27

UNIVERSIDAD NACIONAL AUTONOMA DE MÉXICO

FACULTAD DE CIENCIAS

RETROPSEUDOGENES HUMANOS, UN ANÁLISIS GENÓMICO Y EVOLUTIVO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G O

P R E S E N T A:

SERGIO BARBERÁN SOLER

DIRECCIÓN DE ESTUDIOS PROFESIONALES
FACULTAD DE CIENCIAS
SECCION ESCOLAR

DIRECTOR DE TESIS:
DR. VÍCTOR MANUEL VALDÉS LÓPEZ

2002

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

M. EN C. ELENA DE OTEYZA DE OTEYZA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunico a usted que hemos revisado el trabajo escrito:
"Retropseudogenes humanos, un análisis genómico y evolutivo"

realizado por Sergio Barberán Soler

con número de cuenta 9332544-4 , quién cubrió los créditos de la carrera de Biología

Dicho trabajo cuenta con muestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dr. Víctor Manuel Valdés López
Propietario	Dr. Arturo Carlos II Becerra Bracho
Propietario	Biol. Alfonso José Vilchis Peluyera
Suplente	Dra. Luisa Alvarina Alba Lois
Suplente	M. en I.B.B. Claudia Andrea Segal Kischinevzky

Consejo Departamental de Biología

Dra. Patricia Ramos Morales

FACULTAD DE CIENCIAS
U.N.A.M.

DEPARTAMENTO
DE BIOLOGÍA

Gracias a:

Víctor Valdés, Luisa Alba, Alfonso Vilchis,
Claudia Segal y Arturo Becerra

A mi madre, a mi padre, a mis hermanos
y sobre todo a Stella

INDICE

Resumen 1
Introducción 2

Antecedentes y Justificación 10
Proteínas de choque térmico 10
Retropseudogenes 11

Objetivos 16

Metodología 17
Basic Local Alignement Search Tool 17
Resultados del Blast 18
Método de análisis de marcos de lectura de Shepherd 20
Alineamientos pareados y múltiples 22
Datación de las secuencias 23

Resultados y Discusión 25
Identificación de los genes hsp10 y hsp60 en el genoma humano 25
Búsqueda de los pseudogenes 28
Historia de los pseudogenes 29
Análisis de los pseudogenes 32
Gráficas de Shepherd de los pseudogenes 36
Alineamientos pareados con el cDNA y datación de los pseudogenes 38
Calibración del Blast para establecer los limites de la detección 41
Gráficas de las dataciones 42
Sustituciones por sitio 45
Alineamiento múltiple 45
Árboles filogenéticos de los pseudogenes 47
Búsqueda del EPF 47

Conclusiones 49

Bibliografía 51

Anexo I 55

Anexo II 62

Glosario 65

RESUMEN:

El genoma humano está compuesto de 3,000 millones de pares de bases. De éstos, solamente el 2% está compuesto por secuencias que codifican para proteínas. El restante está comprendido por DNA no codificante. Dentro de éste hay secuencias muy parecidas a los genes que, a diferencia de estos, no son funcionales. Estas copias son denominadas pseudogenes. Estos se clasifican en diversos grupos dependiendo de su origen. Los que fueron originados por la inserción en el genoma de un cDNA creado por la transcriptasa reversa a partir de un mRNA, son denominados retropseudogenes. Estos retopseudogenes tienen la característica de carecer de intrones, ya que provienen de un mRNA citoplasmático maduro. Se ha estimado que en el genoma humano hay entre 23,000 y 30,000 copias de retropseudogenes. En este trabajo de tesis se identificaron y analizaron los retropseudogenes del grupo de las chaperonas moleculares Hsp10 y Hsp60. Para su búsqueda, se utilizó el borrador de la secuencia del genoma humano que se encuentra disponible en la base de datos del NCBI. Se identificaron 24 retropseudogenes del gen de la Hsp10 y 15 de la Hsp60, distribuidos en 14 cromosomas. La mayoría de estos retropseudogenes muestran evidencias de haber sido originados por procesos de retroinserción y no por duplicaciones génicas. Estos retropseudogenes fueron datados por diferentes métodos, lo que permitió observar que, en el caso de la Hsp10 no fueron originados uniformemente a lo largo del tiempo, sino en eventos discontinuos. Esto nos permite sugerir que las tasas de formación de los pseudogenes pueden reflejar eventos evolutivos peculiares.

INTRODUCCIÓN:

Desde que el ser humano empezó a criar animales ha sido obvio que cada óvulo fertilizado contiene oculto un plan o diseño que guía el desarrollo del organismo. En los tiempos modernos la ciencia de la genética ha crecido alrededor de la premisa de los elementos invisibles que contienen esta información: los **genes**, los cuales son distribuidos a cada célula hija cuando estas se dividen. Para lograr esto, la célula tiene que hacer una copia de sus genes para dar un juego completo a cada una de las células hijas. Los genes contenidos en los espermatozoides y en los óvulos llevan entonces la información hereditaria de una generación a la siguiente.

Para finales del siglo XIX los biólogos reconocieron que los encargados de llevar esta información de la herencia son los **cromosomas**, que se vuelven visibles en el núcleo cuando la célula empieza a dividirse. Pero la evidencia de que el DNA en estos cromosomas es la molécula de la cual los genes están hechos, vino mucho después, por medio de estudios realizados en bacterias. En 1944 se descubrió que al adicionar DNA purificado de una cepa de bacterias a otra, se obtenían en esta segunda cepa características provenientes de la primera. En ese entonces se pensaba que solamente las proteínas tenían la suficiente complejidad conformacional para llevar a cabo la **transformación**, este descubrimiento llegó como una sorpresa, y no fue aceptado por la mayoría de los científicos hasta la década de los 50's. Hoy en día, la idea de que el DNA lleva la información genética en su larga cadena de nucleótidos es tan fundamental para el pensamiento biológico que es difícil imaginar el enorme hueco que este descubrimiento llenó.

Actualmente, las investigaciones nos llevan a estudiar la totalidad de la información genética de un organismo. A este programa genético completo de una célula se le denomina el genoma (palabra que deriva de la unión de genes y cromosomas); éste en muchos casos está dividido en dos componentes: el genoma nuclear y el genoma mitocondrial (o de los cloroplastos en el caso de las plantas), estando la gran mayoría del material contenido en los cromosomas que se encuentran en el núcleo de la célula y una parte mucho menor en la mitocondria. La parte que aquí nos interesa es el genoma nuclear de los organismos.

Con la publicación del primer borrador de la secuenciación del genoma humano por el Consorcio Internacional del Proyecto del Genoma Humano en febrero de 2001, se abrieron nuevas puertas en la investigación genómica. Las bases de datos públicas en donde se encuentra el borrador de la secuencia del genoma humano nos permiten llevar a cabo un análisis genómico que antes hubiera sido imposible realizarlo, debido a lo costoso de la secuenciación.

El tamaño del programa genético de los eucariontes varía enormemente, así como también varía el número de cromosomas en los que éste está dividido. Una de las características que llaman la atención, es que las **regiones codificantes** solamente representan una pequeña proporción del DNA total. Es por esto que, al ver el tamaño del genoma de un organismo no podemos deducir el número de genes que éste contendrá, ya que no existe una relación directa entre el número de genes y tamaño del genoma.

Los genomas eucariontes varían en tamaño, desde los más pequeños de 10 megabases (10^6 pares de bases) hasta los más grandes de 90,000 Mb (Brown 1999). El tamaño del genoma tiene cierta correlación con la complejidad morfológica del organismo -aunque no está determinado solamente por esto- teniendo los eucariontes superiores los genomas más grandes. Esta correlación no está determinada solamente por la complejidad del organismo ni por el número de genes que éstos contengan, sino por la cantidad de **DNA repetitivo**, de manera que los genomas más grandes son los que mayor cantidad de DNA repetitivo contienen y en muchos casos no son los organismos más complejos, morfológicamente hablando. La cantidad total de DNA en un genoma haploide de un organismo es conocido como el "valor C" y a esta diferencia entre el tamaño de los genomas y la complejidad de los organismos se le ha denominado "paradoja del valor C", ya que los eucariontes superiores tienen un genoma mucho más grande de lo que sería necesario para codificar todas sus proteínas.

Al igual que el tamaño del genoma, una característica que también varía mucho es el número de cromosomas, teniendo todos los eucariontes por lo menos dos cromosomas. Su número no tiene una correlación directa con el tamaño del genoma. Cabe mencionar como ejemplo que la levadura *Saccharomyces cerevisiae* tiene cuatro veces más cromosomas que la mosca de la fruta *Drosophila melanogaster* (Brown, 1999), a pesar de

que el tamaño del genoma es mucho mayor en *D. melanogaster* (165 Mb vs. 12 Mb de *S. cerevisae*).

Los genomas eucariontes tienen varias peculiaridades al compararlos con los genomas procariontes. Por ejemplo, la integridad de los genes está interrumpida por los **intrones**. Asimismo se pueden encontrar varias o múltiples copias de una secuencia en particular, también se encuentran largos segmentos de DNA que no codifican para ninguna proteína y grandes segmentos de DNA repetitivo.

El método mas utilizado para predecir el potencial codificante de una secuencia de DNA es el de identificar las regiones que cuenten con marcos de lectura abiertos y de esta manera calcular a *grosso modo* cuántos genes se codificarán en este genoma. Este análisis es usado cuando se dispone de la secuencia genómica completa de un organismo. Pero debido a que los genes eucariontes no son continuos y que no todos los marcos de lectura se transcriben, este método sólo muestra el potencial codificante de un genoma pero no nos permite calcular el del número exacto de genes que contiene. Otro método utilizado consiste en extrapolar los datos obtenidos de una pequeña región al genoma completo y de esta manera se puede suponer el número de secuencias codificantes de un genoma. Un análisis alternativo para calcular el número de genes de un genoma es analizando el **proteoma**, ya que en él están representados todos los genes estructurales.

Se ha calculado que en el genoma humano menos del 2% está compuesto de secuencias que codifican proteínas (los exones); el 98% restante está compuesto de DNA no codificante y de éste, el 45% es DNA repetitivo que proviene de secuencias de DNA "parásito" conocidas como elementos móviles repetidos (Dennis, 2001), este DNA repetitivo ha sido llamado "DNA basura". Curiosamente, mucho de este contenido repetitivo de nuestro genoma representa remanentes ancestrales de viejos **transposones**, en contraste con los genomas de la mosca de la fruta y del ratón en donde se encuentra un gran número de elementos más jóvenes y activos. Este porcentaje de DNA repetitivo no tiene precedentes en otros organismos, ya que en el genoma bacteriano sólo el 1.5% es repetitivo, en la mosca de la fruta el 3%, 7% en *C. elegans* y 11% en *Arabidopsis thaliana* (Dennis 2001).

El DNA altamente repetitivo es una característica común de los eucariontes pero es virtualmente desconocido en los procariontes. Una clase de DNA repetitivo es el llamado **DNA satélite**. El DNA satélite se refiere a pequeñas series de DNA altamente repetido que se encuentran con mayor frecuencia en zonas específicas del genoma, por ejemplo en los **centrómeros**.

Aún cuando sólo el 2% del genoma contiene secuencias codificantes se ha encontrado que algunos genes se encuentran repetidos, en lo que se denomina familias génicas; estas son grupos de genes con secuencias idénticas o similares, un ejemplo de esto es la familia de genes de las globinas, en donde se encuentran algunos genes que son homólogos, pero codifican para proteínas con diferente función específica. Algunos genes de familias génicas se localizan en zonas cercanas entre si mientras que otros, como los genes de la aldosa, se encuentran dispersos en diferentes cromosomas del genoma humano.

Baxevanis (2001) define a la similitud entre secuencias como una cantidad medible que puede ser expresada, como porcentaje de identidad; señala también que la homología por su parte, se refiere a una inferencia obtenida de estos datos, en el sentido de que dos secuencias comparten un ancestro común en su historia evolutiva. Los estudios de la evolución de genes y proteínas involucran la comparación de homólogos – secuencias que tienen un origen común pero pueden o no mostrar una actividad en común. Los genes homólogos se clasifican en ortólogos, parálogos y xenólogos.

Los genes ortólogos son homólogos producidos por la especiación. Representan genes derivados de un ancestro común que se separaron debido a la divergencia de los organismos a los cuales pertenecen. Generalmente tienen la misma función.

Los genes parálogos son homólogos producidos por la duplicación génica. Estos son genes derivados de un gen ancestral común, que se duplicó en el mismo organismo y después divergieron. Generalmente tienen funciones diferentes, o variantes de una misma función.

Los xenólogos son genes homólogos resultantes de una transferencia génica horizontal entre dos organismos. La función de los xenólogos puede ser variable

dependiendo de qué tan significante fue el cambio de contexto para el gen que se movió horizontalmente; generalmente las funciones tienden a ser similares (Baxevanis, 2001).

Se ha reportado que la mayoría de los genomas albergan una gran variedad de secuencias móviles de DNA. Estos elementos transponibles varían en longitud desde unos pocos cientos, hasta miles de pares de bases y están generalmente presentes en copias múltiples en los genomas. Cada uno de estos elementos transponibles es ocasionalmente activado para moverse a otro sitio del DNA en la misma célula por medio de un proceso llamado transposición. La transposición puede ocurrir por varios mecanismos. Una gran familia de elementos transponibles usa un mecanismo que es indistinguible del ciclo de vida de los **retrovirus**. Estos elementos, llamados retrotransposones, están presentes en una gran variedad de organismos (Alberts 1994). El proceso mediante el cual estos elementos se multiplican se denomina **retrotranscripción**, y es mediada por una **transcriptasa reversa** codificada por elementos LINE (vease mas adelante) y finalmente por su inserción en el genoma.

En el genoma humano se han encontrado solamente dos elementos móviles repetitivos que están todavía activos: los LINE (elementos nucleares largos ínter dispersos) y los SINE (elementos nucleares cortos ínter dispersos). De los segundos, los más comunes en el humano son las secuencias **Alu**. Juntos, los LINE y los Alu conforman más del 60% de las regiones repetidas de nuestro genoma (Dennis, 2001). La diferencia entre estos dos elementos radica en que los LINE tienen la maquinaria necesaria para auto replicarse –llevar a cabo la retrotranscripción- mientras que los elementos Alu necesitan de las secuencias LINE para replicarse; aún con esto, los elementos Alu son muy abundantes, contando con cerca de 1 millón de copias en el genoma humano. Se ha calculado que la mayor parte de los elementos transponibles aparecieron en el linaje de los mamíferos hace más de 100 millones de años, mucho antes de que los mamíferos con placenta (euterios) evolucionaran. Algunos tipos de transposones aumentaron su número, como los LINE1 y los Alu. Otros, parece que no encontraron un ambiente adecuado. Por ejemplo, solamente se encuentran algunos remanentes de otro grupo de elementos transponibles llamados retrotransposones LTR, que a diferencia del genoma human se encuentran activos en el genoma del ratón. (Dennis, 2001)

Otro tipo de secuencias repetidas son los transposones de DNA, que han marcado nuestra evolución genómica con dos etapas de actividad: antes y después de la evolución de los placentados (Dennis, 2001). Estos transposones de DNA son menos comunes que los retrotransposones. Probablemente hay menos de 1,000 en el genoma humano. Estas secuencias no necesitan de la transcriptasa reversa puesto que no se transcriben a RNA. Su actividad se basa en movilizar su secuencia entre distintas localizaciones en el genoma en el genoma. Estos mecanismos de corte e inserción, son mediados por enzimas codificadas por el propio transposón.

Otro mecanismo que incrementa el tamaño del genoma son las duplicaciones. Estas han jugado un papel muy importante en la evolución del genoma humano. Se calcula que cerca del 5% del genoma ha sido originado por la duplicación de grandes bloques de mas de 10,000 pares de bases (Lander *et al*, 2001). Una duplicación provoca que una copia de un gen se coloque en un nuevo sitio, en donde puede adquirir una función fisiológica diferente. Se ha propuesto que las regiones duplicadas han contribuido mucho a la expansión de las familias génicas en los humanos. La familia de genes que codifican para los receptores olfatorios son un buen ejemplo. Se han reportado cerca de 1,000 genes olfatorios repartidos en el genoma de los mamíferos, que debido a su similitud se infiere que descienden de un ancestro en común, además de que hay varios grupos de estos genes repetidos a lo largo de todo el genoma.

Una posible explicación del por qué las familias génicas presentan tanta diversidad en las secuencias, sostiene que, después de una duplicación génica, ocurren mutaciones que provocan que las secuencias de los genes diverjan gradualmente, lo cual les va otorgando algunas peculiaridades funcionales que los hace selectivamente importantes. Por otro lado, si estas mutaciones dan por resultado que la secuencia se vuelva inactiva, esta se convierte en un pseudogen. Estos son genes que se han vuelto no funcionales. De cualquier manera, las duplicaciones génicas son una gran fuerza en la evolución genómica. Puede ser incluso el mecanismo predominante para la evolución de nuevas funciones de los genes (Ohno, 1970). Hay que señalar que aún siendo un importante mecanismo para la creación de nuevas funciones, se ha calculado que la tasa de pérdida de genes (por su conversión a pseudogenes u otras causas) después de las duplicaciones debe ser por lo menos de un orden de magnitud mayor que la de

divergencia de estos genes (Wagner, 1998). Cuando los genes se inactivan por mutaciones se crea un pseudogen convencional, por ejemplo una mutación que genere un codón de termino en la región codificante produce una proteína truncada que es incapaz de llevar a cabo su función bioquímica. Ohno (1985) determina que los genes duplicados tienen una vida media que va desde 1 hasta 50 millones de años (Ma), dependiendo de la secuencia y de la localización en el genoma. Se ha pensado que esto se debe a que no son necesarias dos copias de la mayoría de los genes, ya que con una funcionando se pueden cumplir los requerimientos celulares.

Un segundo tipo de pseudogenes no son originados por el decaimiento evolutivo sino como una consecuencia no "intencionada" de los mecanismos de retrotransposición de elementos de tipo LINE, en donde la transcriptasa "equivocadamente" utiliza un mRNA citoplasmático maduro como molde y crea una copia (**cDNA**) que se reinserta en el genoma. Como consecuencia, la copia carece normalmente de los elementos de expresión como los promotores y al ser originados de un mRNA maduro carecen también de los intrones, algunos conservan remanentes de lo que fue el sitio de poliadenilación. Este proceso, derivado de la retrotransposición, es el que genera las llamadas retrosecuencias, que son secuencias de mRNA procesados que son insertados en el DNA. A estas retrosecuencias se les llama retrogenes o genes procesados, cuando son funcionales y retropseudogenes si no son funcionales. Un retrogen es una secuencia funcional que produce una proteína idéntica o casi idéntica a la producida por el gen de donde derivó, pero con la diferencia de que no contiene intrones. Hay muchas razones por las cuales es poco probable que un mRNA que fue retrotranscrito mantenga su funcionalidad. Primero, el proceso de la retrotranscripción no es a prueba de errores y pueden suceder muchas mutaciones entre el templado de RNA y el cDNA; segundo, a menos que el gen procesado fuera derivado de un gen transcrito por la RNA polimerasa III, usualmente no contiene las secuencias regulatorias, ya que éstas residen en las regiones no transcritas, por lo cual es muy probable que sea inactivo aunque su región codificante esté intacta. En ciertos casos los retrotranscritos se ubican cerca de otro promotor, estas casualidades son las que llaman la atención; y por último, un gen procesado puede ser insertado en una locación genómica que no sea adecuada para su

expresión, por lo que muchos autores sostienen que las retrosecuencias están inactivas desde su inserción (Brown, 1999).

