UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
INSTITUTO DE BIOTECNOLOGÍA

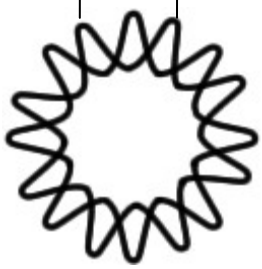
POSGRADO EN CIENCIAS BIOQUÍMICAS

Identificación *in silico* de nuevos sitios de entrada internos para el ribosoma en genomas eucariontes

Tesis que para obtener el grado de Maestra en Ciencias presenta:

BIÓL. VIRIDIANA AVILA MAGAÑA

Asesor de Tesis
DR. ENRIQUE MERINO PÉREZ



Cuernavaca, Morelos **OCTUBRE 2011**

ÍNDICE

IDENTIFICACIÓN *in silico* DE NUEVOS SITIOS DE ENTRADA INTERNOS PARA EL RIBOSOMA EN GENOMAS EUCARIONTES

Resumen

El inicio *bona fide* de la síntesis de proteínas de la mayoría de las proteínas eucariontes, depende del reconocimiento por parte del ribosoma de la estructura en el extremo 5´ cap, nucleótido 7-metilguanosina en el extremo 5' de los mRNAs. Bajo condiciones de estrés celular, se ha reportado que un subconjunto de mRNAs, inician la traducción de manera cap independiente, a través de la formación de estructuras secundarias que son capaces de reclutar al ribosoma *per se* y a otras proteínas, dichas estructuras estables se denominan sitios internos de entrada para el ribosoma (IRES, por sus siglas en inglés).

A pesar de que comúnmente los IRES forman estructura secundaria estable de RNA no se ha identificado ningún tipo de motivo estructural conservado, aunado a que el grado de estructuración puede variar considerablemente. Debido a lo anterior, el criterio de energía libre mínima para la identificación de IRES en genes específicos, tiene una limitada predictibilidad. Como prueba de ello podemos observar que el número de IRES celulares identificados hasta la fecha es muy reducido, y su identificación se ha restringido al genoma humano, a pesar de que actualmente se cuente aproximadamente con un centenar de genomas eucariontes secuenciados.

Asumiendo que la diversidad de organismos cuyos contextos celulares son similares a los contextos en los que se han descrito IRES celulares en humano, podemos suponer que la regulación de la expresión de dichos genes podría también llevarse a cabo mediante IRES. Por lo tanto, es necesario implementar nuevas estrategias *in silico* para la detección global de IRES en genomas eucariontes. En el presente trabajo proponemos que la genómica comparativa de genes ortólogos que potencialmente poseen IRES, puede incrementar notablemente la capacidad predictiva.

El objetivo de esta tesis es identificar nuevos IRES celulares en la región no traducida 5' en los mRNAs de genomas eucariontes secuenciados. Para lograr lo anterior desarrollamos una nueva metodología basada en la predicción estructural de nuevos IRES celulares en genes de mamíferos ortólogos a los genes en los que previamente se han descrito IRES. Para proponer candidatos *bona fide*, definimos dos aproximaciones principales de estudio: i) el análisis probabilístico y ii) el análisis de la función celular específica en la que los posibles IRES actuarían durante la traducción cap-independiente. De esta manera, el análisis conjunto de ambas aproximaciones metodológicas nos permitió identificar a 50 posibles genes a presentar un IRES celular en sus respectivas regiones de regulación 5´.

Introducción

La traducción es un proceso sofisticado que requiere de una extensa maquinaria biológica que involucra ribosomas, tRNAs y una variedad de enzimas. El estudio comparativo de genomas ha hecho posible el estimar la cantidad mínima de información genética necesaria para ensamblar esta maquinaria de síntesis proteica; así por ejemplo en el genoma más pequeño descrito a la fecha, *Mycoplasma genitalium,* el cual codifica para 480 proteínas, cerca del 25% de ellas participan en el proceso de síntesis de proteínas (Mathews *et al*., 2000). Además de esta gran variabilidad de genes involucrados en la traducción, la cantidad de recursos que intervienen en dicho proceso también es muy grande. Por ejemplo, una levadura en crecimiento contiene cerca de 200, 000 ribosomas que ocupan del 30-40% de volumen celular (Warner J.R. 1999).

Mediante estudios de genómica comparativa y perfiles proteómicos, se ha observado que para diversos genes existe una baja correlación entre los niveles de mRNA y las concentraciones intracelulares de sus proteínas correspondientes, lo cual indica que el control post-transcripcional es más importante en la regulación de la expresión genética de lo que usualmente se había asumido (Holcik y Sonenberg, 2005). La traducción es el paso final en el flujo de la información genética, y en organismos eucariontes, es a este nivel en donde se provee cierta plasticidad para responder rápidamente a cambios en las condiciones fisiológicas, sin invocar rutas nucleares para la síntesis y transporte de mRNA.

Dado el significado del control traduccional en la decisión de supervivencia celular, es importante entender cómo es que la célula controla la traducción de sus mensajes.

Comúnmente la traducción se divide en tres fases: inicio, elongación y término. Las tres fases se encuentran controladas, sin embargo el paso limitante durante la síntesis proteica es su inicio. La regulación del inicio de la traducción es un punto de control importante durante el estrés celular, la diferenciación e incluso en el desarrollo de algunas enfermedades (Graber, *et al.,* 2006). Presumiblemente este mecanismo evolucionó debido a que es más efectivo controlar el principio de ciertos procesos biológicos, que interrumpirlos posteriormente y en el caso particular de la traducción, hacer frente a las consecuencias de una síntesis proteica aberrante.

Inicio de la traducción en Eucariontes

Aspectos generales del inicio canónico de la traducción

Como se mencionó antes, el inicio de la traducción es el paso crucial limitante en el control de la síntesis proteica. En eucariontes dicha síntesis está espacial y temporalmente desacoplada con la transcripción. Antes de que la traducción ocurra, el transcrito de mRNA debe ser sintetizado en el núcleo, procesado en su extremo 5' con la estructura cap (^{7}mGpppN, en donde N es cualquier nucleótido), empalmado y poliadenilado. El mRNA maduro es exportado en forma de ribonucleoproteína (mRNP) al citoplasma a través de poros nucleares. El inicio de la síntesis proteica resulta en la movilización del mRNA a un monosoma y posteriormente ocurren eventos de inicio en donde se van reclutando más ribosomas al mRNA, dando lugar a un polisoma (Mathews *et al*., 2000).

La mayoría de los mRNAs eucariontes inician su traducción a través del escaneo ribosomal dependiente de la estructura cap, bajo condiciones fisiológicas normales.

La traducción de los mRNAs implica el reconocimiento y el reclutamiento de éstos a la maquinaria de inicio de la traducción, y el ensamblaje del ribosoma al mRNA. Este proceso es mediado por proteínas que se conocen como factores de inicio eucariontes (eIFs), la mayoría de estos son fosfoproteínas, de las cuales han sido caracterizadas al menos 11 o más de ellas (Tabla 1). El estado de fosforilación de estos factores se correlaciona positivamente con la traducción y crecimiento celular, y es modulado en una amplia variedad de circunstancias, afectando la traducción durante el ciclo celular, infecciones virales, después del choque térmico, o en respuesta a factores de crecimiento y hormonas (Mathews *et al*., 2000), lo cual será discutido posteriormente.

La disociación de los ribosomas 80S es necesaria para el inicio de la traducción, la formación del complejo de pre-inicio 43S depende de la disponibilidad del grupo de subunidades de ribosoma disociadas, las cuales se mantienen en ese estado por los factores eIF3 y eIF1A. La subunidad 40S, se asocia con eIF3 y, el eF1A se une después a un complejo ternario –formado por eIF2, Met-tRNAi y GTP- para formar el complejo de pre-inicio 43S. El ensamblaje de este complejo ternario está regulado por eIF2B. (Figura 1).

El reconocimiento de la estructura ^{7}mG cap en el extremo 5' del mRNA está mediado por el complejo de unión a cap, eIF4F, el cual comprende 3 proteínas: eIF4E, eIF4G y eIF4A. El primero de ellos es el factor de unión a la estructura cap, eIF4A es una RNA helicasa, eIF4B y EI4H promueven la actividad de RNA helicasa de eIF4A, el EIF4G es una proteína de andamiaje y participa en la circularización del mRNA, interactuando con la proteína de unión a la cola de poliadeninas (PABP) en el extremo 3' del mRNA (Figura 2a).

Nombre	Subunidades	Masa (kD)	Función
eIF1	1	12.6	Permite el escaneo del ribosoma; desestabiliza complejos de inicio aberrantes; reconocimiento de AUG.
eIF1A	1	16.5	Promueve la unión de Met-tRNA a la subunidad 40S; promueve el escaneo del ribosoma.
eIF2	3	126	Unión dependiente de GTP de Met-tRNA a la subunidad 40S; GTPasa.
eIF2B	5	261	Factor de intercambio de guanina para el factor e1F2.
eIF3	11	700	Disociación del ribosoma; Promueve la unión de Met-tRN a la subunidad 40S.
eIF4A	1	44	ATPasa dependiente de RNA; RNA helicasa.
eIF4B	1	70	Promueve la actividad de helicasa del factor E1F4A, e1F4F.
eIF4E	1	26	subunidad de unión a ^{7}mG cap.
eIF4G	1	154	Une al RNA, PABP, e1F4E, e1F4A, e1F3; provee andamiaje.
eIF4F	3	223	Heterotrímero e1F4E/4A/4G: une ^{7}mG cap; RNA helicasa.
eIF5	1	49	Activa a e1F2 en su actividad de GTPasa.
eIF5B	1	139	Unión de las subunidades del ribosoma; GTPasa.

Tabla 1. Características de algunos de los factores de inicio de traducción en mamíferos. Tomado de Hellen y Sarnow, 2001.

Se sugiere que este modelo de "loop" cerrado favorece el uso y/o el reciclaje de los ribosomas 40S para entrar la siguiente ciclo de inicio inmediatamente después del término de la síntesis proteica, lo cual genera una traducción eficiente (Mignone y Pesole, 2007).

La unión del complejo de pre-inicio 43S al mRNA en mamíferos depende de la interacción entre eIF3 y eIF4G. Se especula que eIF3 recluta a los ribosomas 40S. Una vez que esta subunidad se une al mRNA, se da el escaneo del mRNA en dirección 5' → 3' hasta que se encuentra el codón de inicio, en donde se une la subunidad del ribosoma 60S, formando el complejo de inicio 80S. Las secuencias vecinas al codón de inicio AUG son determinantes en mamíferos; una traducción eficiente ocurre en un contexto en donde existe una secuencia consenso con un contenido en purinas en la posición -3 y +4 (donde A de AUG es +1). Los ribosomas que durante el escaneo encuentran un AUG que no cumple con las características de la secuencia consenso, no se detienen y continúan el escaneo hasta encontrar otro AUG río abajo que tenga un contexto favorable para su traducción (Mathews *et al*., 2000).

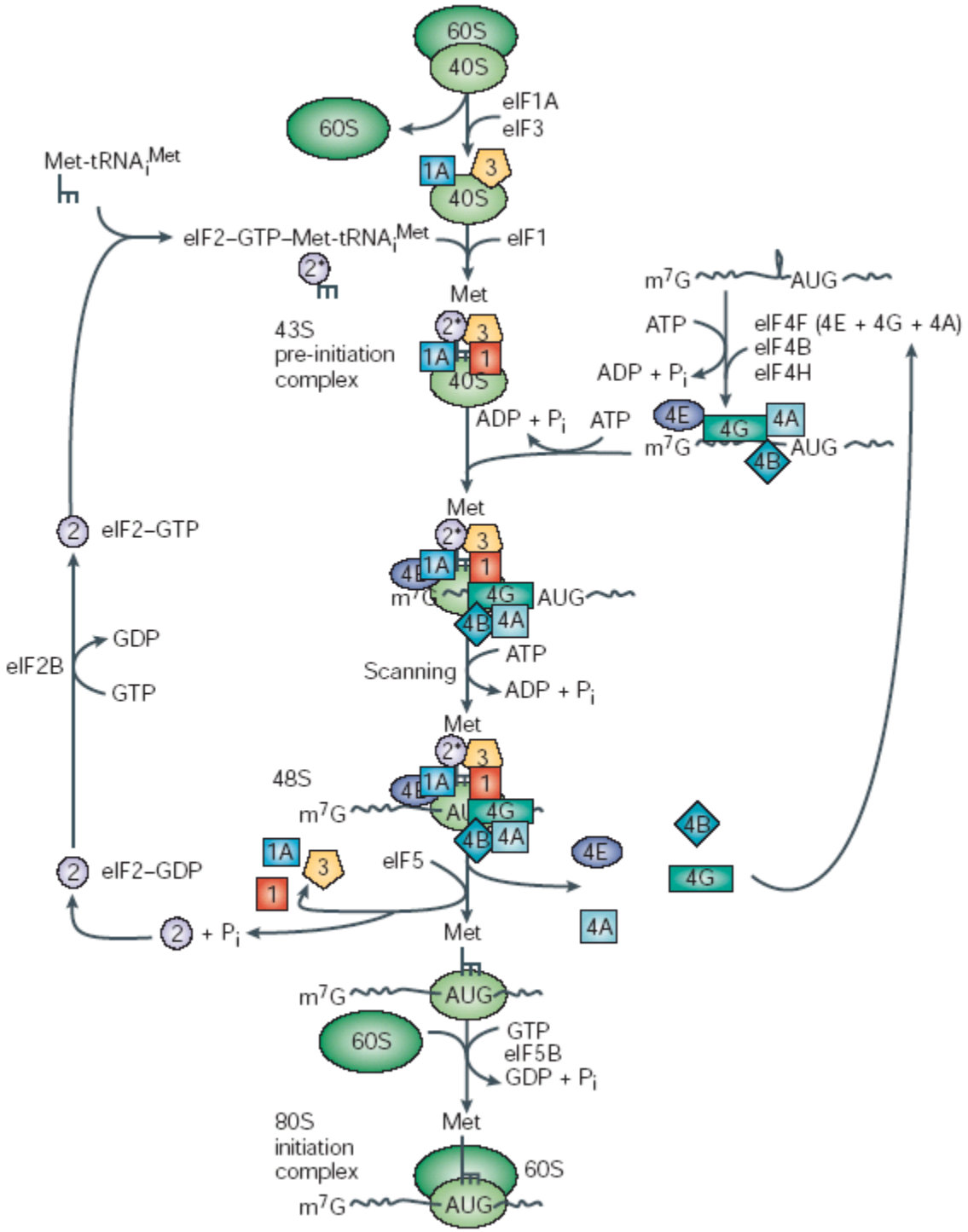


Figura 1. Inicio de la traducción. Existe evidencia genética y bioquímica que demuestra que el proceso canónico de inicio la traducción comienza en la región 5´ no traducida, cuando se une el complejo de inicio que comprende la subunidad ribosomal pequeña (40s), factores proteicos, así como el tRNA de inicio Met-tRNAi. De manera muy general, los mRNAs eucariontes presentan una estructura cap (^{7}mGpppN), en donde N es cualquier nucleótido) al final 5´ del mRNA, la cual les confiere estabilidad y es reconocida por el complejo ribosomal de inicio de la traducción. Una vez que ésto ha ocurrido, el complejo de inicio busca en la región no traducida (UTR) 5´ hasta encontrar el codón de inicio, en donde la subunidad 60s del ribosoma se ensambla y así comienza la síntesis de las proteínas. Tomado de Holcik y Sonenberg, 2005.

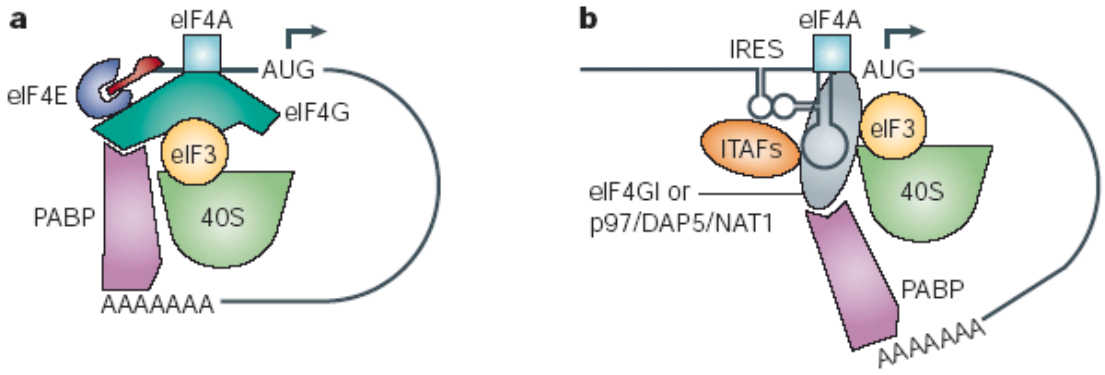


Figura 2. Inicio de la traducción dependiente de cap y dependiente de un sitio interno de entrada al ribosoma. A) Formación de un puente proteico entre 5'm^7GppN y el ribosoma: 'm^7Gcap- eIF4E-eIF4G-eIF3-subunidad ribosomal 40S. La proteína dependiente, eIF4 (azul), se unen a la estructura cap 5'm^7GppN. El extremo con cap del mRNA se une a la subunidad del ribosoma 40S (verde) por una molécula adaptadora eIF4G (verde oscuro), la cual se une a eIF3 (amarillo). eIF4 (cian) es una ATPasa dependiente de RNA y una RNA helicasa, que se encarga de desdoblar las estructuras secundarias en la región no traducida (UTR) 5' del mRNA. La proteína de unión a poli-A (PABP, rosa) circulariza al mRNA a través de la interacción con el extremo 3' UTR (a través de la cola poliA) y 5'UTR (a través del factor eIF4G). B) El sitio de entrada interno al ribosoma (IRES), los factores en *trans* (ITAFs, naranja) y los fragmentos proteolíticos de eIF4GI o p97/DAP5/NAT1, un homólogo distante de eIF4G (gris) estimulan la traducción vía IRES. Sólo los eIFs que son pertinentes para este proceso se indican. Tomado de Holcik y Sonenberg, 2005.

Los factores de inicio que participan en la traducción son liberados después de la formación de un complejo de inicio 48S y son reciclados para otra ronda de inicio. La liberación de eIFs es promovida por eIF5, el cual facilita la hidrólisis de GTP cargado por eIF2, permitiendo la disociación del complejo 48S. eIF5B se requiere para la unión de la subunidad 60S, en donde empieza la elongación de polipéptidos y con ello la traducción (Mathews *et al*., 2000). Este mecanismo de inicio de la traducción se conoce como dependiente de cap (Mokrejs *et al.,* 2006).

La formación del complejo de inicio así como el escaneo del mensajero por parte del ribosoma, requieren una gran cantidad de ATP para llevar a cabo la fase de inicio de la traducción (Mathews *et al*., 2000).

El mecanismo básico de reclutamiento de ribosomas en eucariontes se ha conservado durante la evolución. Todos los organismos eucariontes poseen un complejo eIF4F, el cual consiste al menos de una proteína de andamiaje (eIF4G), una RNA helicasa (eIF4A) y un cofactor de eIF4A, como lo es eIF4B. Sin embargo, diferentes organismos han seleccionado distintas estrategias para regular la actividad de estos factores (Mathews *et al.,* 2000).

Características de la estructura y composición del mRNA relevantes durante el inicio de la traducción

La regulación de la traducción comprende en principio, estrategias múltiples para controlar la producción de proteínas que recaen en la capacidad de que un RNA se traduzca, ello depende de las regiones no traducidas (UTRs, UnTranslated Regions) que se localizan en los extremos 5' y 3'. Los mRNAs eucariontes siempre presentan UTRs (Chatterjee y Pal, 2009).

La presencia de la estructura 5'-cap, estructuras secundarias, múltiples marcos abiertos de lectura (ORFs, Open Reading Frames), múltiples AUGs, sitios de entrada internos para el ribosoma (IRES, Internal Ribosome Entry Sites), el contexto en secuencia del codón de inicio AUG, señales de poliadenilación y motivos en secuencia primaria, son puntos críticos de la regulación de la traducción, aunado a que todos ellos son capaces de interactuar con factores que actúan en *trans* (Figura 3). Todas estas características controlan la síntesis de una proteína empleando algunos mecanismos para alternar la estabilidad del mRNA, la accesibilidad del mensajero a los ribosomas, la circularización del mensajero y la interacción de éste con la maquinaria de traducción (King, *et al.,* 2010).

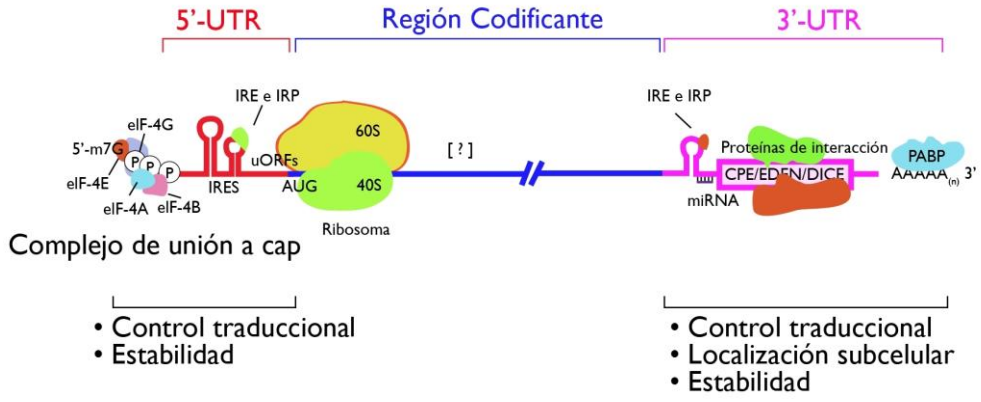


Figura 3. Organización estructural de un mensajero eucarionte y los puntos de posible regulación de la traducción a través de varios factores que actúan en *trans*. 5'-m7G → estructura cap; eIF → factor de inicio eucarionte; CPE → elemento de poliadenilación del citoplasma; EDEN → señal de deadenilación embriónica; DICE → elemento de control diferencial; PABP → proteína de unión a poli-A; [?] → posibles sitios de interacción de factores en *trans* (aún no descritos) en la región codificante. Se señalan las regiones del mensajero, involucradas en la localización subcelular y la estabilidad. Tomado de Chatterjee y Pal, 2009.

La eficiencia de traducción está influenciada por la estructura primaria y secundaria del mRNA. Todos los mRNAs citoplasmáticos presentan la estructura cap, pero algunas estructuras se encuentran enmascaradas por estructura secundaria y no pueden ser reconocidas por EIF4F. Así la accesibilidad de la estructura cap se relaciona con una alta eficiencia de traducción.

La longitud de las 5’ UTRs se relaciona con la eficiencia de traducción. Los mRNAs que codifican para proteínas “house-keeping”, aquellas que son fundamentales para el mantenimiento y viabilidad celular, tales como las enzimas involucradas en el metabolismo central y en la síntesis y procesamiento del DNA, RNA y proteínas, usualmente tienen longitudes de menor tamaño que los mRNAs que codifican para proteínas reguladoras (proteínas que se unen a secuencias regulatorias específicas en el DNA, y encienden o apagan a los genes). Dado lo anterior, los mRNAs de las proteínas “house-keeping” se traducen con una mayor eficiencia (Mignone y Pesole, 2007). La longitud de la 5’ UTR influye sobre la eficiencia de la traducción, determinando la energía necesaria para que un ribosoma se desplace para alcanzar el codón AUG a través de una 5’ UTR estructurada. Los truncamientos o mutaciones en las 5’ UTRs usualmente producen una síntesis proteica aberrante (Chatterjee y Pal, 2009).

Respecto al contenido de GC, existe una diferencia significativa entre los mRNAs de proteínas “house-keeping” y reguladores, ya que estos últimos tienen un alto contenido GC. Los mRNAs “house-keeping” muestran una asimetría alta G/C y A/T, lo cual desfavorece la formación de estructuras estables que puedan reducir la eficiencia traduccional (Mignone y Pesole, 2007).

Los genes que codifican para las proteínas involucradas en procesos celulares importantes, tales como la fertilización, desarrollo, ciclo celular, respuesta a estrés y oncogénesis suelen estar regulados a nivel de la traducción. Dicha regulación ocurre predominantemente a través de las UTRs (Pickering y Willis, 2005) La ocurrencia de 5’ UTRs alternativas en estos genes, se relaciona con la progresión de varias formas de cáncer (Davuluri, *et al.,* 2000).

Mecanismos de regulación del inicio de la traducción durante condiciones de estrés

La célula se encuentra en homeostasis y constantemente contiende con distintos tipos de estrés, impuestos por el ambiente. Algunas condiciones de estrés incluyen los cambios en la temperatura, limitación de nutrientes, estrés oxidativo, irradiación ultravioleta, hipoxia, así como la exposición a varias toxinas y drogas. La exposición celular al estrés desencadena respuestas adaptativas que requieren la expresión coordinada de genes de respuesta a estrés, los cuales afectan la supervivencia celular, la apoptosis, la progresión del ciclo celular y la diferenciación.

La traducción global se reduce en respuesta a la mayoría de los tipos de estrés celular. Lo anterior resulta en un ahorro de energía notable, se consume bastante de ella en el proceso de traducción (se estima que un poco más del 50% de la energía celular, dependiendo del organismo en cuestión, Mathews *e t al.,* 2000).

La atenuación en la traducción también previene la síntesis de aquellas proteínas que pueden interferir con la respuesta al estrés celular, mientras que por otro lado ocurre la traducción selectiva de aquellas proteínas requeridas para la supervivencia celular bajo estrés.

Se han descrito algunos mecanismos en la regulación del inicio de la traducción durante la respuesta a estrés: uno de ellos sucede durante la formación o regeneración del complejo ternario eIF2- Met-tRNAi-GTP y el otro en el reclutamiento del ribosoma a mRNA (Holcik y Sonenberg, 2005).

La fosforilación de los factores de inicio como mecanismo de regulación de la traducción

La compleja naturaleza del inicio de la traducción en eucariontes ofrece un número de posibles blancos de regulación como lo son los factores de inicio, que dependiendo de su estado de fosforilación pueden controlar la traducción globalmente. Las proteínas encargadas de fosforilarlos y desfosforilarlos son respectivamente cinasas y fosfatasas implicadas en condiciones de estrés.

El inicio de la traducción dependiente de cap puede ser inhibido entonces por la fosforilación de IF2 o por hipo-fosforilación de las proteínas que se unen al factor de inicio 4E-BP (eIF4E-binding protein) (King *et al.,* 2010).

La unión de GTP a eIF2 es el paso limitante en el ensamblaje del complejo ternario. eIF2 se recicla cuando es recargado con GTP por un factor intercambiador de guanina (GEF), eIF2B. Sin embargo, cuando eIF2 es fosforilado en el residuo de serina 51 de la subunidad α, se convierte en inhibidor competitivo de eIF2B, el cual deja de intercambiar el GDP por el GTP, previniéndose en primera instancia la formación del complejo ternario, y reduciendo la traducción de la mayoría de los mRNAs. La fosforilación de este factor depende de 4 cinasas en mamíferos, HRI que se activa en condiciones en donde la concentración del grupo prostético hemo es baja o en tratamiento con arsenito, choque térmico por calor o choque osmótico; GCN2 que se activa en respuesta a la limitación de aminoácidos e irradiación UV; PKR es activada por RNA de doble cadena en la respuesta antiviral; PERK se activa en respuesta al estrés en el retículo endoplásmico. Sin considerar el tipo de estímulo, la fosforilación de eIF2α causa efectos idénticos en la traducción: la inhibición de la traducción, aunque ocurre de igual forma la activación de una traducción selectiva (Holcik y Sonenberg, 2005). (Tema central de esta tesis y que se discute en la siguiente sección).

Por otro lado, cuando 4E-BP se encuentra hiper-fosforilado, no se puede unir a eIF4E; sin embargo cuando ocurre el caso contrario y está hipo-fosforilado es capaz de competir en la unión de eIF4E con eIF4G, y consecuentemente secuestra a eIF4E. Así, los niveles de eIF4F llegan a un estado limitante y la traducción dependiente de cap se inhibe (King *et al.,* 2010).

Inicio de la traducción cap independiente

Durante mucho tiempo el inicio dependiente de cap o el modelo de escaneo se consideró la única ruta posible a través de la cual la traducción de los mRNAs eucariontes podría iniciar. Sin embargo, a finales de 1980, estudios sobre la expresión genética viral revelaron una manera alternativa de iniciar la traducción en células eucariontes, que no requiere de la estructura cap como sitio de ensamblaje para los factores de inicio y por tanto no ocurre el reconocimiento de cap por la subunidad 40S del ribosoma, vía el complejo de unión a cap intacto eIF4F (Komar y Hatzoglou, 2011). Un proceso en donde estructuras secundarias dentro del mRNA reclutan *per se* a la subunidad 40S del ribosoma. Este inicio de la traducción se conoce como inicio de la traducción interno; los sitios de unión al ribosoma estructurados, presentes en la 5'UTR del mRNA responsables de este proceso se conocen como sitios internos de entrada para el ribosoma o IRES (por sus siglas en Inglés, Internal Ribosome Entry Sites) (Baird *et al.,* 2006).

IRES virales

Los primeros IRES fueron identificados originalmente en miembros de la familia picornaviridae. Esta familia viral presenta genomas de RNA positivos, en donde los mensajeros se encuentran sin estructura cap y aún así pueden ser traducidos eficientemente por la célula hospedera eucarionte, por lo que en un inicio se especuló que podrían ser traducidos de una manera cap independiente (King *et al.,* 2010).

Subsecuentemente se demostró que las 5' UTRs de gran tamaño y altamente estructuradas de los mRNAs del virus de encafalomiocarditis y poliovirus eran capaces de dirigir el inicio de la traducción reclutando la subunidad pequeña del ribosoma directamente, independientemente de la actividad de eIF4F, ya que ocurría la degradación proteolítica del factor de inicio eIF4G por parte del virus, funcionando como una estrategia para reducir la eficiencia en el inicio de la traducción dependiente de cap en la célula hospedera, y así traducir selectivamente los mRNAs virales (Pelletier y Sonenberg, 1988; Jang *et al.,* 1988).

Los IRES virales muestran un menor requerimiento de los factores de inicio canónico, de manera particular pueden prescindir de los miembros del complejo eIF4F (eIF4E y eIF4G)[1]. Los factores de inicio canónicos que requieren los IRES varían en diferentes mRNAs. La alta conservación estructural del RNA de los IRES entre grupos virales juega un papel determinante en su función, la compleja estructura secundaria y terciaria está involucrada en la interacción directa entre el IRES y algunos factores de inicio canónicos. En algunos casos estas interacciones reducen casi por completo el requerimiento de los factores de inicio. Más aún, en casos extremos, el inicio puede darse sin involucrar a ninguno de los factores, recayendo sólo en las interacciones directas entre el IRES y el ribosoma 40S (Figura 4) (Filbin y Kieft, 2009). A la fecha existen 62 casos de IRES virales reportados en la literatura (Mokrejs *et al.,* 2006).

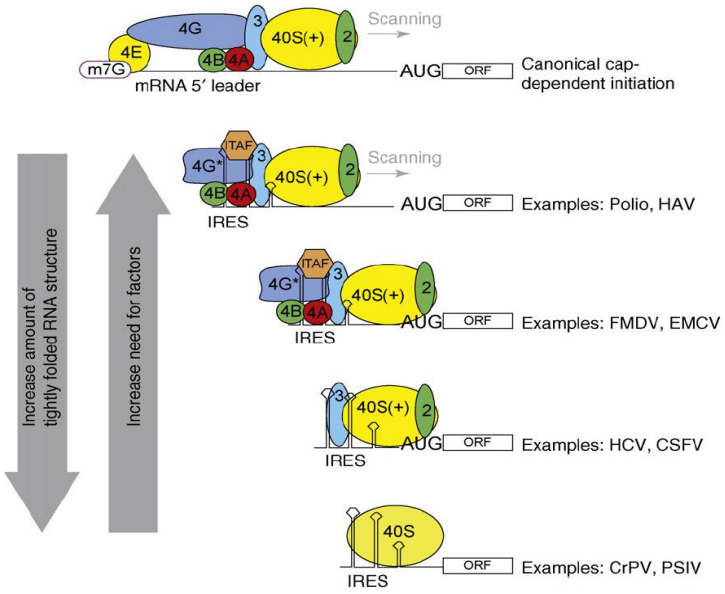


Figura 4. El RNA que conforma a los IRES con las estructuras más estables (flecha de la izquierda) son aquellos que requieren menos factores de inicio de la traducción, caso contrario los IRES que se encuentran menos estructurados requieren una mayor cantidad de ITAFs (IRES specific Trans Acting Factors) y eIFs. Esto ocurre al menos en IRES virales, el grado en que esto puede ser predicho o extendido a IRES celulares aún es desconocido. Tomado de Filbin y Kieft 2009.

Adicionalmente existen otros factores proteicos celulares que modulan (por lo común intensifican) el inicio interno y actúan en *trans*, ITAFs (IRES-specific cellular *trans*-acting factors), los cuales no forman parte de la maquinaria de inicio clásica y son proteínas de unión a RNA que tienen una variedad de funciones celulares adicionales a la promoción del inicio de la traducción interna, sin embargo parecen no estar involucrados en el

[1] Son los mismos virus los que presentan las proteasas para degradar a dichos factores de inicio o desencadenar respuestas celulares (p.ej. apoptosis) en las cuales caspasas u otras proteínas pueden degradar a dichos factores, comprometiendo el inicio dependediente de cap como se tratará más adelante.

inicio de la traducción cap dependiente. Se ha propuesto que la función de estos factores es estabilizar la conformación de los IRES para promover la unión de otros factores y del ribosoma, funcionando como chaperonas de RNA (Filbin y Kieft, 2009) (Figura 2B). De ellos se hablará más detalladamente en secciones posteriores.

IRES celulares

Los IRES celulares se encontraron por primera vez en células infectadas con poliovirus, en donde se demostró que proteasas virales degradaban al factor de reconocimento de la estructura cap eIF4E, lo que comprometía la traducción dependiente de cap y en última instancia causaba prevenía la traducción de la mayoría de los mRNAs celulares, sin embargo se observó la traducción selectiva por un lado de los mRNAs virales, para los cuales ya se sabía que dicha traducción se debía a la formación de IRES en la región 5'UTR de los mRNAs, que eran capaces de reclutar al ribosoma y algunos otros factores de la maquinaria de inicio de la traducción. Por otro lado, en el mismo estudio se reportó la traducción selectiva del mRNA de la proteína celular de unión a la cadena pesada de inmunoglobulina BiP del hospedero; en el mismo estudio se determinó que la región 5'UTR de éste gen presentaba un IRES (Macejak y Sarnow, 1991). En estudios posteriores se empleó la misma estrategia de abatir la traducción dependiente de cap, a través de la infección con poliovirus en células, y así se identificaron a otros IRES celulares (Johannes *et al.*, 1999).

Posteriormente se describieron una serie de IRES celulares en donde la traducción dependiente de cap era menos eficiente en varias condiciones celulares, sin embargo ciertos mRNAs específicos seguían traduciéndose de manera relativamente eficiente (Baird *et al.,* 2007).

Las condiciones en donde se altera la homeostasis de la célula, en las cuales se han descrito IRES celulares, incluyen el estrés (por choque térmico, hipoxia, limitación de nutrientes), el crecimiento celular y diferenciación, ciclo celular, apoptosis, entre otras (Tabla 2). Es durante estas condiciones que el inicio de la traducción dependiente de cap se ve disminuido, debido a la fosforilación de eIF2α y/o la proteólisis de la proteína de andamiaje eIF4G o la proteína de reconocimiento de la estructura cap eIF4E, previniendo la formación del complejo eIF4F activo.

La proteólisis del factor eIF4G, causada por las proteasas virales durante la infección o por caspasas durante la apoptosis, separa físicamente los sitios de unión para eIF4E y eIF3 (los cuales reclutan al ribosoma), por lo cual no ocurre el inicio de la traducción dependiente de cap y consecuentemente se da el inicio independiente de cap, vía un sitio de entrada interno para el ribosoma (Graber *et al.,* 2006).

Se estima que hasta un 10% de los mensajeros celulares tienen IRES, ya que del 3% al 5% de los mRNAs se asocian con los poli-ribosomas cuando se inhibe el inicio de la traducción dependiente de cap en las células infectadas con poliovirus (Johannes *et al.,* 1999) y un 3% de los mRNAs se encontró en los polisomas durante la mitosis (Qin y Sarnow, 2004). Lo anterior implica que el inicio de traducción interno es un mecanismo celular importante para la regulación de la traducción y no sólo una estrategia viral especializada.

Así, la traducción mediada por IRES provee de una estrategia para escapar de la reducción global de la síntesis de proteínas y permite la traducción selectiva de mRNAs específicos que son responsables de contender y responder a las condiciones de estrés y en última instancia de promover la sobrevivencia o la muerte celular (Holcik y Sonenberg, 2005).

Los IRES celulares han sido descritos de manera limitada, no obstante ha ido creciendo la cantidad de mRNAs que poseen estos sitios. Al menos 70 mRNAs celulares que contienen IRES han sido descritos en levadura, mosca, maíz y algunos mamíferos (Mokrejs *et al.,* 2006) (Figura 5) (Tabla 2).

Como se puede observar en la figura 5, existe un sesgo en la identificación de IRES celulares ya que la mayoría de éstos se restringen al humano y a organismos mamíferos, dado el interés que suscitan por el papel que desempeñan en la sobrevivencia celular y más aún la importancia en su relación con algunas enfermedades.

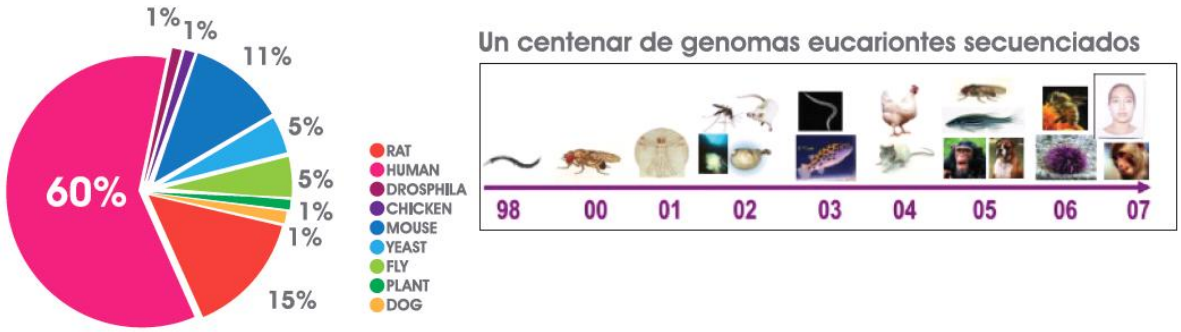


Figura 5. Porcentaje de IRES celulares descritos a la fecha. El organismo en el cual se ha descrito la mayoría de IRES es el humano, se han descrito muy pocos para un grupo reducido de otros eucariontes, pese al crecimiento de genomas eucariontes totalmente secuenciados.

Se sabe que la mayoría de los IRES que han sido reportados se encuentran en mRNAs con 5'UTRs largos, con un alto contenido de GC y altamente estructurados, en algunos casos pueden presentar varios codones de inicio; estas características son suficientes para dificultar el escaneo por lo que los IRES son relativamente ineficientes en dirigir la traducción bajo condiciones fisiológicas que favorecen la traducción dependiente de cap (Holcik y Sonenberg, 2005). El tamaño de los IRES es variable y comprenden aproximadamente desde 100 hasta 1500 nucleótidos.

IRES celular	Organismo	Circunstancia celular general	Vía en la que participa [a]
AML1/RUNX1	*Homo sapiens*	Diferenciación	*Cáncer *Leucemia mieloide crónica *Leucemia mieloide aguda
Antp	*Drosophila melanogaster*	Desarrollo	*Desarrollo del corazón *Especificidad en la identidad de segmentos * Desarrollo del neuroblasto *Desarrollo de la glándula linfoide
Apaf-1	*Mus musculus, Homo sapiens*	Apoptosis	*Vía de señalización p53 *Apoptosis *Enfermedad de Alzheimer *Enfermedad de Parkinson *Esclerosis lateral amiotrófica *Enfermedad de Huntington *Tuberculosis *Cáncer de pulmón
AQP4	*Homo sapiens*	Eclampsia	*Vasopresina-regulada por la reabsorción de agua *Secreción biliar
hAT1R-A,B,C,D	*Homo sapiens*	Limitación de suero fisiológico	*Vía de señalización por calcio *Interacción neuroactiva ligando-receptor
BAG-1	*Homo sapiens*	Choque térmico por calor	*Procesamiento de proteínas en el retículo endoplásmico
BCL2	*Homo sapiens*	Apoptosis	*Procesamiento de proteínas en el retículo endoplásmico *Apoptosis *Adhesión focal * Vía de señalización de la neurotrofina *Sinapsos colinérgica *Esclerosis lateral amiotrófica *Toxoplasmosis *Tuberculosis Vías en cáncer *Cáncer colorectal

			*Cáncer de próstata *Cáncer de pulmón
BiP	*Homo sapiens*	Choque térmico por calor, Hipoxia	*Exportación de proteínas *Procesamiento de proteínas en el retículo endoplásmico *Pocesamiento y presentación de antígenos *Enfermedades por priones
c-IAP1	*Homo sapiens*	Apoptosis	*Proteólisis mediada por ubiquitina *Apoptosis *Adhesión focal *Señalización por receptor similar a NOD *Toxoplasmosis *Vías en cáncer *Cáncer de pulmón
c-jun	*Gallus gallus*	Crecimiento celular, Diferenciación, Oncogénesis	*Vía de señalización MAPK *Vía de señalización ErbB *Vía de señalización Wnt *Diferenciación de osteoclastos *Adhesión focal *Vía de senalización por receptores similares a Toll *Vía de señalización por el receptor de células B *Vía de señalización de neurotrofina *Vía de señalización GnRH *Señalización en células epiteliales durante la infección por *Helicobacter pylori* *Pertussis *Leishmaniasis *Enfermedad de Chagas *Influenza A *Vías en cáncer *Cancer colorectal *Carcinoma renal *Artritis reumatoide
c-myc	*Homo sapiens*	Desarrollo, Apoptosis, Estrés genotóxico, Hipoxia, Oncogénesis, Ciclo celular	*Vía de señalización MAPK *Vía de señalización ErbB *Ciclo celular *Vía de señalización Wnt *Vía de señalización TGF-beta *Vía de señalización Jak-STAT *Vías en cáncer *Cáncer colorectal *Cáncer endometrial *Cáncer de tiroides *Cáncer de vejiga

			*Leucemia mieloide crónica *Leucemia mieloide aguda *Cáncer de pulmón
c-Src	*Homo sapiens*	Oncogénesis	*Vía de señalización ErbB *Shigellosis *Tuberculosis *Vía de señalización VEGF *Uniones Gap *Invasión bacteriana del tejido epitelial
Cat-1	*Rattus norvegicus*	Limitación de aminoácidos	*Transporte de aminoácidos catiónicos
CCND1	*Homo sapiens*	Ciclo celular	*Ciclo celular *Vía de señalización p53 *Señalización Wnt *Adhesión focal * *Vía de señalización Jak-STAT *Sarampión *Vías en cáncer *Cáncer colorectal *Cáncer pancreático *Cáncer endometrial *Glioma *Cáncer de próstata *Cáncer de tiroides *Melanoma *Cáncer de vejiga *Leucemia mieloide crónica *Leucemia mieloide aguda *Cáncer de pulmón *Miocarditis viral
DAP5	*Homo sapiens*	Apoptosis	*Transporte de RNA *Miocarditis viral
eIF4G, eIF4GI, eIF4GII	*Homo sapiens*	Traducción	*Transporte de RNA *Miocarditis viral
ELG1	*Homo sapiens*	Apoptosis	*Apoptosis
ELH	*Aplysia californica*	Sinapsis	*Neurohormona
FGF1	*Homo sapiens*	Desarrollo	*Vía de señalización MAPK *Regulación del citoesqueleto de actina *Vías en cáncer *Melanoma
FMR1	*Homo sapiens*	Sinapsis	*Retraso mental ligada al cromosoma X
Gtx	*Mus musculus*	Diferenciación	*Diferenciación de oligodendrocitos
Hairless	*Drosophila melanogaster*	Ciclo celular , Desarrollo	*Desarrollo de órganos sensoriales
HAP4	*Saccharomyces cerevisiae*	Represión catabólica	*Transporte de electrones en mitocondria

Hif1α	*Mus musculus*	Hipoxia	*Vía de señalización mTOR *Vías en cáncer *Carcinoma renal
hSNM1	*Homo sapiens*	Mitosis, Ciclo celular	*Ciclo celular *Reparación de DNA
Hsp101	*Zea mays*	Choque térmico por calor	*Termotolerancia
Hsp70Aa	*Drosophila melanogaster*	Choque térmico por calor	*Spliceosoma *Procesamiento de proteínas en retículo endoplásmico *Endocitosis
HSPA1A	*Homo sapiens*	Choque térmico por calor	*Spliceosoma *Vía de señalización MAPK *Procesamiento de proteínas en retículo endoplásmico *Endocitosis *Presentación y procesamiento de antígenos *Enfermedades por priones *Toxoplasmosis *Sarampión *Influenza A
Hsp83	*Drosophila melanogaster*	Choque térmico por calor	*Procesamiento de proteínas en retículo endoplásmico *Maduración del ovocito mediada por progesterona
IGF2	*Homo sapiens*	Desarrollo	*Desarrollo gestacional
Kcna4	*Mus musculus*	Sinapsis	*Gradientes de potasio en miorcardio
L-myc	*Homo sapiens*	Desarrollo, Apoptosis, Estrés genotóxico, Hipoxia, Oncogénesis	*Carcinoma de pulmón *Protoncogen
LamB1	*Homo sapiens*	Diferenciación	*Adhesión focal *Interacción con receptor ECM- *Toxoplasmosis *Amibiasis *Vías en cáncer *Cáncer pulmonar
LEF1	*Homo sapiens*	Desarrollo, Hipoxia, Estrés genotóxico	*Vía de señalización Wnt *Uniones adherentes *Melanogenesis *Vías en cáncer *Cáncer colorectal *Cáncer endometrial *Cáncer de próstata *Cáncer de tiroides *Carcinoma de células basales *Leucemia mieloide aguda *Cardiomiopatía aritmogénica ventricular derecha

MNT	*Homo sapiens*	Desarrollo	*Crecimiento celular
MTG8a	*Homo sapiens*	Diferenciación	*Vías en cancer *Leucemia mieloide aguda
c-myb	*Homo sapiens*	Desarrollo, Oncogénesis	*Infección por HTLV-I
MYT2	*Homo sapiens*	Crecimiento celular	*Desarrollo del sistema nervioso
n-MYC	*Homo sapiens*	Desarrollo, Crecimiento celular, Apoptosis, Oncogénesis	*Protoncogen *Neuroblastoma *Vías en cáncer
NDST1,3,4L,4S	*Mus musculus*	Crecimiento celular	*Biosíntesis de glucosaminoglucanos- heparan sulfato *Vías metabólicas
NRF1	*Homo sapiens*	Crecimiento celular	*Enfermedad de Huntington
NtHSF1	*Nicotiana tabacum*	Choque térmico por calor	*Termotolerancia
ODC1	*Rattus norvegicus*	*Ciclo celular	*Metabolismo de arginina y prolina *Metabolismo de glutatión *Vías metabólicas
p27kip1	*Homo sapiens*	Ciclo celular	*Vía de señalización ErbB *Ciclo celular *Sarampión *Vías en cáncer *Cáncer de próstata *Leucemia mieloide crónica *Cáncer pulmonar
p53/p47	*Homo sapiens*	Daño celular, Diferenciación, Ciclo celular	*Vía de señalización p53 *Estrés oxidativo
PDGF2/c-sis	*Homo sapiens*	Diferenciación, Oncogénesis	*Vía de señalización MAPK *Interacción receptor citosina-citosina *Adhesión focal *Uniones Gap *Regulación del citoesqueleto de actina *Vías en cáncer *Carcinoma renal *Glioma *Cáncer de próstata *Melanoma
PIM1	*Homo sapiens*	Diferenciación , Oncogénesis	*Vía de señalización Jak-STAT *Leucemia mieloide aguda
PITSLRE	*Homo sapiens*	Ciclo celular, Apoptosis	*Melanoma *Neuroblastoma
Rbm3	*Mus musculus*	Choque térmico por frío	*Choque térmico por frío
Rpr	*Drosophila*	Choque térmico por	*Apoptosis

	melanogaster	calor, Apoptosis	
Scamper	*Canis familiaris*	Apoptosis	*Apoptosis
TBP1	*Saccharomyces cerevisiae*	Limitación de glucosa	*Factores de transcripción basales
TIF4631	*Saccharomyces cerevisiae*	Crecimiento celular	*Transporte de RNA (Factor de inicio de la traducción 4G)
Ubx	*Drosophila melanogaster*	Desarrollo	*Especificación alas y patas
UNR	*Homo sapiens*	Traducción (ITAF)	*ITAF, traducción cap independiente
Ure2	*Saccharomyces cerevisiae*	Represión catabólica	*Metabolismo de glutatión
Utrn	*Mus musculus*	Desarrollo	*Distrofia muscular
VEGF-A	*Mus musculus*	Hipoxia	*Interacción de receptor citosina-citosina *Vía de señalización mTOR *Vía de señalización VEGF *Adhesión focal *Vías en cáncer *Carcinoma renal *Cáncer pancreático *Cáncer de vejiga *Artitris reumatoide
XIAP	*Homo sapiens*	Apoptosis	*Proteólisis mediada por ubiquitina *Apoptosis *Adhesión focal *Vía de señalización similar a NOD *Toxoplasmosis *Infección por HTLV-I *Vías en cáncer *Cáncer pulmonar
YAP1	*Saccharomyces cerevisiae*	Estrés oxidativo	*Vías de reducción de tioles

Tabla 2. IRES celulares reportados hasta la fecha, en la base de datos IRESite. [a] La vía se reporta según la base de datos KEGG.

IRES celulares como moduladores de la traducción

Los IRES celulares en contraste con los IRES virales median la traducción de los transcritos con una eficiencia menor, lo cual ha dificultado su estudio, aunado a que la traducción mediada por IRES celulares parece estar regulada por una serie sofisticada de múltiples mecanismos de control (Komar y Hatzoglou, 2005).

La creciente evidencia acerca del mecanismo de los IRES celulares indica que éstos tienen dos funciones fisiológicas principales; por un lado generan bajos niveles de inicio de la traducción para aquellos mRNAs con IRES que están muy estructurados (incompatibles con un escaneo eficiente) bajo condiciones fisiológicas en donde la traducción dependiente de cap se encuentra completamente activa y, promueven la traducción eficiente de los mensajeros celulares en una amplia gama de condiciones fisiológicas en donde la traducción dependiente de cap se halla comprometida (Komar y Hatzoglou, 2011).

Se ha considerado que todos los mensajeros deben presentar la estructura cap y son capaces de unirse al complejo eIF4F. Sin embargo prevalece aún la idea de que el escaneo convencional se vería imposibilitado en el caso de las 5´UTRs con IRES por la gran longitud que tienden a presentar, un alto contenido de GC, la estructura secundaria y en algunos casos la presencia de codones de inicio de la traducción río arriba del codón de inicio *bona fide.* Existe al menos un caso descrito en el cual operan ambos mecanismos de inicio de la traducción sobre el mismo mensajero. El mRNA de la neurogranina, una proteína de unión a la calmodulina, puede traducirse por el mecanismo 5´ dependiente de cap y por vía inicio interno (Pinkstaff *et al.,* 2001).

Las distintas condiciones de estrés en el retículo endoplasmático, hipoxia, limitación de nutrientes, mitosis, diferenciación celular entre otras, favorecen la traducción vía IRES, y se considera que éstos se vuelven más competitivos por la cantidad disponible de ribosomas y factores de inicio, incluyendo a los factores canónicos e ITAFs.

El mecanismo de acción de los IRES virales se comprende de una manera más completa, mientras que aún no se entiende bien a bien el mecanismo molecular por el cual los IRES celulares dirigen la traducción.

Estructura de los IRES celulares

La estructura secundaria de algunos IRES celulares se ha determinado utilizando pruebas químicas y enzimáticas de algunos transcritos, incluyendo a c-Myc, Apaf-1 y FGF2, entre otros. Algunos de ellos presentan estructura secundaria compleja que incluye la formación de tallos y asas, así como pseudo-nudos (Komar y Hatzoglou, 2011).

La cercanía filogenética entre grupos virales de la misma familia ha generado una idea errónea acerca de la conservación estructural de IRES virales descritos, causada por una redundancia de secuencias genómicas, y por tanto una sobrestimación en la búsqueda de motivos conservados. En los IRES celulares es importante la formación de estructuras secundarias estables (Filbin y Kieft, 2009). En contraste con los virus, no existe una conservación obvia en la estructura secundaria entre IRES celulares, incluso en IRES celulares que modulan a

genes parálogos (Figura 6); algunos de los organismos modelo en los que han sido descritos por ejemplo son *Saccharomyces cerevisiae* y *Drosophila melanogaster,* aunque para ninguno de éstos se han hecho estudios en donde se encuentren nuevos IRES celulares a partir de la extrapolación en otros genes parálogos.

Para algunos casos, los IRES únicamente se conservan motivos estructurales muy pequeños o estructuras alternativas complementarias al rRNA 18S. Así mismo, esta falta de conservación entre IRES celulares también es evidente en relación a su secuencia primaria de RNA (Baird et al., 2006).

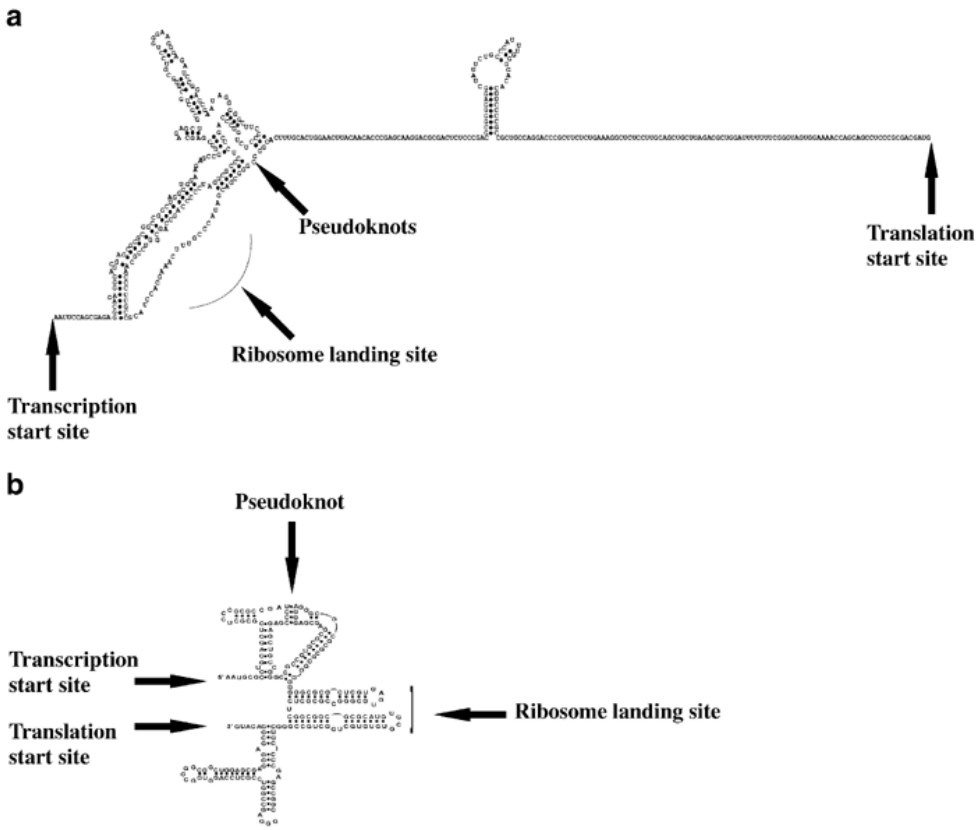


Figura 6. Modelos de estructura secundaria de los IRES c-*myc* y L-*myc*. Se observa que la estructura secundaria es totalmente distinta en ambos casos, siendo que sus genes son parálogos. Tomado de Stonoley y Willis, 2004.

Pese a la falta de conservación en secuencia primaria y estructura secundaria, se ha observado que las secuencias regiones 5´UTR con IRES tienden a estar altamente estructuradas y con un alto contenido de GC (Dinkova *et al.,* 2005); esta compleja estructuración pudiera estar involucrada en la interacción con múltiples componentes de la maquinaria de traducción (factores de inicio de la traducción canónicos, ITAFs, así como la subunidad 40s del ribosoma) (Komar y Hatzoglou, 2005). Más aún, basándose en experimentos de deleción y mutagénesis dirigida, se ha observado que algunas secciones individuales de los IRES celulares son capaces de modular *per se* el inicio de la traducción, aunque no son tan eficientes como el IRES en su totalidad. Se ha propuesto la hipótesis de que la gran mayoría de los IRES celulares están compuestos de varios módulos

pequeños y que el efecto combinado de éstos promueve el inicio interno de la traducción (Stonoley y Willis, 2004). El cómo estos motivos se combinan para promover el inicio interno aún tiene que ser esclarecido.

Los IRES celulares entonces parecen ser mucho más diversos en su estructura de lo esperado, y se cree que los ITAFs están directamente relacionados la inducción de cambios estructurales en los IRES celulares para que estos respondan precisamente a los cambios en las condiciones celulares, lo cual aumenta la complejidad en la comprensión del mecanismo independiente de cap.

Mecanismo de mediación de la traducción a través de IRES celulares

Los IRES celulares participan en múltiples interacciones con los componentes de la maquinaria de inicio de la traducción (factores de inicio canónicos, ITAFs y subunidades ribosomales 40S). Se ha propuesto que dichas interacciones proveen un posicionamiento del codón de inicio en el sitio P del ribosoma, sin que ocurra el escaneo del ribosoma desde el extremo 5' del mensajero (Komar y Hatzoglou, 2005). Sin embargo esta idea continua siendo una hipótesis, ya que no existen estudios extensivos y sistemáticos de la capacidad de los IRES celulares a unirse a la subunidad ribosomal 40S o del requerimiento parcial de los factores de inicio durante el inicio interno.

Se especula que algunos IRES podrían funcionar junto con el ribosoma en un mecanismo de "aterrizaje" (landing), en la vecindad del codón de inicio y escaneo, un mecanismo típico de los IRES de picornavirus. Los IRES de c-Myc, L-Myc y N-Myc utilizan este mecanismo (Spriggs *et al.,* 2009). Asimismo se ha sugerido que algunos IRES celulares pueden operar con una interacción parecida a la que se da entre el ribosoma y la secuencia Shine-Dalgarno, de tal forma que esta interacción podría ocurrir entre el IRES y el rRNA 18S (Komar y Hatzoglou, 2011).

Recientemente el primer caso de un IRES celular que es capaz de unirse al ribosoma directamente se reportó en el mensajero de la cinasa c-Src (Allam y Ali, 2010).

El papel de los ITAFs en la mediación de la traducción a través de IRES celulares

La compleja naturaleza de la regulación de los mRNAs celulares bajo distintas condiciones de estrés sugiere que existen varias maneras por las cuales los IRES celulares modulan el inicio. Todos ellos responden diferencialmente a cada contexto celular en el cual la traducción dependiente se encuentra inhibida. Por ejemplo, durante la apoptosis, el IRES de Apaf-1 (Apoptosis protease-activating factor-1) es activo, en contraste con el IRES de XIAP (X-linked inhibitor of apoptosis protein) que se encuentra inhibido (Holcik y Sonenberg,

2005). El por qué los IRES responden de distinta forma bajo el mismo estímulo se ha explicado en gran parte por la participación de los ITAFs, los cuales son responsables de monitorear los cambios en el metabolismo celular e influenciar la actividad de los IRES (Lewis y Holcik, 2008).

La mayoría de los ITAFs pertenecen al grupo de las ribonucleoproteínas heterogéneas nucleares (HnRNP, A1, C1/C2, I, E1/E2, K y L) que se transportan entre el núcleo y el citoplasma, adicionalmente a su participación en una variedad de actividades tales como el "splicing" (Komar y Hatzoglou, 2005) (Tabla 3).

El mecanismo subyacente a la función de los ITAFs es desconocido pero existen algunas hipótesis, a saber: 1) estos remodelan la estructura espacial para producir conformaciones con mayor o menor afinidad a los componentes de la maquinaria de traducción, 2) construyen o abaten puentes entre el mRNA y el ribosoma adicionales a los que se forman con los factores de inicio, 3) toman el lugar de los factores de inicio canónicos, construyendo puentes entre el mRNA y el ribosoma (Lewis y Holcik, 2008).

ITAF	Blanco
PTB/nPTB	Apaf-1
	Bag-1
	Mnt
	Myb
	MTG8a
	BiP
	IGF1R
PCBP1/2	c-myc
hnRNPK c-	c-myc
La	BiP
	XIAP
Unr	Apaf-1
hnRNPC1/2	XIAP
DAP5	XIAP
	HIAP
	Apaf-1
	c-myc
Fragmento eIF4GI	Apaf-1
	DAP5
ELAV/Hu	P27

Tabla 3. ITAFs celulares y sus IRES blanco descritos hasta la fecha. Tomado de Spriggs *et al.,* 2005.

La concentración intracelular de los ITAFs es importante en la modulación de la actividad de los IRES, pero el mecanismo responsable de regular dicha concentración no ha sido del todo definido. Algunos estudios

recientes señalan que existe una distribución subcelular (núcleo/citoplasma) de ITAFs que determina la actividad de los IRES. Por ejemplo, la relocalización citoplásmica de un ITAF puede tener efectos opuestos en la traducción dependiente de IRES, dependiendo del IRES blanco para el ITAF (Lewis y Holcik, 2008). La acumulación citoplásmica de hnRNP A1 intensifica la traducción vía IRES de FGF2 (factor de crecimiento del fibroblasto 2) pero reprime la traducción vía IRES de XIAP (Lewis *et al.,* 2007).

Hay al menos dos propuestas para explicar los efectos de la compartamentalización de ITAFs. En un modelo, los ITAFs que se encuentran en el núcleo se asocian con el IRES blanco y lo secuestran en el núcleo, lejos de la maquinaria de traducción (Figura 7a). En el otro, los ITAFs en el núcleo se encuentran separados del IRES blanco el cual reside en el citoplasma (Figura7b). Cuando ocurren las señales apropiadas (causadas por estrés u otras condiciones fisiológicas), el complejo ITAF-IRES (primer modelo) o los ITAFs solos (segundo modelo) se translocan desde el núcleo al citoplasma, permitiendo que la traducción de los mensajeros prosiga (Lewis y Hocik, 2008).

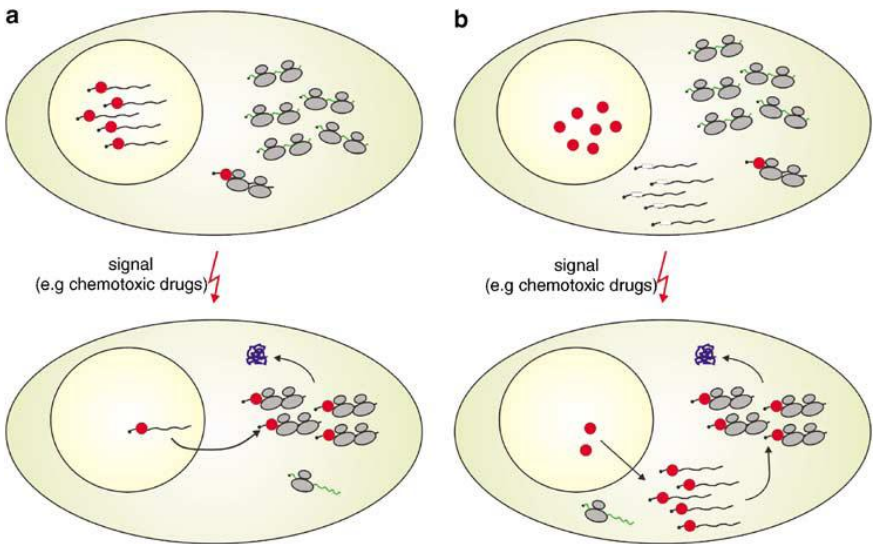


Figura 7. Modelos de localización subcelular propuestos en la regulación de IRES-ITAFs. Los ITAFs son los círculos rojos, la estructura m7Gcap es el circulo pequeño negro, los IRES son los rectángulos blancos y los no IRES son verdes. Tomado de Lewis y Holcik, 2008.

En algunos casos, las modificaciones post- traduccionales de los ITAFs, desencadenadas por el estrés, afectan la localización subcelular y la afinidad de unión por el IRES. La fosforilación de HnRNP A1 afecta la distribución subcelular y su capacidad de modular la actividad de sus IRES blanco: ciclina D1, c-Myc, FGF2, VEGF, XIAP, Apaf-1 y Unr, respectivamente (Lewis *et al.,* 2007).

La expresión genética en primera instancia, y la concentración de los ITAFs entre líneas celulares es variable, lo cual explica por qué aún y cuando se trate del mismo IRES celular, se registra una actividad distinta para la línea celular en cuestión, dando como resultado una actividad tejido específica (King *et al.,* 2010). Por ejemplo, el ITAF neuronal específico nPTB es responsable de intensificar el inicio interno de Apaf-1 en líneas celulares de origen neuronal (Stoneley y Willis, 2004). Así, la regulación de los niveles de ITAFs es importante para un control preciso de la actividad de los diferentes IRES.

Aunque la cantidad de la lista de IRES celulares sigue en aumento, no se ha descrito uno o más ITAFs "universales" para mediar la traducción vía IRES en células eucariontes. Lo anterior sugiere que los IRES son elementos que se regulan diferencialmente bajo distintas condiciones fisiológicas.

Importancia pato-fisiológica de los IRES celulares

Uno de los mayores obstáculos en la elucidación del significado de la mediación de la traducción vía IRES en las distintas vías celulares es la compleja naturaleza de muchas de las proteínas con IRES y los múltiples mecanismos que controlan su expresión. Muchas de estas proteínas (p.ej. c-Myc, Apaf-1, FGF, XIAP, p53, VEGF entre otras) son reguladoras maestras de la sobrevivencia celular, la proliferación y muerte. La expresión de estas proteínas está usualmente controlada por una variedad de mecanismos que operan a diferentes niveles, incluyendo la transcripción, el "splicing", la traducción, la localización proteica y estabilidad. Aunque es común que muchos de dichos mecanismos operen y se enciendan simultáneamente, se ha podido discernir la importancia de los IRES como moduladores de la traducción, y más aún se ha podido identificar el papel directo de los IRES en algunas enfermedades, como se ejemplifica más adelante. También se ha sugerido un papel indirecto de estos elementos, en algunas condiciones patológicas como la diabetes, enfermedades cardiovasculares, desarrollo y progresión de ciertos tipos de cáncer (Komar y Hatzoglou, 2005).

La regulación aberrante de la traducción a través de un IRES se ha identificado en el mieloma múltiple en humano. Los pacientes con este desorden tienen una expansión en el plasma celular de la médula ósea y exhiben osteólisis. Se ha identificado una substitución en un residuo de citosina por un residuo de uracilo (C→U) en el IRES de c-Myc, resultando en un incremento de la traducción (Stonoley y Willis, 2004). Dicha substitución estabiliza la formación de complejos RNA-proteína presentes en el IRES c-Myc, específicamente, aumenta la interacción con el ITAF hnRNPK, lo que resulta en la intensificación del inicio interno de la traducción del transcrito de c-Myc en esta enfermedad (Chappell *et al.,* 2000).

Antecedentes

Mientras que el mecanismo de inicio de traducción mediado por IRES puede explicar como algunos genes celulares se traducen en ausencia de la estructura de cap cuando la maquinaria traduccional es atenuada, la naturaleza del mecanismo molecular que subyace en los IRES aún no se entiende en su totalidad (Baird *et al.,* 2006). No obstante, se han identificado las propiedades comunes a los elementos IRES: a) una 5'UTR mayor a la longitud observada en aquellas 5' UTRs que no presentan IRES b) un contenido relativamente alto de GC c) múltiples tripletes AUG d) múltiples uORFs e) formados por elementos de estructura secundaria estables, f) Carecer de conservación en su estructura secundaria, g) Carecer de conservación en su secuencia primaria, h) estructuras secundarias altamente estructuradas, con una gran estabilidad, i) son estructuras modulares. Es pertinente señalar que no todas estas características se cumplen para todos los IRES celulares descritos a la fecha, si bien son tendencias observadas en la mayoría de ellos. Dado lo anterior y por la gran variabilidad de los diferentes tipos de IRES celulares, resulta evidente que no existe un mecanismo molecular universal del funcionamiento de los IRES celulares y por tal motivo, resulta complejo desarrollar una estrategia general y eficiente para la identificación *in silico* de nuevos IRES celulares.

A la fecha, la mayoría de las estrategias computacionales de identificación de IRES se basan en la combinación de diferentes métodos de predicción de estructura secundaria del RNA y en la presencia de pequeños motivos estructurales de RNA descritos previamente en IRES celulares (Baird *et al.,*2006).

La predicción de estructura secundaria en el RNA se puede determinar de dos formas, la primera se basa en la energía libre mínima del plegamiento termodinámicamente más estable. La segunda forma establece un método comparativo que examina la covarianza en secuencias homólogas con el fin de predecir las posiciones en la secuencia que pudieran ser consistentes con el apareamiento de las bases, inclusive si éstas cambiaran de una secuencia homóloga a otra. Una de las limitaciones de este método es que la predicción de la estructura requiere de una gran cantidad de secuencias con una estructura similar para que sea funcional (Baird *et al.,* 2006).

Uno de los aspectos más discutidos se relaciona con la estructuración de los IRES; se ha señalado que éstos tienden a presentar una energía libre menor y por tanto ser más estables, característica que sería reconocida por el ribosoma durante la traducción independiente de cap, sin embargo, aún no es clara la diferencia entre el nivel de estructuración de los IRES celulares del resto de otros mensajeros del genoma que no presentan IRES.

A través de un análisis estadístico no-paramétrico en donde se construyen árboles de decisión y clasificación (CART, Classification and regression trees) Davuluri *et al.,* (2000) señalan a la energía libre del RNA como la variable con mayor predictibilidad para la clasificación de las 5'UTRs respecto a otras características que también fueron tomadas en cuenta.

Respecto a la metodología experimental para la verificación de IRES celulares, existe un método estándar que consiste en la construcción bicistrónica con dos genes reporteros, en donde la secuencia candidata IRES se inserta entre los dos cistrones reporteros, ambos son transcritos en el mismo RNA. El principio de este ensayo recae en la observación de que los eucariontes no traducen todos los ORFs en un mRNA policistrónico, debido a que el ribosoma es liberado del mensajero después de la traducción del cistrón 5'. Si el inserto a probar causa un incremento en la expresión del cistrón río abajo (cistrón 3') respecto a la expresión del cistrón río arriba (cistrón 5'), se puede considerar como una evidencia para el mecanismo de traducción interna (Baranick *et al.,* 2008). Calculando la proporción de producto de cistrón 3' respecto al producto del cistrón 5', se puede determinar la actividad de IRES de la secuencia insertada (Figura 8).

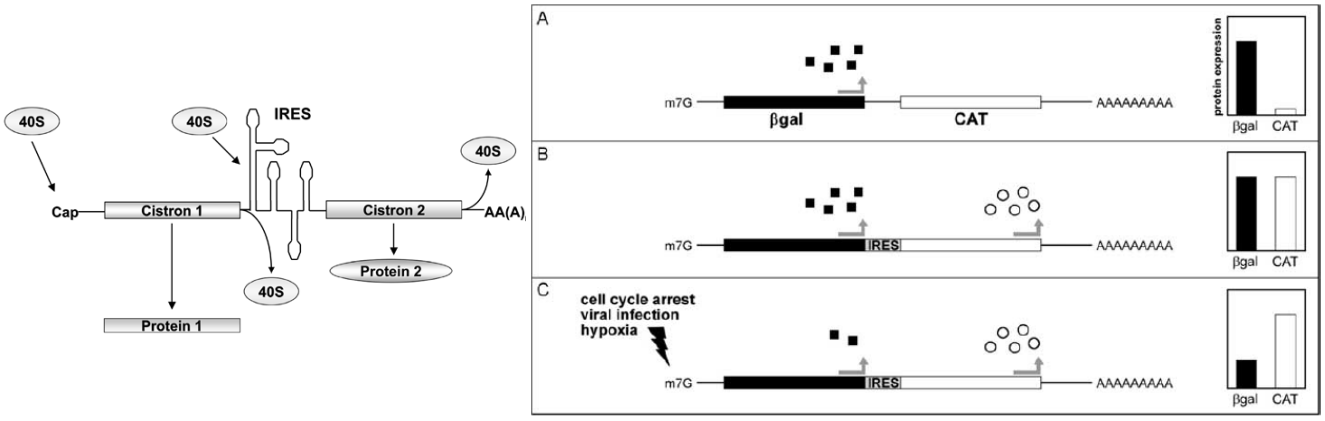


Figura 8. Construcción bicistrónica para caracterizar experimentalmente la actividad del IRES. A) Bajo condiciones fisiológicas normales, el inicio dependiente de cap media la traducción de Beta-Galactosidasa, mientras que la traducción del cistrón río abajo, cloranfenicol acetil trasnferasa (CAT), no ocurre debido a la incapacidad del ribosoma de reiniciar. B) Un transcrito que tiene una secuencia IRES intercistrónica puede mediar la traducción independiente de cap de CAT y la traducción dependiente de cap de Beta- Galactosidasa, lo que incrementa la proporción CAT/Beta-Galactosidasa, si se compara con el plásmido reportero que no contiene un IRES intercistrónico. C) El estrés fisiológico causa una reducción marcada en la traducción dependiente de cap que media la traducción de Beta- Galactosidasa, mientras que la traducción independiente de cap vía IRES es inducida, por lo que aumenta la actividad del cistrón CAT. Tomado de Graber *et al.,* 2006.

Sin embargo el método puede encontrar falsos positivos con una supuesta actividad de IRES debido a la presencia de promotores internos y/o sitios de empalme en el primer gen del arreglo bicistrónico, que afectan aumentando la expresión del siguiente cistrón (Baird *et al.,* 2006).

Dentro de lo que se ha descrito para la identificación masiva de IRES en diversos genomas eucariontes, únicamente se encuentra el trabajo de Graber et al., (2006), en donde se propone abordar el problema desde una perspectiva funcional, dada la poca generalización de IRES que ha resultado de tomar como característica universal la conservación de estructuras altamente estables en el RNA.

Para ello se emplea un microarreglo del estado de traducción (Translate State Array Analysis), el cual se basa en el hecho de que el número de ribosomas asociados a un mRNA particular es proporcional a la eficiencia de traducción, y que el inicio de la traducción es el paso limitante para este proceso. Así, los mRNAs que se traducen activamente tienen múltiples ribosomas asociados a ellos (polisomas) y los mRNAs poco activos tienen un sólo ribosoma asociado (monosoma) o ninguno. Posteriormente se puede centrifugar por gradiente a estos ribosomas y aislar el mRNA de las fracciones polisómicas para realizar un microarreglo, el cual se lleva acabo en distintos contextos celulares.

En un microarreglo con una condición celular particular en donde se ha observado que el inicio de la traducción dependiente de cap está restringido, se esperaría el aislamiento y detección de aquellos mRNA con 5'UTRs que contengan IRES.

Justificación

La detección bioinformática de nuevos IRES celulares no ha sido una tarea sencilla, prueba de ello es la escasa cantidad de IRES descritos en un número limitado de organismos eucariontes. Esta limitada descripción de IRES celulares es más notoria si se considera que actualmente existe una gran cantidad de genomas eucariontes secuenciados para los cuales no existe reportado elementos IRES.

Asimismo, no se ha descrito ninguna metodología bioinformática en la identificación de IRES celulares que considere la existencia de nuevos IRES a partir de las relaciones evolutivas de ortología y paralogía entre distintos genomas eucariontes. Cabe notar que, dentro de lo que nosotros conocemos, solamente existe una propuesta experimental basada en microarreglos para la búsqueda exhaustiva de IRES celulares (descrita anteriormente), aunado a que su estudio ha estado limitado a describirlos individualmente dentro de un genoma, predominantemente el genoma humano.

Por lo anteriormente expuesto, se hace patente la necesidad de implementar nuevas estrategias bioinformáticas, a partir de enfoques evolutivos que permitan identificar de manera global IRES celulares, que no se restrinjan únicamente a la identificación de motivos de secuencias conservadas o de consensos estructurales de su plegamiento, ya que el criterio de energía libre mínima para la detección de IRES en genes particulares tiene una limitada capacidad predictiva.

La genómica comparativa de potenciales IRES en genes ortólogos y parálogos puede aumentar considerablemente dicha capacidad predictiva. Este nuevo enfoque conceptual nos puede permitir identificar, de manera global, nuevos IRES celulares.

Más aún, dada la importancia del inicio de la traducción cap independiente en la sobrevivencia celular, con la identificación masiva de nuevos mensajeros eucariontes que presenten IRES podría ayudar a determinar nuevos agentes terapéuticos.

Hipótesis

El desarrollo de la tesis presente está basado en tres hipótesis centrales:

1. Existe una tendencia a que el mecanismo de inicio de la traducción independiente de cap mediante IRES esté conservado entre los diferentes genes ortólogos de diversos organismos debido a que los contextos celulares (estrés por calor, hipoxia, proliferación celular, entre otras) de dichos genes comúnmente son similares.

2. La tendencia de conservación de IRES celulares en genes ortólogos será mayor en tanto que la distancia filogenética de los organismos que las contengan sea pequeña, y menor si la distancia filogenética de dichos organismos sea grande.

3. La capacidad de identificar IRES celulares en un organismo aumenta al considerar la energía libre de las estructuras secundarias presentes en la región 5'UTR de los genes ortólogos de organismos filogenéticamente relacionados.

Objetivo General

- Identificar *in silico* IRES celulares en genomas eucariontes secuenciados.

Objetivos Particulares

- Identificar potenciales IRES en genes ortólogos a IRES celulares descritos previamente en humano mediante el uso de genómica comparativa, de manera particular en el grupo filogenético de mamíferos.

- Desarrollar un método estadístico que considere al contexto celular bajo el cual opera un IRES celular en particular, así como la probabilidad de encontrar un IRES celular dadas sus relaciones de ortología con otros organismos.

Metodología y Desarrollo

Esquema general de la metodología

En general nuestra premisa de trabajo establece que la probabilidad de que un gen y sus ortólogos en organismos filogenéticamente cercanos, presenten estructuras secundarias estables de RNA en sus secuencias 5'UTRs, es mayor si al menos para uno de ellos se ha descrito la existencia de un IRES. Con el propósito de tener una visión global de la metodología empleada durante la realización de esta tesis, a continuación se describe el esquema general de la misma. Posterior a esta descripción, se detalla cada uno de los puntos que intervienen en ella.

El primer paso de nuestra metodología consistió en obtener las secuencias 5'UTRs de algunos genomas eucariontes. Cada secuencia total 5´UTR del genoma en cuestión se dividió en ventanas consecutivas de longitud constante; a cada ventana de una secuencia 5´UTR se le calculó la energía libre utilizando el programa RNAfold (Hofacker *et al.,* 1994) y se seleccionó aquella con menor energía libre. Se obtuvo una muestra, de los valores de energía libre más negativos por genoma.

Con el fin de determinar si el valor de energía libre de la 5'UTR en estudio era estadísticamente inferior a la energía libre de la región 5´UTR de genes seleccionados al azar, se realizó la estadística descriptiva de la muestra con el cálculo de la media como medida de tendencia central y de la desviación estándar como medida de dispersión, de los valores de energía libre como variable aleatoria.

Las distribuciones obtenidas se analizaron y compararon con la distribución de energía libre de IRES celulares, en el caso de que estuvieran descritos para el genoma en cuestión, para lo cual se consultó la base de datos IRESite (www.iresite.org) (Mokrejs *et al.,* 2006), que contiene todas las estructuras de los IRES celulares y virales que han sido verificados experimentalmente.

En segunda instancia, se identificaron los genes ortólogos a aquellos genes descritos en humano que son traducidos por IRES celulares, con base a la similitud en secuencia de aminoácidos utilizando el programa BLASTp (Altschult *et al.,* 1990) contra algunos genomas mamíferos. Una vez que se identificaron los genes ortólogos correspondientes, se obtuvieron sus secuencias 5´UTRs y se calcularon los valores de energía libre de la misma manera que se hizo para las 5´UTRs genómicas, es decir por ventanas, y considerando la energía libre más negativa de todos ellos. Se realizó una comparación entre los valores de energía libre de los IRES celulares descritos en humano y los valores de energía libre en las 5´UTRs ortólogas y la media de éste último conjunto, con la finalidad de encontrar si existía conservación filogenética alguna de esta variable.

Como se mencionó anteriormente, algunos IRES celulares tienden a presentar estructuras secundarias de RNA con mayor estabilidad que el promedio de las estructuras secundarias que se forman en las regiones 5´UTRs de genes que no tienen IRES. No obstante, esta tendencia no es suficientemente grande para que se puedan identificar IRES celulares en un genoma de estudio exclusivamente basados en su energía libre de plegamiento. Con el objetivo de intensificar la señal basada en la energía secundaria de IRES celulares, para cada IRES celulares de humano se calculó la probabilidad conjunta de que las 5´UTRs de sus correspondientes genes ortólogos presentarán también estructuras secundarias estables. Con tal motivo se verificó que la distribución de energía libre de las regiones 5´UTRs de los genes de un genoma, correspondieran a la de una distribución normal estándar.

Finalmente, con base a los resultados del cálculo de probabilidad conjunta antes mencionada y a la función celular de los genes identificados, se obtuvo un listado de genes candidatos a ser traducidos mediante un IRES.

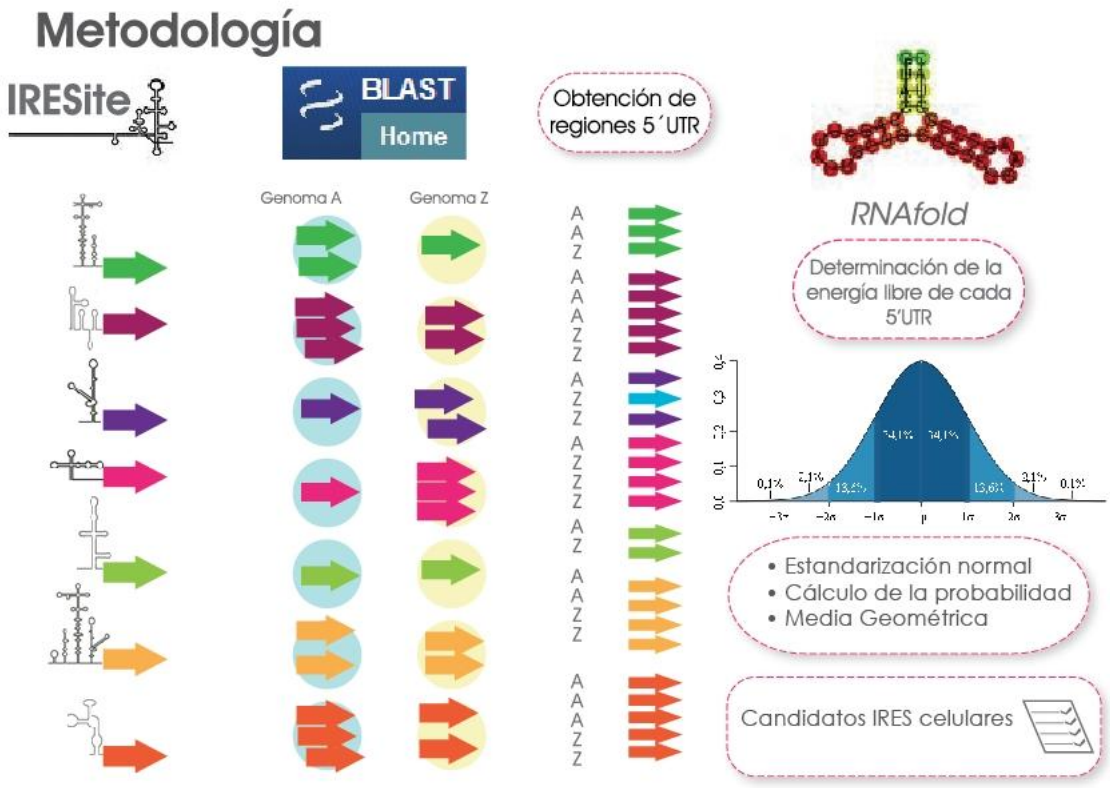


Figura 9. Esquema general del método para la detección de nuevos IRES celulares. Ver texto.

1. Obtención de las secuencias de RNA analizadas

1.1 Obtención de secuencias 5'UTRs

Las secuencias 5'UTRs de 6 genomas eucariontes, *Saccharomyces cerevisiae, Drosophila melanogaster, Schizosaccharomyces pombe, Arabidopsis thaliana* y *Plasmodium falciparum* fueron obtenidos de la base de datos de nuestro grupo (http://www.ibt.unam.mx/biocomputo/base/docu_base.html). Como un control interno, de esta misma base también se obtuvieron las regiones 5'UTRs del genoma *Bacillus subtilis.* Las secuencias 5' UTR de humano, se extrajeron de la base de datos UTRdb (Grillo *et al.,* 2009), que contiene las regiones no traducidas 5' y 3' de una gran número de genomas eucariontes, que se encuentra separada en bases de datos más pequeñas, específicas para un determinado grupo de organismos: mamíferos, invertebrados, plantas, roedores, primates y vertebrados. Aunado a la extracción de las secuencias 5' UTR de humano, se extrajeron las secuencias de otros mamíferos cercanos, de los roedores *Rattus norvegicus* y *Mus musculus;* así como de los primates *Macaca mulatta* y *Pan troglodytes.* (Tabla 4).

Genoma	Secuencias 5' UTR	Regiones codificantes
Bacillus subtilis	4105	4105
Homo sapiens	124, 315 36,055 no redundantes	No se obtuvieron
Mus musculus	28,047 19,323 no redundantes	No se obtuvieron
Rattus norvegicus	18,678 13,747 no redundantes	No se obtuvieron
Macaca mulatta	26,794 17,949 no redundantes	No se obtuvieron
Pan troglodytes	41,902 24,905 no redundantes	No se obtuvieron
Drosophila melanogaster	9,299	17,238
Saccharomyces cerevisiae	4,722	5,875
Arabidopsis thaliana	23,606	29,032
Schizosaccharomyces pombe	5,205	5,044
Plasmodium falciparum	4.768	5,266

Tabla 4. Número de secuencias obtenidas por genoma, 5' UTRs y regiones codificantes, usadas para calcular la distribución de energía libre mínima más negativa.

1.2 Obtención de secuencias codificantes de los organismos de estudio

Las secuencias nucleotídicas correspondientes a las regiones codificantes de todos los organismos de estudio, con excepción de mamíferos, fueron obtenidas de la base local de datos de nuestro grupo (http://www.ibt.unam.mx/biocomputo/base/docu_base.html) (Tabla 4).

1.3 Eliminación de secuencias redundantes originadas por variantes de splicing alternativo

El *splicing alternativo* es un fenómeno en organismos eucariontes a partir del cual se pueden generar distintas variantes de un mismo transcrito. La edición por *splicing alternativo* puede efectuarse tanto en regiones codificantes, como en regiones 5' o 3´ no-codificantes. La secuencia de las regiones 5' UTRs de los genomas mamíferos de nuestro estudio fueron obtenidos de la base de datos UTRdb (Grillo *et al.,* 2009). Cabe mencionar, que para cada una de las variantes de *splicing alternativo* conocidas de un gen (ASPicDB, Alternative Splicing Prediction Data Base, Castrigniano *et al.,* 2008) existe un registro diferente en la base UTRdb (Grillo *et al.,* 2009). Para nuestro estudio, las variantes de regiones 5' UTRs de genomas mamíferos fueron tratadas como entidades independientes entre sí, relevantes para el análisis genómico. Sin embargo, las isoformas de variantes de *splicing alternativo* en regiones codificantes que son iguales en la región 5'UTR, fueron consideradas como redundantes. Si las isoformas redundantes se incluyeran para el análisis de distribución de energía libre, la estadística realizada tendría un sesgo al tener 5' UTRs sobre-representadas. Para eliminar dicha redundancia, se empleó el programa CD-HIT (Cluster Database at High Identity with Tolerance, Li y Godzik, 2006).

CD-HIT toma una base de datos de secuencias en formato fasta y devuelve un conjunto de secuencias representativas no- redundantes. Con base en la secuencia más larga de la base de datos ('longest sequence first') se crea un conjunto de secuencias que remueve a aquellas que se encuentran por debajo de cierto corte de identidad con esta primera secuencia. Adicionalmente en este procedimiento se implementa un método heurístico rápido para encontrar segmentos con alta identidad entre secuencias.

La redundancia de isoformas de 5'UTRs idénticas se eliminó con un valor de corte de 0.9, esto es, que al realizar una comparación pareada, aquellas secuencias que se asemejen en un 90% o un porcentaje mayor, estas serán eliminadas.

2. Cálculo de la energía libre mínima

2.1 Cálculo de la energía libre mínima de estructuras de RNA en las regiones 5'UTR y las regiones codificantes

Los valores de energía libre de secuencias de RNA fueron evaluadas programa RNAfold Viena RNA package (Hofacker *et al.,* 1994). Dicho programa predice cuales son las estructuras secundarias de una molécula de RNA con mayor probabilidad de ser formadas *in vivo* o *in vitro,* con base a sus correspondientes estabilidades termodinámicas.

Debido a que: a) No toda la secuencia 5´UTR es parte del IRES; b) No todas las secuencias 5´UTR son del mismo tamaño y c) No toda la región del 5´UTR se pliega en una sola estructura, el análisis de energía mínima de las estructuras fue efectuada por ventanas discretas que pudieran, en la mayoría de los casos, contener a IRES celulares o a los elementos más relevantes de los mismos. Para ello, las secuencias de estudio fueron divididas en ventanas consecutivas de longitud constante. En vista de que el tamaño de los IRES más pequeños reportados es aproximadamente 50 nucleótidos y que las de mayor tamaño suelen contener varios módulos de estructuras secundarias cuyos tamaños promedio son aproximadamente de 150 nucleótidos, los tamaños de ventana utilizados en nuestro estudio, estuvieron un tamaño dentro del rango 50 a 150 nucleótidos.

Se realizó un programa en el lenguaje de programación Perl que específicamente tomara cada una de las secuencias 5' UTR o regiones codificantes de cada genoma, y las fragmenta en ventanas consecutivas de tamaño predeterminado. Las secuencias 5' UTR son leídas desde el nucleótido +1 (es decir el inicio de la traducción) hasta río arriba, hacia el extremo final 5' en las distintas ventanas que se generan. Denominamos "n" a la variable que define el tamaño de las secuencias 5' UTR, "m" a la posición de la ventana respecto al inicio de la traducción y la variable "v" representa el tamaño de ventana. El análisis para cada secuencia 5' UTR involucra n-v ciclos, en donde la variable "m" tiene un valor inicial de cero y se incrementa en una unidad cada ciclo de análisis hasta llegar a un valor de n-v. El valor de energía libre mínima es calculado para cada n-v ventanas (Figura 10).

En nuestro análisis las ventanas con valores de energía libre más negativas corresponden a as regiones de las secuencia 5' UTR, que pueden plegarse en las estructuras secundarias de RNA más estables. Partiendo del supuesto de que los IRES poseen estructuras secundarias con una gran estabilidad, del conjunto de ventanas en las que fue dividida la secuencia 5' UTR, se seleccionó aquella que tuviera el valor menor de energía libre. Este valor de energía fue asociado al gen localizado inmediatamente río abajo de la secuencia 5' UTR.

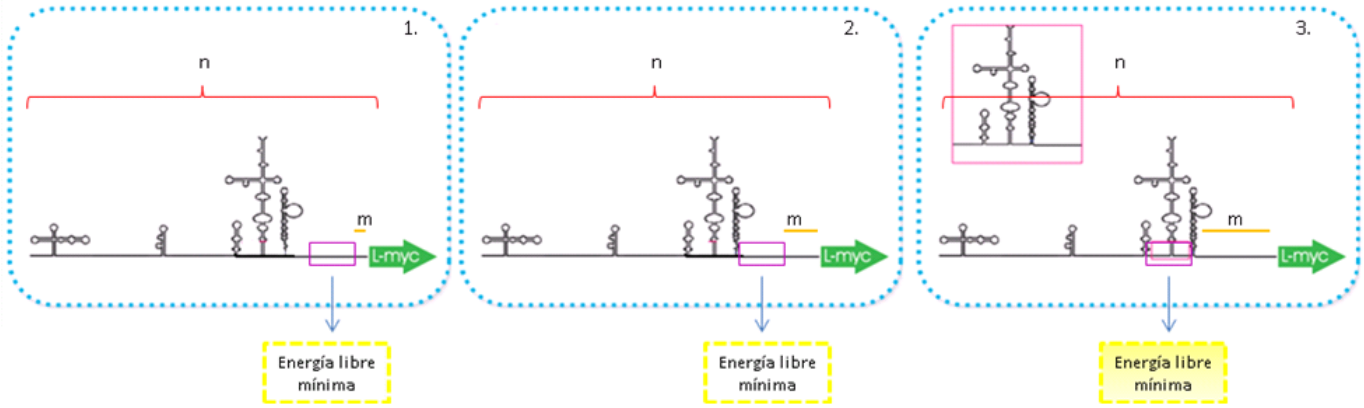


Figura 10. Cálculo de la energía libre mínima. Se obtiene el conjunto de valores de energía libre mínima de una secuencia 5'UTR y se selecciona aquella que sea más negativa. El cuadro lila representa la ventana de análisis. Para cada uno de los cuadros del dibujo, la posición de la ventana de análisis es diferente, siendo la ventana #3 en la que se ha encontrado el valor de energía libre más negativo. El cálculo sigue hasta el extremo 5' de la secuencia que considera n-v, al finalizar el script selecciona la energía libre mínima más negativa, en este caso corresponde a la ventana #3.

2.2 Cálculo de energía libre en el genoma procarionte control: Bacillus subtilis

La atenuación transcripcional o traduccional, es un mecanismo de regulación de la expresión génica en organismos procariontes. Dicha regulación se lleva a cabo mediante estructuras secundarias estables de RNA que se forman en la región 5' no-traducida de los genes regulados (Merino y Yanofsky, 2005). Dado lo anterior, consideramos que el análisis de energía libre del genoma procarionte de *B. subtilis* pudiera constituir un primer control de nuestro protocolo computacional para la identificación de estructuras secundarias estables. La energía libre de las regiones 5' no-traducidas y regiones codificantes se determinó tal y como se describió en la sección anterior, 2.1.

Variables de estudio obtenidas

Considerando todas las variables de nuestro análisis de energía de plegamiento, se obtuvieron los resultados para 10 genomas eucariontes (*Homo sapiens*, *Macaca mulatta, Pan troglodytes, Mus musculus, Rattus norvegicus, Plasmodium falciparum, Drosophila melanogaster, Saccharomyces cerevsiae, Schizosaccharomyces pombe, Arabidopsis thaliana*) un genoma procarionte (*Bacillus subtilis*), de las regiones 5' UTRs, regiones codificantes, en cuatro ventanas de análisis (50, 75, 100 y 125 nucleótidos).

3. Caracterización estadística de la energía libre mínima del RNA

3.1 Caracterización estadística de la energía libre en 5' UTRs y regiones codificantes de los genomas en estudio

Los resultados de la muestra de los valores de energía libre mínima calculados por genoma de las secuencias 5' UTRs y secuencias de regiones codificantes, fueron analizados y representados mediante la estadística descriptiva. El script en lenguaje Perl descrito en la sección anterior también calculó como medida de tendencia central la media y como medida dispersión la desviación estándar muestral con la siguiente fórmula:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \qquad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2},$$

En donde "x" representa cada elemento de la muestra, es decir el valor de energía libre mínima para cada secuencia y "n" es el número de elementos de la muestra, el número total de secuencias por genoma.

La distribución de energía libre para cada conjunto de 5' UTRs o regiones codificantes por genoma para cada ventana de análisis (50, 75, 100 y 125 nucleótidos), se representó gráficamente en histogramas de frecuencias relativas, con una longitud de intervalo de clase constante igual a 4, con una amplitud de 100 y 25 clases. Los valores de energía libre mínima representados en dichas histogramas abarcan valores desde 0 kcal/mol hasta valores muy negativos, aunque por razones prácticas no se representan dichos valores en los ejes de estos histogramas con el símbolo menos (-).

3.2 Caracterización estadística de la energía libre en IRES celulares descritos y predichos

La base de datos IRESite (www.iresite.org) cuenta en su mayoría con IRES celulares descritos en humanos y se encuentra curada a través de la información experimental con la que se cuenta sobre cada uno de ellos; asimismo posee una gran cantidad de información acerca de los IRES: la naturaleza del elemento IRES, origen, tamaño, secuencia, estructura, posición relativa con respecto a las regiones codificantes, los genes reporteros usados en el monitoreo de la actividad del IRES, entre otros aspectos. El sitio muestra las similitudes entre los IRES y secuencias de rRNA, así como la interacción RNA-proteína (Mokrejs *et al.,* 2006). De dicha base de datos se obtuvieron los IRES celulares reportados para *H.sapiens*, en *D. melanogaster* y *S.cerevisiae* (Tabla 5). De la

misma forma que se calculó y representó la distribución de energía libre para las 5' UTRs genómicas, se realizó para el conjunto de IRES celulares en humano. Los IRES celulares que se encuentran caracterizados en esta base de datos para los otros genomas analizados, son unos cuantos, específicamente en *D. melanogaster* y *S. cerevisiae,* de tal modo que para ellos no se construyó una distribución, únicamente se ubicó el valor de energía libre en la clase a la que correspondían en los histogramas de distribución de energía libre para las 5'UTRs.

Por otro lado, la base de datos UTRdb (Grillo *et al.,* 2009) usada para obtener las secuencias 5' UTR en humano, incluye la predicción de algunos motivos estructurales conocidos que se encuentran involucrados en la regulación, entre ellos IRES celulares, la predicción se basa en la implementación de la búsqueda de un patrón estructural descrito por Le y Maizel, 1997.

Tanto los valores de energía libre mínima de las secuencias 5 UTR's de humano predichas por UTRdb (Grillo *et al.,* 2009), como los valores de energía libre mínima de los IRES de humano, fueron usados para construir los histogramas de distribución de energía libre.

4. Conservación de energía libre entre los IRES celulares de humano y sus ortólogos

4.1 Distribución de tamaño de IRES celulares y de 5'UTRs genómicas en Homo sapiens

Los valores de tamaño para cada IRES celular de humano se obtuvieron de la base de datos IRESite (www.iresite.org) (Tabla 5). Asimismo, el tamaño de las secuencias 5'UTRs del genoma humano, así como el contenido de GC se obtuvo con un script hecho en Perl, el cual calculó la distribución de tamaños y la media muestral de dicho conjunto.

4.2 Coincidencia de la ventana de nucleótidos de menor energía libre mínima de las regiones 5'UTRs y elementos IRES descritos

Uno de los supuestos fundamentales de nuestro protocolo de análisis en la búsqueda de nuevos IRES celulares, es que el IRES en cuestión, o parte de éste, puede ser identificado al encontrar la ventana de análisis dentro de la región de su 5'UTR que posea el valor de energía libre más negativo. Con el fin de verificar esta hipótesis, la secuencia nucleotídica de cada IRES de humano reportado se comparó con las secuencias de las ventanas de menor energía libre comprendidas dentro de sus correspondientes regiones 5'UTRs. La comparación de las secuencias se llevó a cabo usando el programa BLASTn (Altschult *et al.,* 1990).

Homo sapiens	
BCL2	Human B-cell leukemia/lymphoma 2 (bcl-2) proto-oncogene mRNA encoding bcl-2-alpha protein
HSPA1A	Homo sapiens heat shock 70kDa protein 1A (HSPA1A)
AQP4	Human mercurial-insensitive water channel mRNA, form 1
CCND1	Homo sapiens cyclin D1 (CCND1)
INS-IGF2	Homo sapiens insulin-like growth factor 2
FGF1	Homo sapiens fibroblast growth factor 1 (acidic) (FGF1)
MNT	Homo sapiens MAX binding protein (MNT)
XIAP	Homo sapiens X-linked inhibitor of apoptosis (XIAP)
LAMB1	Homo sapiens laminin, beta 1 (LamB1)
INR	Homo sapiens insulin receptor
MYCN	Homo sapiens v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian) (MYCN)
RUNX1	Homo sapiens runt-related transcription factor 1 (acute myeloid leukemia 1; aml1 oncogene) (RUNX1)
CSDE1	Homo sapiens cold shock domain containing E1, RNA-binding (CSDE1)
AGTR1	Homo sapiens angiotensin II receptor, type 1 (AGTR1)
CDC2L1	Homo sapiens coiled-coil domain containing 21 (CCDC21) (p58/PITSLRE)
UNR	Homo sapiens cold shock domain containing E1, RNA-binding (CSDE1)
MYB	Homo sapiens v-myb myeloblastosis viral oncogene homolog (avian) (MYB)
BIRC2	Homo sapiens baculoviral IAP repeat-containing 2 (BIRC2)
eIF4G1	Homo sapiens eukaryotic translation initiation factor 4 gamma
Apaf-1	Homo sapiens apoptotic protease activating factor 1 (Apaf-1) mRNA
RUNX1T1	Homo sapiens runt-related transcription factor 1 (cyclin D-related) (RUNX1T1)
FMR1	Homo sapiens fragile X mental retardation 1 (FMR1)
LEF1	Homo sapiens lymphoid enhancer-binding factor 1 (LEF1)
MTG8a	Homo sapiens runt-related transcription factor 1; translocated to, 1 (cyclin D-related) (RUNX1T1)
SRC	v-src sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog (avian)
NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)
Drosophila melanogaster	
Hsp83	Heat shock protein 83
rpr	reaper cell death protein
Saccharomyces cerevisiae	
TIF4631	eIF4G protein
YAP1	Basic leucine zipper (bZIP) transcription factor required for oxidative stress tolerance
SPT15/ TBP	TATA-box binding protein (TBP) component of TFIID and TFIIIB

Tabla 5. IRES celulares con los que se trabajó en este proyecto obtenidos a través en IRESite (ver texto).

4.3 Obtención de secuencias de aminoácidos ortólogas a las secuencias de aminoácidos del genoma humano

Las secuencias de aminoácidos de las regiones codificantes de los genomas de *Homo sapiens* (Humano), *Mus musculus* (Ratón), *Rattus novergicus* (Rata), *Macaca mulatta* (Macaco), *Pan troglodytes* (Chimpancé), *Equus caballus* (Caballo) y *Canis lupus familiaris* (Perro) se extrajeron de la base de datos de nuestro grupo (http://www.ibt.unam.mx/biocomputo/base/docu_base.html). Actualmente existen diferentes aproximaciones metodológicas para identificar proteínas ortólogas entre dos organismos, entre las cuales se encuentran la identificación de los mejores-hits-bidireccionales (Tatusov *et al.,* 1997) o bien, estudios con árboles filogenéticos (Fang *et al.,* 2010). Considerando que para la mayoría de organismos de nuestro estudio no han sido secuenciados en su totalidad, consideramos que el ortólogo más probable de un gene humano en un organismo particular, es aquel que obtiene el valor de expectancia más significativo dentro de una comparación de secuencias hecha con el programa BLASTp del gen humano en cuestión, contra el proteoma del organismo en estudio. En adición a lo anterior, nuestro criterio para definir al ortólogo más probable, el valor de expectancia (valor de e) se exigió sea menor a 1e-5.

Aunado a la búsqueda de ortólogos por BLASTp, ésta se corroboró con la base de datos KEGG GENES Database Entry, que contiene el catálogo completo KO (KEGG Orthology, Kanehisa *et al.,* 2004). Adicionalmente, una vez que se tuvieron las secuencias ortólogas se realizó un alineamiento global con ClustalW (Chenna *et al.,* 2003) de éstas, para establecer a ortólogos bona fide.

4.4 Obtención de regiones 5'UTRs ortólogas

A través de la relación de ortología establecida entre proteínas obtenida en la sección anterior, se identificaron las secuencias 5' UTRs ortólogas a las secuencias 5' UTRs de humano -extraídas en la sección 1.1-, y se les asignó el valor de energía libre mínima calculado anteriormente para cada una de ellas en la sección 3.1, en ventanas de 50, 75, 100 y 125 nucleótidos.

4.5 Análisis de la energía libre de las 5'UTRs ortólogas y de los IRES descritos en H. sapiens

Se tomaron los valores de energía libre mínima existentes para los IRES celulares de humano calculados en los distintos tamaños de ventana en la sección 4.2. Asimismo, fueron tomados al azar los valores de energía libre mínima de 20 regiones 5' UTRs de genes de humano, estos constituyen el grupo control.

Los valores de energía libre de las regiones 5'UTRs ortólogas a los IRES de humano, así como los valores de energía libre para los genes control se ubicaron dentro de la distribución de energía libre de 5'UTRs del

genoma humano. Aunado a lo anterior, se calculó la media muestral de energía libre para los valores de las 5'UTRs ortólogas y también se ubicó en la distribución del genoma humano.

5. Predicción de nuevos IRES celulares

5.1 Estandarización normal de la energía libre mínima de las secuencias 5'UTRs genómicas

A partir de cualquier variable aleatoria que se distribuya de manera similar a una distribución normal con los parámetros de media y desviación estándar, ésta última diferente de cero, se puede llevar a una variable aleatoria normal estándar haciendo la siguiente transformación:

$$Z = \frac{X - \mu}{\sigma}$$

A este proceso se le llama estandarización. En donde "X" es el valor a estandarizar, μ es la media muestral y σ la desviación estándar.

Para poder estandarizar normalmente, primero se corroboró que las distribuciones de energía libre mínima de las regiones 5'UTRs de los genomas de *Mus musculus* (Ratón), *Rattus novergicus* (Rata), *Macaca mulatta* (Macaco), *Pan troglodytes* (Chimpancé), *Equus caballus* (Caballo) y *Canis lupus familiaris* (Perro) se distribuyeran de manera similar a una normal, a través de una comparación entre los cuantiles de la muestra (5´UTRs genómicas) y los cuantiles de la distribución normal teórica, mediante la función qqnorm de R, el cual es un lenguaje y entorno de programación para análisis estadístico y gráfico.

Se estandarizaron las distribuciones de energía libre mínima de las regiones 5'UTRs a través de un script en lenguaje de programación Perl que calculó el valor de Z para cada valor de energía libre mínimo.

5.2 Cálculo de probabilidades normales con base a la distribución de los valores de energía mínima de secuencias 5´UTRs

Las distribuciones normales estándar de los valores de energía libre mínima de secuencias 5´UTRs para cada genoma son importantes porque a partir de ellas se pueden calcular la probabilidad (área bajo la curva) de que alguna secuencia 5´UTR tenga un valor de energía estable y con base en ello proponer que exista un potencial IRES celular. Aún más, a partir de las distribuciones de energía libre mínima de varios genomas, se puede calcular la probabilidad de encontrar un potencial IRES celular dada la probabilidad conjunta de

encontrar estructuras estables de RNA en las secuencias 5´UTRs del gen humano y en las secuencias 5´UTRs de sus correspondientes genes ortólogos, lo cual aumenta considerablemente la predictibilidad en la búsqueda de nuevos IRES celulares.

La probabilidad normal se calculó en base al valor de Z de cada región 5´UTR considerando la distribución de la energía libre mínima de todas las regiones 5'UTRs de sus genomas correspondientes. La probabilidad normal se calculó como el área bajo la curva considerando el intervalo de probabilidad de 0 a 1, con la función *pnorm* del lenguaje de programación R (R Development Core Team, 2009). (Figura 11). Los valores de probabilidad cercanos al cero, representan a aquellas regiones 5'UTRs altamente estructuradas, mientras que aquellas que se encuentran próximas al valore de probabilidad de 1, representan a regiones pobremente estructuradas. El valor de probabilidad de 0.5, corresponde al valor promedio de la distribución de energía libre de las 5'UTRs.

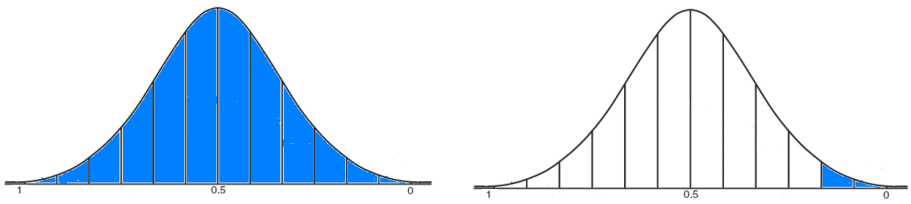


Figura 11. Cálculo de la probabilidad normal. A través de los valores de Z de la distribución de energía libre mínima de las secuencias 5UTR, se calculó el área bajo la curva (azul). El histograma izquierdo corresponde a la de una región 5'UTR pobremente estructurada, por lo que la probabilidad de encontrar una región 5'UTR igual o mayormente estructurada es muy grande (valores de *p* cercanos a 1). De manera contraria, el histograma derecho corresponde a la de una región 5'UTR altamente estructurada, por lo que es poco probable (valores de *p* cercanos a cero) encontrar al azar una región igual o mayormente estructurada.

5.3 Cálculo de la media geométrica de los valores individuales de probabilidad de los genomas de estudio

La hipótesis de nuestro proyecto plantea que la identificación de IRES celulares puede ser más precisa si en lugar de considerar solamente la energía mínima de dicha región, se consideran de manera conjunta las energías mínimas de las regiones 5´UTR de sus correspondientes genes ortólogos. Como se mencionó anteriormente, a partir de un valor de energía mínima se evaluó la probabilidad de encontrar otra región 5´UTR con menor o igual valor de energía mínima dentro de un genoma específico. Con el objeto de obtener una métrica que considere de manera conjunta las probabilidades correspondientes a una región 5´UTR de humano junto con las de sus correspondientes genes ortólogos, se decidió emplear la media geométrica en lugar de la

media aritmética, ya que la media geométrica es más sensible a la presencia de valores pequeños (cercanos a cero) en el conjunto de datos que lo que es el cálculo de la media aritmética. Dichos valores de probabilidad cercanos a cero, representan regiones muy estructuradas y por ende con potencial de ser IRES. Por tanto, definimos a la probabilidad conjunta como la media geométrica de los valores de probabilidad individual. La media geométrica de dichos valores se define de la siguiente manera:

Sea el conjunto $\{a_i\}_{i=1}^{n}$

$$G(a_1, \ldots, a_n) \equiv \left(\prod_{i=1}^{n} a_i \right)^{1/n}.$$

$$G(a_1, a_2) = \sqrt{a_1 a_2}$$
$$G(a_1, a_2, a_3) = (a_1 a_2 a_3)^{1/3},$$

En donde "a_1 …a_n", es cada valor de probabilidad de la secuencia 5'UTR de humano y sus secuencias 5'UTR ortólogas. Así, la media geométrica es la raíz enésima del producto de dichos valores.

Adicionalmente se identificaron los valores de probabilidad del IRES celular de humano y los valores de probabilidad para cada 5´UTR ortóloga al IRES en cuestión. La media geométrica se calculó si y sólo si existían al menos 4 valores de probabilidad correspondientes a los ortólogos y siempre y cuando existiera el valor de probabilidad en humano.

Una vez obtenida la probabilidad normal en humano y la probabilidad conjunta considerando los valores de probabilidades de las regiones 5'UTR de humano y sus ortólogos (media geométrica), se ubicaron dichas probabilidades para los IRES en humano previamente descritos, con el fin de sustentar la factibilidad acerca de las relaciones funcionales y evolutivas de los IRES y sus ortólogos; así como probar que es posible obtener nuevos candidatos a presentar IRES celulares a través de la metodología establecida.

Resultados y Discusión

1. Caracterización estadística de la energía libre de las regiones 5'UTRs de los genomas analizados

1.1 Los valores de la energía libre mínima de las regiones 5'UTRs tienden a tener una distribución normal.

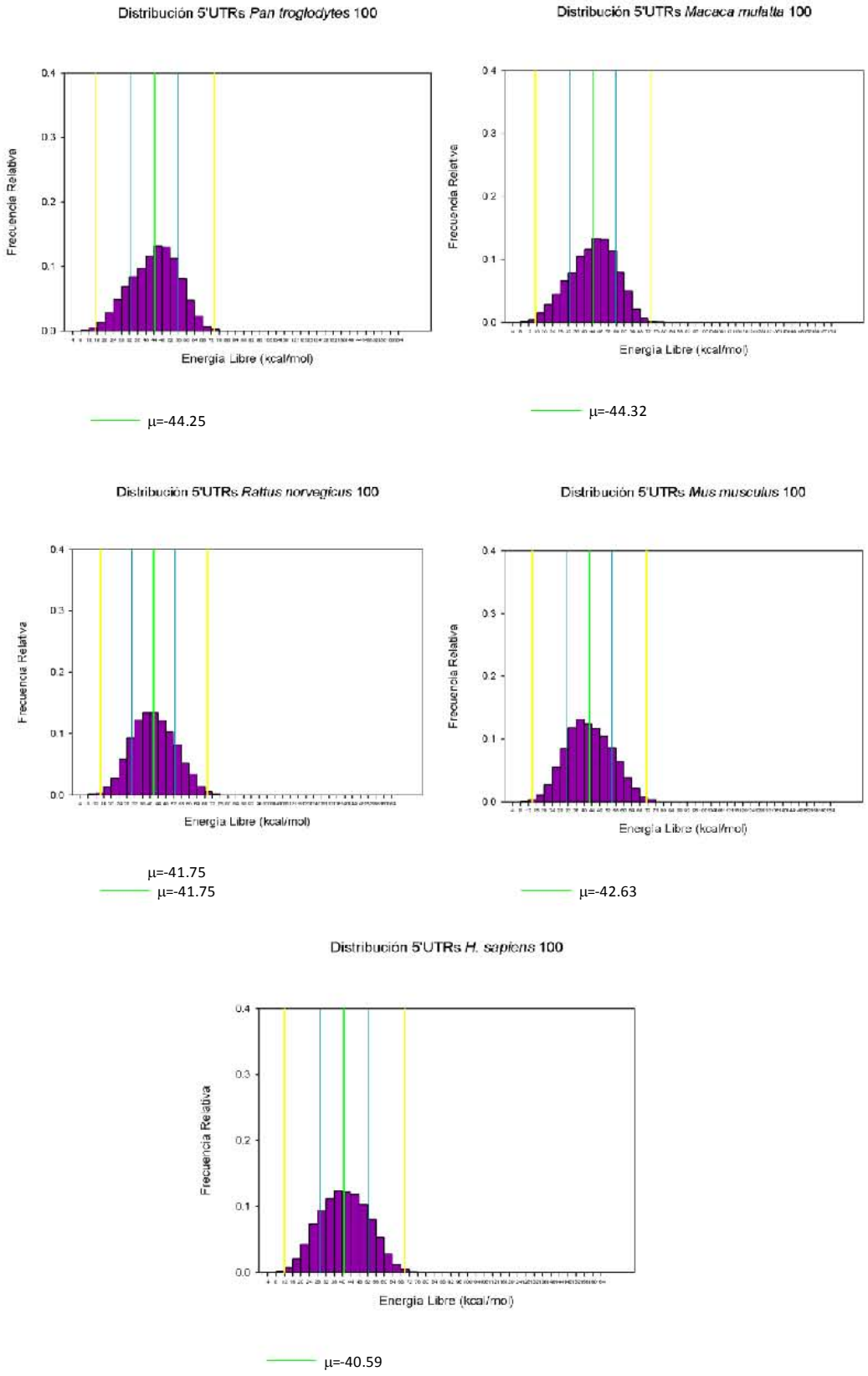
Como se mencionó en la sección de *Metodología y Desarrollo,* el análisis de energía mínima de las regiones 5'UTRs se realizó utilizando ventanas de análisis de longitud constante, siendo la ventana de mínima energía, la representativa de la región 5'UTR. Para todos los genomas analizados, la distribución de energía de estos valores tiende a ser normal (Figura 12). Algunas de las distribuciones se encuentran ligeramente desviadas hacia el extremo de valores de energía libre mínima negativa (por ejemplo en roedores), lo que se denomina asimetría negativa, sin embargo esta asimetría negativa no fue observada en todos los genomas analizados. Las regiones 5' no-traducidas de *Bacillus subtilis*, son la excepción, sin embargo como se mencionó anteriormente, es común que en bacterias las regiones con valores de energía libre más negativa, formen estructuras secundarias, capaces de regular la expresión genética a nivel de transcripción y/o traducción (Merino y Yanofsky, 2005).

1.2 Energía libre mínima de las regiones 5'UTRs vs. Energía libre mínima de las regiones codificantes.

Al comparar la distribución de energía libre en las regiones 5' no-traducidas contra las regiones codificantes de los genomas con los que se contaba para dicho análisis, se puede observar que las regiones codificantes tienen una mayor energía de plegamiento (Figura 13). No es claro aún a que se deban dichas diferencias entre distribuciones. La energía de plegamiento del RNA depende tanto de la composición de nucleótidos, como del orden de estos, por lo que es posible determinar el contenido de GC y la composición de dinucleótidos. En el caso de la distribución de energía libre en *S. cerevisiae*, se sabe que el hecho de que la distribución para regiones codificantes tenga una tendencia a valores más negativos respecto a la de 5' UTRs se debe al orden de los nucleótidos (Ringnér y Krogh, 2005). Sin embargo no ha surgido una explicación y/o hipótesis basada en la funcionalidad biológica de que las regiones codificantes se encuentren mayormente estructuradas.

Desafortunadamente, la falta de secuencias de nucleótidos correspondientes a la región codificante de los distintos genomas mamíferos en este estudio, no nos permitió realizar un estudio comparativo de los valores de energía mínima respecto al obtenido en sus secuencias 5'UTRs. Sería interesante explorar en un futuro ambas distribuciones y dar una explicación biológica en términos de si existe alguna presión de selección que mantenga el comportamiento de las distribuciones que se observen.

Figura 12 ***PANEL A***



PANEL B

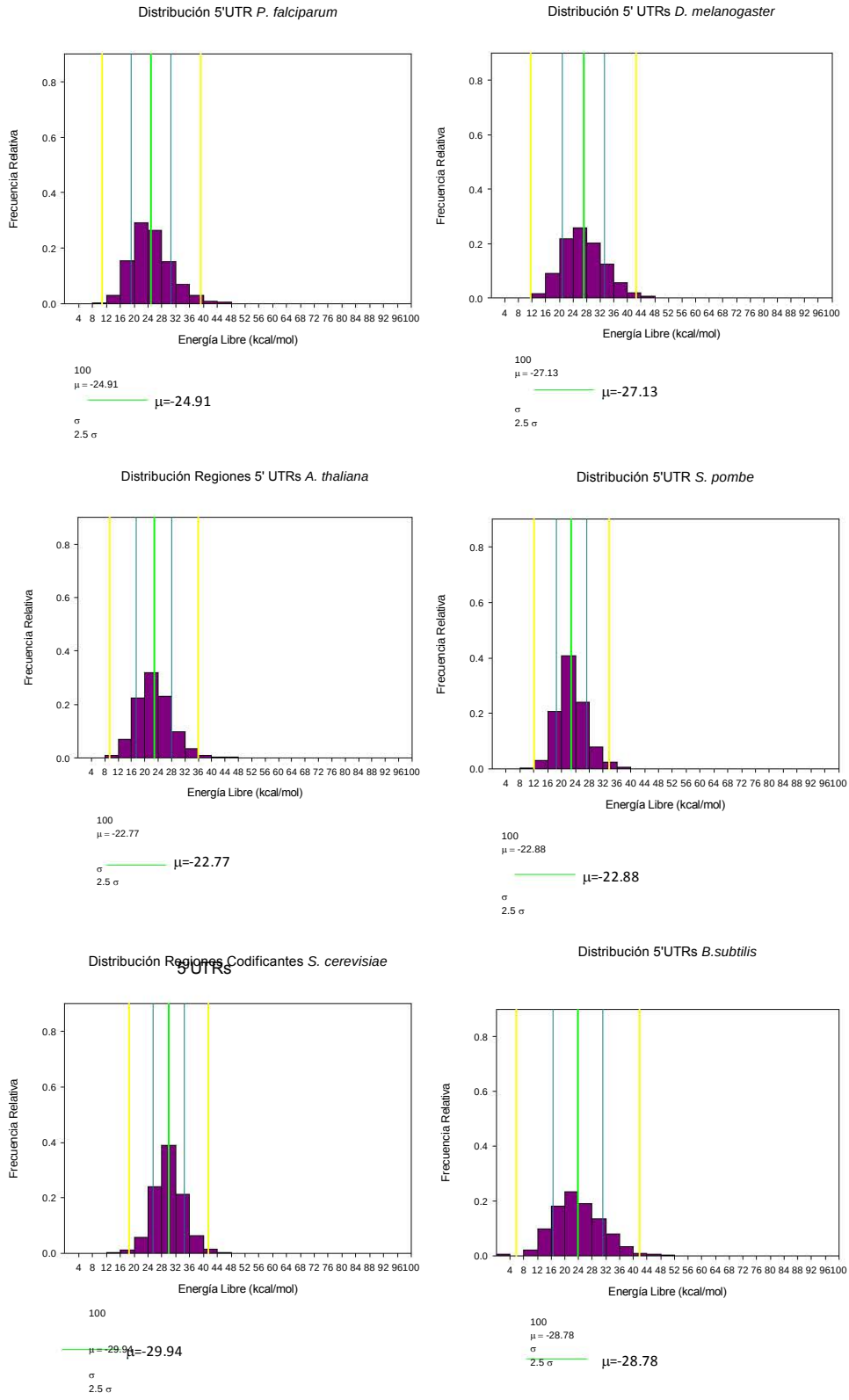


Figura 12. Distribución de la energía libre mínima de las regiones 5´UTR de los genomas en estudio. Panel A. Distribución de energía libre mínima de 5'UTRs de genomas mamíferos. *P.troglodytes, M.mulatta, R.norvegicus, M.musculus* y *H.sapiens.* Panel B. Distribución de energía libre mínima de 5'UTRs de otros genomas eucariontes. *P. falciparum, D.melanogaster, A.thaliana, S.pombe, S.cerevisiae* y del genoma bacteriano de *B.subtilis.* Todos los histogramas corresponden a análisis realizados con ventanas de 100 nucleótidos. Se señalan los valores de la media (línea verde) y la de una (líneas azules) y 2.5 (líneas amarillas) desviaciones estándar.

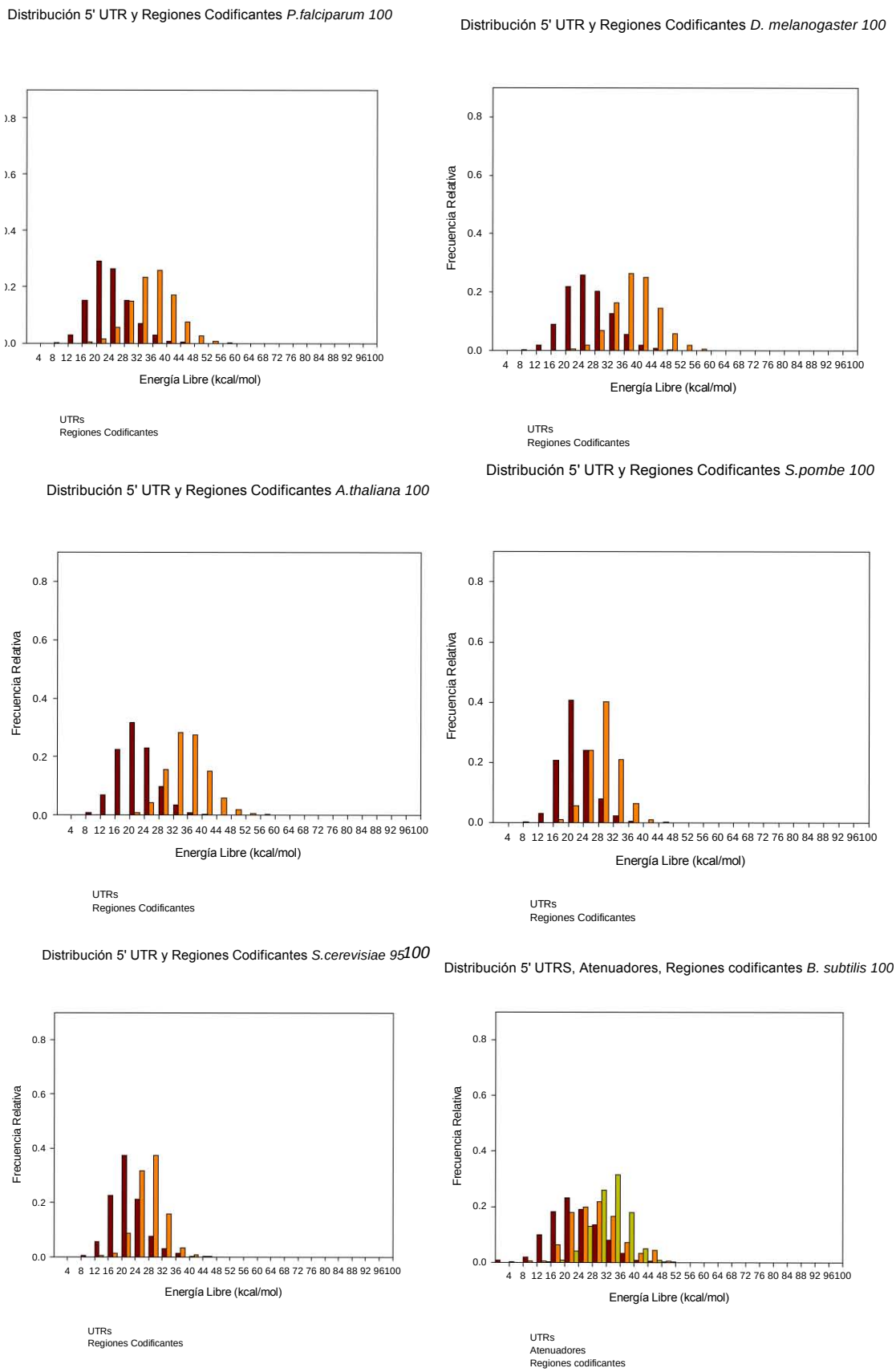


Figura 13. Distribuciones de la energía libre mínima de las regiones 5´UTR (color rojo) y regiones codificantes (color naranja, color verde para *B. subtilis*) de los genomas en estudio. De manera general, las regiones codificantes tienden a estar más estructuradas que las regiones 5´UTR para los genomas en estudio. La bacteria *B. subtilis* presenta la misma tendencia al comparar las regiones codificantes con las regiones 5´UTR, sin embargo al comparar las regiones en donde se presentan atenuadores transcripcionales (color naranja, en este caso) contra las regiones 5´UTR se observa que dichas regiones se encuentran más estructuradas, debido a que dichas estructuras son capaces de regular la expresión genética de algunos genes, bajo condiciones metabólicas particulares.

1.3 Comparación genómica de la distribución de energía libre minima de las regiones 5'UTRs.

Al comparar la distribución de energía libre en las secuencias 5' UTRs en una ventana de 100 nucleótidos para los distintos genomas, se observan claras diferencias entre sus valores más negativos. Sorprendentemente para *H. sapiens, P. troglodytes, M. mulatta, R. norvegicus* y *M. musculus* se reportan valores de estructuras altamente estables, de hasta aproximadamente -100 kcal/mol, mientras que para *S. cerevisiae* los valores de energía libre más bajos que presenta la distribución de energía libre se encuentran cerca de -48 kcal/mol, de manera similar ocurre para los otros genomas no-mamíferos, el valor más negativo de dichos genomas se reporta en *P. falciparum* con un valor de energía libre de -58 kcal/mol. Más aún comparando los valores de la media muestral entre genomas, podemos distinguir claramente que los genomas de mamífero presentan valores de energía libre más negativos (μ ~-40 kcal/mol) (Figura 12, Panel A). Estas observaciones nos permiten separar en dos grupos los distintos genomas, empezando por el grupo de los genomas mamíferos descritos anteriormente, y el grupo de los genomas de *P. falciparum, D. melanogaster*, *S. cerevisiae, S. pombe* y *A.thaliana.*

Ringnér y Krogh (2005) calcularon la energía de plegamiento de las regiones 5' UTR del genoma de *S. cerevisiae*, encontrando una tendencia en las regiones no-traducidas 5' a presentar valores de energía libre altos que favorecen estructuras con plegamientos débiles, aunado a que las estructuras con una energía menor eran más bien casos aislados, en donde los genes de estas 5'UTRs tenían una función molecular desconocida.

Dada esta distribución de energía libre con estructuras poco estables, sería factible que los IRES celulares en levadura presentan valores de energía libre relativamente bajos, lo cual se observa en los histogramas. Al ubicarlos, estos se distribuyen muy cercanos a la media muestral a no más de una desviación estándar de la misma (Figura 14).

Por otro lado, Holcik y Xia (2009) recientemente reportaron que los IRES celulares encontrados en *S. cerevisiae* y *D. melanogaster* exhiben una estructura secundaria pobre, sugiriendo la existencia de un mecanismo compartido de inicio de la traducción independiente de cap, que recae en un segmento de RNA desestructurado. Le y Maizel (1997) encontraron esta propiedad de estructura secundaria poco estable en el IRES del gen Antp en *D. melanogaster*, discutiendo sobre la posibilidad de que este tipo de IRES celulares requieran de otras proteínas, iTAFs para estabilizar la estructura secundaria.

Si bien es cierto lo anterior, se han reportado muy pocos IRES celulares para estos genomas, tres elementos en levaduras y dos elementos en mosca, por lo que tampoco se descarta la posibilidad de que exista otra clase de IRES que puedan estar más estructurados.

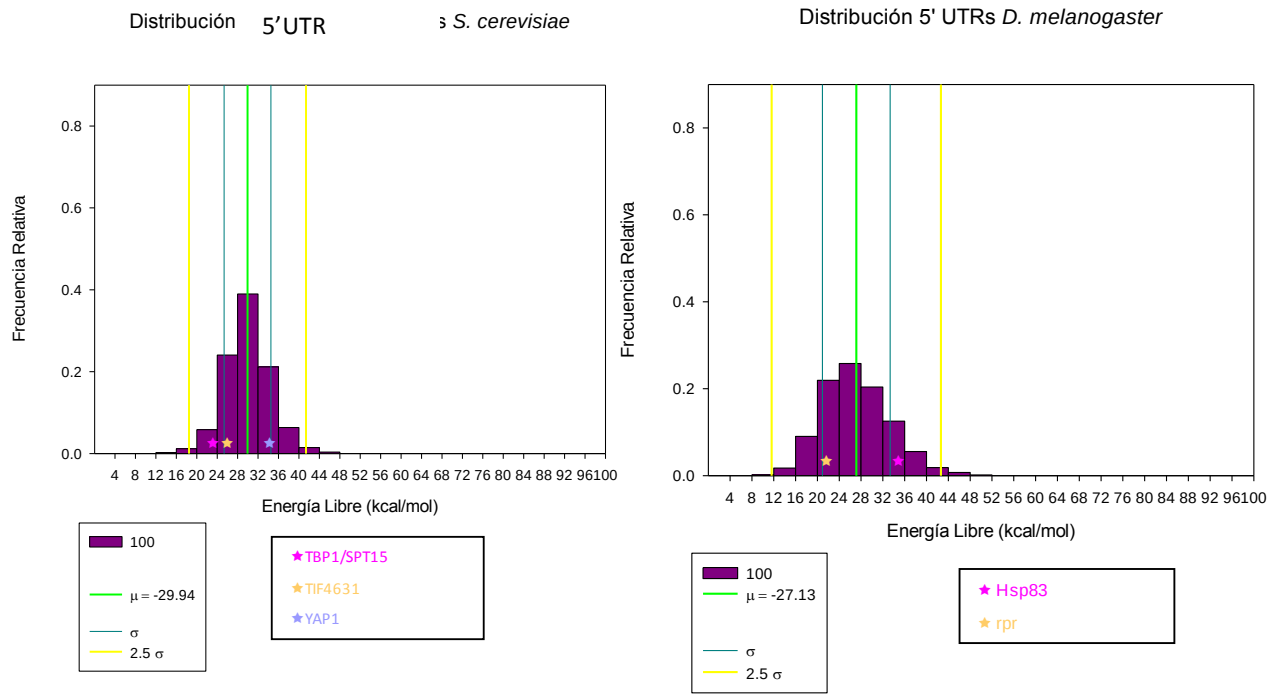


Figura 14. Ubicación de los valores de energía libre mínima de los IRES celulares de *S. cerevisiae* y *D. melanogaster*. Se ha propuesto que los IRES en levadura y mosca son segmentos de RNA desestructurados. En esta figura se ubican los valores de energía libre de los IRES reportados, en la distribución de energía libre de las 5´UTRs. Los histogramas corresponden a análisis realizados con ventanas de 100 nucleótidos. Se señalan los valores de la media (línea verde) y la de una (líneas azules) y 2.5 (líneas amarillas) desviaciones estándar.

Dado lo anterior, en este proyecto se decidió trabajar con genomas de organismos que presentaran valores de energía libre más negativos puesto que se encuentran en su mayoría estructurados, ya que nuestra hipótesis recae en parte en el uso de esta propiedad para identificar nuevos IRES celulares. Así pues, tomamos en cuenta a organismos filogenéticamente cercanos a humano, considerando a algunos otros mamíferos, primates y roedores (*Macaca mulatta, Pan troglodytes, Rattus norvegicus* y *Mus musculus*) ya que son los organismos modelos cuyas secuencias y anotaciones son las más completas, pese a que no estén totalmente terminados.

2. Caracterización estadística de la energía libre de los IRES celulares presentes en humano

2.1 Consecuencias del aumento en el tamaño de ventana y características generales de los IRES celulares en humano.

Una las consecuencias observadas al incrementar el tamaño de las ventanas que se usaron durante el análisis para todos los genomas, es que los valores de energía libre en regiones 5' no-traducidas y codificantes tienen una distribución mejor definida como normal y la varianza de los datos es mayor (Figura 15). Esto puede

ocurrir debido al espacio de secuencia que tienen para plegarse es mayor y en consecuencia la variabilidad de estructuras secundarias puede incluir desde estructuras poco estables hasta estructuras secundarias de mayor tamaño con valores de energía libre muy negativos.

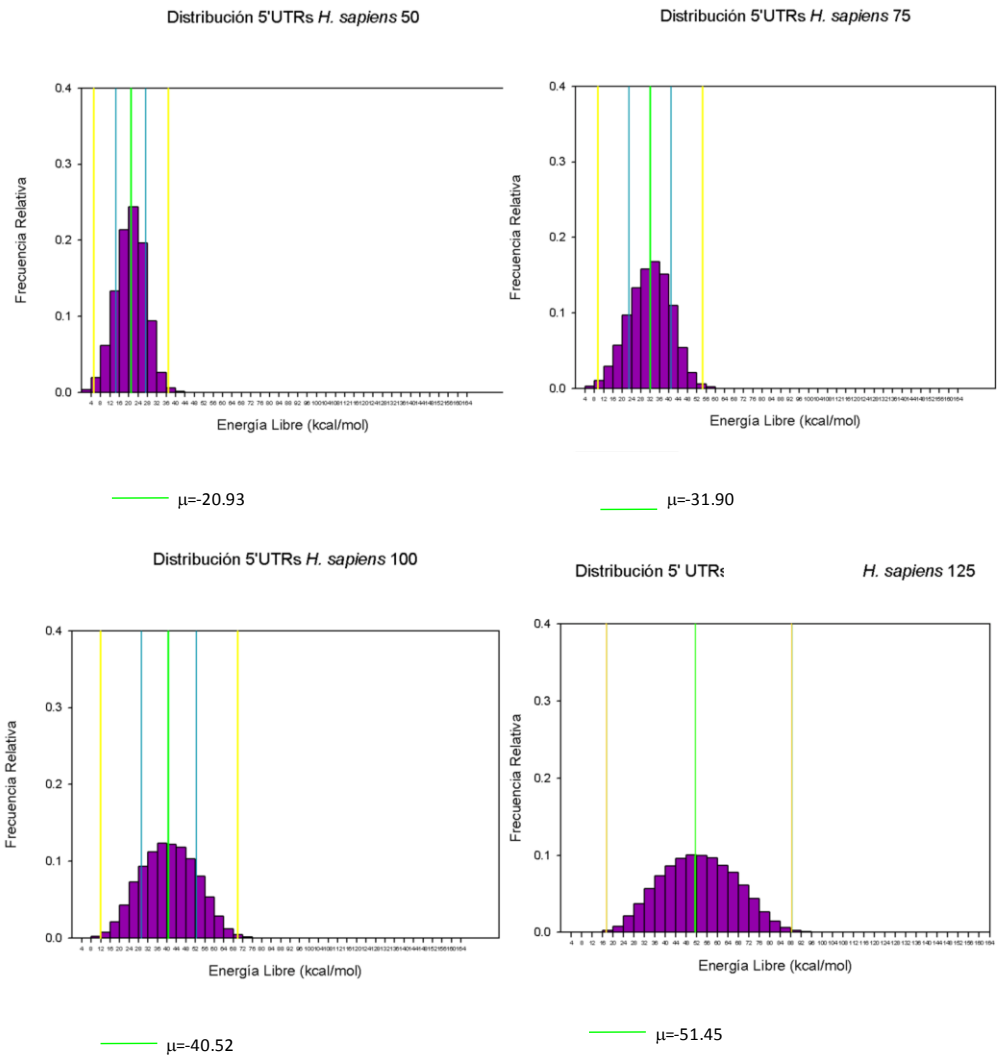


Figura 15. Consecuencias del aumento del tamaño de la ventana de análisis en la distribución de los valores de energía libre mínima. Conforme aumenta el tamaño de ventana la distribución de los valores tienden a tener más claramente una distribución normal y la varianza de los datos es mayor. Se pueden encontrar nuevos valores de energía libre más negativos. Esta figura sólo muestra dicho fenómeno para el genoma humano, sin embargo se presenta de manera similar en los demás genomas eucariontes analizados. Todos los histogramas corresponden a análisis realizados con ventanas de 125 nucleótidos. Se señalan los valores de la media (línea verde) y la de una (líneas azules) y 2.5 (líneas amarillas) desviaciones estándar.

Aunado a lo anterior, al aumentar el tamaño de ventana en nuestro análisis, esperábamos encontrar una distribución normal para las 5'UTRs genómicas, así como un mejor descriptor del tamaño a usar, en la búsqueda de nuevos IRES celulares. El tamaño de la secuencia de IRES descritos en humano comprende desde aproximadamente 50 hasta 1600 nucleótidos, con una media de 391 nucleótidos (Figura 16).

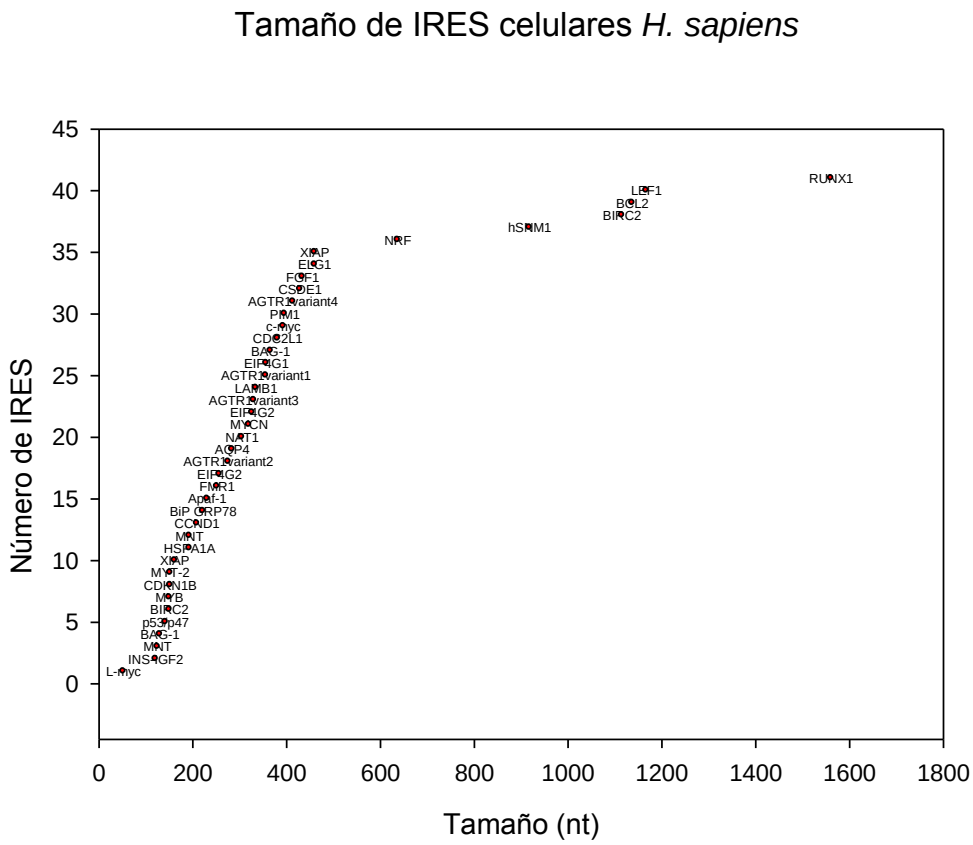


Figura 16. Tamaño de IRES celulares en *H.sapiens.* El tamaño de IRES comprende aprox. desde los 50 hasta los 1560 nucleótidos. La media es de 391 nucleótidos, la mayoría de ellos se encuentra entre los 200 y 400 nucleótidos.

Pese a que los IRES celulares presentan comúnmente grandes tamaños (μ= 391 nt) (Figura 16), en nuestro análisis de IRES en humano y en los otros genomas de mamíferos se decidió utilizar ventanas de 100 y 125 nucleótidos que son menores a este valor promedio del tamaño de IRES, debido a que el RNA es plegado a medida que se sintetiza (Pan y Sosnick, 2006), se van formando estructuras secundarias que conforman a su vez módulos, así ocurre la organización estructural modular presente en los IRES (Chapell *et al.,* 2000). A partir de lo anterior, se tomaron ventanas de longitud pequeña (100-125 nts), ya que consideramos que ventanas de mayor longitud pudieran no representar adecuadamente el fenómeno de plegamiento *in vivo.*

Dado lo anterior, al analizar la energía libre a través de ventanas de distintos tamaños, en algunos casos nos permitirá observar el elemento IRES en su totalidad (aquellos < 200 nucleótidos) y en otros estaremos encontrando módulos o sub-estructuras de distintos tamaños, que conforman al IRES completo. De hecho al llevar a cabo las comparaciones de la secuencia de ventana de cierto tamaño para cada IRES celular en humano, con la secuencia completa del elemento IRES, se identificó que estas ventanas, correspondían a segmentos de los IRES en humano *bona fide.* (Datos no mostrados).

La comparación entre los tamaños de IRES celulares y las regiones 5' no traducidas, nos demuestra que la mitad de las 5'UTRs genómicas son grandes en tamaño, como para presentar un IRES de aproximadamente 300 nucleótidos (Figura 17).

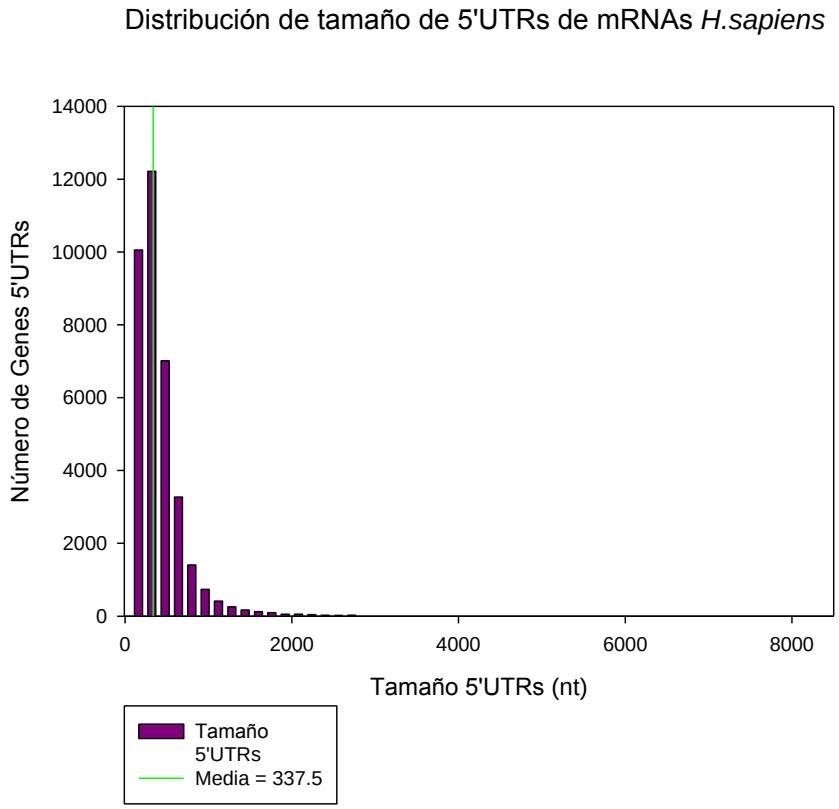


Figura 17. Distribución de tamaño de 5'UTRs de mRNAs en humano. El tamaño de las 5'UTRs comprende desde 10 hasta 1800 nucleótidos aproximadamente. La media de esta distribución es de 338 nucleótidos.

Respecto a la posible correlación entre el contenido de GC y los valores de energía libre, se observa que si bien la estructura secundaria requiere de cierto contenido de GC, no es una determinante para estructurarse, por ejemplo el IRES celular BIRC2/C-IAP1 presenta un bajo contenido de GC y aún así el valor de energía libre mínimo es lo suficientemente alto, como la mayoría de los otros IRES de humano. Otro de los

casos en los que sí se cumple que el IRES tenga un alto contenido de GC y este altamente estructurado es el de FMR1 (Tabla 6). En términos generales, los IRES presentan valores de energía libre mínima en un rango de -40 a -73 kcal/mol y un contenido de GC variable para cada IRES celular en cuestión y no parece existir una correlación entre ambas variables.

IRES	Involucrado en:[a]	Tamaño de 5' UTR (nucleótidos)	Tamaño del IRES (nucleótidos)	%GC	Energía libre [b] (kcal/mol)
RUNX1	-Vías en cáncer -Leucemia mieloide aguda -Leucemia mieloide crónica	7272	1561	50.2	-54.5
Apaf-1	-Vía de señalización p53 -Apoptosis -Enfermedad de Alzheimer -Enfermedad de Parkinson -Esclerosis lateral amiotrófica (ALS) -Enfermedad de Huntington -Tuberculosis -Cáncer de pulmón	5152	583	62.8	-58.2
AQP4	-Reabsorción de agua, regulada por vasopresina -Secreción biliar	1667	284	44	-40.1
AT1R	-Vía de señalización por calcio -Interacción neuroactiva ligando-receptor -Contracción de músculo suave vascular -Sistema renina-angiotensina	2405	414	60.1	-54.1
BCL2	-Procesamiento de proteínas en el retículo endoplásmico -Apoptosis -Adhesión focal -Vía de señalización de la neurotrofina -Sinapsis colinérgica -Esclerosis lateral amiotrófica (ALS) -Toxoplasmosis -Tuberculosis -Vías en cáncer -Cáncer colorrectal -Cáncer de próstata -Cáncer de pulmón	5086	1137	46.8	-40.9
BIRC2/ C-IAP1	-Proteólisis mediada por ubiquitina -Apoptosis -Adhesión focal -Vía de señalización similar a NOD	3742	1115	31.5	-58.4

	-Toxoplasmosis -Infección por HTVL-I -Vías en cáncer -Cáncer de pulmón				
CCND-1	-Ciclo cellular -Vía de señalización p53 -Vía de señalización Wnt -Adhesión focal -Vía de señalización Jak-STAT -Sarampión -Infección por HTVL-I -Vías en cáncer -Cáncer colorrectal -Cáncer de próstata -Cáncer de pulmón -Cáncer de páncreas -Cáncer endometrial -Glioma -Cáncer de tiroides -Melanoma -Cáncer de vejiga -Leucemia mieloide aguda -Leucemia mieloide crónica -Miocarditis viral	4288	209	67.9	-42.7
DAP5	-Metabolismo de cafeína -Metabolismo de drogas -Vías metabólicas	3911	305	57	-40.3
EIF4G1	-Transporte de RNA -Miocarditis viral	5018	357	57.7	-59.6
FGF1	-Vía de señalización MAPK -Regulación del citoesqueleto de actina -Vías en cáncer -Melanoma	1011	434	56.5	-45.3
FRM1	-Retraso mental	4397	252	81	-64.6
Hsp70	-“Spliceosoma” -Vía de señalización MAPK -Procesamiento de proteínas en retículo endoplásmico -Endocitosis -Presentación y procesamiento del antígeno -Enfermedades por priones -Toxoplasmosis -Sarampión -Influenza A	2427	193	63.2	-35.9
IGF2	-Vías metabólicas -Diabetes	1255	121	66.1	-45
LAMB1	-Adhesión focal -Interacción con el receptor ECM	5845	335	68.4	-49

	-Toxoplasmosis -Amibiasis -Vías en cáncer -Cáncer de pulmón				
LEF1	-Vía de señalización Wnt -Uniones adherentes -Melanogenesis -Vías en cáncer -Cáncer colorrectal -Cáncer endometrial -Cáncer de próstata -Cáncer de tiroides -Carcinoma de células basales -Leucemia mieloide aguda -Cardiomiopatía arrítmica ventricular derecha	3594	1167	66.1	-72.5
MNT	-Apoptosis	4841	193	73.1	-47.2
MTG8	-Vías en cancer -Leucemia mieloide aguda	3463	199	62.8	-59.7
MYB	-Infección por HTVL-I	3313	150	68	-59.4
n-MYC	-Apoptosis -Vías en cáncer	2604	320	70.9	-56.1
P27kip1	-Vía de señalización ErbB -Ciclo celular -Sarampión -Vías en cáncer -Cáncer de próstata -Leucemia mieloide crónica -Cáncer de pulmón	2403	152	61.8	-50.8
PIM-1	-Vía de señalización Jak-STAT -Leucemia mieloide aguda	2673	396	75.3	-54.7
PITSLREp58	-Ciclo celular	2471	381	57.5	-52.6
UNR	-Unión a RNA	4115	429	42.2	-51.8
XIAP	-Proteólisis mediada por ubiquitina -Apoptosis -Adhesión focal -Vía de señalización similar al receptor NOD -Toxoplasmosis -Infección por HTVL-I -Vías en cáncer -Cáncer de pulmón	8751	460	26.3	-47.8

Tabla 6. Características generales de los IRES de humano. [a] La vía se reporta según la base de datos KEGG. [b] La energía libre mínima se calculó en una ventana de 100 nucleótidos.

2.2 Distribución de la energía libre mínima de los IRES celulares descritos y los IRES predichos en humano.

La distribución de energía libre mínima de los IRES celulares descritos (Mokrejs *et al.,* 2006) y predichos (Grillo *et al.,* 2009) en humano, parece corresponder a una distribución normal (Figura 18), sin embargo no es posible determinar dicha distribución debido a que el número de IRES celulares para humano que se han descrito, así como las secuencias 5'UTR correspondientes a dichos IRES con las que contábamos, son muy pocos. Es claro que la identificación de nuevos candidatos a IRES celulares nos permitirá hacer comparaciones estadísticas adecuadas entre muestras que se distribuyen de manera similar, en este caso de manera normal. Por ejemplo, se podría calcular la sensibilidad, la especificidad y la precisión de las predicciones que se establezcan para encontrar nuevos IRES celulares.

Por otro lado, si comparamos a la distribución de energía libre de las 5'UTRs de *H. sapiens* y la de los IRES celulares de humano descritos, observamos que la distribución de los IRES se encuentre al menos a una desviación estándar de la media muestral de las 5'UTRs (Figura 18 y Figura 20).

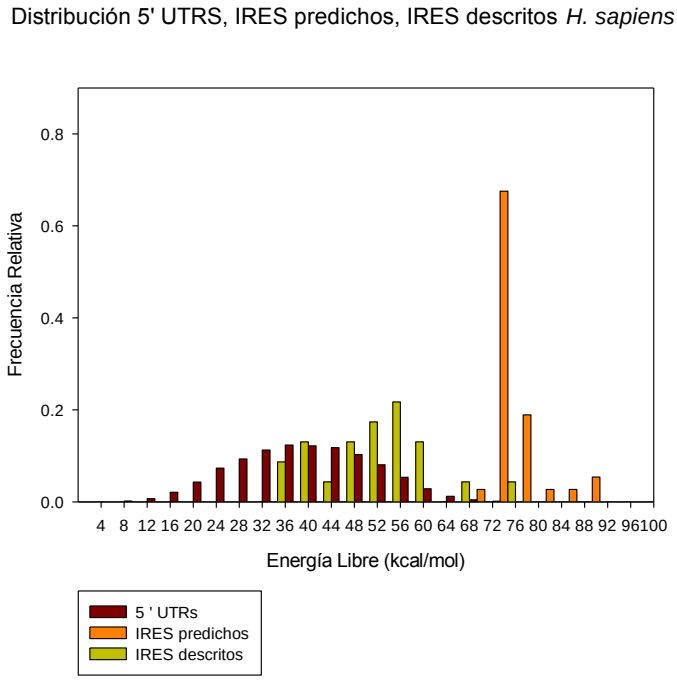


Figura 18. Distribución de energía libre de los IRES celulares descritos y predichos respecto a la distribución de energía libre de las regiones 5'UTR. La energía libre de los IRES descritos y predichos en humano es menor a la energía libre de las 5'UTRs del genoma.

Nuestra caracterización de los valores de energía libre en regiones 5'UTR y en IRES, en términos generales indican que no es posible distinguir con precisión a un IRES celular de humano del resto de las regiones 5'UTRs, pero sí podemos concluir que en promedio, los IRES celulares de humano tienden a estar más estructurados que el promedio de las regiones 5'UTRs que no contienen IRES. Es decir, en promedio los IRES celulares de humano tienden a ser más estructurados que las regiones 5'UTRs que no los contienen, pero este sesgo no es suficientemente grande para que sean identificados estadísticamente.

En la presente tesis, consideramos como hipótesis la posibilidad de magnificar esta señal estadística al considerar que los genes ortólogos a genes de humano regulados por IRES estructurados, tiene una alta posibilidad de también estar regulados por IRES estructurados. Es decir, que la conservación de la regulación traduccional por IRES refleje finalmente la relación evolutiva establecida en términos funcionales, en donde los organismos contienden con condiciones celulares similares bajo el uso de mecanismos moleculares análogos. La idea anterior se discute y prevalece en la sección siguiente, en donde se plantean una serie de pasos para llegar a un método que nos permitiera describir nuevos IRES celulares en otros genomas de mamífero.

En base al criterio de energía de potenciales estructuras secundarias en las regiones 5'UTRs y a la arquitectura de tallo-asa en forma de Y (Figura 19) presente en los IRES BiP y FGF2 (Le y Maizel, 1997), se ha intentado predecir la presencia de IRES celulares en el genoma humano (Grillo *et al.,* 2009). El conjunto de dichos IRES predichos se muestran en la Figura 18 con base a la energía libre de sus correspondientes estructuras secundarias.

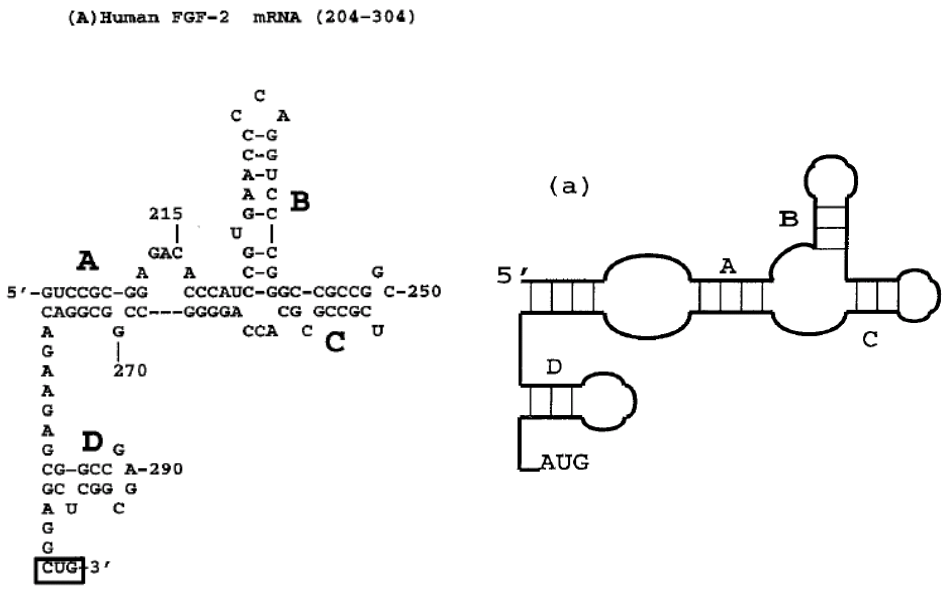


Figura 19. Representación del motivo estructural empleado para predecir IRES celulares. A) Motivo estructural descrito en el IRES FGF2 de humano. a) Representación esquemática del motivo básico, presente en algunos IRES celulares. Tomado de Le y Maizel, 1997.

Como era de esperarse, existe una clara tendencia de estos potenciales IRES a ser ubicados en el extremo de valores energía libre más negativos, más allá de 2.5 desviaciones estándar del valor promedio del conjunto de valores de energía obtenido del universo total de secuencias 5'UTRs. No obstante, cabe mencionar que la predicción de IRES celulares basada en la aparición de dicho motivo, no es suficiente para describir nuevos IRES celulares, ya que un poco más del 20% de las 5'UTRs del genoma humano lo presentan.

3. Conservación de la energía libre mínima en los IRES celulares de humano y las secuencias 5'UTR ortólogas

3.1 La energía libre mínima del IRES Apaf-1 en humano se conserva en genomas ortólogos.

Al obtener la distribución de energía libre mínima para el conjunto de IRES celulares y 5'UTRs de humano, se llevó a cabo un análisis comparativo, para probar la hipótesis y parte de la eficacia del método, al seleccionar un IRES humano, descrito previamente y comprobar para las distintas ventanas, con base en la energía libre de su estructura secundaria, que los IRES ortólogos a éste, tienden a distribuirse de manera similar al mismo en un histograma de frecuencia de los valores de energía de las regiones 5'UTRs de su correspondiente genoma. Cabe mencionar que el gen *Apaf-1* en humano codifica para el factor APAF-1 (pro-Apoptotic Protease-Activating Factor-1), el cual es esencial para la activación de la caspasa-9, que es la caspasa que inicia la apoptosis (Holcick y Sonenberg, 2005). Es importante mencionar que la traducción de *Apaf-1* por un IRES no solamente se ha caracterizado experimentalmente en humano, sino que también se ha descrito para el gen ortólogo en ratón (*Mus musculus*), lo cual enriquece nuestra hipótesis y nos permite llevar a cabo el análisis de distribución de energía libre entre IRES ortólogos. Otros genes ortólogos de APAF-1 se encuentran presentes en *Macaca mulatta, Pan troglodytes* y *Rattus norvegicus.*

Como control de nuestro estudio, se obtuvo la 5'UTR de un gen escogido al azar, cuyo valor de energía fuese cercano a APAF-1 y para el cual no existiera descripción de ser traducido vía IRES. El gen escogido fue CSRNP2, cysteine-serine-rich nuclear protein 2 en humano (298 nt), y sus ortólogos en *Mus musculus* y *Rattus norvegicus.*

Al ubicar los valores de energía libre dentro de la distribución, se puede apreciar que como se esperaba, los ortólogos de APAF-1 se encuentran de manera cercana a APAF-1 de humano para ventanas de 50, 100, 150 y 200 nucleótidos.(Figura 20). Por el contrario, para el gen control CSRNP2, la distribución de sus regiones 5'UTR con base a los valores de energía libre es aleatoria, respecto de la 5'UTR de humano, ya que en las ventanas de análisis de 100 a 200 nucleótidos, ambos ortólogos, el de ratón y rata, se encuentran distantes a la del humano (Figura 20).

En la Tabla 7, se muestran los valores de energía libre de las estructuras más estables en la región 5'UTR del gen APAF-1 de humano y de sus correspondientes ortólogos utilizando diferentes tamaños de ventana de análisis, incluyendo la que abarca a la totalidad de los IRES reportados.

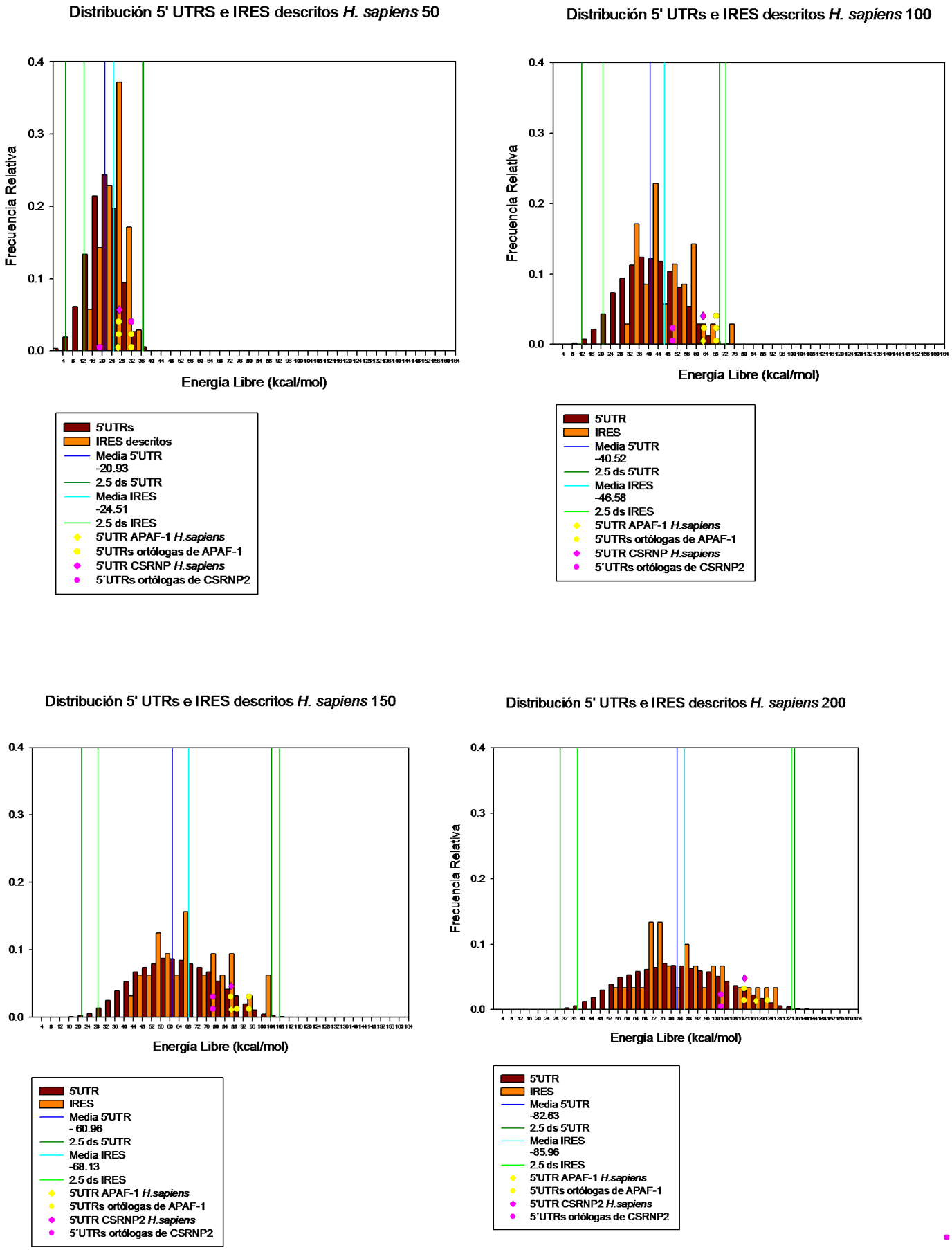


Figura 20. Distribución de los valores de energía libre de regiones 5'UTRs e IRES en humano. Análisis de la distribución de APAF-1 y sus ortólogos. Los valores de energía libre de las regiones 5'UTRs ortólogas a *Apaf-1* se encuentran próximos al valor de energía libre de la región 5'UTR de *Apaf-1* en humano, mientras que las regiones 5'UTRs ortólogas de CSRNP2 se distribuyen lejanamente del valor de energía libre d de e la región 5'UTR de CSRNP2 en humano.

Los resultados de dicha tabla muestran que los valores de energía libre son muy similares entre sí y podrían agruparse, esto sucede incluso con la comparación directa de todo el elemento IRES APAF-1 de humano, con las 5'UTRs ortólogas de los otros organismos, así como en la comparación de los valores de energía libre en una ventana de 100 nucleótidos (Tabla 7). Se observa que los ortólogos roedores, son más cercanos entre sí, así como los ortólogos primates lo son. Ambos grupos presentan un valor de energía libre cuya distribución se encuentra en la cola negativa, al igual que APAF-1 de humano (Figura 20).

Gen/Organismo	Energía libre (Kcal/mol)	Energía libre (Kcal/mol) en una ventana de 100 nts	Tamaño (nt)
APAF-1 *H. sapiens*	-372.20	-58.2	742
Apaf-1 *M. musculus*	-302.10	-71	617
Apaf-1 *R. norvegicus*	-302.10	-68.4	617
APAF1 *P. troglodytes*	-280. 00	-56.2	577
APAF1 *M. mulatta*	-280	-62.8	577

Tabla 7. Valores de energía libre calculados para APAF-1 de humano y sus ortólogos.

3.2 Determinación de genes ortólogos a aquellos de humano traducidos por IRES celulares.

El enfoque empleado en la sección anterior fue usado como evidencia a favor de nuestra hipótesis que sostiene que IRES celulares en humano con valores de energía libre muy negativos, tenderían a tener genes ortólogos potencialmente regulados por IRES estructurados. Con el propósito de ampliar la exploración y confirmar dicha hipótesis con el resto de los IRES celulares de humano, se obtuvieron los potenciales genes ortólogos a partir de las comparaciones de las secuencias de aminoácidos de sus correspondientes proteínas, tal y como se describe en la sección de la Metodología y Desarrollo, una vez identificados los potenciales genes ortólogos, se obtuvieron las secuencias de sus correspondientes regiones 5'UTRs.

A pesar de de disponer de una buena cantidad de secuencias genómicas de mamíferos, éstas no se encuentran de manera completa por lo que es difícil determinar cuál es el ortólogo *bona fide* y/o si ese ortólogo es único, de tal modo que se tomó el mejor hit como el ortólogo más probable.

Aún y cuando este criterio es razonablemente adecuado para determinar a los ortólogos más probables, es necesario emplear otros para consolidar la ortología entre las secuencias. Para ello se hicieron alineamientos globales (Ver Anexo 1). En dichos alineamientos, se puede observar que existe una alta conservación entre los residuos de las distintas secuencias de aminoácidos, por lo que se corroboró que pudieran tratarse de secuencias ortólogas *bona fide.* Adicionalmente se consultó la base de datos KEGG orthologs con el mismo fin, y se comprobó que efectivamente varios de nuestros ortólogos están anotados como los ortólogos inmediatos.

A pesar de que para algunas secuencias 5'UTR, se tiene más de una variante debido al procesamiento por *splicing* alternativo o bien por la existencia de genes parálogos, únicamente se obtuvo una secuencia 5'UTR por cada gen ortólogo y correspondió a la secuencia de mayor tamaño para cada genoma mamífero. No descartamos la posibilidad de que dicho grupo pueda ser ampliado con secuencias que no fueron consideradas, y que los resultados para dichas secuencias puedan diferir del obtenido con la secuencia que nosotros seleccionamos. No obstante, en términos estadísticos, pensamos que nuestra hipótesis puede ser probada en nuestro conjunto de secuencias de estudio.

3.3 La energía libre de los IRES en humano tiende a conservarse en los genomas ortólogos.

Al obtener las secuencias 5'UTRs ortólogas de los genomas mamíferos y calcular la energía libre de las mismas y comparar dichos valores con el valor del IRES de humano, se observa que existe cierta tendencia a encontrar los valores de energía libre muy cercanos al IRES de humano, más aún en algunos casos es claro que la media de los valores de energía de las 5'UTRs ortólogas es igual al valor del IRES original o inclusive corresponden a valores de estructuras secundarias más estables que las del valor original (más negativo) (Figura 21). No obstante hay que considerar que tenemos un grupo muy reducido de ortólogos que corresponden a organismos cercanos filogenéticamente, por lo que es posible especular que la tendencia de conservación de IRES celulares será mayor, y a medida que la distancia filogenética de los organismos de estudio sea mayor, la conservación de los valores de energía libre irá disminuyendo.

Por otro lado, se puede apreciar que para algunos cuantos casos de valores de energía libre de IRES y sus ortólogos, la energía libre no se agrupa y presentan valores dispersos, por ejemplo, el valor de energía del gen UNR, para el cual los valores de energía libre se encuentran alejados entre sí independientemente del tamaño de las ventanas de análisis (50, 75, 100 y 125 nts) (Figura 21).

Estos resultados son consistentes con la idea de que la energía libre de los IRES celulares pudieran estar conservados en organismos mamíferos filogenéticamente cercanos debido a una selección funcional; y más aún que estas relaciones evolutivas nos permitirán enriquecer y describir un número mayor de IRES en genes ortólogos aún en organismos más alejados en la escala filogénetica, así podríamos dilucidar la historia evolutiva del mecanismo molecular de regulación que subyace en los distintos organismos para contender con las mismas condiciones celulares, tales como el estrés, limitación de nutrientes, proliferación celular entre otros.

Es importante destacar, a partir de estos resultados, la observación de que no todos los IRES de humano se encuentran altamente estructurados, si bien existe una tendencia a estarlo, existen algunos cuyos valores de energía libre mínima son muy cercanos a la media (AQP4, BCL2, NAT1,HSP70).

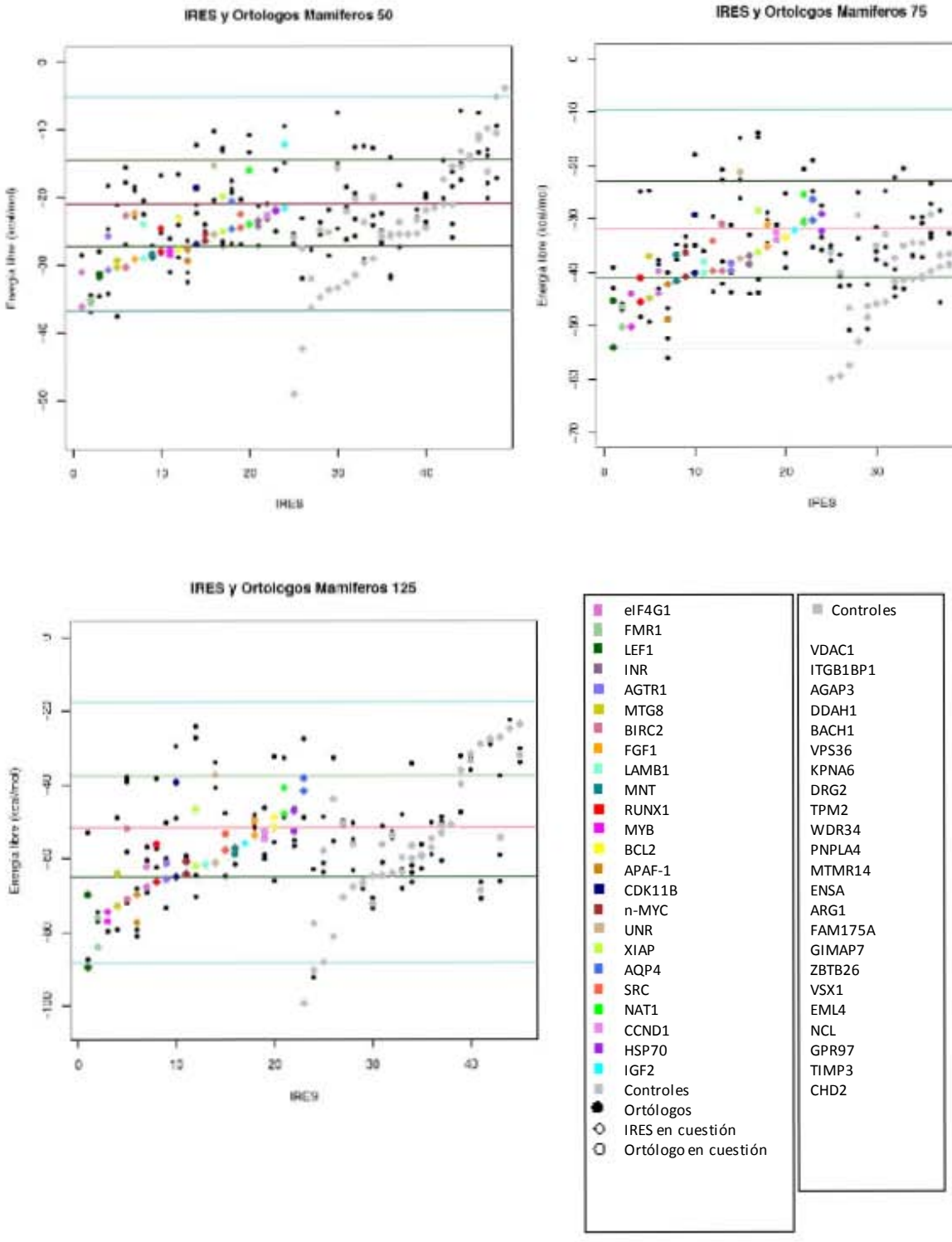


Figura 21. Distribución de energía libre de los IRES celulares descritos y predichos respecto a la distribución de energía libre de las regiones 5'UTR. Cada diamante de color representa un IRES particular de humano. El diamante representa el valor de energía libre para el IRES de humano descrito, los puntos negros los valores de energía libre de las 5'UTRs ortólogas, y el círculo de color la media de los valores de energía libre de las 5'UTRs ortólogas. La línea roja es la media de distribución de energía libre de todas las 5'UTRs del genoma humano, las primeras líneas verdes corresponden a 1 desviación estándar, y las siguientes líneas azules a 2.5 desviaciones estándar de la distribución de energía libre de las 5'UTRs del genoma humano.

Respecto a los valores de energía libre de las 5'UTRs de los genes control, podemos observar que comúnmente se distribuyen de manera cercana a la media muestral del genoma, lo cual se explica al considerar que una distribución normal concentra la mayoría de los valores (68%) cercanos a la media con una desviación estándar de dispersión. Asimismo la probabilidad de que el valor de energía libre mínima de una 5'UTR de un gen tomado al azar, se distribuya en los extremos de la campana de valores de frecuencias, es menor.

La distribución de los valores de energía libre mínima de los ortólogos control, en la mayoría de los casos es aleatoria y parecen no seguir un patrón de agrupación. Lo anterior es más notable para aquellos valores de energía libre de los controles con valores de energía más negativos, en donde los valores de energía libre de los ortólogos son considerablemente mayores.

Otro subconjunto de los valores de energía libre de las 5'UTRs de genes control son aquellos que exhiben un valor de energía cercano al cero y corresponde a regiones no estructuradas. Para éstos, los valores de energía libre de sus ortólogos correspondientes tienden a desplazarse hacia la media muestral del genoma humano cuyo valor es -51.50 kcal/mol. Esto se explica por las mismas razones descritas anteriormente, ya que en una distribución normal los valores de energía libre tenderán a agruparse en la región central de la misma.

Cabe destacar que para el grupo de genes control que se encontraban cerca de la media del genoma humano y sus ortólogos cercanos, la base de datos UTRdb (Grillo *et al.,* 2009) predice la presencia de un elemento IRES y para algunos de ellos la aparición de tractos de polipirimidinas, basados en el motivo estructural que se ha comentado antes, descrito por Le y Maizel, 1997. La aparición de dicho motivo no es necesariamente una prueba de que efectivamente se trate de un IRES. En este sentido, se sabe por ejemplo que el IRES FGF2 presenta un tracto de polipirimidinas, el cual es reconocido por la proteína PTB, la cual actúa como un ITAF durante la traducción cap independiente (Morris *et al.,* 2010). Por lo tanto, una de las propuestas es que para nuevos genes que sean posibles candidatos a tener un IRES, se hace patente la necesidad de analizar los motivos de estructura secundaria (tractos de polipirimidinas, secuencias complementarias a rRNA 18S, tractos de unión de UNR, entre otros) que han sido descritos para algunos IRES celulares.

En conjunto, los resultados obtenidos para el grupo control de valores de energía libre de 5'UTRs de humano tomados al azar, muestra que no existe una tendencia clara a que dichas regiones tiendan a agruparse con sus respectivas regiones 5'UTRs ortólogas en base a sus valores de energía libre. De manera contraria al de los genes traducidos por IRES celulares, dichos genes son traducidos normalmente mediante el reconocimiento de la estructura cap y no existe una presión de selección que mantenga un tipo particular de estructura secundaria en sus regiones 5'UTRs.

3.4 Conservación de los valores de energía libre de las 5'UTRs con IRES celulares y sus 5'UTRs ortólogas.

La tendencia de agrupación de los valores de energía libre de las 5'UTRs de los ortólogos, discutida en las secciones anteriores, podría deberse a alguno de los siguientes motivos: *i)* a que las regiones 5'UTRs de los

distintos genomas de mamífero sean en lo general muy similares porque no hayan tenido el tiempo suficiente de divergir; *ii)* a la presencia de ciertas regiones cuya secuencia primaria sea conservada por motivos funcionales.

Para explorar las distintas posibilidades anteriormente expuestas, se tomaron tres IRES al azar: APAF-1, MNT y CCND1. Se alinearon las 5'UTRs ortólogas de otros mamíferos y se ubicaron dentro del alineamiento los residuos de la ventana de análisis de 200 nucleótidos, ya que el tamaño de esta ventana comprende a las otras más pequeñas que pudieran corresponder a subestructuras del IRES. En el Anexo 2 se observan las estructuras secundarias que se forman.

En el caso de APAF-1, los IRES ortólogos se pueden agrupar en dos conjuntos: *i)* el conformado por los primates, y *ii)* el de los roedores (Anexo 2). Dentro de cada uno de estos grupos, las 5'UTRs ortólogas son casi idénticas, tanto en secuencia como en su estructura secundaria. No obstante esta división, los valores de energía libre de todos los ortólogos analizados, independientemente de su grupo, son muy parecidos entre sí.

En los casos de MNT y CCND1, las regiones 5 UTR's ortólogas son un poco más variables entre sí, pero en todos los casos, con un valor de energía libre cercano a la energía libre del IRES en humano.

Estos ejemplos muestran que los IRES celulares ortólogos tienden a tener valores de energía libre similares pese a que en algunos casos no exista una gran similitud en sus secuencias.

4. Predicción de nuevos IRES celulares

4.1 Desarrollo de un método probabilístico en la predicción de nuevos IRES celulares.

La caracterización de energía libre de los IRES analizados en la sección anterior muestran que las regiones 5'UTRs de genes ortólogos para los que se han descrito IRES celulares en humano, tienden a tener valores de energía libre similares entre si, pese a que en algunos casos no exista una gran similitud en sus secuencias.

Como parte de nuestra hipótesis, sugerimos que dicho fenómeno, se debe a que la presión de selección que opera para mantener el mecanismo de regulación traduccional independiente de cap, y partimos de premisa de que éste se ha conservado evolutivamente en organismos relacionados, ya que estos tienen que contender con circunstancias similares, tales como el estrés calórico o programas de desarrollo celular. Como hemos mencionado anteriormente, la propiedad de algunos IRES celulares de estar estructurados es tan sólo una tendencia y no una propiedad *sine qua non.* Por tal motivo, desarrollamos un método probabilístico que reduce la incertidumbre, al tomar en cuenta la relación entre la probabilidad de los valores de energía más estables de las 5'UTRs ortólogas de los organismos mamíferos de estudio y la probabilidad de la energía libre de las 5'UTRs en humano de manera conjunta, para determinar con mayor grado de certeza la probabilidad de encontrar nuevos genes candidatos a presentar un IRES en su región 5'UTR.

4.2 Comparación entre cuantiles de la distribución de la energía libre de las 5'UTRs y de una distribución teórica normal para el cálculo de la probabilidad normal.

El método probabilístico para la búsqueda de nuevos IRES celulares desarrollado en esta tesis, se basa en la determinación de la media geométrica de los valores de probabilidad calculados a través de una distribución normal estándar. La probabilidad normal se calculó partiendo del supuesto de que la distribución de la energía libre del RNA de las regiones 5'UTR, tiende a distribuirse como una normal. Esta premisa fue corroborada con la prueba comparativa entre los cuantiles de nuestras distribuciones y los cuantiles de una distribución normal teórica, para los genomas mamíferos de *H. sapiens, M. mulatta, P. troglodytes, M. musculus, R. norvegicus, C. lupus familiaris* y *E. caballus*, se distribuyen de manera normal (Figura 22). La obtención de una recta diagonal nos indica que las dos distribuciones son parecidas, y por tanto los valores de energía libre de las 5'UTRs se distribuyen de manera similar a una normal (X~N, esto es, la variable aleatoria tiende a distribuirse como una normal) para la mayoría de los puntos. Cabe notar que en los extremos, los valores se alejan de la normalidad (Figura 22). Consideramos que en estos casos, pudiera existir una selección positiva para que la frecuencia de regiones estructuradas en las regiones 5'UTRs de los genomas, sea mayor a la frecuencia esperada exclusivamente al azar.

4.3 La media geométrica de los valores de probabilidad considera la probabilidad conjunta de las 5'UTRs del genoma humano y las 5'UTRs ortólogas en otros organismos.

Nuestra estrategia estadística para la identificación de IRES celulares consiste en evaluar la probabilidad conjunta de que un IRES celular y sus ortólogos tengan una energía libre menor a la del valor promedio de energías libres de las regiones 5'UTR de sus correspondientes genomas. Esta probabilidad conjunta, denominada *P*, fue evaluada como la media geométrica de las probabilidades independientes *p*, de cada IRES en cuestión. De esta manera, el valor de probabilidad *p* asociado a la energía libre de un IRES en particular, representa el área bajo la curva de la distribución de frecuencias de los valores de energía libre de un genoma, mientras que el valor *P* considera de manera conjunta los valores de probabilidad de los valores de energía libre del IRES en cuestión y de todos sus ortólogos de los genomas en estudio.

4.4 Comparación de la probabilidad individual p y la probabilidad conjunta P de IRES celulares en humano

Con el fin de establecer si la identificación de IRES celulares humanos en base a su energía libre pudiera resultar más eficiente al considerarlo de manera conjunta con el valor de energía libre de las regiones 5'UTRs de sus correspondientes ortólogos, los valores de probabilidad *p* y *P* fueron comparados.

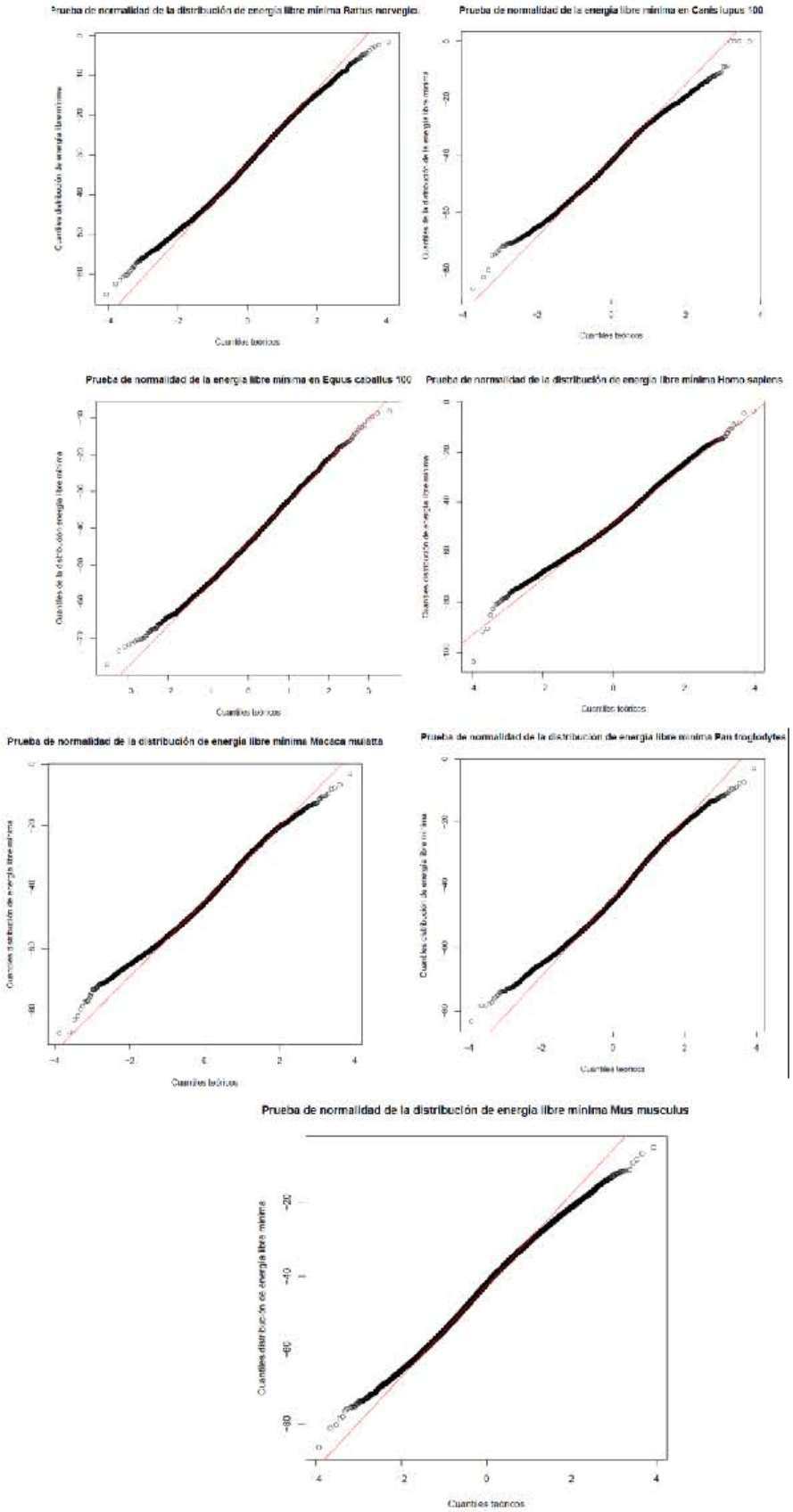


Figura 22. Prueba de normalidad. Comparación entre los cuantiles de una distribución normal teórica y la distribución de energía libre de las regiones 5'UTR para cada genoma de mamífero. En los valores extremos, la frecuencia de una distribución real se aleja a la de una distribución normal.

Cabe recordar que los valores de probabilidad individual *p* para cada uno de los IRES celulares en un genoma particular, fueron obtenidos considerando la distribución normal de los valores de energía libre de las regiones 5'UTRs del genoma en cuestión, mientras la probabilidad conjunta *P*, se evaluó como la media geométrica de los valores *p* de genes ortólogos en los diferentes genomas de mamíferos incluidos en nuestro estudio (*H. sapiens, P. troglodytes, M. mulatta, R. norvegicus* y *M. musculus*). En la Figura 23 se muestran los valores de dichas probabilidades.

En esta figura se puede apreciar que los valores de probabilidad individual *p* de los IRES celulares de humano, se distribuyen a todo lo largo de la recta, mientras que la probabilidad conjunta *P* de dichos IRES celulares son generalmente inferiores. Adicionalmente, es también posible observar casos de IRES celulares cuya identificación a partir del valor de probabilidad individual *p* sería incierta, pero cuya probabilidad conjunta *P* permitiría identificarlos con mayor certeza. Tales son los casos de los IRES de LAMB1 ($p = \sim 0.9$, $P = \sim 0.4$) y MNT ($p = \sim 0.55$, $P = \sim 0.38$).

Consideramos que el valor que ofrece cálculo de la probabilidad conjunta *P*, no tan solo consiste en "amplificar" la señal estadística para la identificación de IRES celulares, si no también constituye una evidencia indirecta del fenómeno subyacente de conservación del mecanismo de regulación de la traducción cap-independiente en diferentes organismos, ya que la media geométrica disminuiría drásticamente el valor de *P* si y sólo si los valores de *p* de las 5'UTRs conservan entre sí la similitud de ser extremadamente pequeños.

Tal y como se esperaba, existen algunos casos de IRES celulares en humano en donde la probabilidad conjunta *P* puede ser mayor a la probabilidad individual *p*. No obstante, en la mayoría de estos casos la diferencia no es tan significativa. Tal es el caso el IRES de XIAP ($p = \sim 0.5$, $P = \sim 0.75$), cuyo valor *P* aumenta respecto al valor *p*.

Con el objetivo de contar con un control negativo en nuestro estudio, se tomaron de manera azarosa 40 regiones 5'UTR del genoma humano y se calcularon sus respectivos valores de probabilidad individual *p* y conjunta *P*, y se ubicaron sus valores correspondientes (Figura 23B, C y D). De manera general, se observa que el aumento o el decremento en los valores de la media geométrica *P* en relación a la probabilidad normal *p* es aleatorio, y no parece haber una tendencia marcada a que la media geométrica disminuya. Este comportamiento aleatorio en los grupos control es debido a que la mayoría de ellos carecen de elementos de regulación dependientes de segmentos de RNA estructurados en sus regiones 5'UTRs.

Tomando en cuenta lo anteriormente descrito, consideramos que la probabilidad conjunta *P* calculada como la media geométrica de las probabilidades de genes ortólogos de poseer estructuras estables de RNA en sus regiones 5'UTR, es un buen descriptor para identificar IRES celulares, tal y como se describe y discute en secciones siguientes.

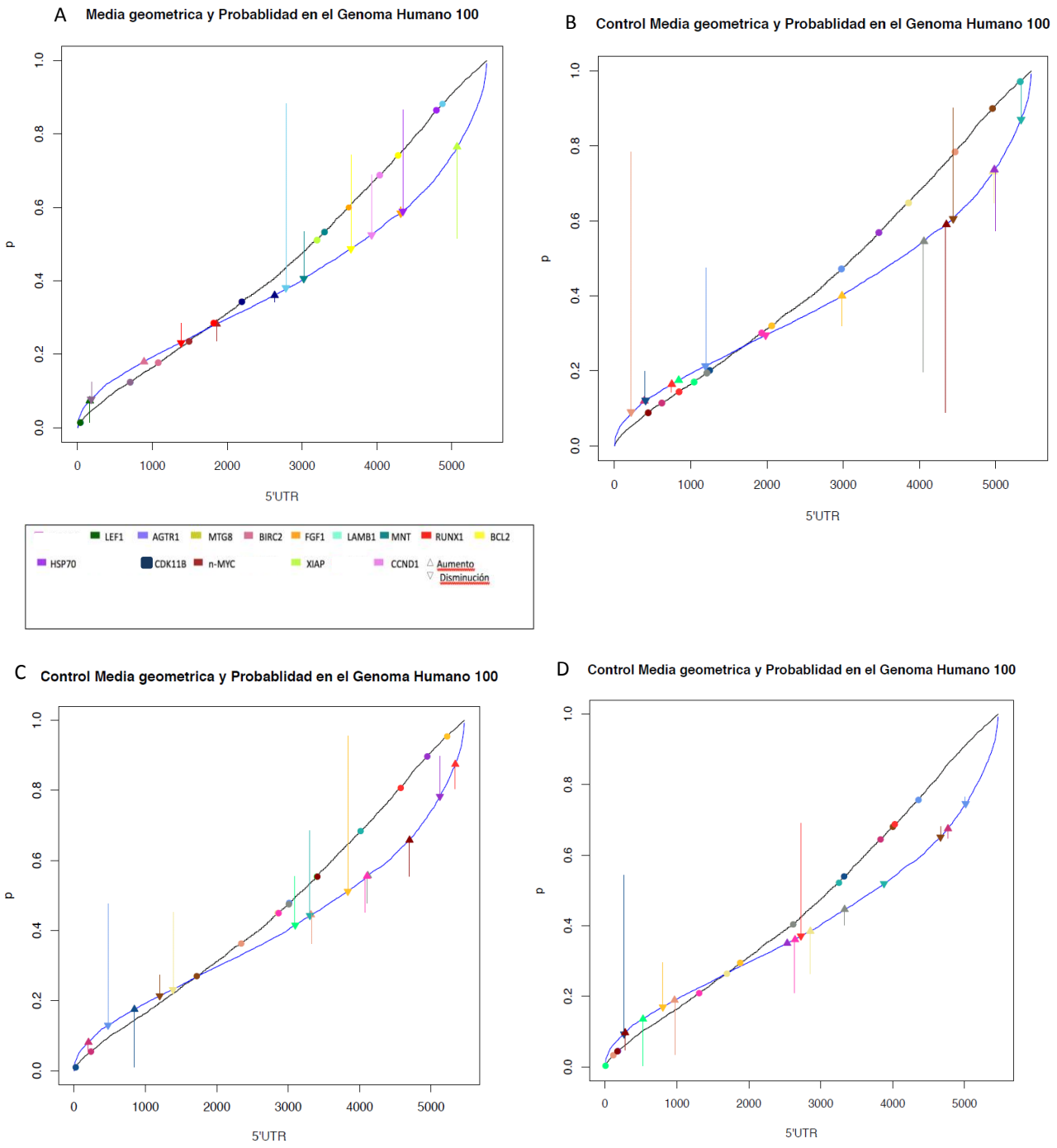


Figura 23. Comparación entre la probabilidad individual *p* y conjunta *P* de IRES celulares de acuerdo a sus correspondientes valores de energía libre. La línea negra representa la distribución de la probabilidad normal obtenida a través de la estandarización de la energía libre de las 5'UTRs considerando únicamente el genoma humano. La línea azul, representa la distribución de los valores, de probabilidad conjunta *P* evaluados como la media geométrica de los valores de la probabilidad normal asociada a la energía libre de las 5'UTRs, considerando el genoma humano y sus ortólogos en *P. troglodytes, M. mulatta, R norvegicus* y *M. musculus*. En la primera gráfica (A) se representa cada valor de los IRES celulares descritos en humano dentro de ambas distribuciones. Los círculos representan el valor de probabilidad individual *p* asociada a cada IRES celular de humano, los triángulos a los valores de probabilidad conjunta *P*, calculada con la media geométrica de los valores *p* de IRES de humano y valores asociados las regiones 5'UTRs de sus ortólogos. Las flechas hacia arriba representan un aumento en el valor de *P* respecto a la *p* en humano, y el triángulo hacia abajo, que el valor de *P* disminuyó respecto a la *p* en humano. Las tres últimas gráficas (B, C, D) son gráficas de control construidas a partir de los datos de probabilidad de individual *p* y conjunta *P* de regiones 5'UTRs del genoma humano tomadas al azar.

4.5 Las regiones 5'UTRs cuyos valores de probabilidad conjunta P son cercanos a cero son potenciales candidatos a presentar IRES celulares

De manera similar al análisis de distribución de los valores de energía mínima efectuado en las secciones anteriores (Figuras 12-15, 18, 20) se evaluaron las distribuciones normalizadas de los valores de las probabilidades conjuntas *P* de cada una de las regiones 5'UTRs en humano. El resultado de esta distribución es mostrado en la figura 24C.

Los datos de dicho histograma fueron divididos en dos conjuntos dependiendo del valor de la energía libre mínima de las regiones 5'UTRs de los genes de humano. El histograma de la figura 24A fue construido a partir de la probabilidad conjunta *P* de regiones 5'UTR de genes del humano cuyos valores de energía libre fueron mayores a la energía libre promedio de todas las regiones 5'UTRs en humano y presenta una distribución asimétrica positiva. Los valores de este histograma corresponden a segmentos de RNA pobremente estructurados. De manera contraria, el histograma de la figura 24B, fue construido con base a la probabilidad conjunta *P* de regiones 5'UTRs de genes de humano cuyos valores de energía libre fueron mayores a la energía libre promedio de dichas regiones y presenta una distribución asimétrica negativa. Los valores de este histograma corresponden a segmentos de RNA estructurados con mayor posibilidad de poseer IRES celulares. En cualquiera de ambos histogramas, la mayoría de los valores *P* fueron menores a los de sus correspondientes valores individuales de *p* de humano. Esto se explica por la propiedad de la media geométrica de ser muy sensible a valores pequeños de los elementos que la constituyen. Esta propiedad de la media geométrica, puede ser particularmente importante en nuestro análisis estadístico ya que resalta el hecho de que alguna o varias de las regiones ortólogas 5'UTR presenten valores de *p* pequeños, lo cual implica una conservación funcional del fenómeno biológico de hallar elementos estructurados de RNA en las regiones 5'UTR de los genes cuya traducción es independiente de cap.

Ésta hipótesis es consistente con lo reportado por Davuluri et al.,(2000) quienes realizaron un análisis de árbol de regresión y clasificación (CART) para agrupar las secuencias 5' UTRs de humano, de acuerdo a su energía libre y concluyeron que, es posible que en el grupo en el cual se encuentran valores de energía menores a -50 kcal/mol y a 2 desviaciones estándar de la distribución de energía libre, se encuentren de manera experimental y/o computacional nuevos IRES celulares.

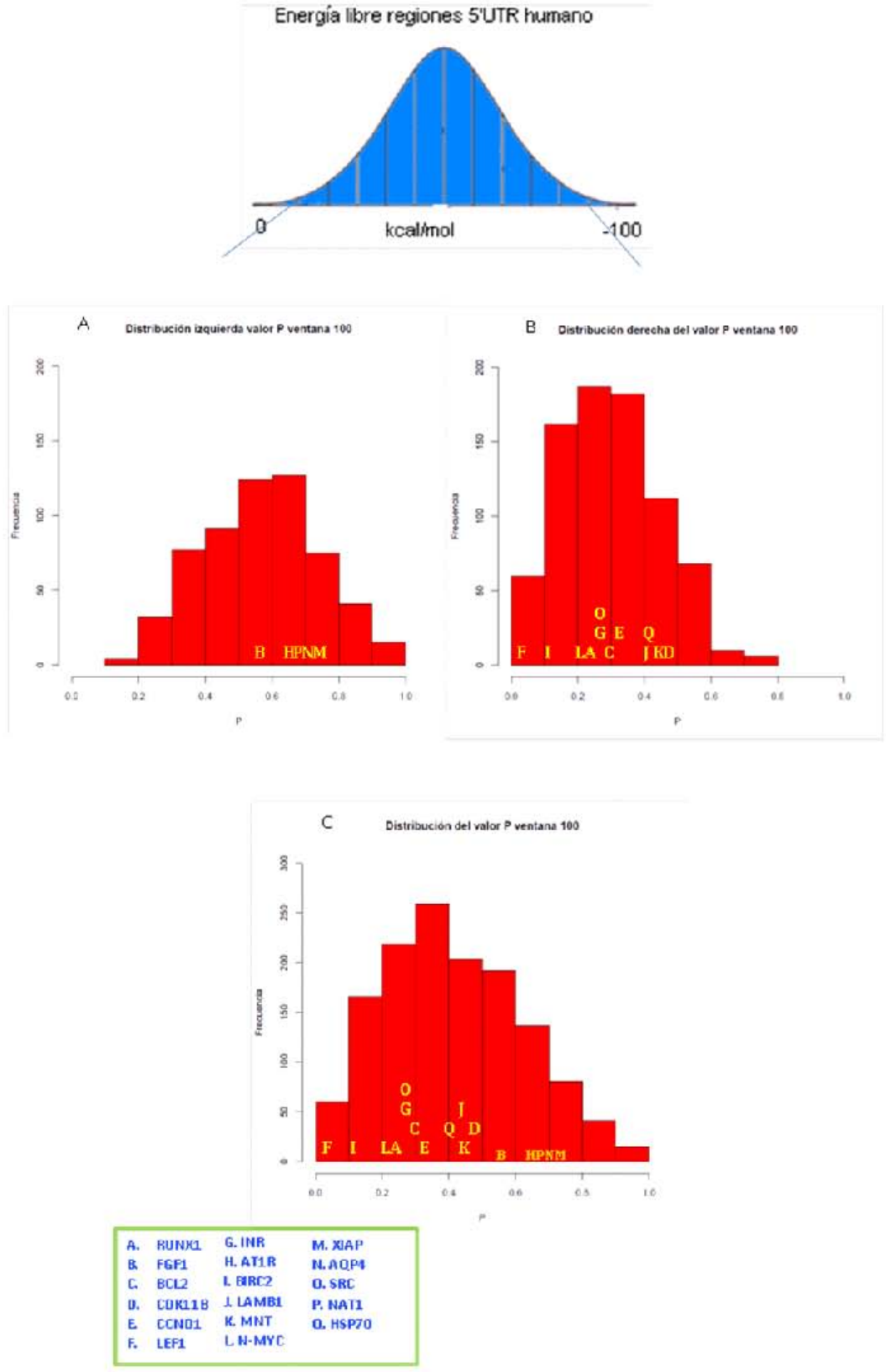


Figura 24. Distribución del valor *P* calculado con siete genomas mamíferos. La distribución de los valores *P* asociados al lado derecho de la distribución de energía libre de las 5'UTRs de humano (histograma azul) (Figura 24 B) (estructuras secundarias más estables que la media), presentan una asimetría negativa, y tienden a distribuirse hacia valores más pequeños de *P*, mientras que para la distribución de los valores *P* provenientes del lado izquierdo de la distribución de la energía libre de las regiones 5'UTRs en humano (histograma azul) (Figura 24 A) (estructuras secundarias menos estables que la media), se distingue una asimetría positiva. La distribución final del valor *P*, considerando ambas distribuciones del valor *P* (derecha e izquierda) se muestra en la Figura 24 C y presenta asimismo una asimetría negativa. Las letras dentro de los histogramas muestran la ubicación de los valores *P* para los IRES descritos en humano. Se observa que la mayoría de ellos se encuentran asociados a la distribución derecha de la energía libre de las 5'UTRs.

4.6 Análisis de los valores de probabilidad conjunta P de los IRES celulares presentes en humano

Una de las restricciones impuestas para el cálculo del valor *P* y el análisis de las regiones 5'UTRs consistía en la existencia de al menos 4 genes ortólogos en nuestros genomas mamíferos modelo. Debido a que la anotación de regiones codificantes en algunos casos es incompleta, solamente se pudieron evaluar dos terceras partes del total de IRES celulares de humano descritos. La mayoría de los valores de *P* para dicho conjunto (dos terceras partes), se encuentran distribuidos en el extremo derecho de la distribución del valor *P* asociados a la formación de estructura secundaria estable, mientras que la otra tercera parte se distribuye del lado izquierdo, que representa a regiones 5'UTR pobremente estructuradas (Figura 24). Dado lo anterior podemos aseverar que si bien no todos los IRES se encuentran altamente estructurados, existe un grupo de ellos que tienen un sesgo a presentar una estructura secundaria estable. Aunado a lo anterior, en nuestro análisis probabilístico analizamos un grupo reducido de IRES descritos en humano, debido principalmente a la limitante debida a la anotación incompleta de las regiones 5'UTR ortólogos, por lo que no fue posible calcular en primera instancia la *p* de cada una de ellas, para poder obtener finalmente el valor de *P*.

Es posible que el grupo de IRES celulares que se distribuyen del lado izquierdo de los valores de probabilidad conjunta *P*, es decir aquellos pobremente estructurados, requieran mayormente de la participación celular de los ITAFs, ya que algunos de estos factores funcionan como chaperonas de RNA que remodelan la estructura espacial para producir conformaciones con mayor o menor afinidad a los componentes de la maquinaria de traducción o hacen un puente de interacción entre el RNA (región 5´ UTR) y factores proteicos. En este sentido sería muy interesante probar si existe alguna relación entre este grupo de IRES poco estructurados y el requerimiento de los factores ITAFs, respecto al grupo de IRES cuyos valores de *P* se ubican del lado derecho de la distribución del valor de probabilidad conjunta *P*, para los que podríamos esperar que se requieran menormente dichos ITAFs, a medida que los IRES se encuentren más estructurados. Este tipo de relación entre el requerimiento de ITAFs y el grado de estructuración de los IRES se ha observado con los IRES virales, tal como se describe en la figura 4. Desafortunadamente, aún no se ha llevado a cabo una caracterización exhaustiva de los ITAFs que participan durante la traducción independiente de cap para cada IRES, por lo que aún no es posible llevar a cabo estudios que incluyan dicha información para poder realizar nuestra propuesta de clasificación de los IRES celulares y más aún, llevar a cabo el análisis inverso que permita la posible identificación de estos factores para aislar a grupos de mRNAs que pudieran estar regulados por IRES.

4.7 Selección de nuevos IRES celulares potenciales

4.7.1 Selección de genes candidatos a presentar un IRES en su región 5'UTR a partir de valores de P extremadamente pequeños

Para cada uno de los distintos tamaños de ventana de análisis (50, 75, 100, 125 nucleótidos) se obtuvo una lista final de los valores de media geométrica *P*. Como se mencionó anteriormente, este valor de *P* conjunta, fue calculado como la media geométrica de las probabilidades normales *p* de los genes ortólogos. A esta lista, fueron agregadas la descripción del tipo de funciones celulares de cada uno de ellos. Debido a que se trata de una gran cantidad de información (3,500 valores de probabilidad conjunta), se muestra únicamente la lista final de candidatos a presentar un IRES celular en las Tablas 8 y 9.

Los posibles candidatos a presentar un IRES celular en la región 5'UTR, se obtuvieron de la lista final al establecer un valor de corte de $P \leq 0.15$, teniendo en este subconjunto los valores del extremo inferior entre 0 y 0.15, de la distribución de los valores de la probabilidad conjunta *P* (Tabla 8). Se estableció este valor de corte, debido a que, como se observa en la figura 24 (Distribución derecha), los valores de *P* entre 0 y 0.15, representan un subconjunto enriquecido con valores de *P* correspondientes a regiones 5'UTRs cuyos valores de energía libre son considerablemente diferentes al promedio de valores de energía del conjunto de energía libre de todas las regiones 5'UTRs del genoma humano.

Aunado a los filtros de selección estadísticos, existe el filtro funcional que da sentido biológico a nuestro análisis y considera el fenómeno de encontrar un IRES dado que su contexto celular esté relacionado a la traducción independiente de cap, y que la señal de IRES celular se encuentre en genes en donde el contexto celular relacionado se conserve a lo largo de la escala filogenética, dentro de nuestros genomas modelo. Pese a que el número de genomas de este primer estudio fue pequeño, no descartamos que en un futuro pueda adicionarse la información de otros genomas mamíferos y/o de otros genomas eucariontes.

Después de considerar estas condiciones, nuestro análisis estadístico identificó alrededor de 50 regiones 5'UTRs candidatas a presentar un IRES celular (Tabla 8). Como puede observarse en esta tabla, los candidatos seleccionados se encuentran involucrados en distintas funciones celulares, en donde la traducción independiente de cap es requerida. Dichas funciones incluyen a la apoptosis, control del ciclo celular, diferenciación celular y crecimiento, traducción, respuesta a choque térmico, cáncer, entre otros (Tabla 9).

La clasificación funcional realizada puede resultar ambigua (Tabla 9), ya que algunos de los procesos involucrados se derivan de la desregulación de otro proceso celular lo que resulta en procesos altamente relacionados. Sin embargo esta clasificación funcional permite ilustrar que nuestro método estadístico recupera de manera eficiente a candidatos que están involucrados en condiciones celulares que favorecen la traducción vía IRES. Estamos conscientes de que no todos los genes listados en la Tabla 9 pudieran presentar actividad de IRES durante la traducción independiente de cap. Como se mencionó en la sección 3.4 (inciso *ii*), existe la posibilidad

de elementos de regulación que estén conservados entre diferentes genes ortólogos y que pese a que no sean IRES, contengan motivos en secuencia primaria y secundaria conservados. Tal es el caso del gen DLC (*deleted in liver cancer*) que presenta una isla de CpG en la región promotora, que puede ser metilada para promover el silenciamiento génico y prevenir la transcripción (Kim *et al*., 2003), es posible que el valor *P* corresponda a la energía libre de dicha isla, que forma parte de la 5'UTR, y por tanto no corresponda a la señal de IRES celular.

Algunos de los candidatos de la Tabla 9 fueron seleccionados, para mostrar que éstos son muy probablemente genes parálogos a los genes que presentan un IRES celular descrito y caracterizado. Por ejemplo para el IRES presente en FGF1, recuperamos lo que podría ser su gen parálogo, FGF2.

Esta identificación de genes parálogos a otros experimentalmente caracterizados como IRES es una evidencia adicional de que no es fortuito encontrar elementos estructurados de RNA conservados, con valores tan bajos de probabilidad conjunta *P*, y nos permite proponer la existencia de una presión de selección para mantener elementos de RNA altamente estructurados como mecanismo de regulación de la expresión genética.

Por otro lado la lista de candidatos a presentar IRES incluye a algunos genes que se encuentran involucrados en casi todos los contextos celulares en donde la traducción dependiente de cap es abatida, tal es el caso del factor transcripcional SRF (*serum response factor, c-fos serum response element-binding transcription factor*) que participa en la regulación del ciclo celular, apoptosis, crecimiento y diferenciación celular (Shore y Sharrocks, 1995). Otro ejemplo es la proteína EP300 (*E1A binding protein p300*), para la que la base de datos KEGG, la reporta involucrada en diversas vías,: ciclo celular, vía de señalización *Wnt*, vía de señalización *Notch*, vía de señalización *TGF-beta*, vía de señalización *Jak-STAT*, cáncer prostático y renal, melanogénesis, infección por herpes simple, influenza A, tuberculosis entre otras. Este tipo de candidatos pueden ser examinados de manera experimental.

En resumen, consideramos que la lista de candidatos mostrados en la Tabla 9 puede ser una buena guía para la selección de potenciales IRES celulares que pudieran ser caracterizados experimentalmente. Para aquellos candidatos que presenten uno o más genes parálogos, sugerimos como el más adecuado para su caracterización experimental, a aquel que presente el valor de *P* más bajo.

P	Gene_ID	Nombre del gen		Contribución de los valores p de ortólogos						
0	55206	SBNO1	strawberry notch homolog 1 (Drosophila)	hsa:0.000	rno:0.764	mmu:0.502	mcc:0.079	ptr:0.000	cfa:0.188	ecb:N
0.01	9274	BCL7C	B-cell CLL/lymphoma 7C	hsa:0.048	rno:0.002	mmu:0.004	mcc:0.013	ptr:0.034	cfa:0.021	ecb:N
0.02	4605	MYBL2	v-myb myeloblastosis viral oncogene homolog (avian)-like 2	hsa:0.007	rno:0.048	mmu:0.002	mcc:0.122	ptr:0.004	cfa:N	ecb:N
0.03	339344	MYPOP	Myb-related transcription factor, partner of profiling	hsa:0.031	rno:0.031	mmu:0.042	mcc:0.018	ptr:N	cfa:N	ecb:N
0.04	7040	TGFB1	transforming growth factor, beta 1	hsa:0.016	rno:0.309	mmu:0.005	mcc:0.182	ptr:0.010	cfa:0.119	ecb:N
0.04	1857	DVL3	disheveled, dsh homolog 3 (Drosophila)	hsa:0.015	cfa:0.730	mmu:0.008	mcc:N	ptr:0.030	ecb:0.266	rno:0.009
0.05	56478	EIF4ENIF1	eukaryotic translation initiation factor 4E nuclear import factor 1	hsa:0.033	cfa:0.016	rno:0.012	mcc:0.039	ptr:0.050	ecb:0.144	mmu:0.453
0.06	282679	AQP11	aquaporine 11	hsa:0.052	rno:0.028	mmu:0.038	mcc:0.025	ptr:0.033	cfa:0.698	ecb:N
0.06	10817	FRS3	fibroblast growth factor receptor substrate 3	hsa:0.044	rno:0.363	mmu:0.022	mcc:0.026	ptr:0.028	cfa:0.052	ecb:0.282
0.06	113201	CASC4	cancer susceptibility candidate 4	hsa:0.063	rno:0.023	mmu:0.012	mcc:0.104	ptr:0.096	cfa:0.401	ecb:N
0.07	22800	RRAS2	related RAS (r-ras) oncogene homolog 2	hsa:0.254	cfa:N	rno:0.010	mcc:0.174	ptr:0.193	ecb:N	mmu:0.015
0.07	79870	BAALC	brain and acute leukemia, cytoplasmic	hsa:0.096	rno:0.092	mmu:N	mcc:0.032	ptr:0.083	cfa:0.07	ecb:N
0.07	1032	CDKN2D	cyclin-dependent kinase inhibitor 2D (p19, inhibits CDK4)	hsa:0.026	rno:0.113	mmu:0.056	mcc:0.091	ptr:0.084	cfa:0.272	ecb:0.027
0.07	3297	HSF1	heat shock transcription factor	hsa:0.214	rno:0.136	mmu:0.149	mcc:0.014	ptr:0.082	cfa:0.031	ecb:N
0.07	3622	ING2	inhibitor of growth family, member 2	hsa:0.186	rno:0.046	mmu:0.057	mcc:0.056	ptr:0.096	cfa:0.060	ecb:N
0.08	6722	SRF	serum response factor (c-fos serum response element-bindingtranscription factor)	hsa:0.214	rno:0.020	mmu:0.028	mcc:0.304	ptr:0.017	cfa:0.292	ecb:N
0.08	7407	VARS	valyl-tRNA synthetase (EC:6.1.1.9)	hsa:0.014	rno:0.021	mmu:0.276	mcc:0.075	ptr:0.054	cfa:0.655	ecb:N
0.08	53917	RAB24	RAB24, member RAS oncogene family	hsa:0.249	rno:0.010	mmu:0.018	mcc:0.163	ptr:0.166	cfa:0.064	ecb:0.225
0.08	10395	DLC1	deleted in liver cancer 1	hsa:0.292	rno:0.278	mmu:0.164	mcc:0.198	ptr:0.001	cfa:0.134	ecb:N
0.09	3487	IGFBP4	insulin-like growth factor binding protein	hsa:0.094	rno:0.186	mmu:0.108	mcc:0.058	ptr:0.060	cfa:N	ecb:N
0.09	3131	HLF	hepatic leukemia factor	hsa:0.131	rno:0.129	mmu:0.048	mcc:0.104	ptr:0.084	cfa:N	ecb:N
0.09	3068	HDGF	hepatoma-derived growth factor	hsa:0.072	rno:0.073	mmu:0.092	mcc:0.394	ptr:0.039	cfa:N	ecb:N
0.1	11031	RAB31	RAB31, member RAS oncogene family	hsa:0.174	rno:0.096	mmu:0.046	mcc:0.067	ptr:0.113	cfa:0.092	ecb:0.160
0.1	25822	DNAJB5	DnaJ (Hsp40) homolog, subfamily B, member 5	hsa:0.076	rno:0.061	mmu:0.491	mcc:0.013	ptr:0.014	cfa:0.295	ecb:0.739
0.1	10605	PAIP1	poly(A) binding protein interacting protein 1	hsa:0.182	rno:0.213	mmu:0.052	mcc:0.166	ptr:0.019	cfa:0.144	ecb:N
0.1	22913	RALY	RNA binding protein, autoantigenic (hnRNP-associated with lethalyellow homolog (mouse))	hsa:0.220	rno:0.716	mmu:0.072	mcc:N	ptr:0.009	cfa:0.001	ecb:0.840
0.1	117178	SSX2IP	synovial sarcoma, X breakpoint 2 interacting protein	hsa:0.146	rno:0.035	mmu:0.010	mcc:0.056	ptr:0.173	cfa:0.424	ecb:0.595
0.1	2263	FGFR2	fibroblast growth factor receptor 2 (EC:2.7.10.1)	hsa:0.142	cfa:0.340	rno:0.076	ptr:0.091	mmu:0.067	ecb:0.050	mcc:N
0.1	11100	HNRNPUL1	heterogeneous nuclear ribonucleoprotein U-like 1	hsa:0.112	rno:0.047	mmu:0.061	mcc:0.065	ptr:0.050	cfa:0.141	ecb:0.861
0.11	9988	DMTF1	cyclin D binding myb-like transcription factor 1	hsa:0.0306	rno:0.407	mmu:0.147	mcc:0.049	ptr:0.014	cfa:0.836	ecb:0.146
0.12	53916	RAB4B	RAB4B, member RAS oncogene family	hsa:0.161	rno:0.150	mmu:0.286	mcc:0.075	ptr:0.104	cfa:0.200	ecb:0.027

0.12	5546	PRCC	papillary renal cell carcinoma (translocation-associated)	hsa:0.304	rno:0.045	mmu:0.052	mcc:0.209	ptr:0.121	cfa:0.114	ecb:0.175
0.12	675	BRCA2	breast cancer 2, early onset	hsa:0.081	rno:0.925	mmu:0.037	mcc:0.069	ptr:0.053	cfa:0.329	ecb:N
0.13	7128	TNFAIP3	tumor necrosis factor, alpha-induced protein 3 (EC:3.4.19.12)	hsa:0.270	rno:0.047	mmu:0.061	mcc:0.212	ptr:0.189	cfa:N	ecb:N
0.13	4086	SMAD1	SMAD family member 1	hsa:0.050	cfa:0.832	mmu:0.022	mcc:0.950	ptr:0.128	ecb:0.345	rno:0.013
0.13	128338	DRAM2	DNA-damage regulated autophagy modulator 2	hsa:0.056	rno:0.754	mmu:0.433	mcc:0.053	ptr:0.038	cfa:N	ecb:N
0.13	9219	MTA2	metastasis associated 1 family, member 2	hsa:0.200	rno:0.055	mmu:0.046	mcc:0.102	ptr:0.104	cfa:0.204	ecb:0.578
0.13	1031	CDKN2C	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	hsa:0.061	rno:0.019	mmu:0.834	mcc:0.166	ptr:0.039	cfa:0.351	ecb:0.299
0.13	203245	NAIF1	nuclear apoptosis inducing factor 1	hsa:0.191	rno:0.176	mmu:0.129	mcc:0.104	ptr:0.107	cfa:0.110	ecb:N
0.13	2033	EP300	E1A binding protein p300 (EC:2.3.1.48)	hsa:0.138	rno:0.279	mmu:0.272	mcc:0.104	ptr:0.066	cfa:0.173	ecb:0.064
0.14	7290	HIRA	HIR histone cell cycle regulation defective homolog A (S.cerevisiae)	hsa:0.169	rno:0.031	mmu:0.129	ptr:0.163	cfa:0.106	ecb:0.528	mcc:N
0.14	54468	MIOS	missing oocyte, meiosis regulator, homolog (Drosophila)	hsa:0.035	rno:0.143	mmu:0.149	mcc:0.128	ptr:0.089	cfa:0.761	ecb:N
0.14	1399	CRKL	v-crk sarcoma virus CT10 oncogene homolog (avian)-like	hsa:0.184	rno:0.125	mmu:0.200	mcc:0.207	ptr:0.118	cfa:0.066	ecb:0.131
0.14	2055	CLN8	ceroid-lipofuscinosis, neuronal 8 (epilepsy, progressive with mental retardation)	hsa:0.026	rno:0.577	mmu:0.205	mcc:0.159	ptr:0.041	cfa:0.356	ecb:N
0.14	7991	TUSC3	tumor suppressor candidate 3	hsa:0.110	cfa:0.285	rno:0.188	mcc:0.069	ptr:0.066	ecb:0.277	mmu:N
0.14	1445	CSK	c-src tyrosine kinase (EC:2.7.10.2)	hsa:0.044	cfa:0.512	mmu:0.604	mcc:0.073	ptr:0.028	ecb:0.266	rno:N
0.14	5930	RBBP6	retinoblastoma binding protein 6	hsa:0.060	rno:0.366	mmu:0.392	mcc:0.036	ptr:0.219	cfa:0.127	ecb:0.103
0.15	8079	MLF2	myeloid leukemia factor 2	hsa:0.032	rno:0.595	mmu:0.271	mcc:0.045	ptr:0.514	cfa:0.080	ecb:N
0.15	8900	CCNA1	cyclin A1	hsa:0.074	rno:0.136	mmu:0.119	mcc:0.155	ptr:0.413	cfa:N	ecb:N
0.15	2113	ETS1	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)	hsa:0.350	rno:0.122	mmu:0.271	mcc:0.104	ptr:0.067	cfa:N	ecb:N

Tabla 8. Lista de genes candidatos a presentar un IRES celular en la región 5'UTR de acuerdo al valor *P*. El valor *P* se refiere al valor de probabilidad conjunta *P* calculada con los valores de *p* de *H. sapiens, P. troglodytes, M. mulatta, R. norvegicus* y *M. musculus, E. caballus* y *C. lupus familairis,* considerando la posibilidad de contar con al menos cuatro de los valores *p* de los ortólogos, el valor *p* en humano, siempre fue el valor condicional para calcular el valor de probabilidad conjunta *P.*

Contexto celular	Tipo de proteínas	Genes candidatos
Ciclo celular	Ciclinas	DMTF1, CCNA1
	cinasas dependientes de ciclinas e inhibidore	CDKN2D, CDKN2C
	Factores de transcripción	MYBL2
	Otros	EP300, MIOS
Diferenciación celular/Crecimiento	Factores de crecimiento e inhibidores	TGFB1, FRS3,ING2, IGFBP4, HDGF, FGFR2
	Factores de transcripción	SRF
	Otros	SBNO1, DVL3, DMTF1, SMAD1, EP300, CSK
Cáncer	Oncogenes	RRAS2, RAB24, DLC1, RAB31, RAB4B, CRKL,ETS1
	Factores de transcripción	HLF, TNFAIP3
	Otros	SBNO1, CASC4, BAALC, SSX2IP, PRCC, BRCA2, MTA2, EP300,TUSC3, MLF2
Choque térmico	Chaperonas	DNAJB5
	Factores de transcripción	HSF1
Apoptosis	Factores de transcripción	SRF, NAIF1
	Otros	DRAM1
Traducción	Factores de inicio de la traducción	eIF4ENIF, PAIP1
	Ribonucleoproteínas	RALY,HNRNPUL1
Otros	Asociados a síndromes, enfermedades	MYBL2, EP300, HIRA, CLN8
	Otros	BCL7C, MYPOP, AQP11, VARS

Tabla 9. Lista de genes candidatos a presentar un IRES celular en la región 5'UTR de acuerdo a su función.

Conclusiones

El contexto de una célula eucarionte puede ser altamente dinámico. Hay condiciones bajo las cuales dicho contexto podría ser alterado del estado basal. Algunas de estas condiciones incluyen el estrés (por choque térmico, hipoxia, limitación de nutrientes); crecimiento celular y diferenciación; apotosis. Bajo estas condiciones, la traducción dependiente de cap puede verse comprometida y existe una necesidad inherente de mantener activa la traducción de algunos mensajeros para dirigir a la célula a un estado de recuperación o hacia la apoptosis. Los IRES celulares median la traducción de dichos mensajeros, para proveer a la célula de un alto control espacial y temporal que constantemente se requiere para proporcionar una respuesta apropiada ante dichas condiciones.

La identificación de IRES celulares no ha sido una tarea sencilla, en parte debido a que no existe un mecanismo universal que los describa. Por lo contrario se piensa que es posible la existencia de diferentes clases de de IRES celulares, dado que éstos operan en distintas configuraciones de la maquinaria traduccional en conjunto con los accesorios proteicos ITAFs, y consecuentemente con diferentes formas de ensamblaje de los ribosomas, bajo contextos celulares distintos, a lo que se le agrega a su actividad la línea celular en la que están presentes. Por tanto, no existe un modelo celular que establezca un mecanismo *bona fide* de iniciar la traducción independiente de cap.

De esta tesis y reportes previos en la literatura, podemos aseverar entonces que la palabra “diversidad”, es la que mejor podría describir la naturaleza y estructura de los IRES celulares. Se ha discutido ampliamente la falta de motivos en secuencia primaria o motivos estructurales conservados universalmente, por lo que su identificación se limita a algunos cuantos genes en organismos eucariontes, de manera particular en humano. Pese a dicha ausencia de conservación estructural en los IRES, la mayoría de ellos descritos en humano, requieren de cierta estructura específica estable en el RNA.

La metodología desarrollada en esta tesis, retoma la idea anteriormente expuesta y considera posible la identificación de nuevos IRES celulares, si se implementan nuevos enfoques conceptuales. Dichos enfoques toman en cuenta por un lado la genómica comparativa, y por otra parte el concepto evolutivo de conservación de un mecanismo molecular durante distintos contextos celulares. Dados estos dos enfoques básicos, en nuestro estudio esperábamos encontrar la conservación de la energía libre mínima en regiones 5’UTR de genes ortólogos que presentaran IRES celulares.

Durante el desarrollo de esta tesis, se construyó una metodología estadística que nos permitió intensificar la señal estructural de IRES celulares y evaluar la probabilidad de que nuestras predicciones sean veraces. Lo anterior tomando en cuenta que la señal de IRES potencial es más fuerte al considerar la congruencia entre los contextos celulares y la traducción independiente de cap, en los que podrían operar los IRES potenciales identificados.

Obtuvimos aproximadamente 50 genes candidatos a presentar un IRES en la región 5'UTR, en los siguientes organismos: *Homo sapiens, Macaca mulatta, Pan troglodytes, Rattus norvegicus, Mus musculus, Equus caballus* y *Canis lupus familiaris*. Esperamos que no todos los IRES predichos sean funcionales al momento de realizar un estudio experimental, pero que una gran mayoría presente actividad durante la traducción independiente de cap. La predictibilidad en la búsqueda de nuevos candidatos a IRES aumentó al considerar la probabilidad de encontrar un IRES dada la probabilidad de las 5UTRs ortólogas, a través de la probabilidad conjunta.

El presente estudio contribuye a la descripción de IRES celulares putativos no identificados previamente, bajo un esquema evolutivo, en el cual, los genes que presenta una especie ancestral pueden conservarse en las especies que de ella derivan; cuando esto ocurre, se dice que tales genes son ortólogos. La hipótesis de este estudio extendió este concepto hasta la conservación entre distintos organismos filogenéticamente relacionados, en la regulación de la traducción independiente de cap de los mRNAs vía un segmento estructurado de RNA durante condiciones de estrés o limitantes para la célula.

Perspectivas

Los resultados obtenidos en este estudio abren nuevos planteamientos y preguntas, cuyas respuestas, eventualmente, podrían contribuir a una comprensión más completa de la regulación de la traducción independiente de cap, vía IRES en organismos eucariontes. Estos planteamientos incluyen:

1. Caracterizar experimentalmente, mediante el uso de construcciones bicistrónicas (Graber *et al*., 2006) y perfiles de asociación de ribosomas (Esposito *et al*., 2010), los genes candidatos a presentar un IRES celular en su región 5' UTR obtenidos en la predicción hecha en este estudio.
2. Automatizar la metodología, para que considere las nuevas actualizaciones en la anotación de los genomas con los que se trabajó en la presente tesis, así como la ampliación de la predicción, ya que ésta nos permitiría poder tener un valor de *P* significativo, usando la vasta cantidad de información genómica que actualmente se encuentra disponible.
3. Usar la metodología empleada en este estudio, considerando subgrupos en donde la cercanía evolutiva sea mayor, con el fin de encontrar con una ´probabilidad mayor la señal de IRES celulares. Se ha reportado por ejemplo, que los IRES celulares presentes en *S. cerevisiae,* tienden a presentar valores bajos de energía libre, por lo cual es posible establecer la misma metodología empleada para los genomas mamíferos, pero con el grupo de genomas de levaduras y posiblemente otros genomas fúngicos.
4. A partir del punto anterior, es posible hacer genómica comparativa mediante el uso de árboles filogenéticos que nos permitan dilucidar la conservación de la regulación de la expresión de los genes que presuntamente presentan un IRES celular en la región 5'UTR, y si existen otros mecanismos de regulación genética que sean explotados entre los diferentes organismos para contender con contextos celulares similares.
5. Considerar en el análisis la existencia de motivos en secuencia primaria, en el mRNA a través del uso de programas como MEME (Multiple Em for Motif Elicitation. Timothy *et al.,* 1994) y MAST (Motif Alignment and Search Tool. Timothy *et al.,* 1998), que nos permitan obtener señales adicionales de conservación en los genes candidatos a presentar un IRES en su región 5'UTR, con el fin de aumentar la predictibilidad de nuestra metodología. Aunado a lo anterior se puede implementar la búsqueda de motivos en secuencia primaria y/o motivos estructurales en los IRES que han sido previamente reportados, como lo son los motivos de reconocimiento de los factores ITAFs y posiblemente presentes en las 5'UTRs de los genes que potencialmente pueden presentar un IRES, reportados en este estudio. Sin lugar a duda, la descripción de nuevos ITAFs,

así como de los motivos en secuencia primaria que éstos reconocen, será de gran utilidad para enriquecer con dicha característica, la predictibilidad de nuestro método.

Referencias

1. Allam H., Ali M. (2010). Initiation factor eIF2-independent mode of c-Src mRNA translation occurs via an internal ribosome entry site. *Journal of Biological Chemistry*. **285**: 5713-5725.

2. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. **215**: 403–410.

3. Bailey T.L., Elkan C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. **2**:28-36.

4. Bailey T.L., Gribskov M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*. **14**:48-54.

5. Baird S.D., Lewis S.M., Turcotte M., Holcik M. (2007) A search for structurally similar cellular internal ribosome entry sites. *Nucleic Acids Research.* **35**: 4664-4677.

6. Baird S.D., Turcotte M., Korneluk R.G., Holcik M. (2006) Searching for IRES. *RNA.* **12**: 1755-1785.

7. Baranick B.T., Lemp N.A., Nagashima J., Hiraoka K., Kasahara N., Logg C.R. (2008) Splicing mediates the activity of four putative cellular internal ribosome entry sites. *PNAS.* **12**: 4733-4738.

8. Castrigniano T., D' Antonio M., Anselmo A., Carrabino D., D'Onorio De Meo A., D'Erchia A.M., Licciulli F., Mangiulli M., Mignone F., Pavesi G., Picardi E., Riva A., Rizzi R., Bonizzoni P., Pesole G. (2008). ASPicDB: A database resource for alternative splicing analysis. *Bioinformatics* **24**:1300-1304.

9. Chappell S.A., LeQuesne J.P., Paulin F.E., deSchoolmeester M.L., Stonoley M., Soutar R.L., Ralston S.H., Helfrich, M.H. Willis A.E. (2000) A mutation in the c-myc-IRES leads to enhanced internal ribosome entry in multiple myeloma: a novel mechanism of oncogene de-regulation. *Oncogene*. **19**: 4437-4440.

10. Chappell S.A., LeQuesne J.PC., Paulin F.E.M., deSchoolmeester M.L., Stoneley M, Soutar R.L., Ralston S.H., Helfrich M.H., Willis A.E. (2000). A mutation in the c-myc-IRES leads to enhanced internal ribosome entry in multiple myeloma: A novel mechanism of oncogene de-regulation. *Oncogene*. **19**: 4437-4440.

11. Chatterjee S., Pal J.K. (2009). Role of 5´- and 3´- untranslated regions of mRNAs in human diseases. *Biology of the Cell*. **101**: 251-262.

12. Chenna R., Sugawara H., Koike T., Lopez R., Gibson T.J, Higgins D.G., Thompson J.D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*. **31**: 3497-3500.

13. Davuluri R.V., Suzuki Y., Sugano S., Zhang M.Q. (2000). CART classification of Human 5' UTR sequences. *Genome Research.* **10**:1807-1816.

14. Dinkova T.D., Zepeda H., Martínez-Salas E., Martínez L.M., Nieto-Sotelo J., de Jiménez E.S. (2005). Cap-independent translation of maize Hsp101. *Plant Journal*. **41**: 722-731.

15. Esposito A.M., Mateyak M., He D., Lewis M., Sasikumar A.N., Hutton J., Copeland P.R., Kinzy T.G. (2010). Eukaryotic Polyribosome Profile Analysis. *Journal of Visualized Experiments*. **40**: 1-4.

16. Fang G., Bhardwaj N., Robilotto R., Gerstein M.B. (2010). Getting started in gene orthology and functional analysis. *PLoS.* **6**: 1-8.

17. Filbin M.E., Kieft J.S. (2009). Toward a structural understanding of IRES RNA function. *Current Opinion in Structural Biology.***19**: 1-10.

18. Fitzgerald K.D., Semler B.L. (2009). Bridging IRES elements in mRNAs to the eukaryotic translation apparatus. *Biochimica et Biophysica Acta.***9-10**:518-528.

19. Graber T.E., Lewis S.M., Holcik M. (2006). An approach to whole-genome identification IRES elements. *Current Genomics.***7**:205-212.

20. Grillo G., Turi A., Licciulli F., Mignone F., Liuni S., Banfi S., Gennarino V.A., Horner D.S., Pavesi G., Picardi E., Pesole G. (2009). UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untraslated regions of eukaryotic mRNAs. *Nucleic Acids Research* **38**: 1-6.

21. Hellen C.U.T., Sarnow P. (2001). Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes & Development.* **15**:1593-1612.

22. Holcik M., Sonenberg N. (2005). Translational control in stress and apoptosis. *Nature Reviews Molecular Cell Biology* **6**: 318-327.

23. Holcik M., Xia X. (2009). Strong eukaryotic IRESs have weak secondary structure. *PLoS ONE.* **4**: 1-3.

24. Hofacker I.L., Fontana W, Stadler P.F., Bonhoeffer S., Tacker M., Schuster P. (1994). Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f. *Chemie* **125**: 167-188.

25. Jang S.K., Kräusslich H.G., Nicklin M.J.H., Duke G.M., Palmenberg A.C., Wimmer E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during *in vitro* translation. *Journal of Virology.* **62**: 2636-2643.

26. Johannes G., Carter M.S., Eisen M.B., Brown P.O., Sarnow, P. (1999). Identification of eukaryotic mRNAs that are translated at reduced cap-binding complex eIF4F concentrations using a cDNA microarray. *PNAS*. **96**: 13118–13123.

27. Kanehisa M., Goto S., Kawashima S., Okuno Y., Hattori M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Research*. **32**: 277-280.

28. Kim T.Y., Jong H., Song S., Dimtchev A., Jeong S., Lee J.W., Kim T., Kim N., Jung M., Bang Y. (2003). Transcriptional silencing of the DLC-1 tumor suppressor gene by epigenetic mechanism in gastric cancer cells. *Oncogene.* **22**: 3943–3951.

29. King H.A., Cobbold L.C., Willis A.E. (2010). The role of IRES *trans*-acting factors in regulating translation initiation. *Biochemical Society Transactions.* **38**: 1581-1586.

30. Komar A.A., Hatzoglou M. (2011). Cellular IRES-mediated translation: The war of ITAFs in pathophysiological states. *Cell Cycle.* **10**: 229-240.

31. Komar A.A., Hatzoglou M. (2005) Internal ribosome entry sites in cellular mRNAs: Mystery of their existence. *Journal of Biological Chemistry.* **280**: 23425-23428.

32. Le S.Y., Maizel J.V. (1997). A common RNA structural motif envolved in the internal initiation of translation of celllular mRNAs. **25**: 362-369.

33. Lewis S.M., Holcik M. (2008). For IRES *trans*-acting factors, it is all about location. *Oncogene*. **27**: 1033-1035.

34. Lewis S.M., Veyrier A., Hosszu Ungureanu N., Bonnal S., Vagner S., Holcik M. (2007). Subcellular relocalization of a *trans*-acting factor regulates XIAP IRES-dependent translation. *Molecular Biology of the Cell.* **18**: 1302-1311.

35. Li W., Godzik A. (2006). Cd-hit: a fast program for clusterinf and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658-1659.

36. Macejak D.G., Sarnow P. (1991). Internal initiation of translation mediated by the 5′ leader of a cellular mRNA. *Nature.* **353**: 90–94.

37. