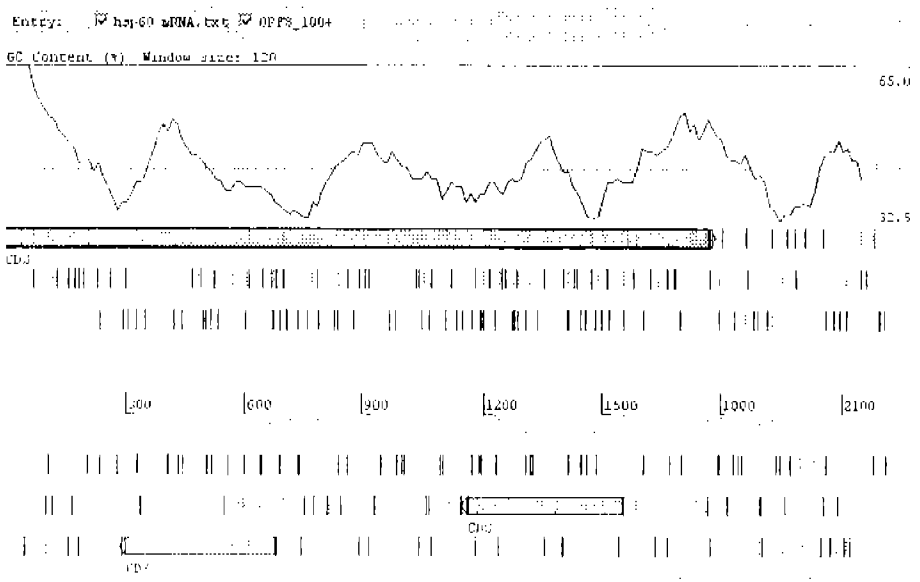


Figura 3. Resultado típico del programa Artemis, usando al mRNA de la *hsp60*. Se muestra la proporción de GC por regiones en la secuencia. Las siguientes tres líneas esquematizan los tres marcos de lectura en dirección 5' → 3' y en los renglones inferiores los tres marcos de lectura en sentido contrario. Las líneas verticales de color rojo identifican un posible codón de inicio, y las líneas negras un codón de término. En cuadros azules se muestran las regiones potencialmente codificantes y su longitud.

Definición práctica de retropseudogen

Los análisis anteriores ayudan a determinar que secuencias son retropseudogenes y cuáles pseudogenes. Algunos autores han generado los conceptos retropseudogen 'intacto' y fragmento pseudogénico. La diferencia estriba en que un retropseudogen 'intacto' conserva >70% del tamaño e identidad con la secuencia funcional y todo fragmento pseudogénico es aquella secuencia con identidad estadísticamente significativa a la secuencia funcional, pero que no cuenta con la longitud mínima estipulada anteriormente (Venter et al. 2001; Zhang et al. 2002). En este trabajo, consideramos que todo fragmento retropseudogénico es un retropseudogen, ya que también pueden existir

retropseudogenes altamente conservados pero de una longitud correspondiente a tan solo un par de exones del gen original, como mínimo.

No siempre es posible la obtención de alineamientos múltiples con este tipo de secuencias, debido a marcadas diferencias de tamaño entre los distintos fragmentos, y a que la distancia genética permitida se excede. En el caso de los retropseudogenes de la hsp10 fue posible realizar un alineamiento múltiple ya que la longitud de las secuencias no involucraba diferencias que excedieran los parámetros de los algoritmos de alineamiento. Por su parte la hsp60, una secuencia con un tamaño al menos 5 veces mayor que la hsp10, cuenta con retropseudogenes de una composición más heterogénea con respecto al tamaño, por lo que no fue posible realizar un alineamiento múltiple con todas las secuencias.

Más allá de la posibilidad real de alinear las secuencias o no, si partimos de la base teórica de que la generación de cada retropseudogen es un evento independiente, un dendrograma con la representación de los alineamientos del total de los pseudogenes no reflejaría necesariamente una relación de edades de los retropseudogenes. En este sentido, la calidad del análisis que podamos hacer de todos los retropseudogenes en conjunto podría ser equívoca, ya que dichos alineamientos y árboles nos agruparían a las secuencias por simple similitud sin que necesariamente exista un sentido evolutivo y biológico, además de que la distancia genética permisible por los algoritmos es claramente violada. Dado lo anterior, en este trabajo optamos por analizar a cada secuencia como un evento independiente, por lo que al momento de realizar los análisis de cada retropseudogen se obtienen tasas de sustitución individual.

La identificación de secuencias duplicadas se llevó a cabo mediante la inspección visual de los alineamientos pareados entre las secuencias de acuerdo al porcentaje de identidad real que existe entre dicha secuencia y el gen original. Además se analizan las repeticiones en los flancos de

aproximadamente 200 pb en cada extremo. Por último, se realizó una inspección de la vecindad genómica independientemente del cromosoma en donde se ubica la secuencia

Para verificar nuestros resultados de búsqueda se consideraron las anotaciones del Ensembl y del NCBI para revisar la existencia de retropseudogenes dentro de las chaperoninas mitocondriales, además de considerar a los miembros de las familias génicas de dichas proteínas. Para la discriminación de la probabilidad de generación de transcritos a partir de las secuencias aquí descritas, utilizamos las bases de datos de Etiquetas de Expresión de Secuencias (ESTs).

Datación de los retropseudogenes

Para realizar la datación de cada retropseudogen se calculó la identidad de cada secuencia con respecto a su gen funcional. Con este propósito, se estima el total de identidades, el total de mutaciones y el total de inserciones y/o deleciones existentes entre ambas secuencias. Para el cálculo de la distancia genética (d) se utilizó el modelo de sustitución de nucleótidos de dos parámetros de Kimura (1980) ya que este modelo cuenta con una corrección para dar el valor *d* tomando en cuenta las tasas de sustitución transicionales y transversionales, asumiendo que la frecuencia de los cuatro nucleótidos y las tasas de sustitución no varían a través de los sitios. El modelo de dos parámetros de Kimura, ha sido considerado por varios autores como la estimación más aproximada de tasas de sustitución en los pseudogenes y retropseudogenes (Li, 1997 y Zhang et al. 2002). Todos los cálculos de dichas distancias genéticas se realizaron en el programa MEGA2 (Kumar et al. 2000). La distancia genética por medio de este método se calcula conforme a la siguiente ecuación:

$$d = -\frac{1}{2}\ln(w_1) - \frac{1}{4}\ln(w_2)$$

Donde,

$$w_1 = 1 - 2P - Q$$
$$w_2 = 1 - 2Q$$

Siendo P y Q las frecuencias de los sitios con diferencias transicionales y transversionales (Nei y Kumar, 2000).

Con la distancia genética es posible conocer la tasa de sustitución por año entre la secuencia funcional y el retropseudogen (Kab), utilizando como grupo externo la secuencia funcional del humano (Kac, Kbc) y considerando un tiempo de divergencia entre el humano y el ratón de 85 millones de años (Barberán-Soler, 2002; Bromham et al. 1999). La fórmula empleada para el cálculo del número de sustituciones por sitio por año fue:

$$r = \frac{Kac + Kbc}{2t}$$

Con el propósito de hacer comparables los resultados de la calibración de nuestro reloj molecular con trabajos realizados por otros autores, decidimos utilizar otras tasas de sustitución empleadas en otros trabajos para el análisis de los retropseudogenes (Zhang et al. 2002; Barberán-Soler, 2002). Usar las tasas de sustitución empleadas en otros análisis no resulta trivial, ya que como estamos trabajando con secuencias no codificantes existe un fuerte debate sobre cómo calcular la tasa de sustitución que existe en este tipo de secuencias (Li, 1981 y Li 1997)

Las tasas de sustitución calculadas en este trabajo (Método I) se obtienen individualmente para cada retropseudogen, calculando la distancia genética (Kab) con respecto al gen funcional en el genoma del ratón, la distancia con respecto al gen funcional en el genoma humano (Kac) que sirve de grupo externo, y finalmente la distancia entre ambas secuencias funcionales en el ratón

y el humano (Kbc). La principal observación metodológica es que para obtener las distancias genéticas de cada alineamiento se considera a cada secuencia como un evento independiente, por lo que se genera una tasa de sustitución por cada uno de los retropseudogenes. Si bien el promedio de sustituciones por sitio, por año calculada de esta forma es de 2.58×10^{-9} contamos con la tasa de sustitución de cada una de las secuencias de los retropseudogenes.

Las tasas utilizadas de otros autores son de 3.5×10^{-9} sustituciones por sitio por año (Li, 1981), 2.55×10^{-9} (Barberán-Soler, 2002), 1.5×10^{-9} (Zhang et al. 2002). Una observación importante es que en los trabajos anteriores se trabaja con promedios globales de tasas de sustitución por lo que podría llegar a despreciarse en alguna forma la dinámica de sustituciones de cada secuencia particular.

5. Resultados y Discusión

La estructura genómica de las chaperoninas mitocondriales.

Se ha determinado en trabajos anteriores la estructura genómica y localización de los genes de las chaperoninas mitocondriales en humano y en rata. En ambos casos se encontró que la estructura de *hsp60* y *hsp10* contiene intrones y se encuentra en una configuración frente a frente con un promotor bidireccional. Por lo cual se espera un arreglo similar en la estructura genómica del ratón (Hansen et al. 2003; Ryan et al. 1997).

Comprobación de la estructura genómica y localización de la *hsp60 y la hsp10* en el genoma del ratón.

El gen *hsp60* tiene el código de acceso **ENSMUSG00000025980** de la base de datos del Ensembl, se localiza en el cromosoma 1, en la cadena 3' – 5' a 55.7 **Mb** del centrómero, abarcando un total de 9.99 kb en dicha región. El transcrito de esta proteína cuenta con 13 exones que comprenden 2220 pb, generando un polipéptido de 574 residuos de aminoácidos.

La media del tamaño del exón es de 170.85 pb, mientras que el intrón promedio de este gen es de 647.17 pb. La estructura de los exones de este gen se muestra en la figura 4

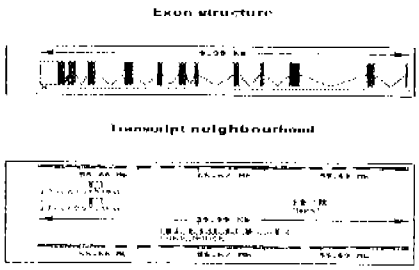


Figura 4. Estructura genómica de la *hsp60*. Tomado de www.ensembl.org

Por otro lado, el gen *hsp10* tiene el código de acceso **ENSMUSG00000038382.** Se localiza en el cromosoma 1, en la cadena 5' – 3' a 55.7 Mb, abarcando un total de 2.87 kb. El transcrito de esta proteína se compone de 4 exones que comprenden 466 pb, con una longitud de 129 residuos de aminoácidos en la traducción.

La media del tamaño del exón es de 116.5 pb, mientras que el intrón promedio de este gen es de 800.67 pb. La estructura de los exones de la HSP10 se muestra en la figura 5.

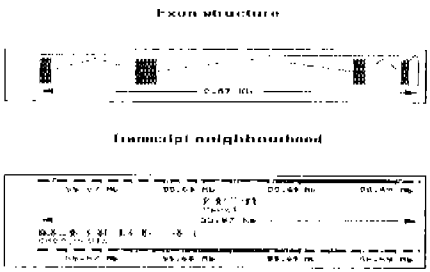


Figura 5. Estructura genómica de la *hsp10*. Tomado de www.ensembl.org

Ambos genes comparten los promotores y se encuentran en un mismo arreglo que en el humano y en el ratón, con una configuración frente a frente y contien intrones en su estructura genómica.

BLAST genómico y caracterización estructural de los retropseudogenes *hsp10 y hsp60*

En el BLAST genómico para la búsqueda inicial de los pseudogenes de *hsp10* y *hsp60* se encontraron un total de 51 hits para *hsp10* y 43 hits para *hsp60* en todo el genoma del ratón (Figura 6a y b).

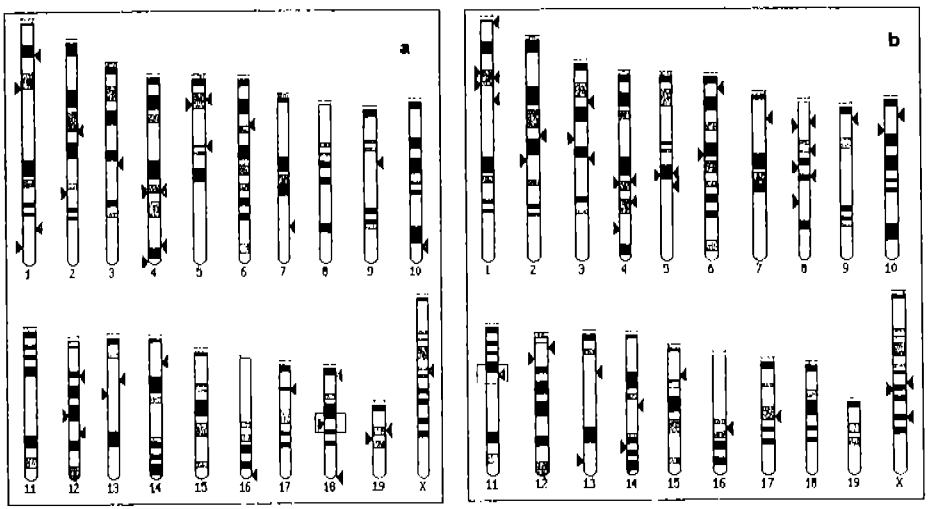


Figura 6. Resultados del BLAST genómico para identificar secuencias de *hsp10* (a), y *hsp60* (b), tomado de http://www.ensembl.org Puntaje de la búsqueda.

Un dato interesante es que en la búsqueda inicial ninguna secuencia, inclusive los cDNAs utilizados como controles, identificaron a los loci de los respectivos genes funcionales. Los cDNAs mostraron una identidad mayor (>95%), el *hsp60* con una identidad del 98.9% con una secuencia del

cromosoma 11, y *hsp10* con un 93.7% de identidad con una secuencia del cromosoma 18. El hecho de que el BLAST detecte la mayor identidad de los cDNA, con secuencias ubicadas en cromosomas distintos, a los del locus funcional, nos dio la pauta para identificar a las primeras dos retrocopias de las chaperoninas.

Del total de aciertos del BLAST ya depurados contamos con 41 secuencias reales para el caso de *hsp60* y con 28 secuencias para *hsp10*, mismas que serán las que se analizarán y se determinará si son retropseudogenes.

Con las secuencias depuradas, se procedió a realizar los alineamientos pareados entre los aciertos del BLAST y el cDNA de la secuencia funcional. En primera instancia se obtuvieron los porcentajes de identidad relativa y posteriormente la identidad real.

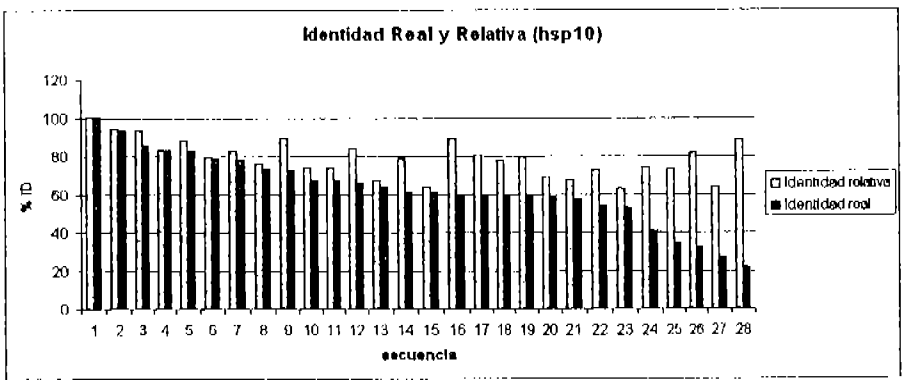


Figura 7. Valores de identidad real y relativas (ver texto) de las alineaciones pareadas de los retropseudogenes de ***hsp10*** con el mRNA funcional, en el genoma del ratón.

En la figura 7 se muestran los valores de identidad del alineamiento pareado para el caso de *hsp10*. La secuencia 1 en la figura es el cDNA, como control y referencia, el cual tiene ambos valores de identidad, relativa y real, del 100%, siendo el control, mientras que el valor más bajo de identidad real se

encuentra cercano al 21%, así como el valor más bajo de identidad relativa de dichas secuencias es del 63%.

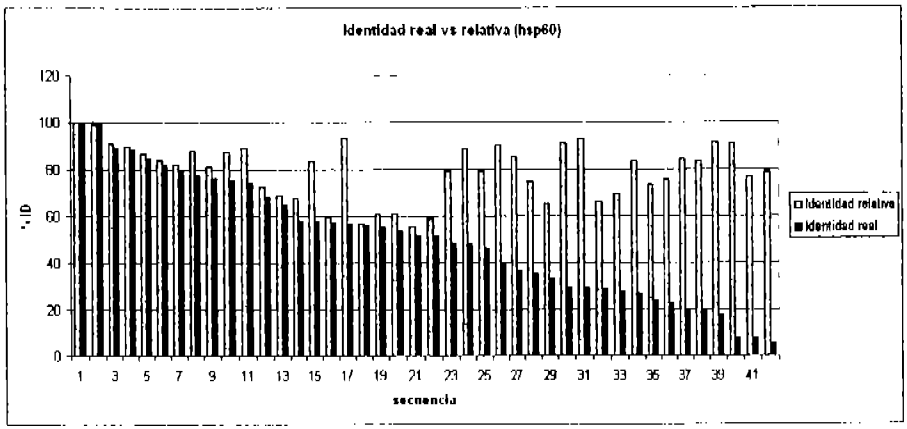


Figura 8. Valores de identidad real y relativas (ver texto) de las alineaciones pareadas de los retropseudogenes ***hsp60*** con el cDNA funcional, en el genoma del ratón.

En el caso de *hsp60* (ver figura 8) también analizamos la identidad relativa contra la identidad real, siendo el cDNA el máximo de identidad contra si mismo, con 100%. En *hsp60* se observa un claro descenso en la identidad real ya que siendo un cDNA funcional de 2,220 pb no se observan secuencias que conserven dicha longitud original además de una conservación de la secuencia en si, contrastante con el hecho de tener valores de identidad relativa por encima del 60% cuando la identidad real no rebasa el 10%.

Lo que expresan las figuras de identidad relativa y real es una característica inherente a nuestras secuencias; en el caso de *hsp10* no existen diferencias tan marcadas entre los porcentajes de identidad relativa y real, lo que significa que el tamaño entre las secuencias y el cDNA no difiere tan drásticamente como en el caso de las secuencias de *hsp60*. Esto puede ir ligado a que es más probable una retrotranscripción completa de un cDNA corto como *hsp10* (460 pb) que a un cDNA largo como *hsp60* (2,220 pb). Partiendo de los resultados anteriores, dedujimos que un alineamiento múltiple de las secuencias

no sería una buena opción con tamaños tan heterogéneos de secuencia. Sin embargo, con la hsp10 que muestra un tamaño más homogéneo de secuencias, nos aventuramos a realizar un alineamiento múltiple de las secuencias para poder inferir en base a éste, posibles duplicaciones y ver si un dendrograma (ver Fig. 8) de dichos alineamientos podría aclarar el panorama de las relaciones de las secuencias entre sí, estando conscientes de que el resultado muy probablemente no cuente con un significado biológico o evolutivo, sino más bien de relaciones de los porcentajes de identidad entre las secuencias.

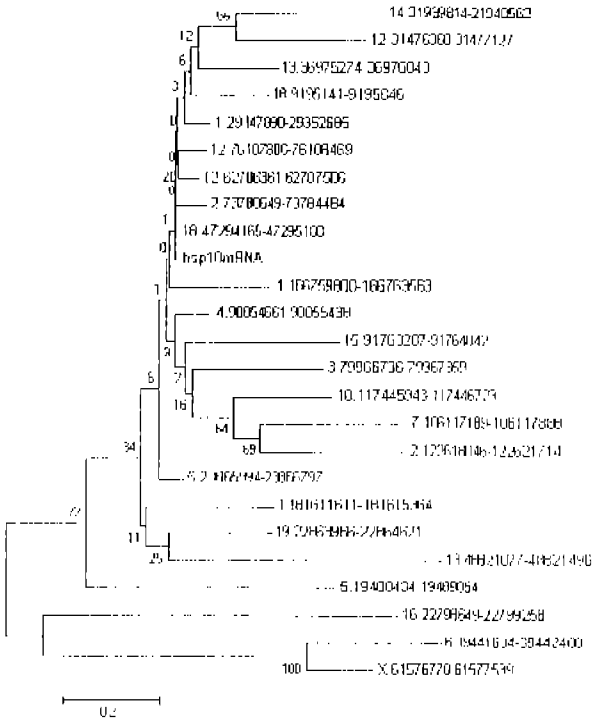


Figura 9. Árbol de Neighbor-Joining de los retropseudogenes ***hsp10*** encontrados en el genoma del ratón. Donde N es el cromosoma en el que se encuentra y P son las coordenadas cromosómicas de inicio y término de la secuencia reportada. Se utilizó el modelo de Kimura de 2 parámetros y un bootstrap de 1,000 repeticiones.

Las secuencias seleccionadas muestran evidencias de retrotranscripción como encontrar copias procesadas de al menos dos exones juntos sin inserciones mayores a 60 pb. Se tomó nota de cuáles exones del gen funcional eran los que se conservaban en las retrosecuencias. Hasta este paso solo se ha desechado una secuencia de *hsp60* que muestra conservada solamente la región 3' UTR, por lo que se elimina de análisis ulteriores. Para desechar duplicaciones génicas entre segmentos, se realizaron alineamientos pareados entre secuencias que presentaban un rango de identidad similar (± 5%), no encontrando pruebas fehacientes de ser duplicaciones génicas en ningún caso.

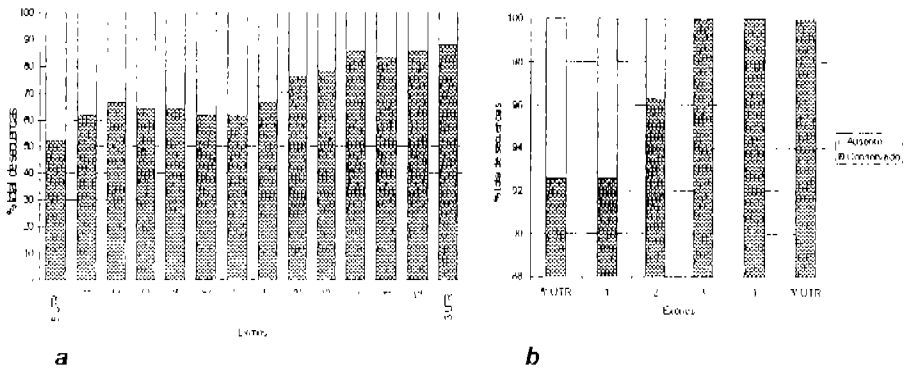


Figura 10. Grado de conservación de los retropsudogenes de *hsp60* (a) y *hsp10* (b) en el genoma del ratón. Se considera la presencia o ausencia del exón o región UTR correspondiente en el alineamiento pareado de la retrosecuencia con el cDNA funcional.

Graficando la conservación por exones, respecto del cDNA funcional, se puede observar en las retrosecuencias de *hsp10* que el extremo 3' se encuentra conservado en su totalidad a partir del exón 3 hasta la región 3' UTR, como se muestra en la figura 10. Por otro lado, si analizamos la conservación de las retrosecuencias de *hsp60,* se observa que existe una tendencia a mantener el extremo 3', ya que el 88% del total de las secuencias muestran intacta esta región, en contraste con la decreciente conservación hacia el extremo 5', donde

el porcentaje de las secuencias que conservan dichas regiones no pasa del 60% del total. Este hecho puede estar ligado a la manera en la que la retrotranscripción actúa, ya que la transcriptasa reversa se ancla en la región 3' del mRNA maduro y durante la retrotranscripción se puede desanclar de la secuencia, por lo que es más probable observar que los retropseudogenes conserven el extremo 3'.

Las diferencias en la forma en la que se está conservando una región entre las retrosecuencias *hsp10* y hsp60 obedecen a la diferencia de tamaños entre ambas secuencias. La *hsp60,* con un tamaño casi 5 veces mayor que *hsp10*, genera una variedad heterogénea de tamaños de retrosecuencias. Sin embargo, el patrón de tender a conservar el extremo 3' es similar en ambos casos, consistente con el mecanismo de retrotranscripción.

Con las identidades entre las secuencias y los cDNAs funcionales se realizó la búsqueda sistemática de elementos repetidos entre las secuencias más cercanas. Sin embargo, no fue posible encontrar algún par o más secuencias que tuvieran pruebas fehacientes de duplicación génica. Asimismo, no encontramos ningún pseudogen no procesado ni de *hsp10* ni de *hsp60*

La lista de retropseudogenes generada en este trabajo, junto con las coordenadas genómicas, porcentajes de identidad real y relativa, además de la conservación por exones con respecto al cDNA funcional se puede consultar en la tabla 1 para *hsp10* y en la tabla 3 para el caso de *hsp60*.

Ψhsp10

Secuencia (cromosoma-secuencia)	Base inicial	Polaridad	Identidad real	Identidad relativa	%GC	5' UTR	1	2	3	4	3' UTR
1-1*	29348090	+	88.93	59.09	37.82	1	1	1	1	1	1
1-3	166760000	+	82.69	78.57	46.09	1	1	1	1	1	1
1-4	181611811	-	80.77	59.09	42.21	1	1	1	1	1	1
2-1*	73780849	+	93.41	85.93	46.37	1	1	1	1	1	1
2-2	122618246	+	69.57	58.87	41.92	1	1	1	1	1	1
3-1	79966936	-	73.18	34.85	42.28	1	1	1	1	1	1
4-1	90054861	-	89.63	72.94	32.42	0	0	0	1	1	1
4-3	136776602	+	64.32	61.26	43.37	1	1	1	1	1	1
5-1	19488634	-	63.14	53.03	32.74	1	1	1	1	1	1
5-2	23866194	+	88.43	82.68	39.13	1	1	1	1	1	1
5-3	57237466	-	78.39	61.26	38.54	1	1	1	1	1	1
6-1	39441834	-	78.00	59.09	44.5	0	0	1	1	1	1
7-1	106117389	-	67.27	64.07	45.78	1	1	1	1	1	1
10-1	117446143	-	74.22	67.32	42.77	1	1	1	1	1	1
12-1	31476560	+	74.05	67.32	40.08	1	1	1	1	1	1
12-2	62706561	+	88.79	22.29	41.68	1	1	1	1	1	1
12-3*	76108006	+	83.51	83.33	43.13	1	1	1	1	1	1
13-1	36975474	+	84.30	66.23	42.55	1	1	1	1	1	1
13-2	48821077	-	63.96	27.27	38.38	1	1	1	1	1	1
14-1	21940014	+	79.30	58.87	36.21	1	1	1	1	1	1
15-1	91763487	-	72.81	53.90	40.98	1	1	1	1	1	1
16-1	22798849	-	82.16	32.90	48.1	1	1	1	1	1	1
18-1	9195341	-	76.23	73.59	38.92	1	1	1	1	1	1
18-2*	47294365	+	94.54	93.72	44.68	1	1	1	1	1	1
18-3	90059788		67.43	57.36	34.8	1	1	1	1	1	1
19-1	22864166	-	74.12	40.91	43.4	1	1	1	1	1	1
x-1*	61576970	+	79.39	79.22	37.17	1	1	1	1	1	1

Tabla 1. Secuencias de retropseudogenes de *hsp10* en el genoma del ratón y sus características generales como el porcentaje de identidad con respecto al cDNA y su caracterización de conservación por regiones del cDNA (1=presencia; 0=ausencia) *Reportado como parte de la familia génica en bases de datos.

Toda secuencia considerada en las tablas 1 y 2, es de un tamaño menor en términos absolutos que la secuencia de cDNA de la cual derivan. Si consideramos que existe evidencia de procesamiento en las secuencias y por el tamaño menor de las mismas con respecto al cDNA, podemos considerar como retropseudogenes a dichas secuencias.

Ψhsp60

Secuencia (cromosoma-secuencia)	Base inicial	Polaridad	Identidad real	Identidad Relativa	%gc	5' UTR	1	2	3	4	5	6	7	8	9	10	11	12	3 UTR
1-1	5072296	+	81.32	78.08	39.68	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1-2	45740418	+	90.42	39.10	36.6	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1-3	49949470	-	88.19	77.70	38.35	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1-4	87415989	+	58.61	51.67	37.64	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2-1	99551657	+	83.43	26.98	37.96	0	0	0	0	0	0	0	0	0	0	1	1	1	1
2-2*	80039757	-	91.21	29.91	43.06	1	1	1	1	1	1	1	0	0	0	0	0	0	0
3-1	33595384	+	60.94	53.56	42.65	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3.2*	62646333	-	81.89	80.05	39	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3-3	78753378	-	89.44	74.37	41.29	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4-1	89368466	+	79.16	48.60	36.69	0	1	1	1	1	1	1	1	1	1	1	1	1	0
4-2	125746302	-	66.15	28.87	46.34	0	0	1	1	1	1	1	1	1	1	1	0	0	0
4-3	87065684	-	72.74	68.65	44.08	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4-4	103852328	-	93.26	58.76	42.41	0	0	0	0	0	0	0	1	1	1	1	1	1	1
5-1	80547098	+	75.57	22.57	42.62	0	0	0	0	0	0	0	0	0	1	1	1	1	1
5-2	90952719	+	76.79	7.75	45.98	0	0	1	1	1	0	0	0	0	0	0	0	0	0
5-3	82620817	+	60.00	57.30	33.97	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6-1	65518403	+	78.88	46.44	43.43	0	0	0	0	0	0	0	1	1	1	1	1	1	1
6-2*	12351904	+	87.35	75.27	41.18	0	0	0	0	0	0	0	1	1	1	1	1	1	1
7-1	21838373	-	73.61	23.87	39.68	0	0	0	0	0	0	0	0	0	0	1	1	1	1
8-2*	8183802	+	61.00	55.32	40.41	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8-3	18630024	-	74.69	35.63	40.04	0	0	0	0	0	0	0	1	1	1	1	1	1	1
8-4	41088691	+	68.93	65.05	40.99	0	1	1	1	1	1	1	1	1	1	1	1	1	1
8-5	21913160	+	78.29	5.36	35.71	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8-6*	61654842		89.95	88.29	41.09	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9-1	12828233	+	89.08	48.15	38.94	0	0	0	0	0	0	0	0	1	1	1	1	1	1
10-1	15615851	-	56.95	56.08	41.39	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10-2	37238418	+	83.47	18.87	37.99	0	0	0	0	0	0	0	0	0	0	0	0	1	1
11-1*	41822615	+	98.92	98.78	43.48	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12-1	20214665	+	93.10	29.19	41.96	0	0	0	0	0	0	0	0	0	0	1	1	1	1
12-2*	117597161	-	83.56	57.70	39.63	1	1	1	1	1	1	1	1	1	1	0	0	0	1
13-1	34493785	+	84.37	18.96	35.33	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13-2	105032749	+	69.50	28.02	44.36	0	0	0	0	0	0	0	0	1	1	1	1	1	1
14-1	60655823	+	67.70	58.06	45.95	0	1	1	1	1	1	1	1	1	1	1	1	1	1
14-2*	92521716	+	91.28	89.10	41.42	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15-1	24282488	-	83.76	82.03	39.13	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16-1	61739599	+	85.43	36.98	40.12	0	0	0	0	0	0	0	0	0	0	1	1	1	1
17-1	47087800	-	91.05	7.79	47.89	0	1	1	0	0	0	0	0	0	0	0	0	0	0
x-1	74292047	-	55.67	51.94	40.29	1	1	1	1	1	1	1	1	1	1	1	1	1	1
x-2	79373485	+	86.78	84.59	38.29	1	1	1	1	1	1	1	1	1	1	1	1	1	1
x-3	101508479	-	65.58	33.51	35.2	0	0	0	0	0	0	0	0	1	1	1	1	1	1

Tabla 2. Secuencias de retropseudogenes de *hsp60* en el genoma del ratón y sus características generales como el porcentaje de identidad con respecto al cDNA y su caracterización de conservación por regiones del cDNA (1=presencia; 0=ausencia) *Reportado como parte de la familia génica en bases de datos.

Basados en el criterio de autores representativos en el área (Devor et al. 2003; Gerstein y Zhang, 2003; Zhang et al. 2002; Mounsey et al. 2002; Harrison et al. 2001) en el presente trabajo encontramos un total de 6 retropseudogenes completos y 34 fragmentos retropseudogénicos para *hsp60*, mientras que *hsp10* cuenta con 4 retropseudogenes completos y 23 fragmentos retropseudogénicos. El criterio bajo el cual se denominan dichas categorías es totalmente arbitrario y solo refleja vagamente la conservación de los retropseudogenes ya que el concepto de retropseudogen completo, según estos autores, representa únicamente un retropseudogen relativamente joven y conservado en una longitud equiparable al mRNA del cual procede. Si consideramos el criterio anterior como válido, despreciamos una enorme cantidad de información de retropseudogenes más antiguos y de 'fragmentos' que pueden revelar incluso más detalles de la historia evolutiva que un retropseudogen completo y joven.

En el presente trabajo se considera bajo la categoría de retropseudogen a todas y cada una de las secuencias que muestran evidencia de retrotranscripción, independientemente de la edad o nivel de identidad con respecto al cDNA del cual proceden. No tenemos otra limitante más que la búsqueda de identidad por parte de los algoritmos como el BLAST y contrastar con las anotaciones realizadas a la familia génica que estamos analizando en las bases de datos.

Análisis de causas y efectos de la generación de retropseudogenes *hsp10* y *hsp60*, en el genoma de *Mus musculus*.

Se ubicaron físicamente las secuencias de los retropseudogenes *hsp10* y hsp60 en sus respectivos loci, para posteriormente tratar de analizar dichas posiciones y ver si muestran algún arreglo aleatorio o no, así como ver si existe alguna relación entre la distancia a regiones como el centrómero y la ubicación de dichas secuencias. En la figura 11 se muestra la ubicación física de las secuencias dentro de cada cromosoma, es conveniente recordar que todos los cromosomas del ratón son acrocéntricos y gráficamente representamos al centrómero como la posición 0 dentro de nuestras figuras.

Como consecuencia de la ubicación cromosómica, surge el interés de resolver si es que los pseudogenes muestran una distribución aleatoria en su distribución a través de todos los cromosomas o si existe un desequilibrio o preferencia por algún cromosoma en particular. Aunque en otros estudios (Harrison y Gerstein, 2002) se haya demostrado una correlación positiva entre el tamaño del cromosoma y la cantidad de retropseudogenes existentes, no se ha analizado de manera particular por cada grupo de retropseudogenes, sino que se trata de una aseveración generalizada. En nuestro caso, partiendo de una hipótesis nula de que los retropseudogenes muestran una distribución aleatoria de inserción en los cromosomas aplicamos una probabilidad esperada de Poisson (Kennedy y Neville, 1982).

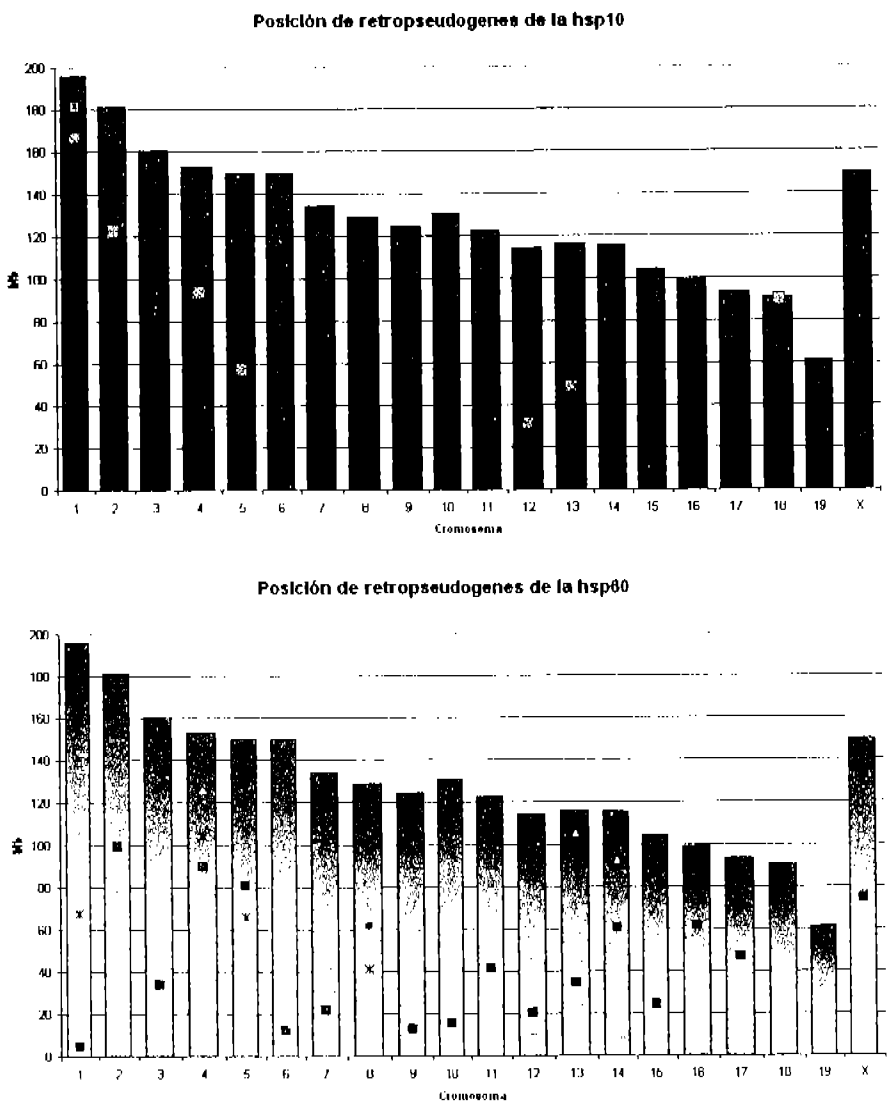


Figura 11. Ubicación en coordenadas cromosómicas de las secuencias reportadas en este trabajo para retropseudogenes *hsp10* y la *hsp60*. Es conveniente aclarar que es posible graficar su ubicación de esta forma debido a que todos los cromosomas en el ratón son acrocéntricos, y se representa el centrómero en la posición 0 y el telómero en el valor máximo por cromosoma.

Retropseudogenes *hsp10*						
Nc Número de copias por cromosoma	0	1	2	3		
Observados	4	9	2	5	**20**	**Σf**
Nc x No (número observado)	0	9	4	15	**28**	**Σfr**
P (Poisson)	0.2465	0.3452	0.2416	0.1127	**1.4**	**Np**
Esperados	4.6853	6.5594	4.5916	2.1427		

Tabla 3. Tabla de contingencia para analizar la aleatoriedad de la distribución de las secuencias de los retropseudogenes ***hsp10*** por cromosoma.

Resolviendo la χ^2 tenemos para *hsp10* una $\chi^2 = 0.7714$ (g.l. = 2, $\chi^2_{(\alpha=0.05)} = 5.99$), y para hsp60 una $\chi^2 = 1.4$ (g.l. = 4, $\chi^2_{(\alpha=0.05)} = 9.488$), por lo que en ambos casos se acepta la hipótesis nula y estamos observando una distribución aleatoria del número de retropseudogenes por cromosoma. Cabe mencionar que en ambos casos realizamos una corrección por frecuencia mínima de clase (Kennedy y Neville, 1982)

Retropseudogenes *hsp60*									
Nc Numero de copias por cromosoma	0	1	2	3	4	5	6		
Observados	2	6	5	2	2	0	1	19	**Σf**
Nc x No	0	6	12	6	8	0	6	38	**Σfr**
P (Poisson)	0.135	0.270	0.270	0.180	0.090	0.036	0.012	2	**np**
Esperados	2.571	5.142	5.142	3.428	1.714	0.685	0.228		

Tabla 4. Tabla de contingencia para analizar la aleatoriedad de la distribución de las secuencias de los retropseudogenes de la hsp60 por cromosoma.

Con los datos generados hasta el momento también podemos analizar si existe alguna región especial dentro de cada cromosoma en la que los retropseudogenes de las chaperoninas se incorporen con mayor frecuencia, ya sea más cercano al centrómero, en el telómero o en alguna región genómica

específica. Para este análisis consideramos un tamaño relativo de todos los cromosomas donde 0 es el centrómero y 1 es el telómero gracias a los cromosomas acrocéntricos del ratón es posible hacer esta aproximación, y correlacionar estos valores relativos con la ubicación de los retropseudogenes. Como se muestra en la figura 12, no existe correlación significativa en ninguno de los casos analizados.

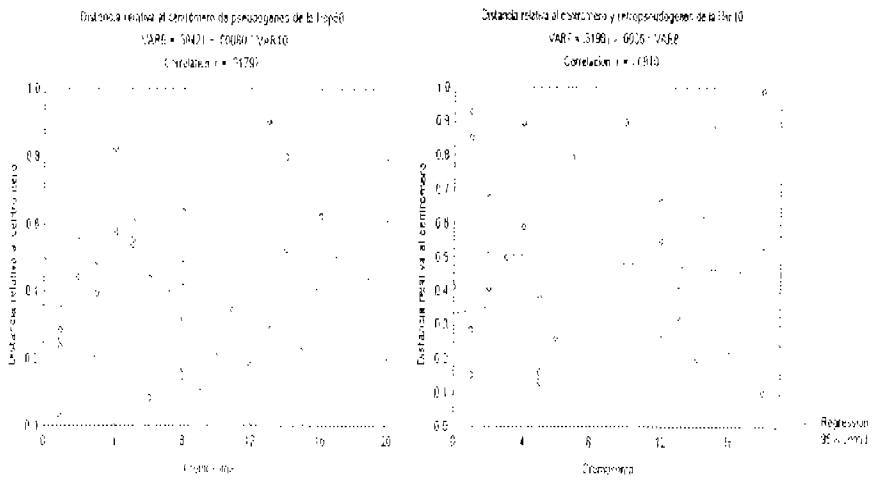


Figura 12. Análisis de la correlación entre la distancia relativa entre el centrómero y los retropseudogenes de la hsp10 (derecha) y la hsp60 (izquierda) por cromosoma.

Calibración y resultados del reloj molecular

Para el cálculo de la edad (ver tablas 5 y 6) de los retropseudogenes de las chaperoninas mitocondriales del ratón, se calibraron cuatro relojes moleculares con las tasas de sustitución descritas en la metodología; cada método es independiente. Fueron evaluadas las diferencias cuantitativas entre la estimación (ver figura 13), para cada una de las diferentes tasas de sustitución, fue evaluada y no existen diferencias significativas sea con las mismas entre el método de Barberán-Soler (2001), de Li (1980) el aquí empleado. Resultan diferencias significativas cuando comparamos cualquiera de los tres métodos antes mencionados con la tasa de sustitución empleada por Gerstein, et al. (2002) siendo una tasa constante propuesta por Li en 1997.

En el método de Gerstein se tiende a calcular las fechas más remotas, a diferencia de los otros tres. Si consideramos el reloj molecular con la tasa de sustitución de dicho método estaríamos hablando de que la mayor parte de los retropseudogenes de estas dos proteínas se generaron antes de la divergencia entre el humano y el ratón, situación que no parece lógica debido al alto porcentaje de conservación de un gran número de estas secuencias, que nos estaría hablando de un origen bastante reciente.

Consideramos que con el método empleado para la calibración de las tasas de sustitución y el reloj molecular en este trabajo, ganamos sensibilidad al momento de analizar cada retropseudogen como un caso particular. Si se trabaja con las estimaciones realizadas a partir de una tasa de sustitución promedio se desprecia toda la información sobre la dinámica de sustitución de cada retropseudogen en particular.

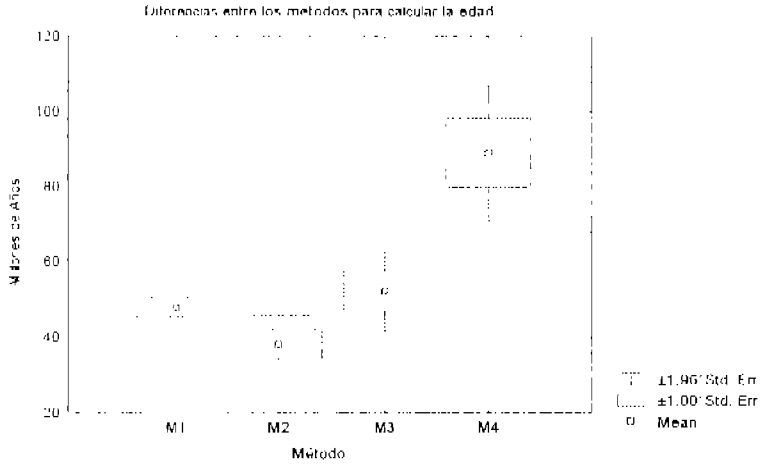


Figura 13. Diferencias cuantitativas entre las 4 tasas de sustitución empleadas para la calibración del reloj molecular

Partiendo del hecho de que no existen diferencias significativas entre los tres primeros métodos (ver Figura 13), de aquí en adelante consideraremos las edades ponderadas a partir del método utilizado en este trabajo (M1) para calcular las tasas de sustitución, considerando una tasa de sustitución independiente para cada retropseudogen.

En el caso de *hsp10* (ver tabla 5), observamos que la edad más reciente para sus retropseudogenes datan de hace 10-12 millones de años, mientras que los más antiguos detectados por esta metodología son de hace 64-67 millones de años. Con el propósito de visualizar claramente las fluctuaciones temporales en la aparición de retropseudogenes, organizamos los datos por categorías de 4 millones de años. El período donde más retropseudogenes se generaron para *hsp10* fue hace 54-57 millones de años (18.5% del total), aunque existen otros picos de generación. Así, en segundo lugar están las categorías que van de 36-39, de 47-49 y de 57-60 millones de años (14.8% del total por categoría) Adicionalmente, se aprecia un pico un poco menor hace 9-12 millones de años (11.1%) y un nivel basal (3.7% por categoría) de generación de pseudogenes en diversos períodos de tiempos intermedios entre las categorías. La categorización por edades de los retropseudogenes de *hsp10* se puede apreciar en la figura 14

	secuencia	Mmus Kab	Hsap Kac	MH Kbc	Tiempo (MA.1)	Tiempo (MA. 2) $r = 3.5 \times 10^{-9}$ (Li, 1981)	Tiempo (MA. 3) $r = 2.55 \times 10^{-9}$ (Barberan-Soler, 2002)	Tiempo (MA. 4) $r = 1.5 \times 10^{-9}$ (Gerstein, 2002)
Ψhsp10	1-1	0.094	0.163	0.114	28.845	13.429	18.427	31.333
	1-3	0.202	0.271	0.152	40.591	28.857	39.599	67.333
	1-4	0.228	0.259	0.095	54.746	32.571	44.695	76.000
	2-1	0.07	0.258	0.181	13.554	10.000	13.722	23.333
	2-2	0.391	0.462	0.17	52.587	55.857	76.649	130.333
	3-1	0.37	0.331	0.137	67.201	52.857	72.532	123.333
	4-1	0.113	0.174	0.1	35.055	16.143	22.152	37.667
	4-3	0.487	0.567	0.19	55.415	69.571	95.468	162.333
	5-1	0.504	0.574	0.118	61.908	72.000	98.800	168.000
	5-2	0.126	0.233	0.178	26.058	18.000	24.700	42.000
	5-3	0.243	0.355	0.191	37.830	34.714	47.636	81.000
	6-1	0.264	0.286	0.118	55.271	37.714	51.753	88.000
	7-1	0.436	0.474	0.176	57.015	62.286	85.470	145.333
	10-1	0.319	0.373	0.181	48.944	45.571	62.534	106.333
	12-1	0.319	0.397	0.16	48.680	45.571	62.534	106.333
	12-2	0.122	0.491	0.52	10.257	17.429	23.916	40.667
	12-3	0.187	0.251	0.184	36.540	26.714	36.658	62.333
	13-1	0.179	0.214	0.11	46.960	25.571	35.090	59.667
	13-2	0.471	0.537	0.105	62.360	67.286	92.331	157.000
	14-1	0.249	0.265	0.11	56.440	35.571	48.812	83.000
	15-1	0.342	0.376	0.117	58.966	48.857	67.043	114.000
	16-1	0.205	0.272	0.112	45.378	29.286	40.187	68.333
	18-1	0.282	0.383	0.192	41.687	40.286	55.281	94.000
	18-2	0.055	0.203	0.197	11.688	7.857	10.782	18.333
	18-3	0.429	0.481	0.141	58.625	61.286	84.098	143.000
	19-1	0.322	0.315	0.11	64.400	46.000	63.122	107.333
	x-1	0.241	0.236	0.302	38.076	34.429	47.244	80.333

Tabla 5. Calibración del reloj molecular para los retropseudogenes de la hsp10.

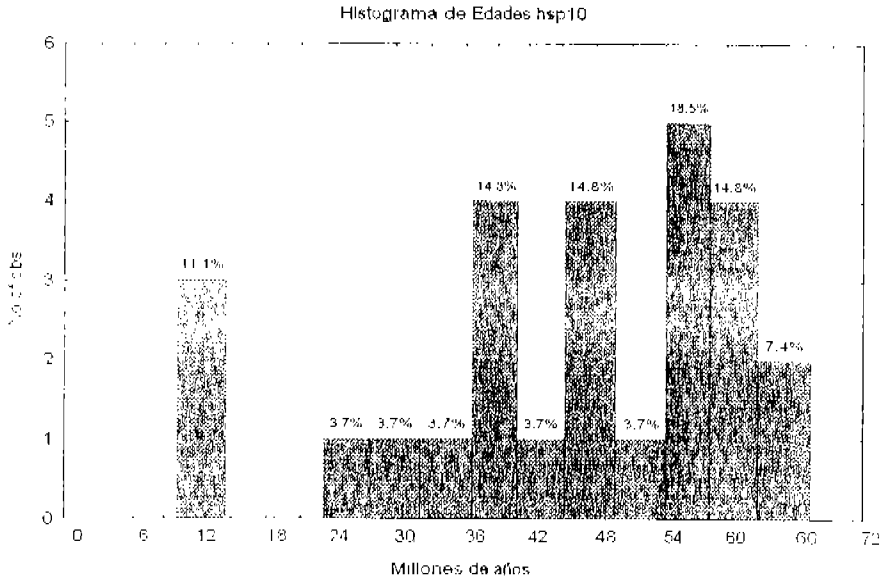


Figura 14. Histograma de edades, en millones de años, de los retropseudogones de *hsp10* en el genoma del ratón.

Por su parte las edades de la *hsp60* (ver tabla 6), oscilan entre los 4 y 68 millones de años. En este caso, la mayor parte de los retropseudogenes se generaron de manera más temprana que en el caso de la hsp10. El mayor pico de generación de retropseudogenes se dio hace 65-68 millones de años (15% del total). Las siguientes categorías con importancia en la generación de retropseudogenes datan de hace 54-61 millones de años considerando que son dos categorías y juntan al 25% del total de retropseudogenes Otro nivel menor de expresión se dio en dos períodos distantes entre si hace 44 y 63 millones de años aproximadamente (10% del total por categoría). En las subsecuentes categorías se observa un nivel basal de generación de retropseudogenes. De manera global y a diferencia de los retropseudogenes de *hsp10*, en este caso se observa una tendencia a una disminución de retropseudogenes a lo largo del tiempo lo cual indica una disminución gradual del nivel de expresión de *hsp60*. Estos resultados se representan gráficamente en la figura 15.

Ψ*hsp60*

secuencia	Mmus Kab	Hsap Kac	MH Kbc	Tiempo (MA,1)	Tiempo (MA, 2) $r = 3.5 \times 10^{-9}$ (Li, 1981)	Tiempo (MA, 3) $r = 2.56 \times 10^{-9}$ (Barberan-Soler, 2002)	Tiempo (MA, 4) $r = 1.5 \times 10^{-9}$ (Gerstein 2002)
1-1	0.218	0.254	0.101	52.197	31.143	42.735	72.667
1-2	0.104	0.176	0.117	30.171	14.857	20.387	34.667
1-3	0.128	0.187	0.099	38.042	18.286	25.092	42.667
1-4	0.595	0.639	0.097	68.716	85.000	116.639	198.333
2-1	0.187	0.257	0.099	44.649	26.714	36.658	62.333
2-2	0.108	0.184	0.094	33.022	15.429	21.171	36.000
3-1	0.553	0.592	0.095	68.421	79.000	108.406	184.333
3.2	0.211	0.257	0.098	50.521	30.143	41.363	70.333
3-3	0.114	0.196	0.1	32.736	16.286	22.348	38.000
4-1	0.245	0.286	0.102	53.673	35.000	48.028	81.667
4-2	0.448	0.569	0.115	55.673	64.000	87.822	149.333
4-3	0.332	0.395	0.1	57.010	47.429	65.083	110.667
4-4	0.072	0.151	0.109	23.538	10.286	14.114	24.000
5-1	0.301	0.347	0.11	55.985	43.000	59.006	100.333
5-2	0.282	0.291	0.076	65.313	40.286	55.281	94.000
5-3	0.575	0.655	0.102	64.564	82.143	112.719	191.667
6-1	0.252	0.284	0.1	55.781	36.000	49.400	84.000
6.2	0.14	0.199	0.186	30.909	20.000	27.445	46.667
7-1	0.322	0.343	0.109	60.553	46.000	63.122	107.333
8-2	0.55	0.603	0.1	66.501	78.571	107.818	183.333
8-3	0.307	0.351	0.094	58.640	43.857	60.182	102.333
8-4	0.403	0.441	0.099	63.435	57.571	79.001	134.333
8-5	0.273	0.278	0.119	58.451	39.000	53.517	91.000
8-6	0.107	0.16	0.102	10.551	15.286	20.975	35.667
9-1	0.119	0.194	0.104	33.943	17.000	23.328	39.667
10-1	0.639	0.717	0.099	66.563	91.286	125.265	213.000
10-2	0.177	0.245	0.097	43.991	25.286	34.696	59.000
11-1	0.011	0.103	0.102	4.561	1.571	2.156	3.667
12-1	0.073	0.142	0.097	25.962	10.429	14.310	24.333
12-2	0.189	0.246	0.104	45.900	27.000	37.050	63.000
13-1	0.178	0.178	0.07	61.008	25.429	34.894	59.333
13-2	0.391	0.47	0.094	58.927	55.857	76.649	130.333
14 1	0.431	0.484	0.101	62.624	61.571	84.490	143.667
14-2	0.093	0.16	0.1	30.404	13.286	18.231	31.000
15 1	0.186	0.241	0.101	46.228	26.571	36.462	62.000
16-1	0.164	0.239	0.098	41.365	23.429	32.149	54.667
17-1	0.095	0.145	0.066	36.270	13.571	18.623	31.667
x-1	0.658	0.76	0.106	64.584	94.000	128.989	219.333
x-2	0.147	0.194	0.098	42.791	21.000	28.817	49.000
x-3	0.467	0.485	0.098	68.087	66.714	91.547	155.667

Tabla 6. Calibración del reloj molecular para los retropseudogenes de la *hsp10*.

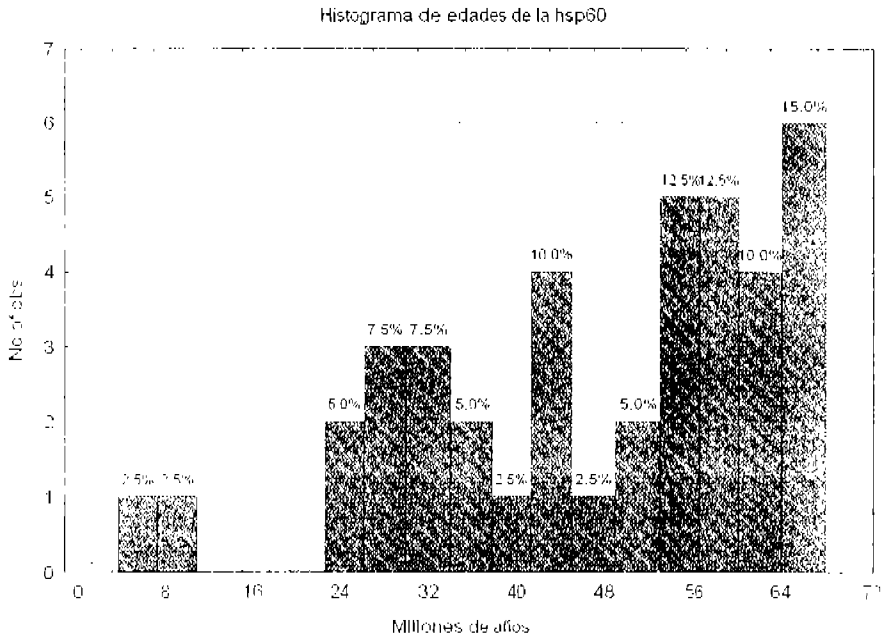


Figura 15. Histograma de edades, en millones de años, de los retropseudogenes de *hsp60* en el genoma del ratón.

El objetivo de datar las secuencias y presentarlas en agrupaciones por edad en un histograma, es tratar de visualizar la existencia de algún patrón de expresión diferencial a través del tiempo y analizar si en una ventana temporal la probabilidad de generar un retropseudogen fue mayor con respecto a otras. Con este tipo de análisis podemos realizar inferencias sobre posibles escenarios evolutivos que hayan requerido un aumento significativo en la síntesis de estas proteínas. En este caso, el aumento de la cantidad de chaperoninas moleculares podría auxiliar en el plegamiento correcto de otras proteínas. Como se ha mencionado, las chaperoninas son proteínas especialmente sensibles a cambios ambientales y a estrés inducido al organismo. Consecuentemente, un cambio en las condiciones de "normalidad" requerirá un aumento en la expresión de dichas proteínas, aumentando la tasa de transcripción y traducción de las mismas. Con el aumento de la tasa de transcripción se hace más probable que un mRNA

maduro de dichos genes se reincorpore al genoma, vía retrotranscriptasa, generando una retrocopia o retropseudogen de si mismo. Las diferencias en los picos de generación presentes entre los retropseudogenes de *hsp10* y *hsp60* pueden deberse, además de dichas condiciones de densidad a que las subunidades que componen el oligómero de las chaperoninas pueden tener funciones cualitativamente distintas y podemos esperar diferencias en la generación de retropseudogenes cuando no necesitamos una coordinación entre ambos genes para poder generar al oligómero funcional.

También se revisó si existía algún tipo de relación entre la edad de los retropseudogenes y sus coordenadas cromosómicas, observándose una ligera, más no significativa, correlación entre una edad menor y una cercanía al centrómero, como se aprecia en la figura 16.

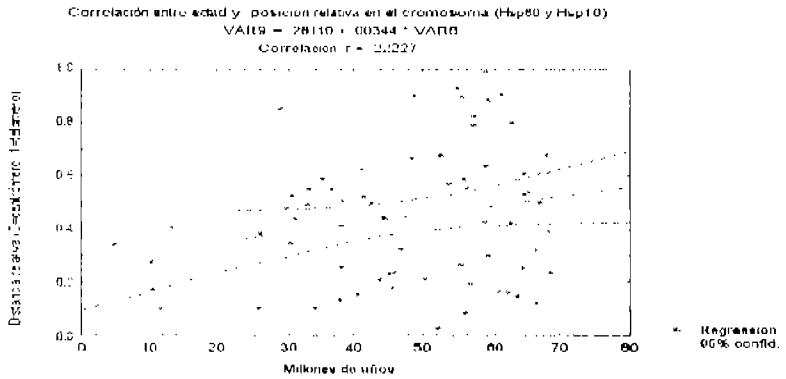


Figura 16. Análisis de la relación entre la edad, en millones de años, y la posición relativa del cromosoma de los retropseudogenes de *hsp10* y *hsp60*, se observa una ligera correlación positiva (r = 0.2222), aunque no os significativa.

El contenido de GC se analiza para ver si existe alguna relación con respecto a la edad de los retropseudogenes y también para ver si las coordenadas cromosómicas están influenciadas por el porcentaje de GC en la secuencia. En ambos casos se muestra una ligera correlación negativa, no

significativa, por lo que el contenido de GC en el caso de estos retropseudogenes no es un factor relacionado con la edad o con la posición dentro de los cromosomas, como se muestra en la figura 17.

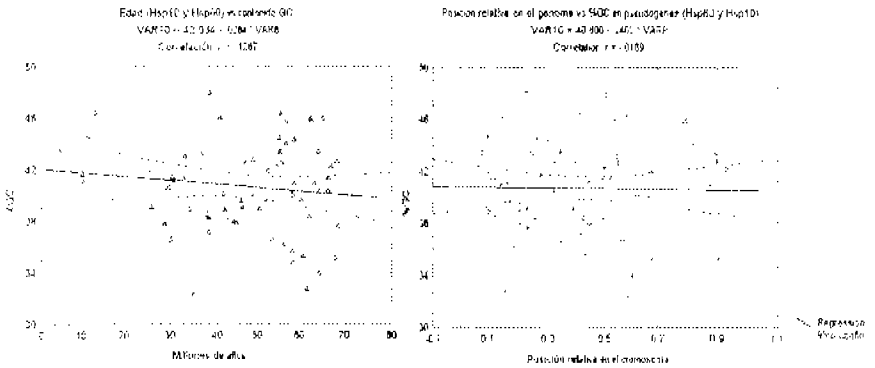


Figura 17. Análisis de la relación entre el contenido de GC y la edad de los retropseudogenes de hsp10 y hsp60 (izquierda), observándose una correlación negativa (r = -0.1287), sin llegar a ser significativa. A la derecha se observa la posición relativa de los retropseudogenes de hsp60 y hsp10 siendo analizadas contra el porcentaje de GC, observándose una correlación negativa (r = -0.0169) sin llegar a ser significativa.

Se considera como un paradigma en el análisis de este tipo de secuencias por diversos autores, que hay factores genéticos y genómicos determinantes de la aparición de retropseudogenes, como el contenido GC del gen funcional del cual provienen los retropseudogenes y la longitud del mismo en pares de bases. Para verificar la validez de dichas aseveraciones, se efectuó una búsqueda de retropseudogenes en la base de datos http://www.pseudogene.org donde tomamos como muestra a los 250 retropseudogenes de humano y ratón con el mayor número de retropseudogenes, tomando nota del porcentaje GC de la secuencia codificante a partir de la cual se originan y la longitud de la misma y tratando de correlacionar dichos valores con el número de retropseudogenes caracterizados. Como se esquematiza en la figura 18, no pudimos establecer una relación causa-efecto entre estas variables y la aparición de retropseudogenes.

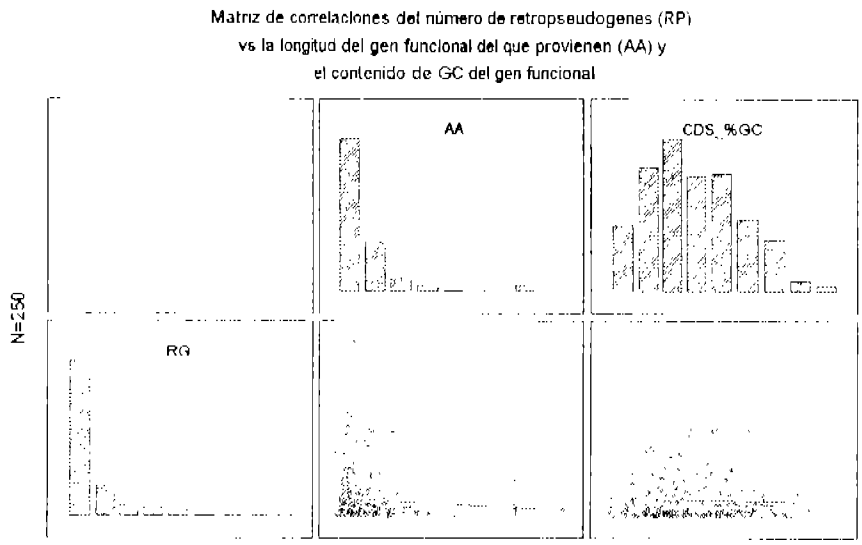


Figura 18. Correlaciones entre el número total de retropseudogenes (RG) contra la longitud en aminoácidos (AA) del cDNA del que derivan y el porcentaje de GC del cDNA original. Realizado a partir de la base de datos de Gerstein (http://www.pseudogene.org) contando con una N = 250 No se observa ninguna correlación significativa.

Si comparamos el patrón de generación de retropseudogenes con otros organismos es posible generar hipótesis evolutivas. Como se demostró que no existen diferencias significativas entre nuestro método para calibrar el reloj molecular y el utilizado por Barberán-Soler (2002), es posible comparar los resultados de ambos trabajos directamente, sin necesidad de estandarizar, por lo que estamos en posibilidad de comparar el patrón de generación de retropseudogenes de las chaperoninas mitocondriales entre los genomas del humano y el ratón.

Se observa que en el caso de la *hsp10,* el ratón ha tenido picos de generación de retropseudogenes mayores, en términos globales, en comparación con el humano, aunque también se observa que en los últimos 10 millones de años el hombre ha obtenido una intensa generación de

retropseudogenes de la *hsp10* y en el murino se observa un decaimiento de la generación de retropseudogenes en el mismo período (Ver Figura 19a).

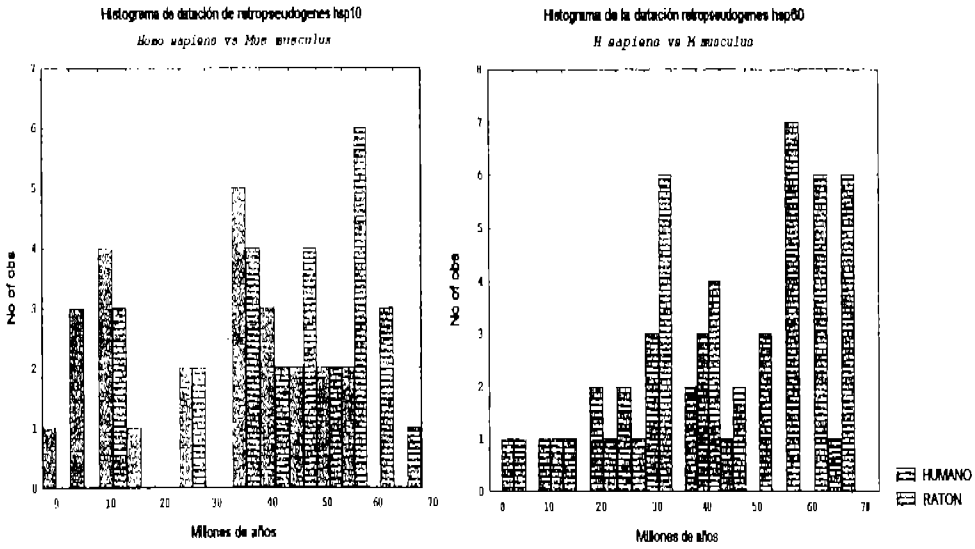


Figura 19. Comparación de las edades de los retropseudogenes de *hsp10* y *hsp60* presentes en el genoma del ratón y del humano.

En la generación de retropseudogenes de la *hsp60,* la tasa en el ratón es significativamente mayor que en el humano. Con un pico de generación de entre hace 30 y 60 millones de años, con un descenso claro en la actividad de retrotranscripción desde hace 20 millones de años, pero con un nivel basal de generación compartido con humanos (Ver Figura 19b). Por lo anterior, volvemos a corroborar que a diferencia de los retropseudogenes de *hsp10,* podemos constatar una disminución gradual del nivel de expresión de *hsp60* a lo largo del tiempo.

Potencial de transcripción de los retropseudogenes *hsp10 y hsp60*

En el presente trabajo trataremos de ahondar en otros aspectos de los retropseudogenes, como el analizar su potencial de transcripción y las implicaciones que este hecho conlleva. Para medir el potencial de transcripción no existe otro método más confiable que realizar la búsqueda experimental *in vivo* o *in vitro* para comprobar dicha expresión. Aquí utilizamos los métodos *in silico* para detectar el potencial de expresión de las secuencias aquí reportadas.

Mediante el programa Genscan (Burge y Karlin, 1997) es posible rastrear en una secuencia específica, la existencia de señales como promotores y otros rasgos como señales de procesamiento, en todo marco de lectura posible. Además, podemos obtener la posible secuencia de residuos de aminoácidos que codificaría. Otro análisis realizado, fue la búsqueda extensiva de ORFs en los 6 marcos de lectura posible con el programa Artemis v6.0 (Rutherford et al 2000), aunque como ya se mencionó en metodología, consideramos únicamente a los ORFs probables en los 3 marcos de lectura con polaridad 5' → 3'. Con dicho análisis se obtienen las secuencias de los posibles ORFs.

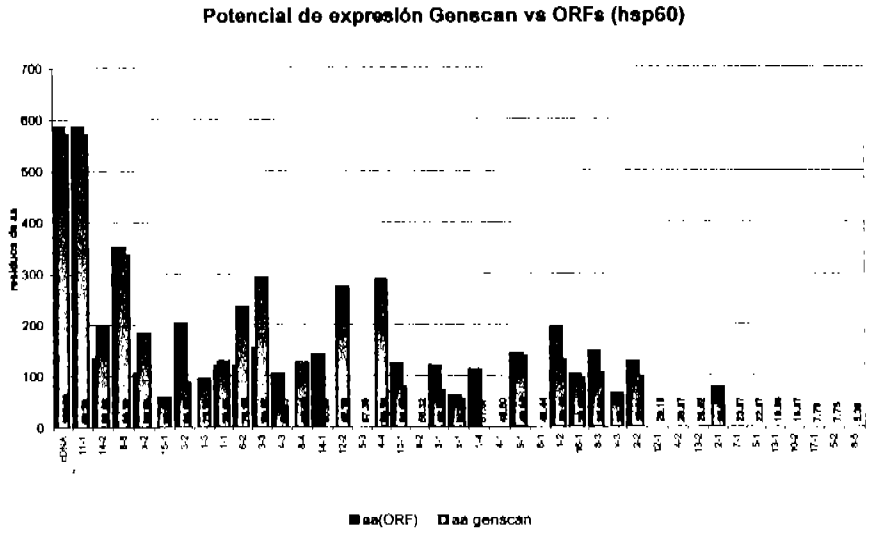
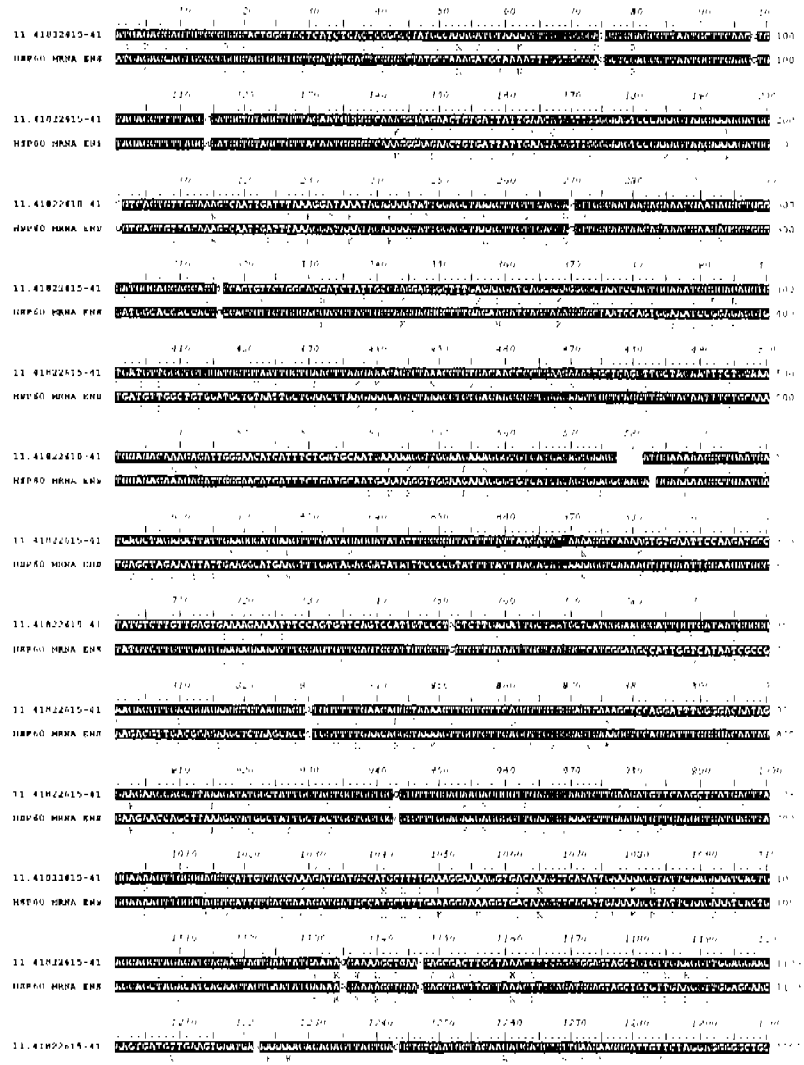


Figura 20. Estimaciones de la longitud de la posible secuencia codificante de los retropseudogenes de la hsp60 por la longitud de ORF y por Genscan.

En el caso de las retrosecuencias de *hsp60*, el cDNA se utilizó como control y Genscan es el método que estima exactamente la longitud del transcrito de este gen con 574 residuos de aminoácidos, mientras que con la búsqueda de ORFs se tiende a sobreestimar un poco el tamaño del transcrito, en el caso del mRNA, de 588 residuos de aminoácidos (ver Figura 20).

El retropseudogen ψ(*hsp60*)11-1 es el que tiene la mayor probabilidad de transcripción y si lo hiciera sería de 573 residuos de aminoácidos, uno menos que el gen funcional, al mismo tiempo que este retropseudogen muestra una identidad del 98.9% a nivel de nucleótidos con el cDNA, la más alta de todos los retropseudogenes reportados para *hsp60* y se estima que se generó hace unos 4.5 millones de años (ver Figura 21). Es posible que dicha retrocopia sea funcional ya que existen diversos reportes de ESTs que confirman su presencia (Kempe et al. 2004; Mootha et al. 2003; Danial et al. 2003; FANTOM Consortium y RIKEN Genome Exploration Research Group, 2002; Lotscher y Allison, 1990;

Venner y Gupta, 1990). Para entender si la secuencia ψ(*hsp60*)11-1 se encuentra bajo presiones de selección utilizamos el coeficiente R que calcula la tasa entre transiciones/transversiones obteniendo un valor global de R = 2.0, lo que nos indica que existe presión de selección, además de un valor R = 0.7 para las sustituciones en terceras posiciones, lo cual es indicativo de que la mayor cantidad de sustituciones son en la tercera posición.



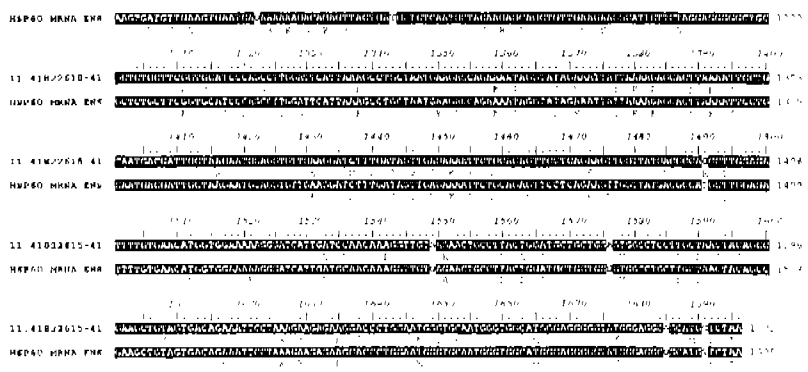


Figura 21. Alineamiento pareado entre el cDNA de *hsp60* y ψ(*hsp60*)11-1.

En los casos restantes de ψ*hsp60*, además de la anotación de los ESTs actuales, observamos que si los retropseudogenes fueran expresados serían de un tamaño menor que el gen funcional del cual proceden. Se puede llegar a pensar que 14 retropseudogenes ψ(*hsp60*) han perdido el potencial de transcripción y traducción, al menos de una secuencia de tamaño similar a la Hsp60, ya sea por corrimientos de marcos de lectura o codones de término prematuros, tanto por los resultados bioinformáticos como por la comparación de dichos resultados con las bases de datos.

En el caso de ψ*hsp10*, se encuentran cerca de 17 secuencias que perdieron totalmente la capacidad de codificar para algún polipéptido, al menos no detectable por los métodos aquí empleados (ver Figura 22), ni por ESTs. De las 11 secuencias restantes con posible capacidad de codificación, según métodos bionformáticos, cuentan con un tamaño menor o igual al mRNA de origen. De las 11 secuencias solo se tiene apoyo experimental de expresión para una secuencia: ψ*(hsp10)*18-2.

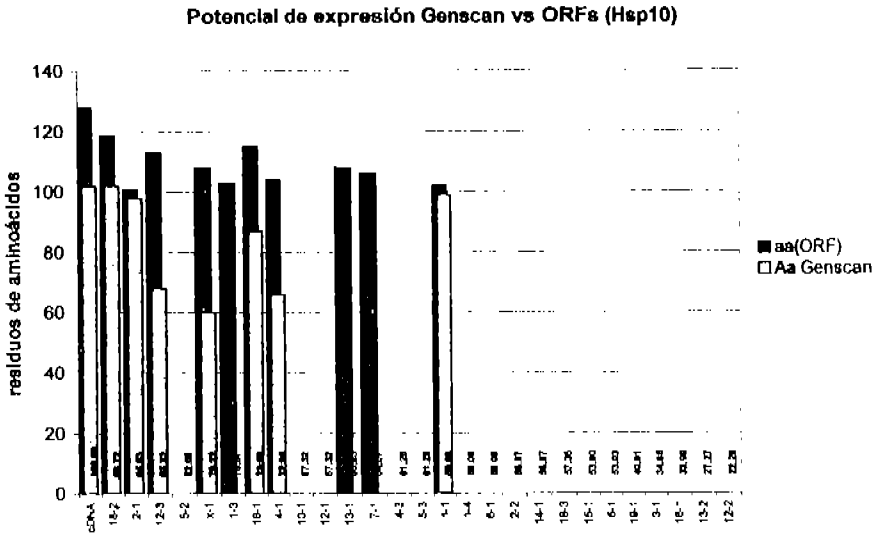


Figura 22. Estimaciones de la longitud de la posible secuencia codificante de los retropseudogenes de hsp10 por la longitud de ORF y por Genscan.

La *ψ(hsp10)*18-2 es la secuencia con el más alto índice de identidad (93.72%) al cDNA, con una edad aproximada de generación de 7.857 millones de años de antigüedad. Se estima una longitud aproximada de 119 a 102 residuos de aminoácidos, recordando que la longitud del cDNA es de 128 residuos, para el caso de *hsp10*.

La secuencia *ψ(hsp10)*18-2 está presente en las anotaciones de la base de datos de Ensembl como perteneciente a la familia génica de las chaperoninas mitocondriales, con la notación Hspe-rs1, misma que aparece con la entrada AF247846 en Genbank y que coincide con la secuencia reportada por Summers et. al, 1996, como el Early Pregnancy Factor (*EPF*), además de contar con ESTs que confirman la expresión de dicha secuencia (Mootha et al. 2003; FANTOM Consortium y RIKEN Genome Exploration Research Group, 2002; Fletcher et al. 2001). Como ya se mencionó, el *EPF* es una secuencia sin intrones con

mínimas diferencias a nivel de nucleótidos con la secuencia de *hsp10* con 93.72% de identidad, considerando 5'UTR y 3'UTR. En la figura 23 se esquematiza el alineamiento entre la Secuencia Codificante (CDS) del cDNA de la *hsp10* y el de la secuencia *ψ(hsp10)*18-2.

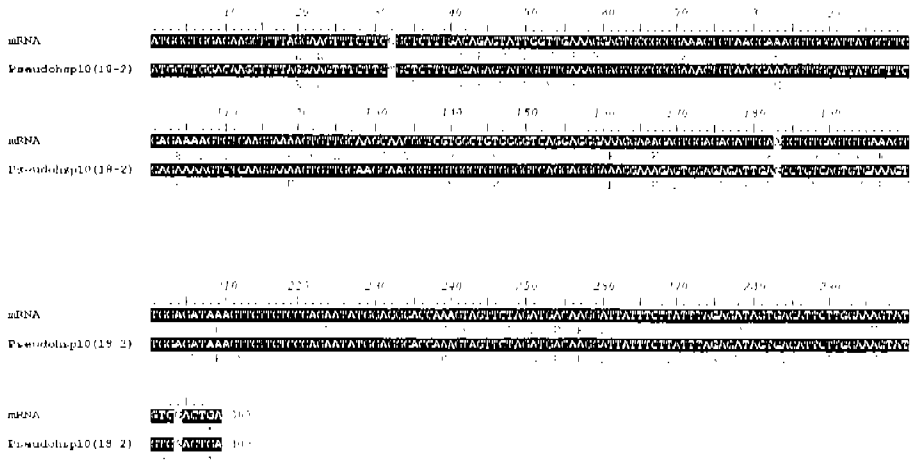


Figura 23 Alineamiento pareado entre el cDNA de *hsp10* y *ψ(hsp10)18-2*.

Partiendo de los resultados anteriores, podemos constatar que la secuencia *ψ(hsp10)*18-2 es el Early Pregnancy Factor, por ser una secuencia que en su estructura genómica no cuenta con intrones y con un 93.7% de identidad a nivel de nucleótidos, a nivel global de transcrito (Fletcher et al. 2001); podemos considerarlo como un retrogen que se origina a partir de un evento de retrotranscripción de un mRNA maduro de la *hsp10* y la reincorporación posterior de dicha secuencia al cromosoma 18 del ratón, de forma aleatoria conjugándose con un evento fortuito en el que la secuencia fue retroinsertada junto a un promotor potencial o un elemento "enhancer", generando un producto polipeptídico con una función alternativa y por lo cual su función no fue deletérea.

Cuando hablamos de retrogenes, tenemos que delimitar si son parálogos o bien parte de una familia génica al interior del genoma. Es difícil determinar la edad del evento de retrotranscripción por los métodos aquí empleados ya que la retrosecuencia es funcional y conserva su papel o ha adquirirído uno nuevo, de esta forma conserva presiones de selección desde el inicio del evento de retrotranscripción. La presión de selección deja de ejercer en las secuencias no funcionales y un proceso de deriva génica es el que gobierna la evolución de las secuencias en dicha situación, por lo que la tasa de sustitución se eleva en comparación con sus parálogos funcionales.

El efecto de la liberación de la presión de selección sobre los retropseudogenes se puede analizar bajo la lógica de que un evento de retrotranscripción reciente guarda una identidad total con del mRNA del cual deriva, dicha identidad va decreciendo en función del tiempo del evento de retrotranscripción. Si partimos de la hipótesis de que un retropseudogen joven guarda más probabilidad de transcribir un gen similar al del mRNA del que diverge, podemos aseverar que dicha probabilidad descenderá proporcionalmente a la tasa de mutación a la cual se encuentre sujeta. Podemos observar en la figura 24 la correlación negativa (r_1=-0.6433, p=0.000; r_2=-0.4146, p=0.008) que existe entre el potencial de codificación detectado por métodos bioinformáticos, contra la edad de los mismos; en otras palabras, los retropseudogenes jóvenes tienen mayor probabilidad de transcribirse y generar un polipéptido funcional.

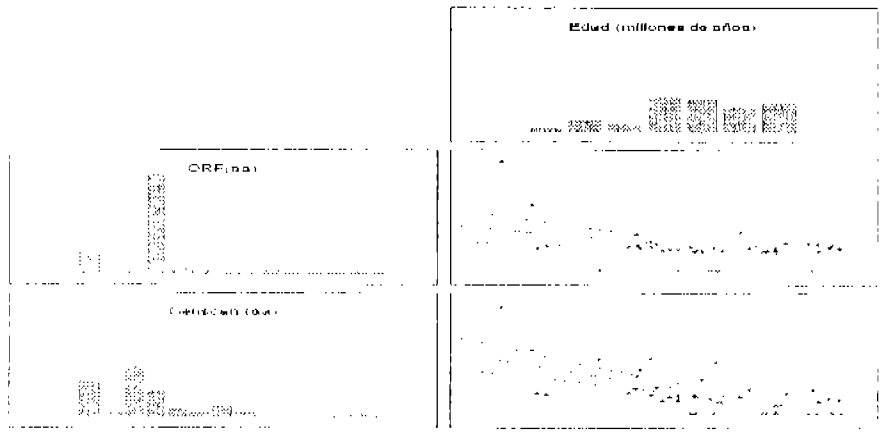


Figura 24. Correlación entre la edad de los retropseudogenes y el potencial de transcripción predicho por métodos bioinformáticos. Se observa una correlación negativa significativa en ambos casos (r_1=-0.6433, p=0.000; r_2=-0.4146, p=0.008).

Con base en el potencial de los retropseudogenes jóvenes para expresarse, se plantea un modelo de generación de familias génicas por este mecanismo. A mayor nivel de expresión de un gen tendremos un aumento en la concentración de su mRNA y la probabilidad de que una transcriptasa reversa lo copie a cDNA y se reincorpore al genoma de manera aleatoria. Por probabilidad, una retrosecuencia puede reincorporarse en el genoma en vecindad con promotores constitutivos o inducibles, por lo que podremos ver la generación de transcritos con un alto grado de identidad con su secuencia de origen. Ahora bien, esto tiene que ser un proceso inmediato a la retrotranscripción, de otra manera las mutaciones y la deriva génica actuarán sobre la secuencia agregando o quitando partes de la secuencia que originarán corrimientos del marco de lectura, paros prematuros en la traducción, etc.

En el caso concreto del EPF en el ratón, podemos verificar que es una secuencia que potencialmente se está transcribiendo y traduciendo con funciones cualitativas distintas a las de Hsp10, por lo que la edad calibrada para el EPF puede ser un artefacto ya que es probable que la secuencia exista de forma previa a la separación del linaje de los homínidos y los roedores, ya que

se ha reportado que dicho gen también se encuentra presente en el humano (Athanasas et al. 1989) por lo que es más probable una generación previa a la separación de los linajes, que una generación independiente en ambos linajes.

Con el EPF identificado en el ratón, realizamos una búsqueda en el genoma humano para tratar de identificar el locus ortólogo. Se consideran dos candidatos probables para ser el EPF en el humano, el primero se encuentra localizado en el brazo largo del cromosoma 16 con una identidad a nivel de nucleótidos del 81.22% y la otra secuencia en el brazo largo del cromosoma 5 con solo un 49.6% de identidad en nucleótidos. Estas aseveraciones se basan en una búsqueda con tBLASTx, y constatado con ESTs (ver Figura 25).

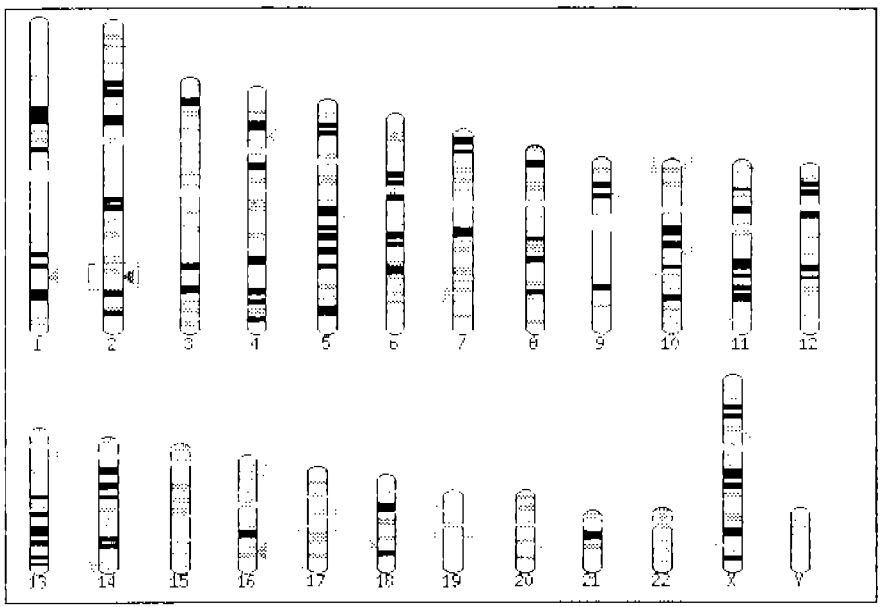


Figura 25. Resultados del tBLASTx para la búsqueda del EPF en el genoma humano, tomado de www.ensembl.org

¿Por qué considerar una secuencia con cerca del 50% de identidad? El candidato más aceptable por homología resulta ser la secuencia presente en el cromosoma 16 del genoma humano, ya que incluso hay ESTs que confirman su expresión, sin embargo, partiendo de que es muy probable que el EPF se generara previo a la divergencia entre el linaje humano y del ratón podemos pensar que dichas secuencias se encontrarían en sintenia en los dos genomas. El EPF caracterizado en el ratón se encuentra alrededor de 50 Mb del centrómero del cromosoma 18 y la sintenia de dicha región corresponde al cromosoma 5 de humanos (ver Figura 26), por lo que inicialmente se pensaría encontrar al ortólogo en dicha región, sin embargo en su lugar, encontramos a un retropseudogen.

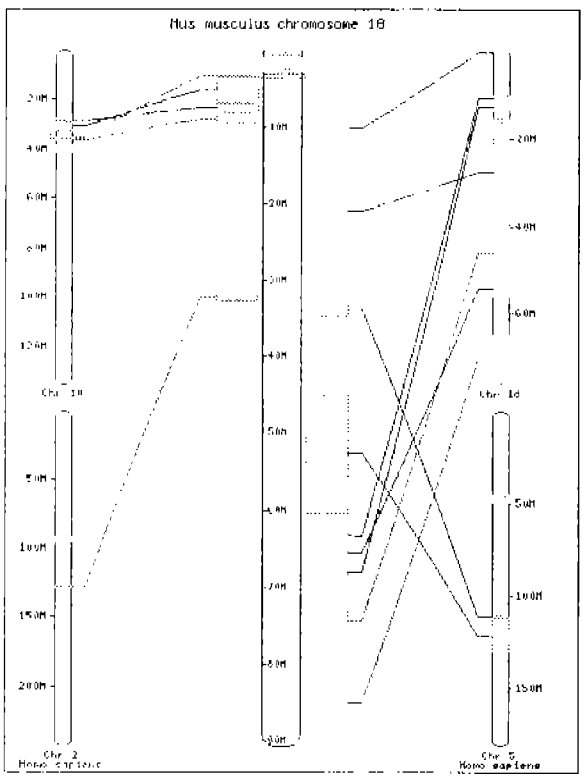


Figura 26. Mapa de Sintenia del cromosoma 18 del genoma del ratón con el genoma humano, el EPF se localiza en la región de 50Mb del genoma del ratón, con su respectiva región sinténica en el cromosoma 5 humano. Tomado de http://www.ensembl.org

La hipótesis que se plantea es que quizá originalmente la secuencia del EPF se encontraba en el cromosoma 5 humano, pero un evento ancestral de duplicación génica es el que generó a la secuencia presente actualmente en el cromosoma 5, para posteriormente darse un evento de traslocación o transposición que colocó al EPF en el cromosoma 16 en humanos.

Otro posible escenario plantea la posibilidad de que el EPF en el ratón es muy reciente y la función del EPF en el humano sea cubierta por el mismo locus de *hsp10*. Esto se deriva de un alto grado de homología entre la secuencia funcional del *hsp10* y el *EPF* en el humano y en el ratón, con una tasa de sustitución de casi el doble de la humana se generará esta retrosecuencia, creando una ganancia de función. También se piensa esto ya que las búsquedas extensivas del EPF en humanos (Barberán-Soler, 2002) no han determinado precisamente un retrogen con las características del EPF en ratón. Por lo que quizá el enfoque práctico para resolver dicha disyuntiva sea comprobar en base a ESTs la capacidad codificante de los posibles candidatos para codificar el EPF en el humano.

6. Conclusiones y perspectivas

La evaluación de retropseudogenes vistos como fósiles moleculares nos permite entender la dinámica de la generación de los mismos y ver que lejos de seguir un patrón aleatorio de generación en una escala temporal tenemos picos de generación de estas secuencias.

Interpretando los picos de generación temporal de los retropseudogenes, de las chaperoninas mitocondriales, tenemos la hipótesis que sugiere que la generación no aleatoria de dichas secuencias refleja cambios evolutivos en los patrones de expresión de dichos genes y pueden correlacionarse con respuestas adaptativas del organismo ante el ambiente.

Las diferencias entre los picos de generación de retropseudogenes *hsp10* y *hsp60* entre el ratón y humano son un reflejo de cómo varía la respuesta adaptativa entre dichas especies. El análisis de los retropseudogenes en los distintos genomas puede brindar pistas sobre los niveles de expresión génica en tiempos paleontológicos, pudiendo determinar las diferentes respuestas de genes que se sobreexpresan por inducción o presión medio ambiental, entre distintas especies y evaluar el papel que juegan dichas diferencias de expresión génica en procesos evolutivos.

Otra interpretación evolutiva trascendente es la posibilidad de que no siempre un evento de retrotranscripción y subsecuente reincorporación al genoma genera un retropseudogen. Aunque la retrotranscripción y la reincorporación son aleatorias, algunas veces podemos tener una retrocopia que se hospeda en vecindad con algún promotor potencial y seguir siendo funcional, transcribirse y traducirse, con lo que podemos aseverar que este es un mecanismo poco estudiado para la formación de familias génicas y la adquisición de nuevas funciones divergentes del gen de origen.

6. Conclusiones y perspectivas

La evaluación de retropseudogenes vistos como fósiles moleculares nos permite entender la dinámica de la generación de los mismos y ver que lejos de seguir un patrón aleatorio de generación en una escala temporal tenemos picos de generación de estas secuencias.

Interpretando los picos de generación temporal de los retropseudogenes, de las chaperoninas mitocondriales, tenemos la hipótesis que sugiere que la generación no aleatoria de dichas secuencias refleja cambios evolutivos en los patrones de expresión de dichos genes y pueden correlacionarse con respuestas adaptativas del organismo ante el ambiente.

Las diferencias entre los picos de generación de retropseudogenes *hsp10* y *hsp60* entre el ratón y humano son un reflejo de cómo varía la respuesta adaptativa entre dichas especies. El análisis de los retropseudogenes en los distintos genomas puede brindar pistas sobre los niveles de expresión génica en tiempos paleontológicos, pudiendo determinar las diferentes respuestas de genes que se sobreexpresan por inducción o presión medio ambiental, entre distintas especies y evaluar el papel que juegan dichas diferencias de expresión génica en procesos evolutivos.

Otra interpretación evolutiva trascendente es la posibilidad de que no siempre un evento de retrotranscripción y subsecuente reincorporación al genoma genera un retropseudogen. Aunque la retrotranscripción y la reincorporación son aleatorias, algunas veces podemos tener una retrocopia que se hospeda en vecindad con algún promotor potencial y seguir siendo funcional, transcribirse y traducirse, con lo que podemos aseverar que este es un mecanismo poco estudiado para la formación de familias génicas y la adquisición de nuevas funciones divergentes del gen de origen.

Con la secuenciación, en camino, de distintos genomas eucariontes, podremos analizar un conjunto de datos que nos permitan dilucidar patrones de generación de retropseudogenes, analizar estos fósiles moleculares y ver si existe algún patrón en la generación de retropseudogenes en una escala temporal que nos permita ver con más claridad las posibles implicaciones evolutivas. El refinamiento de las técnicas de búsqueda bioinformáticas, junto con la depuración de ESTs, nos permitirá tener caracterizados de manera total a los pseudo y retropseudogenes de genomas completos.

6. Glosario

bp (pares de bases): En una cadena doble de ácidos nucleícos, una purina y una pirimidina en diferentes cadenas que interactúan mediante puentes de hidrógeno, siendo lo más común una GC o AT.

Cap: estructura en el extremo 5' de los mRNAs eucariontes; añadida después de la transcripción, por la unión 5' – 5' del trifosfato terminal de una 7-metil-guanosina.

cDNA: DNA sintético transcrito a partir de un molde de RNA específico mediante una enzima específica.

Codón: triplete de nucleótidos que representa un aminoácido o una señal de término.

Consenso, secuencia: secuencia idealizada en la que cada posición representa la base más común cuando varias secuencias son comparadas.

Corrimiento de marco de lectura: La inserción o deleción de uno o más nucleótidos causando una disrupción del marco de lectura de la traducción

Cromosoma: estructura compuesta de una larga molécula de DNA y proteínas asociadas que tiene parte (o toda) de la información genética de un organismo. Especialmente evidente en las células sufriendo mitosis o meiosis, cuando cada cromosoma se compacta y se vuelve visible.

Deleción: se generan por la pérdida de una secuencia de DNA, las regiones a cada uno de los lados se unen.

DNA repetitivo: secuencias idénticas o parecidas que se presentan cientos o miles de veces en el genoma, no tienen que estar adyacentes.

DNA satélite: consiste de muchas repeticiones agrupadas (idénticas o similares) de una pequeña unidad repetitiva.

EST: Etiqueta de expresión de secuencia, una secuencia parcial codificante aislada aleatoriamente de una biblioteca de cDNA, utilizado en la identificación y mapeo de secuencias codificantes, para el descubrimiento de nuevos genes mediante identidad.

Exón: Proviene de la fusión de las palabras "expression region". Se refiere a una región de un gen eucarionte que se traduce a una secuencia de residuos de aminoácidos.

Intrón: región no codificante de un gen eucarionte, que se transcribe a una molécula de RNA, pero luego es cortada durante su procesamiento.

Locus: posición en un cromosoma en donde reside un gen en particular.

Mutaciones neutrales: cambios de nucleótidos en una secuencia con respecto a otra, que no cambian la funcionalidad del producto polipeptídico del gen.

ORF: Marco de lectura abierto. Sección de una secuencia de DNA que comienza con un codón de inicio y termina con un codón de paro de la traducción.

polyA, sitio: Sitio de unión de la secuencia de poliadenilación durante el procesamiento del mRNA.

Promotor: Región regulatoria en el DNA, usualmente hacia 5' de un gen, actuando como sitio de unión para factores transcripcionales.

Retrotranscripción: la síntesis de DNA con un templado de RNA; realizado por la enzima transcriptasa reversa ya sea *in vivo* o *in vitro*.

Sitio donador: El sitio en el pre-mRNA que corresponde al extremo 3' de un exón y al 5' de un intrón.

SNP: (Polimorfismo de nucleótidos sencillos), un polimorfismo causado por el cambio de un solo nucleótido. La mayoría de la variación génica se debe a estos SNPs.

Traducción: La síntesis de una proteína a partir de un molde de mRNA.

Transcripción, sitio de inicio (TSS): La posición de un gen donde la síntesis del mRNA comienza. El primer nucleótido transcrito se denota como +1.

Transcripción: La síntesis de RNA mediante un molde de DNA.

Transcriptasa reversa: enzima presente en los retrovirus, que a partir de una cadena sencilla de RNA hace una copia de DNA, endógeno de genomas mamíferos en secuencias altamente repetitivas como las LINE, que codifican para su propia transcriptasa reversa.

Transposón: segmento de DNA que se puede mover de una posición en el genoma a otra.

UTR: Región No Traducida, región del mRNA que no se traduce.

7. Referencias

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. 2004. Molecular Biology of the Cell. 4th Edition. Garland Publishing.

Altschul SF, Gish, W Miller, Myers EW y Lipmasn DJ 1990. Basic local alignment search tool J. Mol. Biol. 215:403-410.

Anfinsen CD, Principles that govern the folding of protein chains. Science 181:223-230, 1973.

Athanasas S, Quimm KA, Wong T-Y, Rolfe BE, Cavanagh AC et al. 1989. Passive immunisation of pregnant mice against early pregnancy factor (EPF) causes loss of embryonic viability. J Reprod Fertil 87, 495-502.

Bangsborg J, Hoiby N, Hindersson P. 1991. Sequence analysis of the Legionella micdadei groELS operon. FEMS Microbiol. Lett. 61: 31- 38.

Bromham L, Phillips MJ y Penny, D. 1999. Growing up with dinosaurs: molecular dates and the mammalian radiation. Trends Ecol. Evol. 14 : 113-118.

Burge C, and Karlin S. 1997. GENSCAN. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78-94

Coates A. 1996. Chapter X: Immunological Aspects of Chaperonins. The Chaperonins. RJ Ellis Editor. AP NY. 323 pp.

Danial NN, Gramm CF, Scorrano L, Zhang CY, Krauss S, Ranger AM, Datta SR, Greenberg ME, Licklider LJ, Lowell BB, Gygi SP y Korsmeyer SJ. 2003. BAD and glucokinase reside in a mitochondrial complex that integrates glycolysis and apoptosis. Nature 424(6951), 952-956.

Devor EJ, Moffat-Wilson K. 2003. Molecular and temporal characteristics of Human Retropseudogenes. Hum. Biol. 75:661-672.

Devor, E.J. 2001. Molecular archeology of an SPI00 splice variant revisited: dating the retrotranscription and Alu insertion events. Genome Biology. 2001, 2(9):1-6

Dickson R, Larsen B, Viitanen PV, Tormey MB, Geske J, Strange R, Bemis LT. 1994. Cloning, expression, and purification of a functional nonacetylated mammalian mitochondrial chaperonin 10. J. Biol. Chem. 269(43): 26858-26864.

Ellis RJ.1996. The Chaperonins. Academic Press, San Diego, Cal. USA.

Fletcher BH, Al Cassady, KM Summers, A Cavanagh. (2001). The murine chaperonin 10 gene family contains an intronless, putative gene for early pregnancy factor, Cpn10-rs1. Mammalian Genome 12:133-140.

Gonçalves, I., L. Duret and D. Mouchiroud. (2000). Nature and Structure of Human Genes that Generate Retropseudogenes. Genome. 10:672-678.

Griffiths AJF, Miller JF, Suzuki DT, Lewontin R, Gelbart WM. 2000. An Introduction to genetic analysis. W.H. Freeman and Company. (http://www.ncbi.nih.gov/Books)

Habich C, Kempe K, Burkart V, Van Der Zee R , Lillicrap M., Gaston H. and Kolb H. 2004. Identification of the heat shock protein 60 epitope involved in receptor binding on macrophages. FEBS Lett. 568 (1-3), 65-69

Hall, TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41:95-98.

Hansen JJ, Bross P, Westergaard M, Nyholm Nielsen M, Eiberg H, Borglum A, Mongensen J, Kristiansen K, Bolund L, Gregersen N. 2003. Genomic structure of the human mitochondrial chaperonin genes: HSP60 and HSP10 are localised head to head on chromosome 2 separated by a bidirectional promoter. Hum Genet 112 : 71-77.

Harrison P y M Gerstein. 2002. Studying Genomes Through the Aeons: Protein Familias. Pseudogenes and Proteome Evolution. J. Mol. Biol. (2002)318: 1155-1174.

Harrison P, Echols N y M Gerstein. 2001. Digging for dead genes: an analysis of the characteristics of pseudogene population in the *Caenorhabditis elegans* genome. Nucl. Ac. Res. (29):3 818-830

Hemmingsen SM , Tilly K , Ellis RJ , Hendrix RW , Georgopoulos C. , Dennis DT , Vandervies SM , Woolford C. 1988. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. Nature 333: 330- 334

Kaufmann SHE. 1994. Heat shock proteins and autoimmunity: A critical appraisal. Intl. Arch. Allergy Immunol. 103, 317-322.

Kennedy J y Neville A. 1982. Estadistica para Ciencias e Ingenieria. 2ª edición. HARLA. México D.F.

Kumar S, K Tamura, I Jakobsen, and M Nei (2001) MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA.

Langer T y W Neupert. 1996. Chapter IV: Chaperonin-Mediated Folding and Assembly of Proteins in Mitochondria. The Chaperonins. RJ Ellis Editor. AP NY. 323 pp.

Li WH, and D Graur. 1991. Fundamentals of Molecular Evolution. Sinauer Associates, Inc. Sunderland Massachussets

Li WH. 1997. Molecular Evolution, Sinauer Associates, Inc., Snderland Massachussets.

Li, W.H., T. Gojobori and M. Nei. 1981. Pseudogenes as a paradigm of neutral evolution. Nature. Vol 292:237-239.

Lotscher E y JP Allison. 1990. Nucleotide and deduced amino acid sequence of a murine cDNA clone. Nucleic Acids Res. 18(23):7153.

Martin J, Langer, Horwich AL and Hartl FU. 1992. Prevention of protein denaturation under heat stress by the chaperonin hsp60. Science 258, 995-998.

Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, Bolouri MS, Ray HN, Sihag S, Kamal M, Patterson N, Lander ES y Mann M. 2003. Integrated analysis of protein composition, tissue diversity and gene regulation in mouse mitochondria. Cell 155 (5), 629-640.

Mounsey A, Bauer P y IA Hope. 2002. Evidence suggesting that a fifth of annotated *Caenorhabditis elegans* genes may be pseudogenes. Gen. Res. 12:770-775.

Mouse Genome Sequencing Consortium (MGSC). 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Ophir R y D Graur. 1997. Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene 205 (1997) 191-202.

Otwinowski Z, Boisvert D , Hegde R , Braig K , Sigler PB , Horwich AL, Joachimiak A. 1994. The crystal structure of the bacterial chaperonin GroEL at 2.8 A. Nature 371: 578- 586.

Prasad TK, Stewart CR. 1992 cDNA Clones encoding *Arabidopsis thaliana* and *Zea mays* mitochondrial chaperonin HSP60 and gene expression during seed germination and heat-shock. Plant Mol. Biol. 18: 873- 885

Rutherford K, J Parkhill, J Crook, T Horsnell, P Rice, M-A Rajandream and B Barrell. 2000 Artemis: sequence visualisation and annotation." Bioinformatics 16 (10) 944-945.

Ruud P, Fodstad O, Hovig E. 1999. Identification of a novel cytokeratin 19 pseudogene that may interfere with reverse transcriptase polymerase chain reaction assays used to detect micrometastatic tumor cells. Int. J. Cancer 80 119–125.

Ryan MT, Herd SM, Sberna G, Samuel MM, Hoogenraad NJ, Hoj PB (1997) The genes encoding mammalian chaperonin 60 and chaperonin 10 are linked head-to-head and share a bidirectional Promoter. Gene 196:9–17

Summers, KM, Murphy RM, Webb GC, Peters GB, Morton H, et al. (1996). The human early pregnancy factor/chaperonin 10 gene family. Biochem Mol Med 58, 52-58.

The FANTOM Consortium and the RIKEN Genome Exploration Research. 2002. Analysis of the mouse transcriptome based on functional annotation. Nature 420, 563-573.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24:4876-4882

Valdés-López V, Vilchis-Peluyera A, Alba-Lois L, Rodarte-Murguía B, Segal-Kischinevzky C, Rodriguez-Ponce B y LD Alcaraz-Peraza. 2004. *Evolución genómica: aparición y divergencia de retropseudogenes.* Scientific American ed. Latinoamérica. 32-33.

Venner TJ y Gupta RS. 1990. Nucleotide sequence of mouse HSP60 (chaperonin, GroEL homolog) cDNA. Biochim. Biophys. Acta 1087(3): 336-338.

Venter JC et al. (2001). The Sequence of the Human Genome. Science, Vol. 291, pp 1304-1351.

Young DB. 1990. Chaperonins and the immune response. Semin. Cell. Biol. 1, 27-35.

Zhang Z y M Gerstein. 2003. Identification and characterization of over 100 mitochondrial ribosomal protein pseudogenes in the human genome. Genomics 81:468-480.

Zhang Z, Harrison P y M Gerstein. 2002. Identification and Analysis of Over 2000 Ribosomal Protein Pseudogenes in the Human Genome. Gen. Res. 12:1466-1482.