Los RNA's eucariontes -excepto los que son producidos en la mitocondria y en los cloroplastos que en realidad tienen un origen procarionte- son modificados en el núcleo inmediatamente después de la transcripción. Hay dos modificaciones generales, que son la adición de una estructura "**cap**", compuesta de un residuo de 7-metil guanosina unido a un trifosfato en el extremo 5', y la adición de una cantidad importante de residuos de adenina (poly-A) en el extremo 3'. Por otro lado, los mRNA's de los eucariontes son generalmente mono **cistrónicos**, lo cual significa que cada mensajero traduce solamente un polipéptido (Alberts, 1994). Como se menciono anteriormente, los genes eucariontes son en su mayoría interrumpidos; por lo cual necesitan ser procesados para su adecuada expresión. El transcrito primario tiene la misma organización que el gen, y en algunos casos se le llama pre-mRNA. El proceso por el cual se remueven los intrones se denomina procesamiento ("splicing" en inglés). Este procesamiento ocurre en el núcleo, junto con las modificaciones antes nombradas. De esta manera los intrones son removidos del RNA nuclear por un sistema que reconoce pequeñas secuencias **consenso** conservadas en los límites entre el intrón y el exón. Esta reacción requiere de un gran aparato de procesamiento, que consiste en un arreglo de proteínas y ribonucleoproteínas que funcionan como un gran complejo llamado el "spliceosoma". En el sitio exacto del corte del intrón este complejo reconoce, en primer lugar, el sitio GT-AG, lo que significa que el intrón inicia con GT y termina con AG. Otra de las características conservadas en algunos intrones es la de la secuencia de lazo o "lariat" que consta de una pequeña secuencia de 7 **pb**, involucrada en el procesamiento del intrón; esta secuencia lazo se encuentra entre 18 y 40 pb arriba del sitio 3' de procesamiento. La secuencia consenso lazo es: UACUAAC.

ANTECEDENTES Y JUSTIFICACIÓN:

Proteínas de choque térmico (Heat Shock Proteins)

En este trabajo de tesis, se efectuó un análisis de dos tipos de genes reportados en el borrador del genoma humano: la proteína del choque térmico 10 (HSP10 por sus siglas en inglés) y la proteína de choque térmico 60 (HSP60); estas dos proteínas forman una subclase del grupo denominado proteínas de estrés. La razón por la cual se escogieron estas proteínas es por que han servido como modelos de trabajo experimental en nuestro grupo y tenemos mucha experiencia respecto a sus características estructurales y funcionales. Por otro lado, desde hace tiempo se han reportado varias secuencias de pseudogenes de la hsp60 específicamente.

Las proteínas de estrés se presentan en una amplia gama de pesos moleculares por lo cual se clasifican en: proteínas de alto peso molecular (HMW-Hsps) alrededor de 100-90 kDa; Hsp70's alrededor de 70kDa; Hsp60's de 60kDa y las proteínas de bajo peso molecular (LMW-Hsps) de alrededor de 10-30 kDa (Lindquist, 1986). Las proteínas HSP10 y HSP60 son universales y han sido descritas en todos los organismos en los que han sido buscadas (bacteria, arquea y eucaria). Asímismo, se presentan en prácticamente todos los tejidos y todos los tipos celulares de organismos pluricelulares. Vale la pena señalar que las Hsps que se presentan en cada organismo pueden variar, pero las HSP60 y HSP10 casi siempre están presentes (Lindquist, 1986). Otro aspecto importante es que estas proteínas se sobre-expresan por efecto del estrés; esto significa que normalmente existe un nivel basal de expresión. Dicho de otro modo, forman parte del repertorio cotidiano de genes que se expresan de manera constitutiva (house-keepping genes) (Ang *et al.*, 1991).

El plegamiento de una proteína es un evento de importancia central en biología. Si bien es cierto que la secuencia de aminoácidos posee en potencia la información para determinar la estructura nativa de la proteína, en la célula el plegamiento de los polipéptidos puede llegar a requerir de la ayuda adicional de proteínas cuyo papel es colaborar a que esta estructura se logre. Estas proteínas de gran importancia celular son las hsps, que son denominadas en muchas ocasiones chaperonas moleculares. El concepto de chaperonas moleculares, fue planteado de manera formal por Ellis, quien las definió

como un grupo de proteínas no relacionadas que son capaces de promover el plegamiento y el ensamblaje correctos de otras proteínas, pero que no forman parte ellas mismas de las estructuras funcionales finales (Ellis, 1987).

Las chaperonas moleculares se encuentran a todo lo largo de la escala filogenética y muchas de ellas son clasificadas como proteínas de estrés, aunque como quedó señalado anteriormente, también se ha demostrado que tienen una función esencial bajo condiciones normales. Mas recientemente las chaperonas moleculares han sido definidas como proteínas que unen y estabilizan confórmeros inestables, ayudan en el transporte de proteínas a compartimentos subcelulares específicos e inclusive participan en su degradación. Las chaperonas moleculares no contienen información estérica para especificar el plegamiento correcto. Más bien, ha sido propuesto que su mecanismo de acción consiste en prevenir interacciones incorrectas, dentro y entre polipéptidos en estado no-nativo, de manera que se incrementa el rendimiento (cantidad final de polipétidos plegados), pero no la tasa (velocidad) de la reacción de plegamiento.

La mayor parte del conocimiento que tenemos de las chaperonas moleculares, ha provenido principalmente de la investigación de las hsp70y la hsp60. Desde hace casi tres décadas se demostró que en *Escherichia coli*, la hsp60 se encontraba en el mismo ***locus*** genético que la hsp10 (Tilly 1981). Las chaperoninas del tipo de las hsp60, están constituidas de dos anillos, los cuales están formados por siete subunidades, cada una de alrededor de 60 kDa (de ahí su nombre). Estos anillos moleculares cooperan con un cofactor, la hsp10 que está constituida por un sólo anillo heptamérico de subunidades de alrededor de 10 kDa. Este complejo encierra una cavidad central donde se ubica el sitio de unión y plegamiento del polipéptido.

Retropseudogenes

En principio, la aparición de retrogenes o retropseudogenes puede ser considerado como un "accidente molecular". No obstante, en el genoma humano existe evidencia de que este "accidente" ocurre frecuentemente, tal cual lo indica el gran número de estos elementos. Asímismo, la naturaleza probabilística de estos eventos se manifiesta en el mismo hecho de que las retrocopias de genes específicos están dispersas por todo el

genoma. Más adelante en el capítulo de resultados, se muestran los datos obtenidos en este sentido. Por otro lado, está la interrogante de si cualquier mRNA tiene la misma posibilidad de ser usado como templado por la retrotranscriptasa o no.

En este sentido se han postulado algunas características que tienen en común los genes que producen mas retropseudogenes (Gonçalves *et al.*, 2000):

a) Genes con alto nivel de expresión

b) Genes conservados evolutivamente

c) Transcritos de mRNA cortos (vale la pena notar que el promedio del tamaño de los mRNA es de 1000 pb)

d) Genes con un bajo contenido de GC

La primera característica se debe a que los genes que más retropseudogenes tienen deben de ser altamente expresados en las líneas germinales para que tengan más probabilidades de ser heredados. El que sean cortos y pobres en GC podría estar relacionado a una mayor eficiencia de la transcripción reversa y la transposición. Desde luego que podrían existir otros factores aún no caracterizados como podría ser la longitud del polyA, pues existe evidencia de que en esta secuencia se ancla la transcriptasa reversa e inicia la retrotranscripción. Se ha reportado que en el genoma humano, deben existir entre 23,000-33,000 copias de retropseudogenes (Gonçalves *et al.*, 2000). Este rango es debido a la incertidumbre del número de genes que contiene el genoma humano, ya que el cálculo fue realizado extrapolando el número de retropseudogenes encontrados para los genes que cumplen con las características antes mencionadas.

Para que un retroelemento se fije en el genoma, la inserción tiene que ocurrir en una célula germinal. Por esta razón los genes que dan origen a los retropseudogenes tienen que ser expresados por lo menos en alguna etapa en el tejido germinal. Por su parte la probabilidad de que un mRNA de origen a un retroelemento depende de la eficiencia de los siguientes pasos: transcripción reversa, transferencia al núcleo e integración en el cromosoma; algo igualmente importante es el nivel de expresión de este gen, ya que entre mas mRNA's existan en las células germinales es más probable que sea tomado como substrato de la retrotranscriptasa reversa.

Dhellin *et al* (1997), demostraron que la transcriptasa reversa involucrada en la transcripción reversa de los genes que dan origen a retropseudogenes es probablemente parte de una LINE. Asimismo, se ha establecido que al igual que los genes que dan origen a los retropseudogenes, los elementos LINE de los mamíferos son pobres en GC (Korenberg & Rykowsky, 1988).

Trabajando con diferentes especies y subespecies de ratones de laboratorio Tanooka *et al* (2001), encuentran que la tasa de mutación para los pseudogenes de la p53 es entre 20 y 30 veces mas rápida que para el gen funcional. Con base en la teoría del reloj molecular, que afirma que la tasa de mutación es constante durante el tiempo, Friedberg *et al* (2000) calculan la edad de diferentes pseudogenes de primates. Las mutaciones en los pseudogenes son selectivamente neutrales en todas las posiciones y las diferencias en la tercera posición no deben de ser tan pronunciadas en comparación con la primera y la segunda. Friedberg acepta que es cuestionable el uso de secuencias de genes funcionales para calcular la edad de origen de un pseudogen ya que estos no tienen ninguna presión que evite que se fijen las mutaciones, a diferencia de las secuencias funcionales.

Los únicos trabajos reportados con los pseudogenes del grupo de las hsp60 y hsp10 son de Gupta (1982) y Venner (1990). En 1990 Venner *et al*, reportan el gen de la hsp60 en humanos junto con algunos pseudogenes. Todas estas secuencias son reportadas con la peculiaridad de no tener intrones, inclusive la secuencia del gen. En ese reporte no se publican las regiones regulatorias de ninguna de estas secuencias. Se reportan 10 pseudogenes que no tienen un marco de lectura abierto lo suficientemente grande como para poder ser el gen de la hsp60, pero que tienen una buena similitud con el cDNA. Por su parte Gupta *et al* (1982) reportan que el gen de la hsp60 solamente tiene una copia en el genoma de ratón, y reporta también la secuencia de algunos pseudogenes de la hsp60 en ese mismo genoma. En contradicción con estros trabajos, posteriormente se ha reportado que los genes de la HSP60 y HSP10 si tienen intrones (Hansen *et al* 1999). Por otro lado los retropseudogenes reportados por Venner y Gupta no han podido ser localizados en el genoma humano y en esta tesis ninguno de los retropseudogenes de HSP60 que localizamos corresponde a aquellos reportados por Venner y Gupta.

A las secuencias repetitivas en el genoma humano se les ha descrito como DNA "basura" y se tratan como secuencias sin interés. Sin embargo, actualmente son consideradas una fuente extraordinaria de información sobre los procesos biológicos. Las repeticiones constituyen un rico registro paleontológico, que contiene pistas cruciales sobre los eventos y fuerzas evolutivas. Como marcadores pasivos, proveen muestras para estudiar procesos como la selección y mutación. Es posible reconocer generaciones de repeticiones que se originaron al mismo tiempo y seguir sus destinos en diferentes localidades genómicas o en diferentes especies; esto nos ayuda a determinar las diferencias evolutivas entre cada *locus* del genoma, o entre dos especies diferentes. Como agentes activos, las repeticiones han moldeado el genoma provocando rearreglos, creando nuevos genes, modificando genes existentes y modulando el contenido de GC (Lander *et al*, 2001). Por lo anterior, en el presente trabajo nos hemos planteado el objetivo de analizar una pequeña parte de este DNA "basura" para describir algunas de las características de una historia en particular de la evolución genómica.

Por otro lado, algunos eventos de retroinserción dejan copias procesadas funcionales de genes sin intrones. Venter *et al* (2001) sugieren que hay más de 200 de estos genes en el genoma humano. Posiblemente uno de estos casos sea el del factor temprano del embarazo (EPF por sus siglas en inglés).

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 secretada por células normales, transformadas y neoplásicas durante el crecimiento y la división celular; también se sabe que es requerida para la continuación del crecimiento celular. Aparte de estas funciones regulatorias en el crecimiento celular el EPF tiene una actividad inmuno-moduladora (Fletcher *et al*, 2001).

En 1996, Summers *et al* reportan la secuenciación del cDNA (obtenido de una librería humana de cDNA) del EPF de humano. En el mismo trabajo reportan una homología entre las secuencias de la chaperonina 10 (hsp10) y este factor temprano del embarazo (EPF) recientemente secuenciado, con lo cual sugieren que se trata de la misma familia génica. Usando la **clona** de la secuencia del EPF como sonda, realizaron un

Southern blot genómico en el DNA humano, obteniendo como resultado señales positivas significativas de esta secuencia en ocho diferentes cromosomas.

Se han reportado grandes diferencias en la localización celular y en la función de las dos proteínas (EPF y hsp10), lo cual sugiere grandes diferencias en su regulación (Summers *et al*, 2001). Para explicar este comportamiento existen diferentes posibilidades. Una de ellas es que sean transcripciones alternativas del mismo gen y otra sería que sean genes diferentes los que codifican para cada una de estas proteínas. Como se sabe que existen diferencias en por lo menos dos aminoácidos entre las proteínas hsp10 y EPF, es mas probable que se trate de dos genes diferentes.

En el mismo reporte de Summers *et al* (2001), identifican el gen del EPF en el genoma del ratón. Reportan que se trata de una secuencia sin intrones, que tiene tres diferencias a nivel de nucleótidos con respecto a la de la hsp10. Esto les indica que se trata de genes diferentes, además de que en el genoma del ratón ya se ha reportado la secuencia del gen de la hsp10 con intrones (Ryan *et al*, 1997). Hasta mayo de 2002 no existe ningún reporte en la base de datos Genbank del gen del EPF en el genoma humano, por lo cual en este trabajo se pretende encontrar la posible localización de este gen.

OBJETIVOS:

- Establecer la localización exacta de los genes de la hsp10 y la hsp60 en el genoma humano.
- Identificar el número y las características de los retropseudogenes de las dos hsp en la secuencia del genoma humano
- Determinar la posibilidad de que algunos de éstos retropseudogenes fueran funcionales en algún momento.
- Datar las secuencias de los retropseudogenes, así como determinar el posible origen de éstos, por medio de árboles filogenéticos de las secuencias.
- Calcular la tasa de generación de pseudogenes para estas dos familias.
- Identificar la secuencia o secuencias que tengan posibilidades de ser el gen del EPF.

METODOLOGÍA:

Basic Local Alignment Search Tool (BLAST)

Las búsquedas en el genoma humano fueron realizadas en el programa BLAST: "Basic Local Alignment Search Tool" (Altschul *et al,* 1990). Se utilizó la base de datos "genome" en donde se encuentran concentrados los datos del Proyecto de Secuenciación del Genoma Humano, el subprograma utilizado fue el "blastn" que hace una búsqueda nucleótidos *vs.* nucleótidos.

El blast tiene las siguientes opciones de búsqueda:

- **"Expect"**: este parámetro asigna el número de secuencias que pueden ser encontradas por azar, acorde al modelo estocástico de Karlin y Altschul (1990). De esta manera si la probabilidad estadística dada al acierto es mas grande que el valor de Expect, el acierto no será reportado como tal. Así los valores bajos del Expect, llevan a menos probabilidades de que un acierto sea reportado.
- **"Filter"**: con esta opción se desvanecen segmentos de la secuencia que tienen una baja complejidad composicional, esto determinado por el programa DUST de Tatusov y Lipman (en preparación). El filtro así elimina reportes significativos estadísticamente pero sin interés biológico. Asimismo está en desarrollo un filtro que elimine los reportes de las regiones repetitivas como son LINE's y SINE's.
- **"G"**: se refiere al valor que se le otorga a insertar un gap al momento de alinear las secuencias y obtener después la calificación del acierto. El valor predeterminado es 5.
- **"E"**: valor negativo que se le otorga a la longitud del gap (medida en bases); de esta manera un gap de 3 bases tienen un valor negativo para la calificación de la secuencia de 3 veces el valor de E. El valor predeterminado es de 2.
- **"Q"**: penalidad otorgada a las diferencias en el alineamiento. El valor predeterminado es -3.

- **"R"**: valor otorgado a una similitud en el alineamiento. El valor predeterminado es 1.
- **"W"**: tamaño de la palabra. La palabra es considerada un grupo de bases de longitud N que tienen que ser 100% similares por lo menos una vez entre las dos secuencias para que sean reportadas como similares. El valor predeterminado es de 11.
- **Megablast**: utiliza un algoritmo mas codicioso para la búsqueda, diferente que el del blastn, el **NCBI** indica que es útil para buscar secuencias que varían poco debido a la secuenciación o a otros errores similares.

Los parámetros anteriores son opciones del análisis para buscar las secuencias; el programa BLAST tiene más opciones pero se utilizan en la presentación de los resultados y no alteran el algoritmo de búsqueda de secuencias similares. Cabe mencionar que algunos parámetros como el "W" tienen un límite para hacer una búsqueda en la base de datos del genoma, así el programa no acepta búsquedas con un tamaño de palabra menor a 7, y utilizando el Megablast, solamente se permite un tamaño de palabra de 8.

En cada proceso de los resultados en que fue usado el blast, únicamente se nombraran los parámetros que sean diferentes a los predeterminados por el programa.

Resultados del BLAST

Al terminar la búsqueda, el blast presenta los resultados, mostrando el número de aciertos ("Hits") que encontró y una gráfica (como la figura 1) en donde se observan las diferentes secuencias encontradas con respecto a la secuencia buscada ("query"); los colores de las secuencias encontradas representan el valor que el blast les determinó, después de evaluar cada uno de los nucleótidos de acuerdo con la matriz y/o con los parámetros determinados como opcionales (Q, R, G y E). Después el programa presenta una tabla con las secuencias de los diferentes cromosomas donde se encontró por lo menos un acierto en la búsqueda; también da la opción de mostrar un mapa de todos los cromosomas con los aciertos que se encontraron en cada uno, como en la figura 5. El otro

punto importante en cualquier resultado del blast, son los alineamientos pareados de cada secuencia de la búsqueda, lo cual proporciona los siguientes datos:

```
>ref|NT_023097.8|Hs5_23253 Homo sapiens chromosome 5 working draft sequence segment
      Length = 806620

 Score = 3975 bits (2005), Expect = 0.0
 Identities = 2180/2237 (97%), Gaps = 1/2237 (0%)
 Strand = Plus / Minus
```

la primera línea corresponde al número de referencia de la secuencia donde hubo un acierto, así como el nombre, que en este caso es del cromosoma 5. A continuación se muestra la calificación otorgada en "bits", que para este caso es 3975, entre paréntesis se muestra la calificación pero sin considerar los valores que se le otorgaron a cada uno de los parámetros, por lo cual el valor que se puede comparar con otros aciertos es el de los "bits". Después marca el valor de "Expect" (definido antes) que se calculó para esta secuencia, las identidades encontradas, los gaps utilizados para el alineamiento y el sentido de las dos secuencias; en este caso se indica que la secuencia buscada está en sentido positivo ("plus") y la encontrada en negativo ("minus"), y es como se muestra a continuación:

```
Query: 1      cctcactcgccgccgacgacctgtctcgccgagcgcacgccttgccgccgccccgcagaa  60
              +||||||||||||||||| |||||||||||| ||||| ||| ||| |||||||| |||||
Sbjct:210253cctcactcgccgccgacaacctgtctcgccgcgcgcatgccctgcagccgccccacagaa 210194
```

de manera que mientras la secuencia buscada va en un sentido la encontrada va en el otro. Al final del resultado se muestran los parámetros que fueron utilizados para la búsqueda:

```
Database: Homo sapiens genomic contig sequences
Posted date:  Feb 8, 2002 10:36 AM
Number of letters in database: -1,436,328,255
Number of sequences in database:  2991

Lambda    K     H
1.37   0.711    1.31

Gapped
Lambda    K     H
1.37   0.711    1.31

Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 5,310,274
Number of Sequences: 2991
Number of extensions: 5310274
```

Number of successful extensions: 542
Number of sequences better than 1.0e-02: 21
length of query: 2242
length of database: 2,858,639,037
effective HSP length: 22
effective length of query: 2220
effective length of database: 2,858,573,235
effective search space: 6346032581700
effective search space used: 6346032581700
T: 0
A: 0
X1: 6 (11.9 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 25 (50.1 bits)

Para obtener las secuencias de los aciertos a las búsquedas (los pseudogenes), se utilizó el Map View del NCBI, en donde se pueden escoger los nucleótidos del genoma humano que se quieren obtener, y de esta manera se pueden bajar las secuencias en formato Fasta.

Método de análisis de marcos de lectura de Shepherd

Usando métodos estadísticos, Shepherd (1984) demuestra que las bases en una secuencia codificante se presentan bajo la regla RNY, donde R son las purinas, Y las pirimidinas y N cualquiera de las dos; de esta manera cuando examinamos la fase de lectura de una secuencia, podemos ver que los codones verdaderos presentan preferencialmente esta distribución, lo cual nos permite inferir si la secuencia es codificante o no. Staden (1984), reporta que al usar los codones preferenciales las secuencias cumplen con esta proporción de nucleótidos.

El programa de análisis de Shepherd es parte del paquete informático PC Gene de Intellegenetics Inc. and Genofit, Ver-6-01 (1989).

Para el análisis, una secuencia de nucleótidos de una longitud determinada, L, es convertida a una serie de purinas R, y pirimidinas Y, y a partir del primer nucleótido, la secuencia es examinada en los tres marcos de lectura posibles en una determinada polaridad para definir cuál de las tres fases muestra la menor desviación de una secuencia patrón formada únicamente de tripletes RNY. El programa divide la secuencia bajo análisis en grupos de 3 bases -tripletes- analizando los primeros tres nucleótidos

para la fase de lectura 1, los nucleótidos 2, 3 y 4 para la fase de lectura 2 y los nucleótidos 3, 4 y 5 para las fase de lectura 3. Por iteración, el programa analiza toda la longitud dada, registrando cuál marco de lectura se ajusta con mayor aproximación al serial RNY. El marco de lectura que presenta la menor desviación respecto a este codón de referencia es registrado y graficado en el punto medio de la longitud L, indicando que esta fase de lectura es la que muestra el menor número de desviaciones respecto al codón serial. El procedimiento se repite después de avanzar S número de bases (paso de avance) siempre en la dirección 5'- 3' a lo largo de la secuencia seleccionada.

Un punto importante de señalar es el hecho de que en la selección de la extensión del paso de avance S, mientras más pequeño sea este paso, mayor será la sensibilidad del análisis, ya que al mantener sin variación la longitud de análisis (ventana) y avanzar una unidad (tres pares de bases es la mínima unidad de paso de avance en el programa de Shepherd), el cálculo se vuelve a repetir, obteniéndose un mayor detalle de cada sección de la secuencia; esto es debido a que cada ventana se sobrelapa con la sección previa en toda su longitud, menos la unidad de avance (Shepherd, 1981, 1990; Staden, 1990). De esta manera, si escogemos una sección de longitud L de 60 nucleótidos y un paso de avance S de 3 nucleótidos, cada triplete sucesivo será inspeccionado desde la base 1 a la base 60 (fase 1), desde la base 2 hasta la base 61 (fase 2) y desde la base 3 a la base 63 (fase 3); una vez que el programa haya analizado los primeros 60, avanzará hasta el nucleótido 63 y realizará nuevamente la inspección de 60 nucleótidos, comenzando este nuevo análisis a partir del nucleótido 3 de la sección previa, y por iteración, hasta el final de la secuencia.

En los casos en donde se presentan valores iguales mínimos de desviación en dos ó tres fases de lectura respecto al serial RNY, el cómputo se lleva a cabo sobre una extensión mayor pero aún centrado sobre el mismo punto medio de la sección previa.

Los resultados son presentados en forma gráfica: la secuencia es representada en el eje de las X y los resultados analíticos son graficados en el eje de las Y, en el cual se registran los tres marcos de lectura, con sus respectivas señales de terminación si es que están presentes. La gráfica indica cuál de los tres marcos de lectura presenta la menor desviación del serial RNY en cada triplete a lo largo de la secuencia, señalando con un punto dicha posición; si una fase de lectura muestra la menor desviación en muchas

posiciones consecutivas (tripletes contiguos), los puntos producirán una línea continua en el marco de lectura correspondiente, indicando la fase con mayor probabilidad de codificación potencial. Por el contrario, si los puntos muestran alternancia entre 2 ó 3 fases, la gráfica mostrará una serie de picos y mesetas, indicando desviaciones respecto al formato RNY entre 2 ó 3 marcos. En otros casos, una fase de lectura puede mostrar una preponderancia de la seriación RNY -línea continua- punteada por desviaciones hacia los otros marcos de lectura (estos picos parásitos se pueden interpretar como eventos mutacionales).

Con base en observaciones de los patrones de gráficas obtenidas de secuencias codificantes, se ha propuesto que las regiones codificantes muestran en general pocos cambios de marcos de lectura, y cuando esto sucede, se observa principalmente en genes con intrones, en donde los exones pueden ser leídos en distintas fases, mientras que en regiones no codificantes no existe una preferencia hacia una fase de lectura determinada en la cual se conserve el formato RNY (Stormo, 1987; Shepherd, 1990). La cuantificación del porcentaje del serial RNY se llevó a cabo midiendo la extensión del **ORF** (marco de lectura abierto) en la cual se presenta el formato RNY, considerándose el total de bases como el 100%. De esta manera, los ORFs analizados presentaron porcentajes variables del serial RNY, lo cual, según el criterio de Shepherd, está en relación con su potencial de codificación

Alineamientos pareados y múltiples

Los alineamientos tanto múltiples como pareados fueron hechos utilizando el programa Clustal W 1.6 (Thompson, *et al.*, 1994). Este algoritmo hace el alineamiento en tres etapas: primero, da un valor a la similitud de una pareja de secuencias con respecto a una matriz de evaluación de nucleótidos, con lo cual representa la distancia entre éstas; después mide la similitud de todas las secuencias basado en la similitud de los pares y crea un dendograma al agrupar las secuencias basado en la similitud entre todas; por último crea un alineamiento múltiple empezando con las secuencias más similares y dejando al final las menos similares, basado en el dendograma. Estos dendogramas fueron graficados usando el programa TreeView que solamente nos otorga una interfase

gráfica que interpreta los resultados del Clustal W, pero no realiza ningún cálculo con las secuencias.

El manejo en general de las secuencias y la búsqueda de los marcos de lectura abiertos fueron hechos con OMIGA 2.0 de Oxford Molecular Ltd., que también sirvió como interfase para el uso del Clustal W.

Datación de las secuencias

Para datar las secuencias se utilizaron tres métodos diferentes que se describen a continuación:

Método 1.

Friedberg, *et al* (2000), reportan que es posible usar las diferencias que tienen los pseudogenes con la secuencia de dónde se originaron (el gen) y su homólogo en otra especie para determinar su origen. Este método puede ser usado para cualquier secuencia codificante de la que se quiera conocer su tiempo de divergencia. Con este método estamos asumiendo que los pseudogenes mantienen la tasa de mutación que tiene el gen, aunque ya no sean funcionales. Para llevar a cabo este método necesitamos las diferencias de los pseudogenes con el gen de humano, de los pseudogenes con el gen de ratón, y la diferencia entre el gen del ratón y el del humano. Friedberg, *et al* (2000), hacen la siguiente anotación "... el uso de genes *bona fide* para calcular el origen de los pseudogenes es cuestionable ya que estos no son secuencias funcionales...", pero aún con esta anotación, estos autores defienden el uso de este método.

Método 2.

Li (1991) señala que como los pseudogenes no están sujetos a ninguna restricción funcional, se espera que tengan una tasa de mutación más alta. Incluso, indica que acorde con la hipótesis de las **mutaciones neutrales**, la tasa de sustituciones de nucleótidos se espera sea más grande para genes menos importantes funcionalmente o para partes de genes, que para los genes o partes que son más importantes funcionalmente –el problema reside en definir sin ambigüedades qué genes son importantes- , ya que es mas probable

que los últimos estén sujetos a selecciones purificantes negativas. Este autor reporta para los pseudogenes una tasa de mutación de 3.55×10^{-9} substituciones por sitio por año, lo que es igual a 0.00355 substituciones por sitio por millón de años y es con esta tasa que se calcula la edad de los pseudogenes. Hacen notar en este trabajo que esta tasa puede no ser aplicable para otros pseudogenes que los usados en el cálculo y que estas tasas dan una idea de las diferencias entre las distintas regiones de DNA (Li, 1981).

Método 3.

Tomando como hipótesis de trabajo que los pseudogenes tienen una tasa de mutación más grande que las secuencias codificantes, queda claro que es más confiable usar secuencias no codificantes como referencias para establecer la edad de éstos. Por otro lado, Li (1981), reporta que es mejor usar secuencias homólogas ya que es más probable que tengan tasas de mutación parecidas. Intentando cumplir con estros preceptos, decidimos buscar secuencias no codificantes, que estuvieran relacionadas con las hsp, y que se pudieran identificar en el humano y en otro grupo que se conociera el tiempo de divergencia; para esto, se usó un intrón de la hsp10 de 625 pb de longitud que está presente tanto en el gen de la rata como en el del humano. Asumimos que los roedores se separaron de la línea de los primates hace 85 Ma (Li, 1997). Estos datos nos permitirán datar los pseudogenes de las hsp'sn respecto a una secuencia no codificante de esta misma familia.

Los parámetros en común para los tres métodos fueron: el descartar las inserciones y/o deleciones de las secuencias en el conteo de las substituciones (Li, *et al*, 1981) y en todos los casos sólo se tomó en cuenta la región comprendida entre el codón de inicio y el codón de termino, por lo cual el tamaño (L) es el que entra en esta región, y (N) las substituciones que se encontraron en esta región.

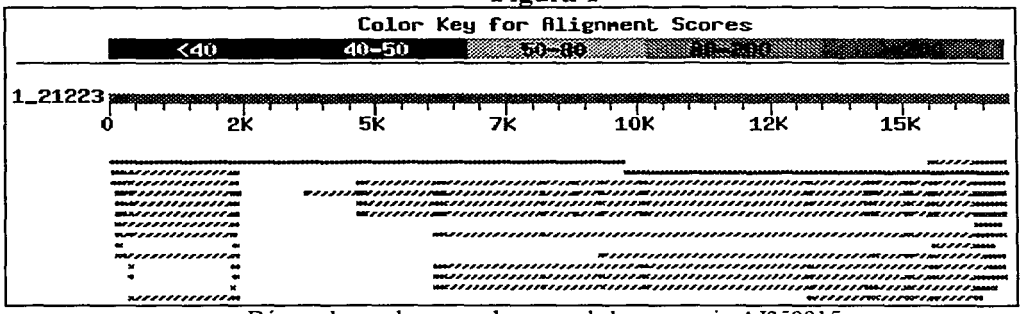
RESULTADOS Y DISCUSION:

Identificación de los genes hsp10 y hsp60 en el genoma humano

Con el fin de buscar la secuencia de los genes de la hsp10 y la hsp60 en el genoma humano, se realizó una búsqueda mediante el blastn (Expect = 0.01 y filtro default) de la secuencia con número de acceso AJ250915 de 16,986 bases (Hansen, *et al,* 2000) que contiene los genes hsp10 y hsp60 así como el promotor para ambos genes. Los resultados de esta búsqueda se pueden ver en la figura 1. Las primeras dos líneas contínuas representan el resultado con el mayor grado de similitud (99%), que muestra la concordancia de la secuencia AJ250915 en el genoma humano, en el cromosoma 2 en la región 2q33.1. En la gráfica original, el color de las líneas representa la puntuación que el blastn le otorga a cada una de las secuencias, siendo el rojo más de 200 puntos, mediante la matriz de sustitución. Las otras líneas están separadas ya que las fracciones del genoma humano en el blastn todavía se encuentran separadas una de la otra, debido a que al secuenciarlo se generaron diferentes fracciones que todavía no están empalmados en su totalidad, aunque estas dos secuencias están juntas en la misma región. Cabe notar que esta búsqueda fue realizada para ubicar el gen y por eso se utilizó la secuencia completa (con todo e intrones) pero aún así, todos los resultados truncados que se pueden ver en la gráfica son secuencias que representan homólogos de los exones de los dos genes repartidos en otros cromosomas. La identificación de una sola secuencia con intrones indica que el resto de las secuencias detectadas corresponden a copias "procesadas" sin intrones. A la secuencia obtenida se le llamará de aquí en adelante como "Secuencia 1".

Se realizó un alineamiento pareado entre la secuencia AJ250915 y la Secuencia 1 para determinar las diferencias entre las dos, utilizando el Clustal W versión 1.6 (Gap: Existencia: 10, Extensión: 5), (Thompson, *et al.*, 1994). La Secuencia 1 tiene 15 diferentes nucleótidos con respecto a la AJ250915. Tiene además 7 inserciones de una base cada una, y 10 deleciones también de una base cada una; todas ellas estan localizadas en intrones y solamente tiene una diferencia en un exon que no cambia el

Figura 1



Búsqueda en el genoma humano de la secuencia AJ250915.

aminoácido codificado por este triplete. Estas diferencias se pueden deber a errores en alguna de las dos secuenciaciones o a la presencia de algunos **SNP's** (polimorfismos sencillos de nucleótidos) en estas regiones, ya que se han calculado que existe un SNP cada 1,000 nucleótidos aproximadamente (Lewis, 2002). En la figura 2 se muestra el diagrama de las diferentes partes funcionales de la Secuencia 1. Los cuatro exones a la izquierda del promotor son los correspondientes a la hsp10, y los 12 del lado derecho son los de la hsp60. Solamente existe un promotor para los dos genes ya que este tiene una función bidireccional.

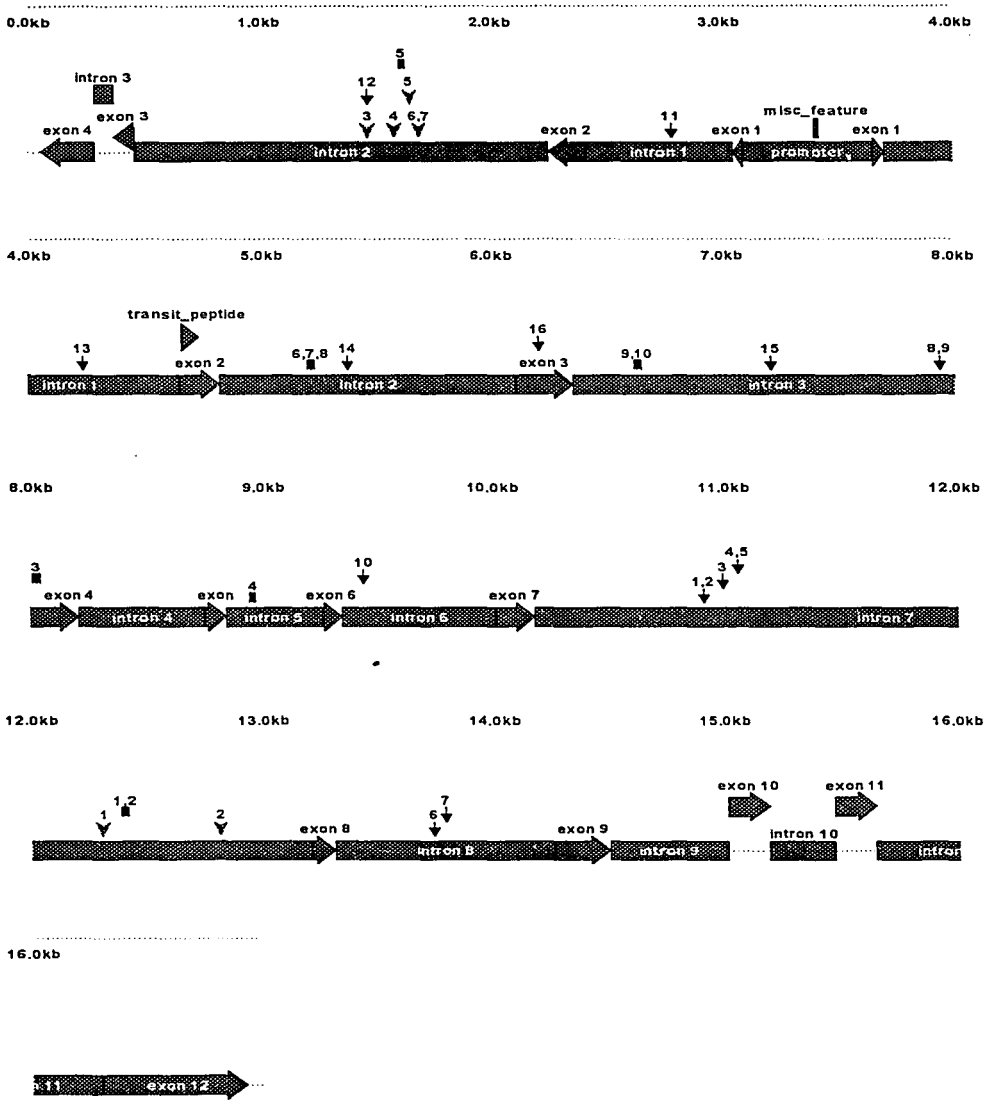
A partir de la Secuencia 1, se empalmó *in silico* el cDNA de la hsp10 y de la hsp60, uniendo todos los exones usando los límites que se reportan en la AJ250915, resultando en dos secuencias: hsp10 de 538 pb y hsp60 de 2242 pb.

En la tabla 1 se muestra la proporción de nucleótidos encontrada en los cDNA de los genes. La hsp10 presenta un contenido de GC de 38.5% y la hsp60 41.75%, con lo cual podemos decir que estos dos genes son pobres en contenido de GC. Recordando a Gonçalves *et al* (2000), esta es una de las características de los genes que generan mas retropseudogenes.

Tabla1

	HSP10		HSP60	
A	179	33.27%	705	31.45%
C	90	16.73%	389	17.35%
G	117	21.75%	547	24.4%
T	152	28.25%	601	26.81%

Figura 2



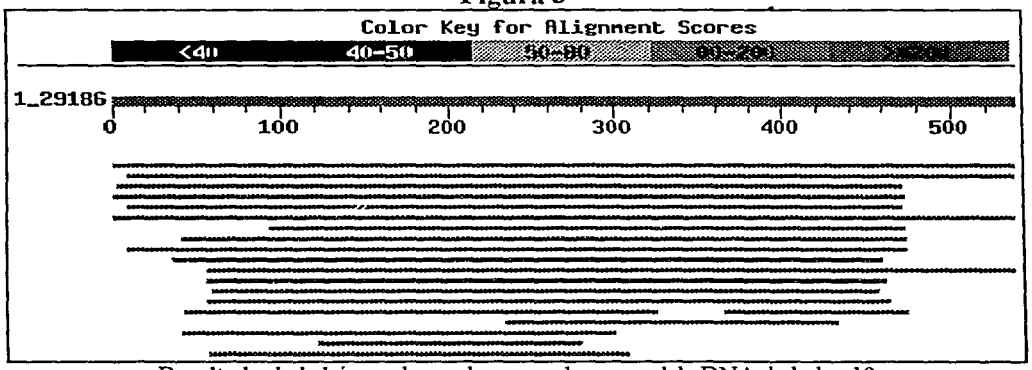
El Diagrama de la Secuencia 1. Las flechas indican cambios que tiene esta secuencia en comparación con la AJ250915, las cabezas de flechas son inserciones, y los cuadros deleciones.

Búsqueda de los pseudogenes

Usando los cDNA's obtenidos de la Secuencia 1, se usó el blastn (Expect: 0.01 y filtro: default) para localizar las secuencias que pudieran tener una similitud importante con estos genes. En la Figura 3, se presentan los "hits" obtenidos en el blastn para el cDNA de la hsp10 y en la Figura 4 para la hsp60. La búsqueda de la hsp60 dió como resultado 42 hits, y la de la hsp10 31 hits. Estos valores no indican en definitiva el número de pseudogenes, ya que debido al algoritmo del Blastn, algunos "hits" corresponden a regiones del "query" separadas por algunos nucleótidos y que fueron unidos para generar una secuencia continua, ya que se encontraron más pseudogenes de la hsp10 (24) que de la hsp60 (15). El hecho de que el HSP10 tenga más pseudogenes podría deberse a su tamaño menor.

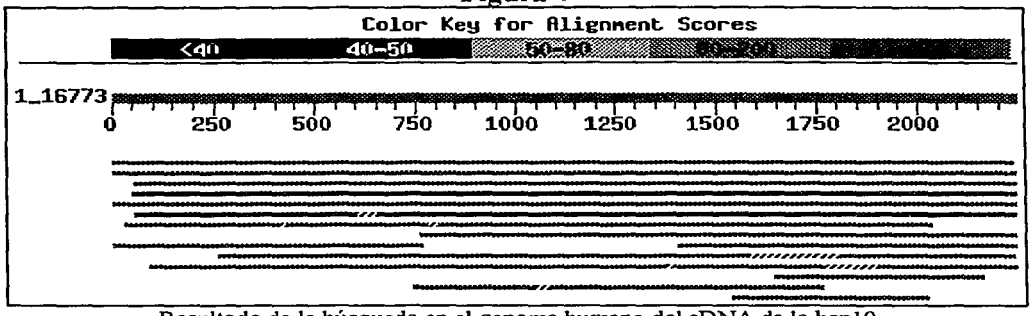
Como se puede ver en las figuras 3 y 4, la mayoría de los aciertos no corresponden a la longitud total de la secuencia buscada, ya que al momento de la retrotransposición es muy probable que no se copien todos los nucleótidos y que por lo tanto queden incompletos, o que después, al ser no funcionales presentan algunas inserciones o deleciones que provoquen que no estén completos.

Figura 3



Resultado de la búsqueda en el genoma humano del cDNA de la hsp10

Figura 4



Resultado de la búsqueda en el genoma humano del cDNA de la hsp10

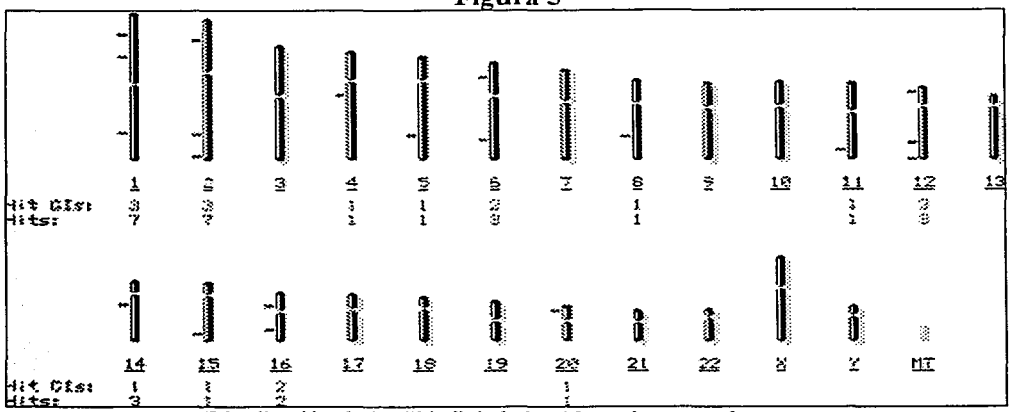
En las figuras 5 y 6 se muestra la distribución en el genoma humano de los "hits" para los dos genes. Los números en la primera línea representan el número del cromosoma. Los números en la fila de "Hit GIs" son los diferentes segmentos del genoma humano en donde se encontró alguna similitud con la búsqueda, y los números de abajo en la fila "Hits" representan cada uno de los "hits". Como se puede ver, en muchos casos hay más de un hit por cada región, esto se debe a que en algunos casos el blast reconoce un pseudogen como dos aciertos, ya que al tener diferentes similitudes en diferentes regiones, el algoritmo los interpreta como aciertos diferentes pero en realidad son una secuencia continua. Solamente hay seis cromosomas que tienen pseudogenes de las dos proteínas, el 2, 5, 6, 8, 12 y el 13, y ocho cromosomas que solamente tienen pseudogenes de una de las dos proteínas, lo cual nos indica que al momento de copiarse no van de la mano un gen con el otro; este resultado nos sugiere que se duplicaron por eventos independientes.

Historia de los pseudogenes

Varios de los pseudogenes encontrados no se presentaban en el genoma en forma continua, esto es, en algunos casos estaban divididos por inserciones de mayor o menor tamaño. Para el análisis evolutivo, se unieron estas partes con el fin de que las inserciones

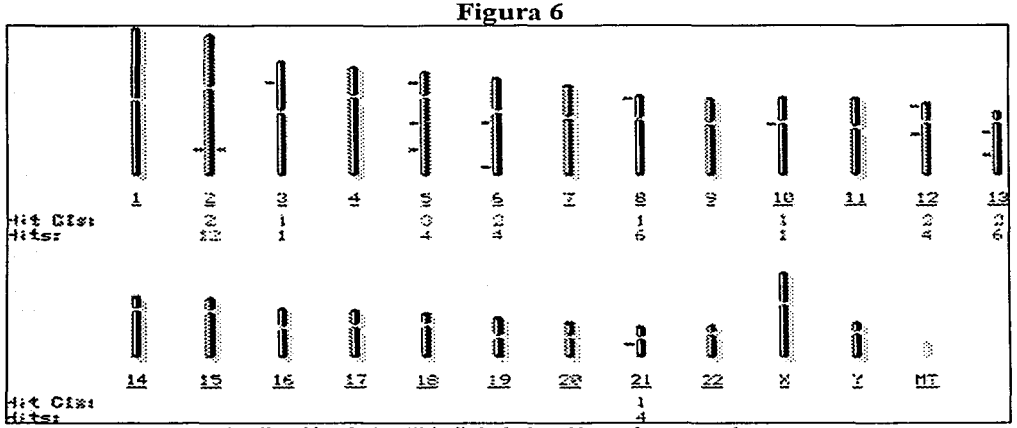
sin importar su origen, no influyeran en la datación de las secuencias, ya que como algunos autores lo han sugerido (Li, *et al,* 1981) es mejor no utilizar las inserciones ni las

Figura 5



Distribución de los "hits" de la hsp10 en el genoma humano

Figura 6



Distribución de los "hits" de la hsp60 en el genoma humano

deleciones en la datación, ya que estos son eventos que no están relacionados directamente con la edad de las secuencias y en muchos casos tienen más que ver con la localización genómica que con la edad de las secuencias. En la tabla 2, se presentan los pseudogenes de la hsp10, indicándose el tipo de arreglo así como las peculiaridades de

cada uno. En la tabla 3 se representan los de la hsp60. En estas tablas solamente se presentan los pseudogenes con peculiaridades y o arreglos.

Tabla 2

Observaciones del arreglo hecho a cada uno de los pseudogenes de la hsp10.

1adHSP10 (456 pb)	Dividido originalmente en dos aciertos, el 1a y el 1b. Esta división es provocada por el algoritmo del blast y no por inserciones.
1bcHSP10 (456 pb)	Es exactamente el mismo caso que el 1ad, y de hecho son 100% similares el 1ad *vs* 1bc, lo que sugiere una duplicación reciente. El 1ad y el 1bc están separados en el cromosoma 1 por 224 kb.
1efHSP10 (379 pb)	Entre el 1e y el 1f hay una inserción de 255 pb, que fue retirada para unir al pseudogen.
1ghHSP10 (247 pb)	Entre el 1h y el 1g hay una inserción de 275 bases, que también fue retirada para unir las dos partes.
2abHSP10 (466 pb)	El 2a y el 2b tienen una inserción entre ellos de 302 pb, que fue retirada para unir las dos partes.
4abHSP10 (434 pb)	División provocada por el algoritmo del blast.*
5abHSP10 (420 pb)	División causada por el algoritmo del blast.
6aHSP10 y 6bHSP10 (416 pb c/uno)	Tienen 100% de similitud entre ellos, seguramente son una duplicación. Están separados por 17 Mb en el cromosoma 6.
8abHSP10 (398 pb)	Tienen una inserción de 414 pb entre los dos segmentos encontrados, el 8a y 8b. Se retira esta inserción y se unen en un sólo pseudogen
16bcHsp10 (399 pb)	Tienen una inserción de 30 bases entre las dos secuencias (16c y 16b), que fue retirada para unir las dos partes.

Los paréntesis de la columna izquierda indican el tamaño de cada uno de los pseudogenes al final de los arreglos. * Estas secuencias son aciertos diferentes debido a que el algoritmo del blast los divide en dos partes, aunque son secuencias continuas en el genoma humano.

Tabla 3

Observaciones del arreglo hecho a cada uno de los pseudogenes de la hsp60

05egfcdHSP60	Dividido por el blast*
06abHSP60	Dividido por el blast

08aceHSP60	Dividido por el blast
08bdHSP60	Dividido por el blast
11baHSP60	Dividido por el blast
12cdbHSP60	Esta dividido en tres fragmentos, el 12c está separado del 12d por una inserción de 300 pb, y el 12d esta separado del 12b por una inserción de 321 pb, se quitaron las inserciones y se unieron los tres. 12b, 12d y 12c.
13baHSP60	Dividido por el blast.
13dcfgeHSP60	Entre estos 5 aciertos, los únicos divididos por una inserción son el 13d y el 13c que tienen 309 pb, entre ellos.
21bcdaHSP60	Entre estos 4 aciertos, los únicos divididos por una inserción son el 21c y el 21d que tienen 300 pb, entre ellos

* Estas secuencias son aciertos diferentes debido a que el algoritmo del blast los divide en dos partes, aunque son secuencias continuas en el genoma humano.

Entre las peculiaridades de los "hits", figura la encontrada en el pseudogen 16aHSP10, que se ubica en una región donde el NCBI tiene reportado el gen de la Cadherina 3. Al analizar más detalladamente esta región, se pudo ver, como se muestra en la figura 7, que el pseudogen se encuentra en un intrón del gen de la Cadherina, lo cual indica que es un pseudogen anidado. Esto nos demuestra que los retropseudogenes no tienen restricciones en el sitio de inserción más que las que afecten la adecuación del organismo y resulten en una mutación selectivamente negativa.

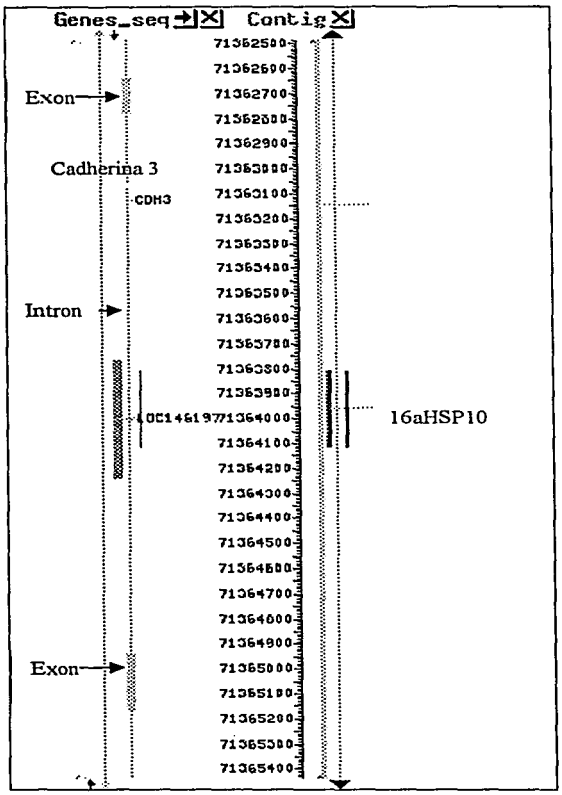
Por otro lado, se identificaron tres de las inserciones que separan a los pseudogenes en diferentes segmentos, estas tienen una identidad promedio de 75% con respecto a la secuencia Alu reportada por Blakey (2000).

Análisis de los Pseudogenes

Para poder identificar si las secuencias obtenidas de la búsqueda de las secuencias homólogas de los dos genes son pseudogenes o secuencias funcionales, se realizaron dos análisis. En el primero, se buscaron los marcos de lectura abiertos (ORF). Como un ejemplo en la figura 8 se presenta la búsqueda de los marcos de lectura abiertos en la

secuencia del pseudogen 12HSP60. Este análisis fue realizado con el Omiga 2.0 de Oxford Molecular Ltd. En la tabla 4, se pueden ver los diferentes tamaños de los ORF's

Figura 7



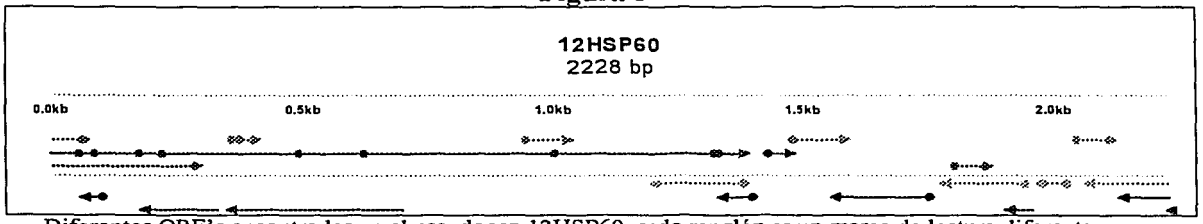
La columna de la izquierda representa al gen de la Cadherina 3. Mientras que la columna de la derecha es el pseudogen 16aHSP10 que como se ve esta en uno de los intrones de la Cadherina 3. Las dos columnas representan la misma región genómica, lo único que cambia son las anotaciones. La numeración se refiere al número de nucleótidos del transecto genómico presentado.

de todos los pseudogenes de la hsp60 y como se observa, en este caso ningún pseudogen tiene un ORF del mismo tamaño que el del gen, por lo que podemos suponer que ninguno tiene capacidad codificante, o por lo menos no la equivalente para la hsp60. Solamente el pseudogen 12hsp60 tiene un ORF de gran tamaño (1344 pb) por lo cual es posible que

codificara para una proteína truncada pero similar a la hsp60, aunque también podría ser un pseudogen nuevo que todavía mantiene bien conservada parte de su secuencia.

En la figura 9 se grafican los tamaños de los ORF de los pseudogenes de la hsp10 así como de su cDNA. Es notable esta gráfica, ya que hay tres pseudogenes con un ORF del mismo tamaño o incluso mas grande que el presentado por el cDNA de este gen. En la figura 10 se muestra el ORF del pseudogen 1adHSP10 que es ligeramente más

Figura 8



Diferentes ORF's encontrados en el pseudogen 12HSP60, cada renglón es un marco de lectura diferente, los círculos son ATG y los triángulos codónes de término.

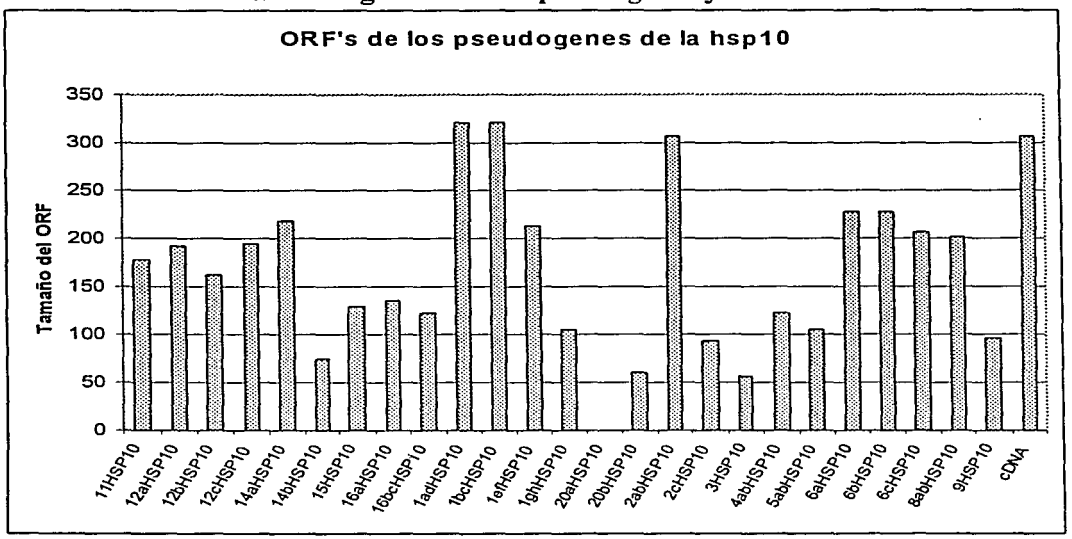
Tabla 4
Tamaño de los ORFs de los pseudogenes de la hsp60 y el cDNA

Secuencia	Tamaño del ORF mas grande
cDNA	1719
03HSP60	879
05bHSP60	485
05egfcHSP60	219
05HSP60	774
06abHSP60	174
06cHSP60	357
08aceHSP60	423
08bdfHSP60	423
10HSP60	129
11baHSP60	339
12cdbHSP60	429
12HSP60	1344
13baHSP60	399
13dcfgeHSP60	339
21bcdaHSP60	297

grande que el del cDNA. La traducción *in silico* de éstas dos secuencias y su alineamiento nos indica que tienen un 84% de identidad en aminoácidos. De esta manera aparentemente si las secuencias 1adHSP10 y el 1bcHSP10 se transcribieran y tradujeran podrían ser secuencias funcionales. Por otro lado, el 2abHSP10 tiene un ORF del mismo

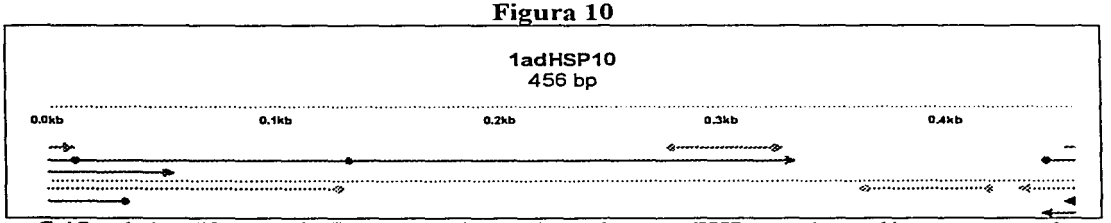
tamaño del cDNA, sin embargo hay que recordar de la tabla 2 que este pseudogen presenta una inserción que lo parte en dos. A este respecto nos planteamos la hipótesis de que esta inserción podría ser un intrón originado recientemente. Se realizó la búsqueda de las zonas conservadas en los intrones para su procesamiento y no se encontró ni el GT – AG, ni ninguna secuencia regulatoria similar a la secuencia lazo, por lo cual se descartó el catalogar a esa inserción como un intron, por lo que seguramente se trata de una secuencia que al no ser funcional sufrió una inserción que originó el pseudogen 2abHSP10. Todos los otros ORF's son truncados lo cual es un indicio de que no son codificantes.

Figura 9
Tamaño del ORF mas grande de los pseudogenes y el cDNA de la HSP10



Gráfica del tamaño de los ORF más grandes (en pares de bases) y del cDNA de la hsp10.

Figura 10



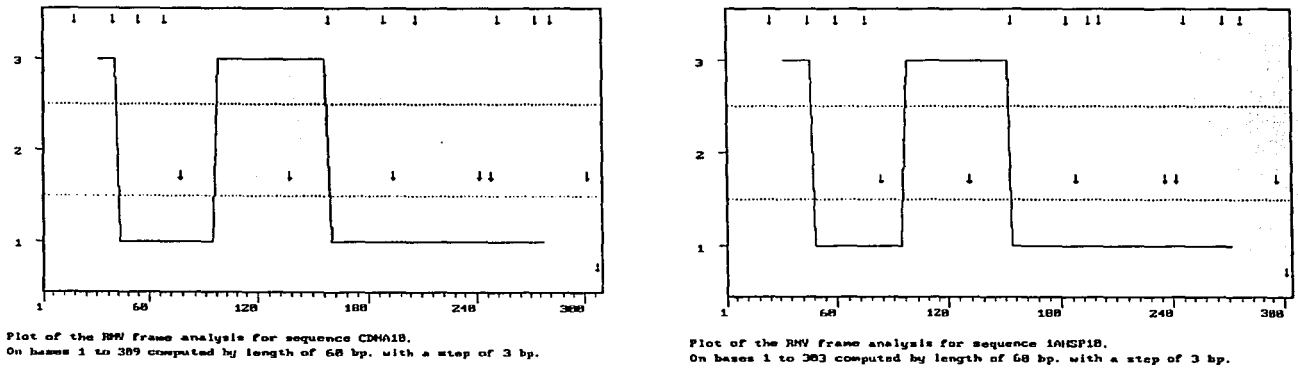
Gráfica de los diferentes ORF's encontrados en el pseudogen 1adHSP10, cada renglón es un marco de lectura diferente, los círculos son ATG y los triángulos codones de término.

Gráficas de Shepherd de los pseudogenes.

Como se explicó anteriormente, el método de Shepherd nos permite predecir probables secuencias codificantes. La figura 11, muestra las gráficas de Shepherd del cDNA de HSP10 y el pseudogen 1adHSP10. Como se puede apreciar estas gráficas son casi idénticas, por lo que se podría predecir que el pseudogen 1adHSP10 conserva la potencialidad de codificar un polipéptido. Como consecuencia el pseudogen 1adHSP10 y el cDNA deben de tener secuencias muy similares.

Al analizar las secuencias, obtuvimos que tienen una identidad del 95% en las secuencias de nucleótidos, y de un 81% a nivel de aminoácidos, el alineamiento a nivel de aminoácidos se presenta en la figura 12. En este alineamiento podemos ver que el pseudogen 1adhsp10 en caso de que fuera traducido, tendría un péptido líder con 10 aminoácidos más que el de hsp10. Asimismo, muestra una **deleción** de cuatro aminoácidos en la región media de la proteína. Adicionalmente, este análisis, nos deja

Figura 11

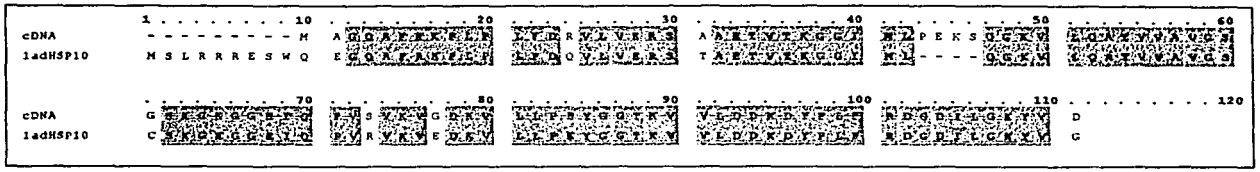


A la izquierda gráfica de Shepherd del cDNA de la HSP10, a la derecha la del pseudogen 1adHSP10. Las flechas indican codones de termino y las líneas continuas muestran el reconocimiento de la proporción RNY en los tres diferentes marcos de lectura, que están indicados en la parte izquierda.

ver que en general, los cambios de aminoácido son de tipo conservativo, excepto el cambio del aminoácido 61, en donde el 1adHSP10 tiene una cisteína. El ORF del pseudogen de la 1adHSP10 tiene su inicio en un codón diferente que el del cDNA, y es por esto que los primeros aminoácidos son diferentes. Por otro lado, como se puede ver en la figura 14, la secuencia de nucleótidos de la 1adHSP10 tiene una deleción de cuatro

bases (de la 51 a la 54) lo cual empalma las dos fases de lectura de manera que la identidad es muy alta de este punto en adelante.

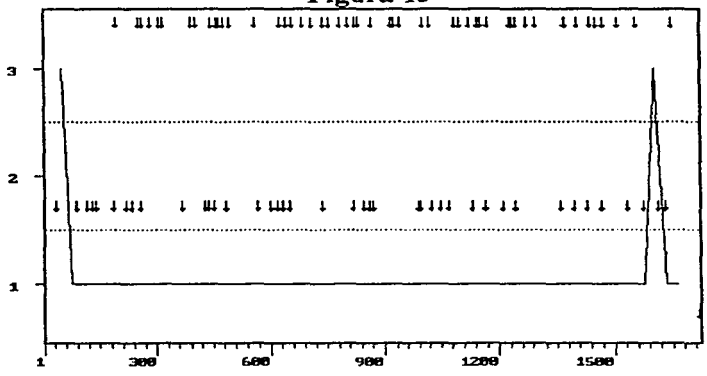
Figura 12



Alineamiento pareado de aminoácidos entre el pseudogen 1adhsp10 y el cDNA de la hsp10

La figura 13 muestra la gráfica de Shepherd del cDNA de la hsp60. Como se puede ver, en el primer marco de lectura se cumple casi a la perfección la regla RNY. Las gráficas de los demás pseudogenes de la hsp60 se muestran en el Anexo I. Las únicas que no se presentan son las del 05bHSP60 y el 10HSP60 ya que son pseudogenes muy cortos para que se pueda apreciar este análisis. La gráfica I.10 nos muestra que el pseudogen 12HSP60 además de tener un ORF de 1344 pb, tiene una buena distribución RNY por lo

Figura 13



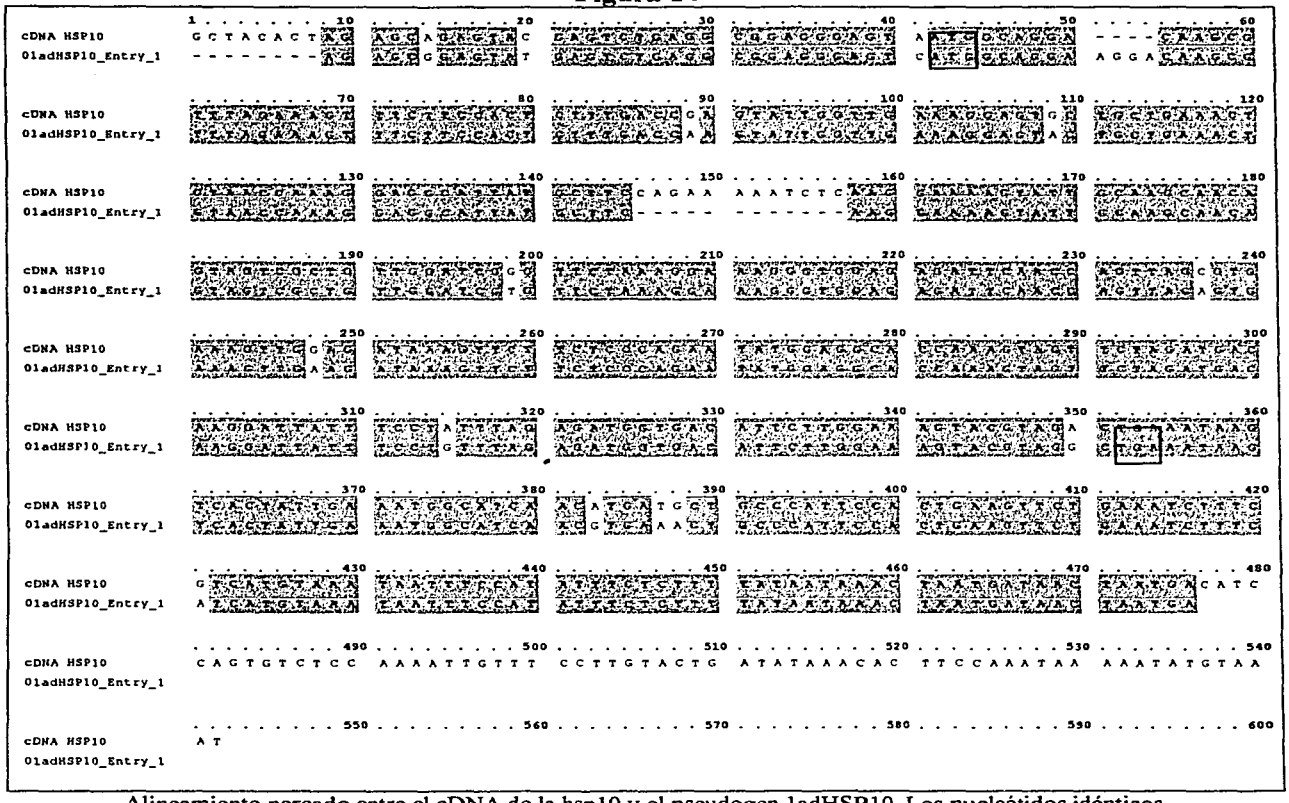
Los parámetros fueron: longitud de búsqueda de 90 pb, con pasos de 30 pb.

cual sería posible que esta secuencia fuera funcional, aunque como ya se dijo, puede ser muy reciente su origen y por eso conserve estas características. Al traducir *in silico* el pseudogen 12HSP60 se puede ver que tiene un 89% de identidad a nivel de aminoácidos con respecto al cDNA.

Alineamientos pareados con el cDNA y datación de los pseudogenes

La figura 14 es un ejemplo de un alineamiento pareado entre el cDNA de la hsp10 y uno de los pseudogenes (el 1adHSP10). Se realizaron alineamientos pareados de todos los pseudogenes *vs* el cDNA con el fin de calcular el número de mutaciones. También se obtuvo de estos alineamientos la posición del codón (primera, segunda o tercera posición) en donde se presentan cada una de las mutaciones. Se utilizó el Clustal W versión 1.4

Figura 14



Alineamiento pareado entre el cDNA de la hsp10 y el pseudogen 1adHSP10. Los nucleótidos idénticos están sombreados y las mutaciones no, resaltados están el codón de inicio y el de término del cDNA, que fue la región que se usó para el análisis de las mutaciones por posición en el codón..

(Gap: Existencia: 10, Extensión: 5) (Thompson, *et al.*, 1994). Hay que recordar que el conteo de las mutaciones únicamente se hizo desde el codón de inicio al de término. Como ejemplo, las secuencias de la figura 14 tienen solamente 7 mutaciones, ya que el

codón de inicio en estas secuencias está en la base 42 y el de término en la 352. En general, los alineamientos pareados de la hsp10 dieron valores de identidad de entre 83% y 97% y los de la hsp60 de entre 82% y 97%.

En la tabla 4 están los diferentes pseudogenes de la hsp10 con su respectiva datación, y en la tabla 5 los de la hsp60. A continuación se presenta un ejemplo de cada uno de los métodos:

Ejemplo del método 1:

- Pseudogen 11HSP10: Tamaño (L) = 290, Substituciones (N) = 32
- P = N/L = 32/290 = 0.1103 (P = índice de similitud)
- Usando la formula de Jukes (1969): d = -(3/4) Ln(1-((4/3)P)
- d = 0.1193 (Kab), esta es la tasa entre el pseudogen y el gen humano.
- Después se calcula la tasa entre el pseudogen y el gen de ratón, en donde lo que cambia es el número de sustituciones (N) = 44
- P = N/L = 44/290 = 0.1517; y usando Jukes (1969) d = 0.1695 (Kac), que es la tasa entre el pseudogen y la secuencia del ratón.
- Por otro lado se calcula la (kbc) entre el gen del ratón y el gen humano; L = 309, N = 29, P = 29/309 = 0.0938 y usando Jukes (1969) d = 0.1002
- Después se calcula el número de sustituciones por sitio por año (r), con la siguiente fórmula: r = Kac + Kbc /2(2T1), siendo T1 el tiempo de divergencia entre el ratón y el humano T1 = 85 Ma.
- r = 0.1695 + 0.1002 / 2(2(85) = 0.0007934
- Luego se usa T2 = Kab/2r, siendo T2 el tiempo de divergencia entre el pseudogen y el gen humano
- T2 = 0.1193 / 2(0.0007934) = **75.21 Ma**
- De esta manera se calculó la edad según el método 1 para cada uno de los pseudogenes tanto de la hsp10 como de la hsp60

Ejemplo del método 2:

- Para este método se utilizó la tasa de substitución que reporta Li para pseudogenes que es de 0.0035 substituciones por sitio por millón de años.

- Se calculó la P = N/L, usando el mismo pseudogen que en el ejemplo anterior P = 32/290 = 0.1103
- Si este pseudogen tiene esta tasa de substituciones actualmente, y damos por hecho que evoluciona a una tasa de substituciones de 0.0035 subst. por sitio por año; entonces T = P / 0.0035, siendo T el tiempo de divergencia, P la tasa que presenta el gen y 0.0035 la tasa que reporta Li (1981).
- T = 0.1103 / 0.0035 = **31.52 Ma**

Ejemplo del método 3:

- Se hizo el mismo procedimiento que en el 2 pero en lugar de la tasa de Li (1981) se utilizó la calculada con los intrones de la hsp10
- Los datos de los intrones son L = 625, N = 271, y el tiempo de divergencia es de 85 Ma.
- Entonces P = 271/625 = 0.4336 en 170 Ma (ya que se separaron hace 85 Ma, entonces esa tasa la alcanzaron en 170 Ma de substituciones por que cada uno acumuló substituciones de manera independiente), por lo que es 0.0025506 substituciones por sitio por millón de años.
- T = 0.1103 / 0.0025506 = **43.26 Ma**

Tabla 4
Datación de los pseudogenes de la hsp10

				Mutaciones	Datación métodos (Ma)		
	Pseudogen	**Cromosoma**	**Tamaño**	**(Identidad)**	**1**	**2**	**3**
1	1adHSP10	1	297	7(95%)	35.0	6.7	9.2
2	1bcHSP10	1	296	7(95%)	33.8	6.8	9.3
3	1efHSP10	1	257	10(95%)	48.2	11 1	15.3
4	1ghHSP10	1	235	16(89%)	80.7	19.5	26.7
5	2abHSP10	2	309	23(92%)	99.2	21.3	29.2
6	2cHSP10	2	282	33(86%)	109.4	33.4	45.9
7	3HSP10	3	278	30(85%)	120.6	30.8	42.3
8	4abHSP10	4	297	45(83%)	113.3	43.3	59.4
9	5abHSP10	5	299	30(86%)	85.0	28.7	39.3
10	6aHSP10	6	308	30(89%)	104.0	27.8	38.2
11	6bHSP10	6	309	30(89%)	104.0	27.7	38.1
12	6cHSP10	6	293	44(84%)	126.5	42.9	58.9
13	8abHSP10	8	280	29(88%)	130.3	29.6	40.6

14	9HSP10	9	260	30(87%)	117.1	33.0	45.2
15	11HSP10	11	290	32(85%)	119.7	31.5	43.3
16	12aHSP10	12	308	11(96%)	25.6	10.2	14.0
17	12bHSP10	12	303	28(90%)	109.4	26.4	36.2
18	12cHSP10	12	290	39(83%)	120.3	38.4	52.7
19	14aHSP10	14	306	6(97%)	29.0	5.6	7.7
20	14bHSP10	14	56	2(96%)	82.9	10.2	14.0
21	15HSP10	15	309	11(96%)	43.6	10.2	14.0
22	16aHSP10	16	308	11(96%)	49.0	10.2	14.0
23	16bcHsp10	16	307	30(88%)	99.2	27.9	38.3
24	20aHSP10	20	271	36(85%)	139.4	38.0	52.1

Tabla 5
Datación de los pseudogenes de la hsp60

				Mutaciones	**Datación métodos (Ma)**		
	Pseudogen	**Cromosoma**	**Tamaño**	**(Identidad)**	**1**	**2**	**3**
1	03HSP60	3	1709	99(93%)	46.7	16.6	22.7
2	05bHSP60	5	234	18(91%)	30.2	22.0	30.2
3	05egfcdHSP60	5	1252	155(86%)	64.3	35.4	48.5
4	05HSP60	5	1722	41(97%)	20.2	6.8	9.3
5	06abHSP60	6	1476	167(87%)	75.8	32.3	44.4
6	06cHSP60	6	391	26(92%)	45.1	19.0	26.1
7	08aceHSP60	8	1732	132(91%)	58.0	21.8	29.9
8	08bdHSP60	8	1722	133(91%)	58.0	22.1	30.3
9	10HSP60	10	128	6(94%)	35.2	13.4	18.4
10	11baHSP60	11	874	146(82%)	97.4	47.7	65.5
11	12cdbHSP60	12	1758	93(93%)	41.7	15.1	20.7
12	12HSP60	12	1763	54(96%)	26.5	8.8	12.0
13	13baHSP60	13	1663	172(88%)	73.3	29.6	40.6
14	13dcfgeHSP60	13	1647	183(87%)	81.1	31.7	43.6
15	21bcdaHSP60	21	1681	139(90%)	63.4	23.6	32.4

Calibración del blast para establecer los límites de la detección.

Dado que el porcentaje de identidad más bajo fue de 82% (que en realidad es un valor alto) nos preguntamos si no existen secuencias más antiguas –que tengan un menor porcentaje de identidad- o si teníamos una limitante del algoritmo. Con este propósito, se realizó una calibración de los parámetros del blastn. Usando un procedimiento al azar, se introdujeron gradualmente cambios para generar secuencias con 90, 85, 80, 75 y 70 % de similitud con respecto al cDNA de la hsp10.

Cada una de estas secuencias fue usada como "query" en el genoma humano mediante un blastn. Por el valor de su similitud con el cDNA se esperaba que todas detectaran secuencias homólogas en el genoma humano. Sin embargo, el valor mínimo de detección fue 80% y solamente en una pequeña región. La evaluación de los diferentes parámetros del blast nos indicó que el que mas influye es el del tamaño de la "palabra" (W). Este valor nos indica el número mínimo de bases contiguas que tienen que tener un 100% de similitud para que el blastn determine que son secuencias homólogas. Este valor esta predeterminado en 11 y solamente se permite disminuirlo a 8 en el Megablast, y a 7 en el Blast normal. Esto indica que para que el blast evalué una secuencia como homóloga, tiene que tener una región de por lo menos 7 pb con 100% de similitud.

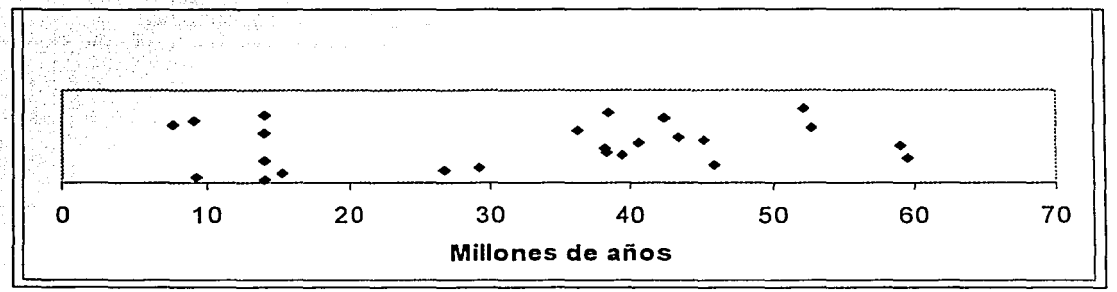
Al utilizar una secuencia de 70% de similitud estamos cambiando aproximadamente una de cada tres bases, por lo que es bastante improbable que quede una zona de 7, y mucho menos de 11 bases con 100% de similitud. El blast establece estos parámetros muy severos ya que debido a la magnitud de la base de datos, si se establecen parámetros menos rigurosos, esto da como resultado muchas secuencias que tienen una similitud por azar y no por ser secuencias homólogas. Todos los métodos de análisis de una u otra manera sufren de esta limitante de especificidad *vs.* sensibilidad (Moreno *et al*, 2001). Además un factor importante es el tiempo de cómputo que el NCBI esta utilizando en cada búsqueda, ya que con parámetros menos severos el tiempo de cada búsqueda aumentaría en gran medida.

En conclusión, por las limitantes metodológicas, no podemos identificar secuencias de pseudogenes con una similitud más baja que los aquí reportados. Como una opción para trabajos posteriores se sugiere hacer el cómputo en sistemas locales, para poder establecer los parámetros deseados sin limitaciones.

Gráficas de las dataciones

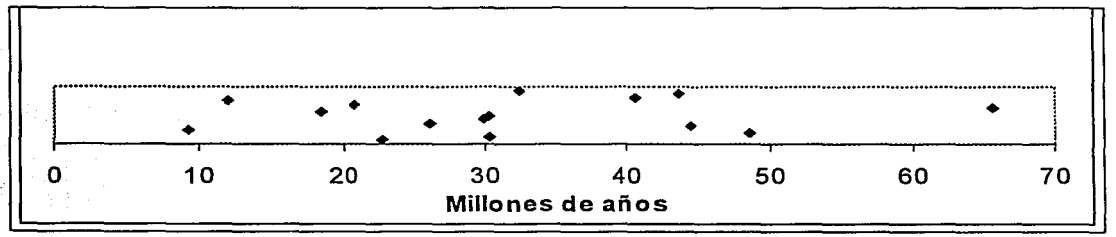
En las figuras 15 y 16 se presentan gráficamente los valores de datación de los pseudogenes calculados por el método 3. En estas, se puede notar la distribución en el tiempo (Millones de años) en el que se generaron los diferentes pseudogenes. Como se puede observar en la figura 15, el origen de los pseudogenes de la hsp10 no está

Figura 15



Distribución en el tiempo de los pseudogenes de la hsp10, usando los datos del método 3.

Figura 16



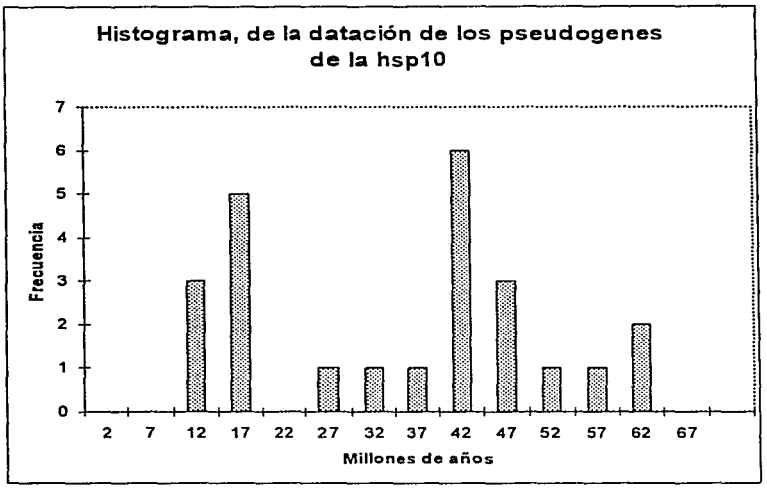
Distribución en el tiempo de los pseudogenes de la hsp60, usando los datos del método 3

uniformemente distribuído en el tiempo, de manera que hay algunos espacios en la escala del tiempo en donde se aprecia una mayor cantidad de aparición de pseudogenes. Por otro lado los pseudogenes de la hsp60 tienen orígenes mas distribuidos a lo largo del tiempo.

Con el fin de facilitar la visualización de los datos obtenidos de la datación de los pseudogenes, se realizó un histograma para ver la distribución de estos a lo largo de los diferentes periodos de tiempo. Los histogramas están representados en las figuras 17 y 18. En la figura 17 se puede ver que los pseudogenes de la hsp10 tienen dos picos en los cuales se originaron más pseudogenes que en el resto del tiempo, correspondientes a los 15 y 40 Ma. Estos datos nos indican eventos episódicos que sugieren momentos en los cuales se aumentó el proceso de formación de los pseudogenes de la hsp10, por lo cual sería interesante en un futuro trabajo determinar si estos eventos moleculares están correlacionados de alguna manera con eventos paleontológicos importantes. La figura 18,

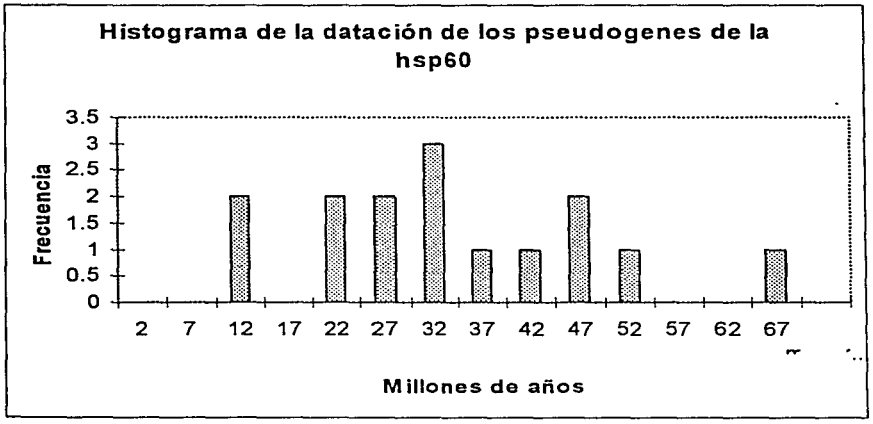
que representa el histograma de los pseudogenes de la hsp60, tiene una distribución más homogénea a lo largo del tiempo.

Figura 17



Histograma de la datación de los pseudogenes de la hsp10 (por el método 3)

Figura 18



Histograma de la datación de los pseudogenes de la hsp60 (por el método 3)

Sustituciones por sitio

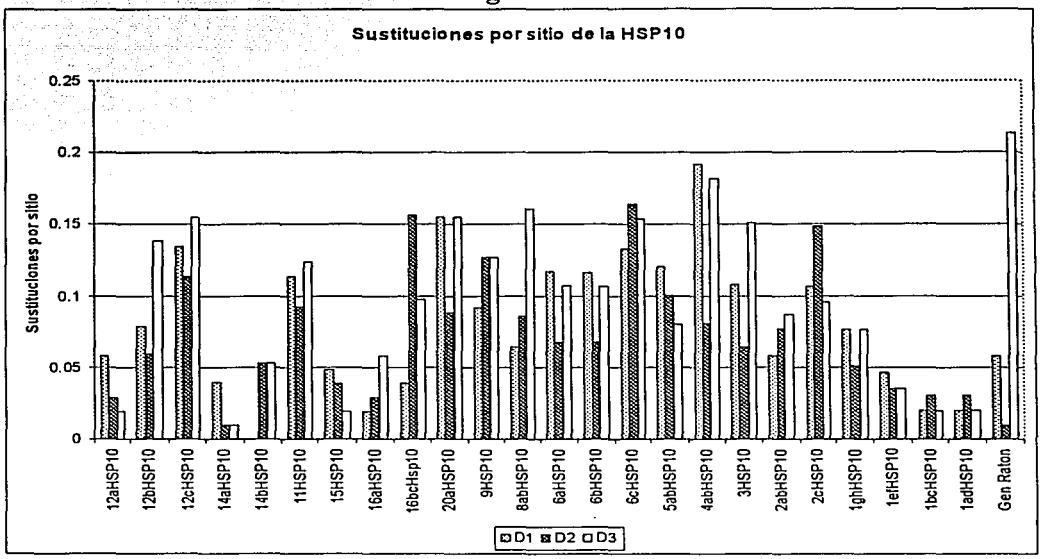
Con el fin de hacer una inferencia acerca de si alguno de los pseudogenes ha sido funcional en algún momento, se realizó un conteo de las mutaciones y en qué base del codón equivalente se presentaban. Para esto, se utilizaron los alineamiento pareados de cada uno de los pseudogenes con respecto al cDNA correspondiente, después se identificaron los codones de inicio y de término en el cDNA y si existían los del pseudogen. Ya con el inicio y término identificados, se compararon codón por codón, para contabilizar en qué sitio (primero, segundo o tercero) se encontraban las mutaciones. En todos los casos se respetó siempre la fase de lectura del cDNA.

En la figura 19, se grafican las mutaciones por sitio en el codón que tiene cada uno de los pseudogenes de la hsp10 y en la figura 20, los de la hsp60. Como se ve en la figura 19, el gen del ratón –que es una secuencia codificante- acumula más cambios en la tercera posición del codón, ya que éstas son en su mayoría cambios silenciosos. La mayoría de los pseudogenes por otro lado no tienen una diferencia tan marcada en las mutaciones de una u otra posición, a excepción del pseudogen 8abHSP10. Los datos de este pseudogen nos indican la posibilidad de que en algún momento fue una secuencia funcional con una presión de selección sobre las mutaciones en primera y segunda posición, para posteriormente convertirse en un pseudogen y perder toda presión de selección. Todos los cálculos fueron realizados por alineamientos pareados entre el cDNA de humano y la secuencia correspondiente.

Alineamiento múltiple

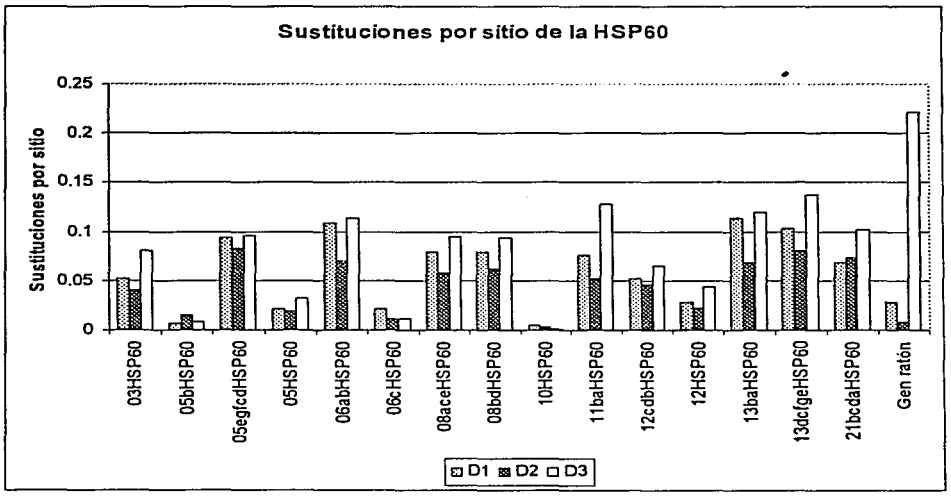
Los alineamientos múltiples de las dos series de pseudogenes fueron realizados utilizando el Clustal W versión 1.4 (Thompson, *et al.*, 1994), Gap: Existencia: 10, Extensión: 1. Este bajo valor en la extensión fue usado para que fuera posible el alineamiento de tantas secuencias, ya que algunas de ellas presentan inserciones y/o deleciones con las cuales no se permitiría el correcto alineamiento. En el anexo II se presenta el alineamiento de la hsp10 y sus pseudogenes, el de la hsp60 no se presenta debido a su gran tamaño, ya que es de más de 15 páginas de extensión.

Figura 19



Gráfica de las sustituciones por sitio de los pseudogenes y los genes de ratón para la hsp10.

Figura 20



Gráfica de las sustituciones por sitio de los pseudogenes y los genes de ratón para la hsp60.

Árboles filogenéticos de los pseudogenes

Con los cálculos de los alineamientos múltiples se hicieron los árboles filogenéticos correspondientes, utilizando el TreeView 1.6.6, el cual utiliza los datos del alineamiento del Clustal W. En general, los árboles filogenéticos no aportan más evidencia sobre la historia de los pseudogenes que la que ya se había obtenido por otros métodos. Como algunos de los pseudogenes tienen muchas mutaciones, la probabilidad de que dos de éstos tengan la misma mutación es alta. Como consecuencia, en algunos casos, en los árboles filogenéticos quedan agrupados los pseudogenes en pequeños grupos o "clusters", que más que representar un origen en común nos muestra la identidad que existe entre éstos.

Por otro lado, lo que si se refleja en los diferentes árboles filogenéticos es la estrecha relación de los pseudogenes duplicados. Este es el caso de los pseudogenes 6aHSP10 y 6bHSP10, del 1aHSP10 y 1bHSP10, del 8bHSP60 y 8aHSP60. Todos estos pares de pseudogenes tienen un muy alto porcentaje de identidad -debida a su reciente duplicación- por lo cual aparecen juntos en los árboles filogenéticos.

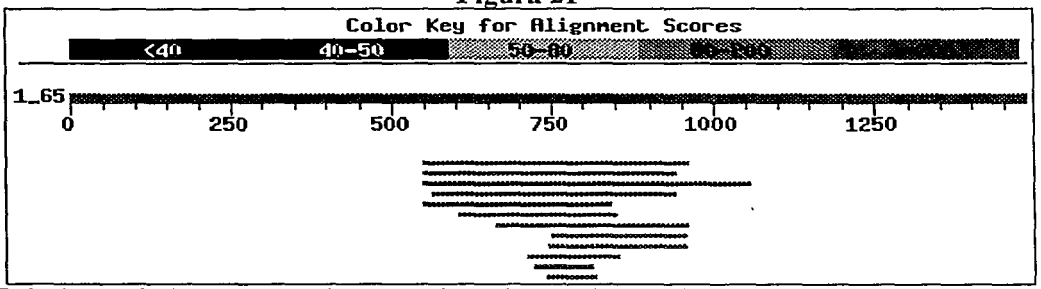
Búsqueda del EPF

Fletcher *et al* (2001), reportaron una secuencia para el EPF (factor temprano del embarazo) del ratón (número de acceso AF247846). En este trabajo se realizo una búsqueda de esta secuencia en el genoma humano, ya que Fletcher y colaboradores reportaton que es homóloga a la secuencia del hsp10. Esta secuencia incluye la región codificante así como las regiones 5' y 3' que podrían corresponder a secuencias regulatorias como el promotor, aunque en el reporte original no se señala la caracterización de estas regiones regulatorias. En la figura 21, se puede ver el resultado de la búsqueda en el genoma humano. Los "hits" se encuentran solamente en la región central, ya que es la que corresponde al ORF del gen, las regiones flanqueantes son posiblemente secuencias que sólo se encuentren en el genoma del ratón.

Todos los resultados de esta búsqueda corresponden a secuencias reportadas en este trabajo como pseudogenes de la hsp10. Los cuatro primeros "hits" corresponden a

los pseudogenes de los cromosomas 14, 16, 15 y 1. Los dos primeros con identidad del 86% y el tercero y cuarto con una identidad del 83% y 82%, respectivamente. Al realizar el análisis de los primeros cuatro "hits" se observo que los primeros tres pseudogenes

Figura 21



Todos los resultados representan las secuencias de los pseudogenes de la hsp10 que se obtuvieron en este trabajo.

no tienen un ORF del tamaño necesario para codificar la proteína del EPF, por lo cual quedaron descartados como secuencias codificantes. El cuarto "hit" encontrado se encuentra en el cromosoma 1 y corresponde al pseudogen 1adHSP10 (que es igual al 1bcHSP10). Estos dos pseudogenes presentan un ORF del tamaño necesario para codificar la proteína del EPF. Con estos datos, planteamos la hipótesis de que este pseudogen podría ser el gen de la EPF. Para aceptar o rechazar esta hipótesis, comparamos las secuencias del EPF y el 1adHSP10 tanto a nivel de nucleótidos como de aminoácidos. En los dos casos tienen una identidad del 82%. Por otro lado, el gen EPF (de ratón) tiene 94% de identidad con el cDNA de la hsp10 del humano. Este alto porcentaje de identidad entre estos dos genes indica que son secuencias bien conservadas en la escala filogenética. Consecuentemente, esperaríamos que el gen del EPF en humano y la hsp10 humana tuvieran una identidad similar. Este razonamiento nos lleva a concluir que el pseudogen 1adHSP10 está muy poco conservado como para tener todavía la función del EPF. Por otro lado, como se puede ver en la figura 11, el pseudogen 1adHSP10 tiene una deleción de 4 aminoácidos con respecto al EPF y al cDNA de la hsp10, lo cual no es un buen indicador de que sean secuencias parálogas funcionales. Su ORF de gran tamaño así como su alta identidad, se pueden deber a que se trata de un pseudogen de reciente formación y no a que sea el gen del EPF.

CONCLUSIONES:

1) Se identificaron 24 pseudogenes del gen de la HSP10 y 15 del gen de la HSP60, lo cual es indicativo de que estos dos genes tienden a generar muchos pseudogenes por transcripción reversa.
2) El pseudogen 16aHSP10 que se encuentra en un intrón de un gen funcional del genoma humano (figura 7), es un buen ejemplo de que los pseudogenes no tienen restricción para el sitio de su inserción mientras no alteren la adecuación del organismo; esto sugiere que mientras no causen una mutación selectivamente negativa se pueden insertar en cualquier sitio del genoma.
3) Algunos de los pseudogenes presentan ORFs que podrían sugerir su posible funcionalidad. Por otro lado, estos ORFs podrían ser los remanentes del marco de lectura del gen de donde se originaron y que debido a su reciente creación no han sido alterados por las mutaciones.
4) La aparición de pseudogenes representa dos diferentes tipos de fenómenos inserciones de *novo* y duplicaciones. En el caso de las primeras se presentan datos de 36 eventos de inserción de pseudogenes. En el caso de las duplicaciones este análisis indica que en la historia del genoma han ocurrido por lo menos tres eventos recientes de duplicación asociados a los cromosomas 1, 6 y 8.
5) La generación de los pseudogenes de la hsp10 indica eventos episódicos no azarosos que sugieren periodos de aumento en la transcripción de estos genes, posiblemente debidos a presiones medioambientales. Dicho de otro modo, la diferenciación filogenética no sólo procede generando presiones adaptativas en las regiones codificantes, sino que mutaciones en las regiones regulatorias pueden aumentar o disminuir el nivel de expresión de genes y tener importantes consecuencias fisiológicas y/o bioquímicas.
6) Se encontraron 8 inserciones con una longitud promedio de 300pb que separan a diferentes pseudogenes en dos partes (tablas 2 y 3). Por otra parte, el Consorcio Internacional de la Secuenciación del Genoma Humano reporta que en el genoma hay insertadas cerca de 1,000,000 de copias de las secuencias Alu que tienen una longitud de entre 100 y 400 pb (Dennis, 2001). Algunas de las inserciones reportadas en esta

tesis tienen una identidad del 75% con respecto a la reportada por Blakey (2000). Esto indica que algunas de estas inserciones seguramente corresponden a secuencias Alu.

7) Como nos indican las figuras 19 y 20, los pseudogenes 8abHSP10 y 11baHSP60 tienen una tasa de mutación en la tercera posición considerablemente mayor, por lo que es posible que estos dos pseudogenes hubieran sido funcionales durante algún tiempo.

8) Nuestros datos no nos permiten establecer con seguridad la localización del gen del EPF. La única secuencia encontrada con potencial codificante suficiente (1adHSP10), muestra un 82% de identidad, a nivel de aminoácidos con la EPF del ratón. Sin embargo, muestra diferencias importantes en la secuencia. De éstas, la más significativa sería la aparición de una cisteina que substituiría a una glicina. Este cambio radical debería de tener un fuerte impacto en la estructura tridimensional de la probable proteína y por lo tanto en su función. Adicionalmente tendría una deleción de cuatro aminoácidos y cambios importantes en el posible péptido líder.

BIBLIOGRAFÍA:

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson. (1994). Molecular Biology of the Cell. Third Edition. Garland Publishing.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipmasn, D.J. (1990). "Basic local alignment search tool" J. Mol. Biol. 215:403-410.

Ang, D., Liberek, K., Skowyra, D., Zylicz, M., and Georgopoulos, G.(1991). Biological role and regulation of the universally conserved heat shock proteins. J. Biol. Chem. 266: 24233-24236.

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.

Ayala, F.J. (1997). Vagaries or the molecular clock. Proc. Natl. Acad. Sci. USA. Vol.94 pp 7776-7783, July 1997.

Baxevanis, A.D., B.F. Ouellette. (2001). Bioinformatics: A practical guide to the analysis of genes and proteins. Second Edition. Wiley-Interscience.

Blakey, S. (2000). Direct Submission to Genbank. Accession: AL078621.

Brown, T.A. (1999). Genomes. Wiley-Liss, New York, USA.

Bustamante, C.D., R. Nielsen and D.L. Hartl. (2002). A Maximum Likelihood Method for Analyzing Pseudogene Evolution: Implications for silent Site Evolution in Humans and Rodents. Mol. Biol. Evol. 19(1):110-117.

Cooper, G.M. (1997). The Cell: A Molecular Approach. Sinauer Associates Inc. Sunderland, Massachusetts USA.

Cortinas, M.N., and E.P. Lessa. (2001). Molecular Evolution of Aldolase A Pseudogenes in Mice. Mol. Biol. Evol. 18(9): 1643-1653.

Dennis, C., and R. Gallagher Editors. (2001). The Human Genome. Nature, Palgrave.

Devor, E.J. (2001). Molecular archeology of an SPI00 splice variant revisited: dating the retrotranscription and Alu insertion events. Genome Biology. 2001, 2(9):research0040.1-0040.6.

Dhellin, O., J. Maestre, and T. Heidmann. (1997). Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. *EMBO. J.* 16:6590-6602.

Doolittle, R.F., (1990). Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. Academic Press, Inc.

Ellis, R.J. (1987). Proteins as molecular chaperones. Nature 328-329.

Ellis, R.J. (1996). The Chaperonins. Academic Press, San Diego, Cal. USA.

Fletcher, B.H., A.I. Cassady, K.M. Summers, A.C. Cavanagh. (2001). The murine chaperonin 10 gene family contains an intronless, putative gene for early pregnancy factor, *Cpn10-rs1*. Mammalian Genome 12,133-140.

Friedberg, F. and A.R. Rhoads. (2000). Calculation and Verification of the Ages of Retroprocessed Pseudogenes. Mol. Phylogenetics and Evolution. 16: 127-130.

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

Gupta, R.S. (1990) Sequence and structural homology between a mouse T-complex protein TCP-1 and the chaperonin family of bacterial (GroEl, 60-65 kDa heat shock antigen) and eukaryotic proteins. Biochem. Inter. 20:833-841.

Gupta, R.S. (1990). Mitochondria, molecular chaperone proteins and the *in vivo* assembly of microtubules. Trends Biochem. Sci. 15: 415-418.

Gupta, R.S. (1995). Evolution of the chaperonin families (Hsp60, Hsp10 and Tcp-1) of proteins and the origin of eukayotic cells. Mol. Microbiol. 15:1-11.

Hansen, L.L., Nielsen, M.N., Thomsen, A., Mogensen, J., Vestergaard, M., *et al.* (2000). Genomic structure and chromosomal localisation of the human hsp60 and hsp10 genes. Direct Submission, Accession AJ250915, NCBI.

Jukes, T.H. and Cantor C.H. (1969). Mammalian Protein Metabolism (ed. Munro, H.N.) 21-123. (Academic, New York, 1969).

Karlin, S. & Altschul, S.F. (1990) "Methods for assessing the statistical significance of molecular seuqence features by using general scoring schemes." Proc. Natl. Acad. Sci. USA 87:2264-2268.

Kumar, S. and S.B. Hedges. (1998). A molecular timescale for vertebrate evolution. Nature. Vol 392, 30 April 1998 pp 917-920.

Lander, E.S., Linton, L.M., Birren, B., Nusbaunm, C., Zody, M. C., Bladwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.

Lewis, R. (2002). SNPs as Windows on Evolution. The Scientist 16(1):16, Jan. 7.

Li, W.H., T. Gojobori and M. Nei. (1981). Pseudogenes as a paradigm of neutral evolution. Nature. Vol 292, 16 July 1981 pp 237-239.

Li, W.H. and M. Tanimura. (1987). The molecular clock runs more slowly in man than in apes and monkeys. Nature. Vol 326, 5 March 1987 pp 93-96.

Li, W.H. and D. Graur. (1991). Fundamentals of Molecular Evolution. Sinauer Associates, Inc. Sunderland Massachusetts.

Lindquist, S. (1986). The heat shock response. Annu Rev. Biochem. 45:39-72.

Nachman, M.W. and S.L. Crowell. (2000). Estimate of the Mutation Rate per Nucleotide in Human. Genetics 156:297-304 (September).

Ohno, S. (1970). Evolution by Gene Duplication. New York: Springer.

Ohno, S. (1985). Dispensable genes. *Trends Genet.* 1:160-164.

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

Ryan, M.T., Herd, S.M., Sberna, G., Samuel, M.M., Hoogenraad, N.J. and Hoj, P.B. (1997). The genes encoding mammalian chaperonin 60 and chaperonin 10 are linked head.to head and share a bidirectinal promoter. Gen 196 (1-2), 9-17.

Shepherd, J.C. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA. 1981 Mar; 78(3):1596-600.

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.

Tanooka, H. *et al.* (2001). p53 Pseudogene dating: Identification of the origin of laboratory mice. Gene 270 (2001) 153-159.

Thompson, JD, Higgins DG and Gibson TJ. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Reserach.* Submitted, June 1994.

Tilly, K., Murialdo, H., and Georgopoulos, C. (1981). Identification of a second *Escherichia coli* groE gene whose product is necessary for bacteriophage morphogenesis. Proc. Natl. Acad. Sci. USA 78:1629-1633.

Vanin, E.F. (1985). Processed pseudogenes: Characteristics and evolution. *Annu. Rev. Genet* 19: 253-272.

Venner, TJ, B. Singh and R. Gupta. (1990). Nucleotide Sequences and Novel Structural Features of Human and Chinese Hamster hsp60 (Chaperonin) Gene Families. DNA Cell Biol. 9:(8) 545-552.

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

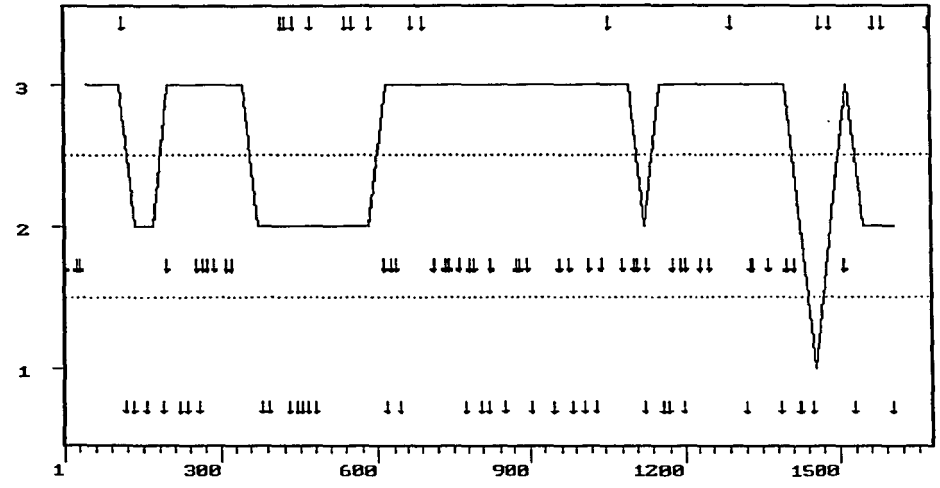
Wagner, A. (1998). The fate of duplicated genes: loss or new function?. BioEssays. 20:785-788.

Wagner, A. (2002). Selection and Gene duplication: a view from the genome. Genome Biology 3(5):reviews1012.1-1012.

ANEXO I

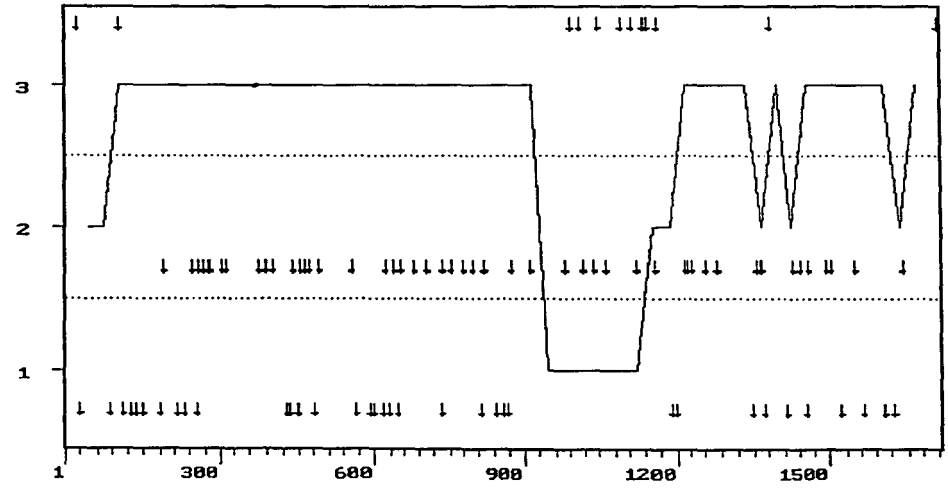
Gráficas de Shepherd de los pseudogenes de la HSP10

Gráfica I.1



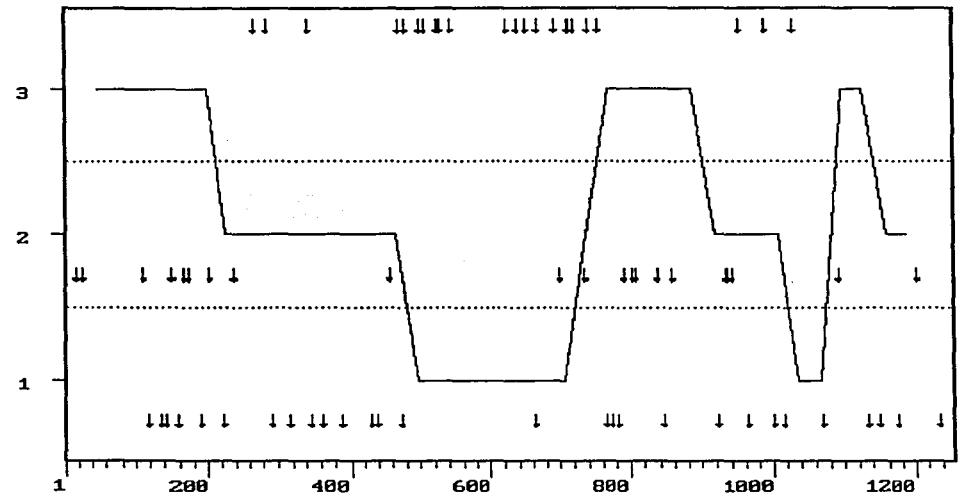
Plot of the RNY frame analysis for sequence 21BHSP60.
On bases 1 to 1673 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.2



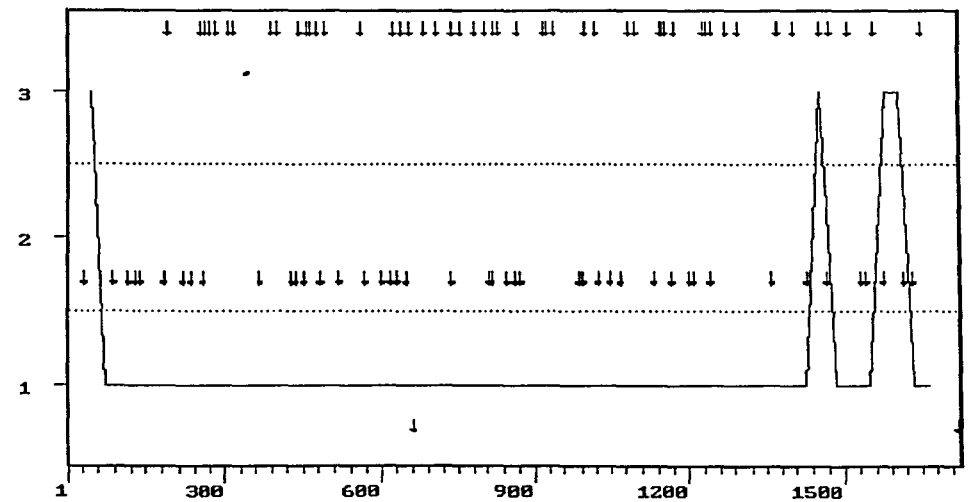
Plot of the RNY frame analysis for sequence 03HSP60.
On bases 1 to 1712 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.3



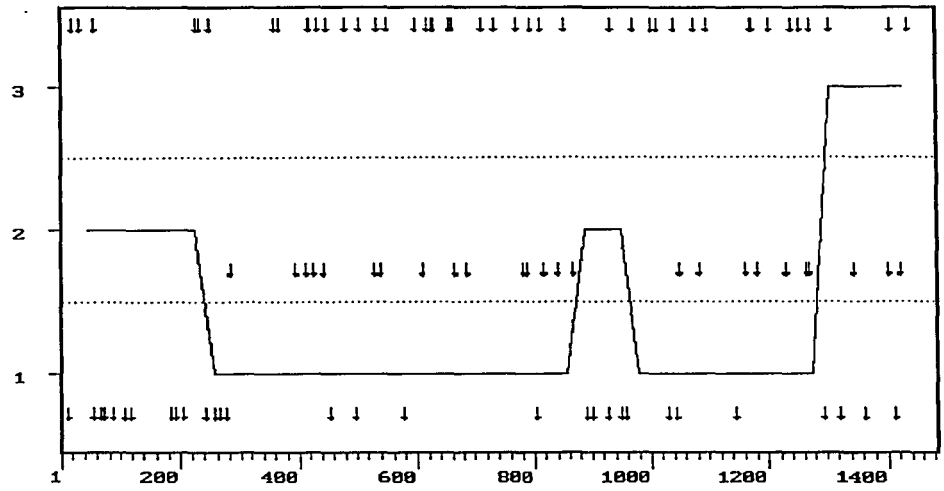
Plot of the RNY frame analysis for sequence 05EHSP60.
On bases 1 to 1252 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.4



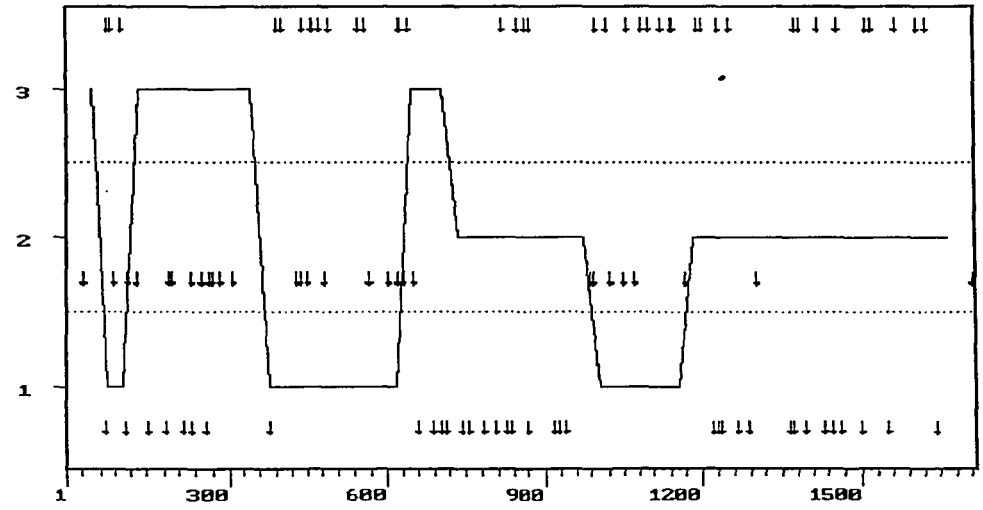
Plot of the RNY frame analysis for sequence 05HSP60.
On bases 1 to 1722 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.5



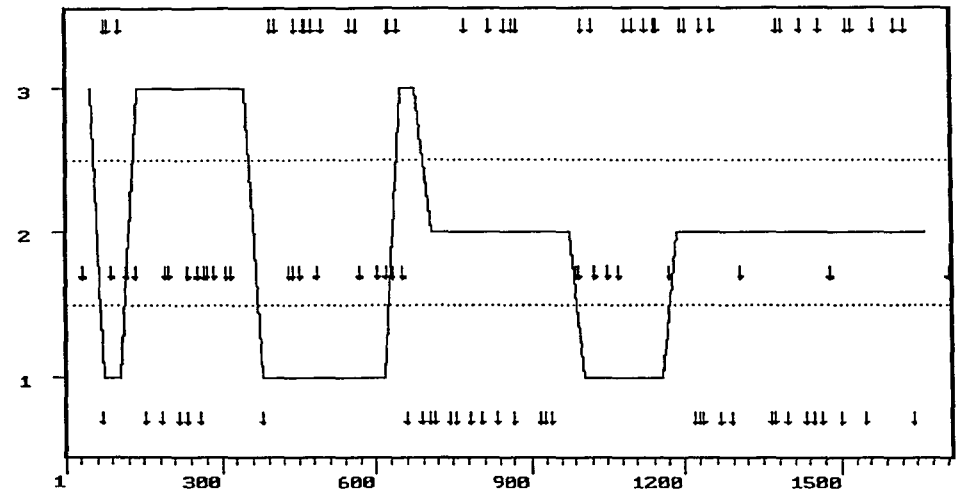
Plot of the RNY frame analysis for sequence 06AHSP60.
On bases 1 to 1483 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.6



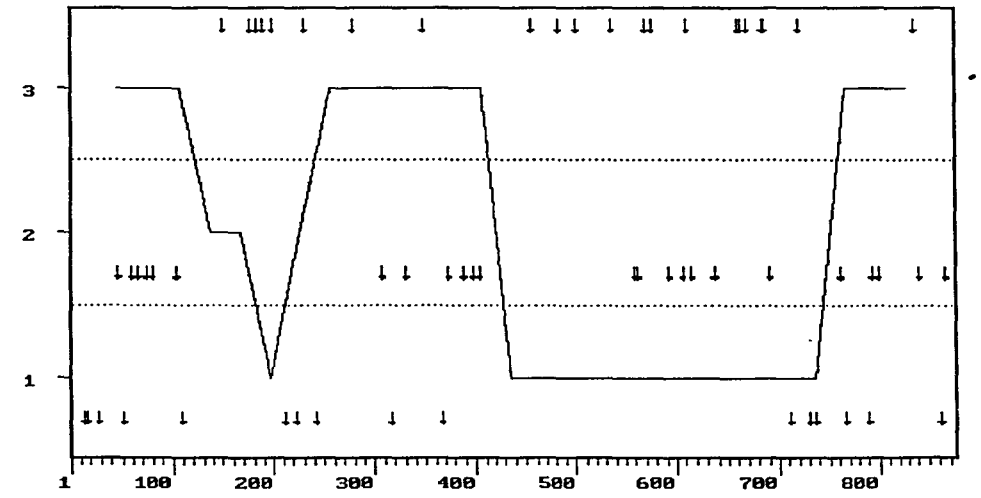
Plot of the RNY frame analysis for sequence 08AHSP60.
On bases 1 to 1714 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.7



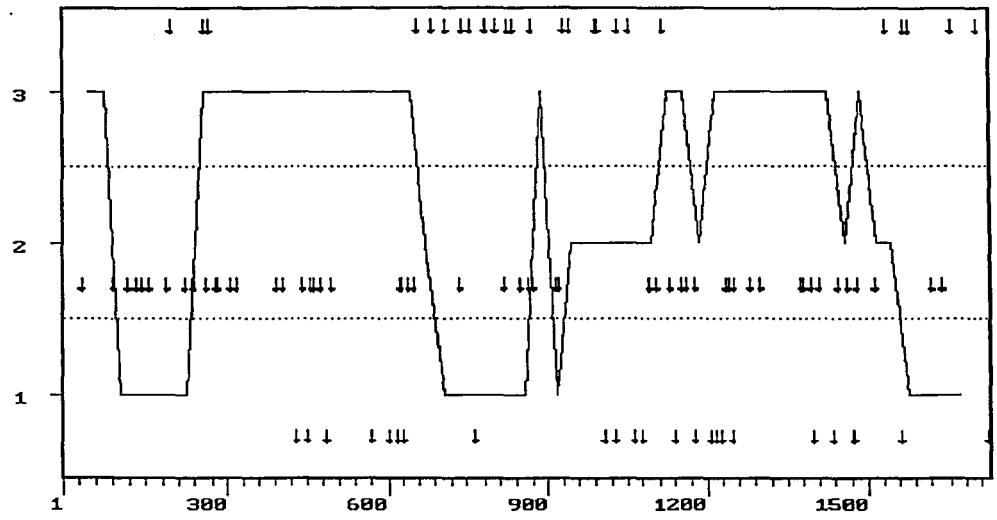
Plot of the RNY frame analysis for sequence 08BHSP60.
On bases 1 to 1714 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.8



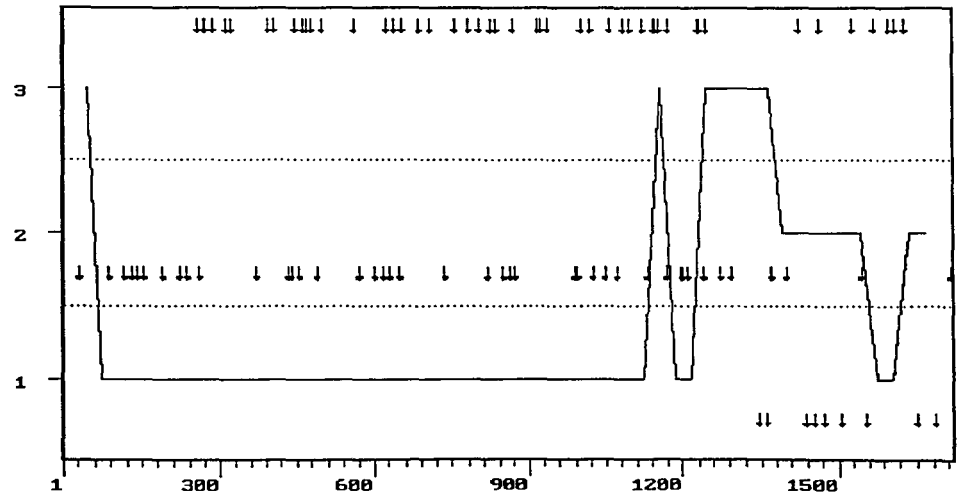
Plot of the RNY frame analysis for sequence 11BHSP60.
On bases 1 to 875 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.9



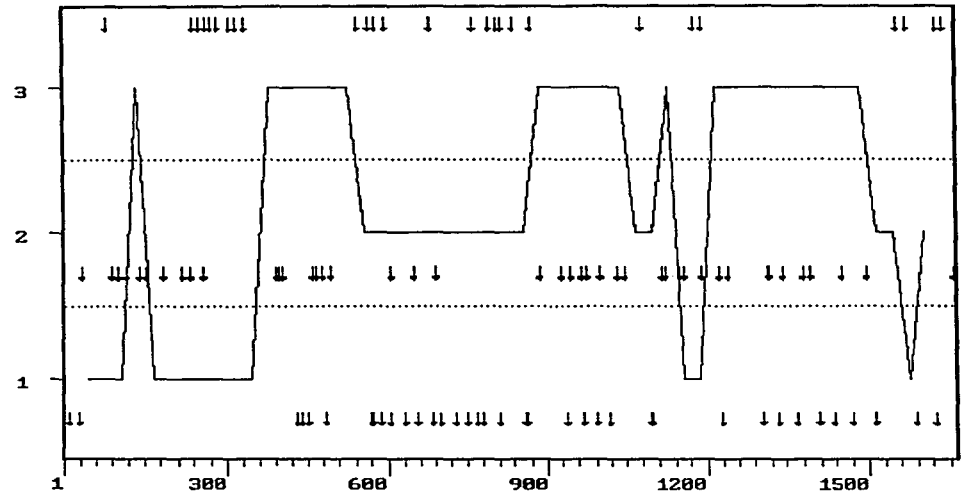
Plot of the RNY frame analysis for sequence 12CHSP60.
On bases 1 to 1725 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.10



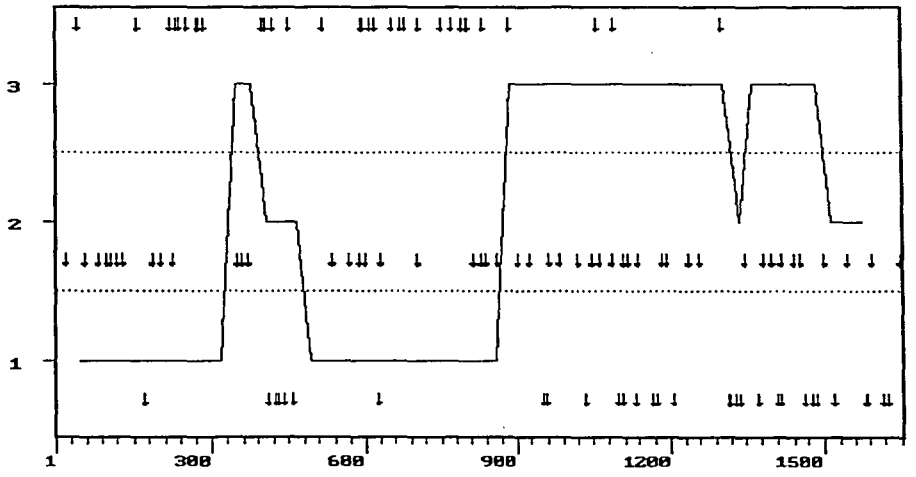
Plot of the RNY frame analysis for sequence 12HSP60.
On bases 1 to 1720 computed by length of 90 bp. with a step of 30 bp.

Gráfica I.11



Plot of the RNY frame analysis for sequence 13BHSP60.
On bases 1 to 1660 computed by length of 90 bp. with a step of 30 bp.

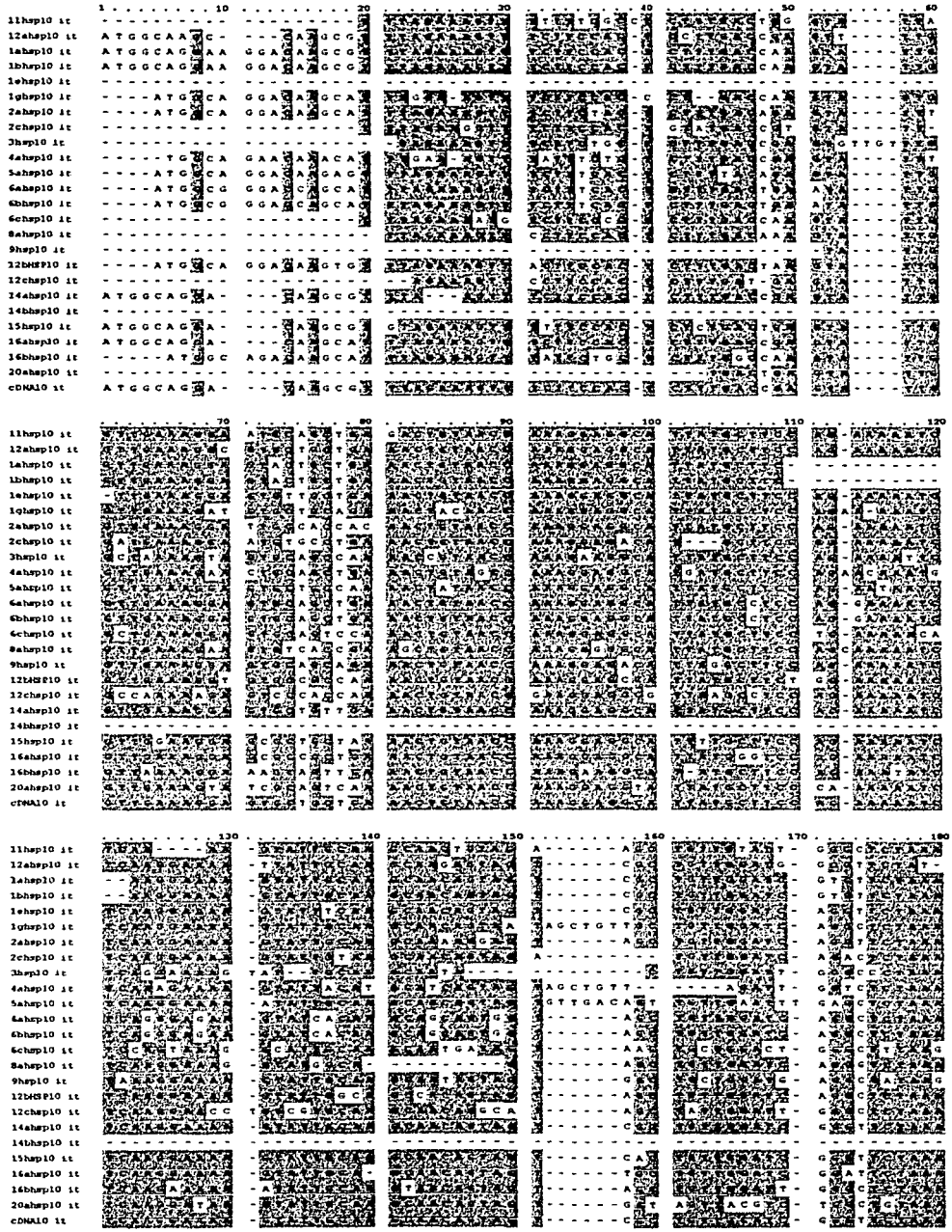
Gráfica I.12

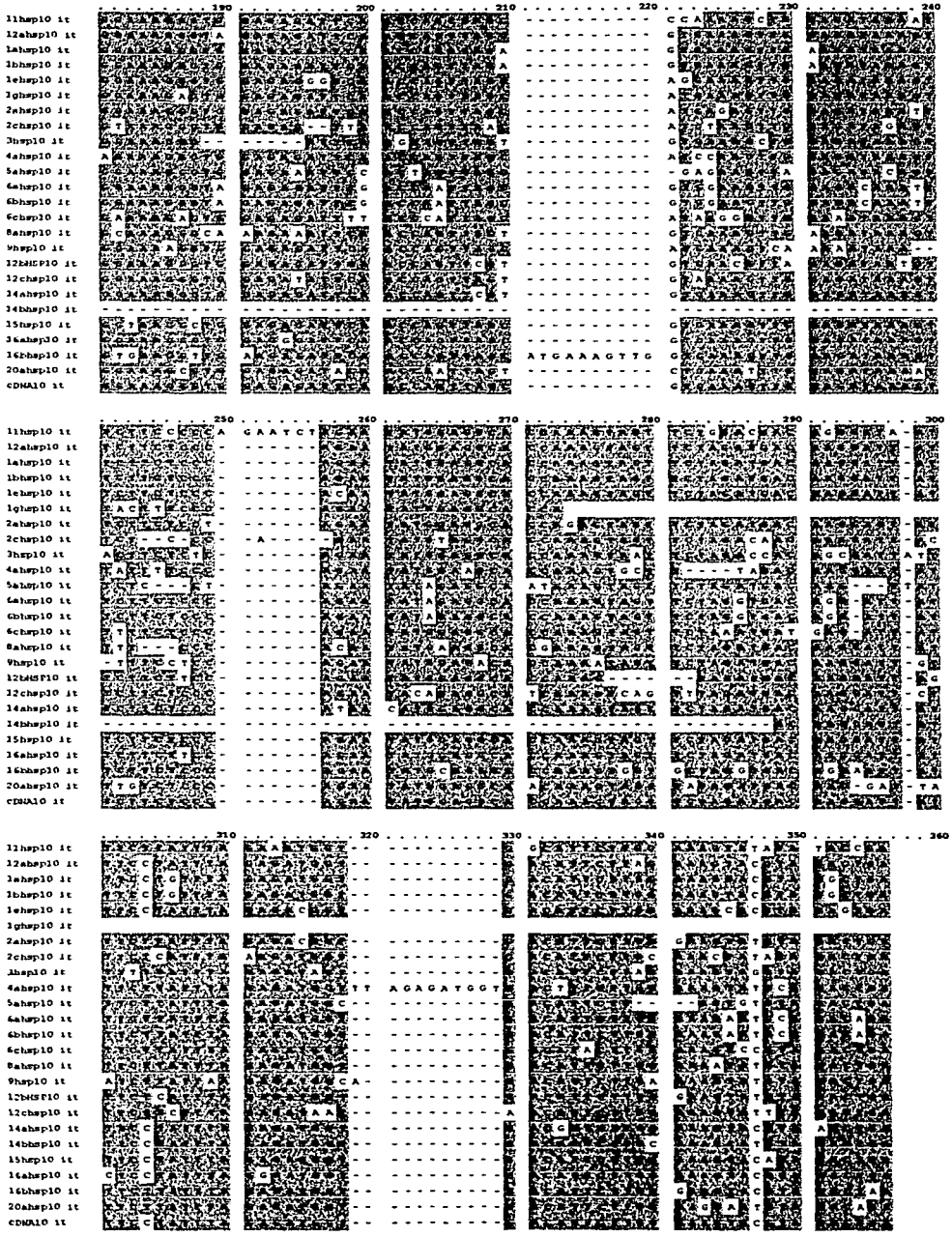


Plot of the RNY frame analysis for sequence 13DHSP60.
On bases 1 to 1651 computed by length of 90 bp. with a step of 30 bp.

ANEXO II

Alineamiento múltiple de los pseudogenes de la HSP 10





Glosario:

- **Alu:** (elementos Alu) grupo de secuencias repetitivas con una alta identidad, dispersas en el genoma humano, cada una de cerca de 300 pb. Todos los miembros tienen sitios de corte de la enzima de restricción *Alu* I (de ahí su nombre)
- **Cap:** estructura en el extremo 5' de los mRNAs eucariontes; añadida después de la transcripción, por la unión 5' – 5' del trifosfato terminal de GTP al extremo terminal del mRNA.
- **cDNA:** DNA complementario; molécula de DNA hecha a partir de un mRNA por lo que carece de intrones.
- **Cistrón:** la unidad genética definida por la prueba cis/trans; equivalente a un gen, ya que comprende la unidad de DNA que representa a una proteína.
- **Centrómero:** región estrecha de un cromosoma mitótico, que mantiene a las cromátidas hermanas juntas.
- **Clona:** gran número de células o moléculas idénticas con un mismo ancestro en común.
- **Codón:** triplete de nucleótidos que representa un aminoácido o una señal de termino.
- **Consenso, secuencia:** secuencia idealizada en la que cada posición representa la base más común cuando varias secuencias son comparadas.
- **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.
- **Gen:** región de DNA que controla una característica heredable, que generalmente corresponde a solo una proteína o RNA. Esta definición incluye toda la unidad funcional, incluyendo la secuencias de DNA codificantes, las secuencias regulatorias y los intrones.
- **Intrón:** región no codificante de un gen eucarionte, que se transcribe a una molécula de RNA pero luego es cortada en el procesamiento del RNA.
- **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.
- **NCBI:** National Center for Biotechnology Information. Pagina electrónica: http://www.ncbi.nlm.nih.gov/
- **ORF:** (Open reading frame), Marco de lectura abierto; región que contiene una serie de tripletes que codifican para aminoácidos sin ningún codón de término; esta secuencia es potencialmente traducible a proteína.

- **Pb:** dos bases nitrogenadas en las cadenas complementarias del DNA unidas por enlaces débiles (A - T / G – C). Habitualmente, se refiere a la longitud de la secuencia
- **Proteoma:** el juego total de proteínas codificadas por el genoma.
- **Región codificante:** DNA a partir del cual se deriva la secuencia traducida del mRNA.
- **Retrotranscripción:** la síntesis de DNA con un templado de RNA; realizado por la enzima transcriptasa reversa ya sea *in vivo* o *in vitro*.
- **Retrovirus:** virus que como material genético tienen RNA, que para replicarse en la célula primero hacen una cadena doble de DNA.
- **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.
- **Southern Blot:** técnica en la que fragmentos de DNA separados por electroforesis, son inmovilizados en una hoja de papel; para después detectar moléculas específicas con una sonda de ácidos nucleicos marcada.
- **Transcriptasa reversa:** enzima presente en los retrovirus, que a partir de una cadena sencilla de RNA hace una copia de DNA.
- **Transformación:** alteración heredable en las propiedades de una célula eucarionte.
- **Transposón:** segmento de DNA que se puede mover de una posición en el genoma a otra.
- **Valor C:** la cantidad total de DNA en un genoma haploide.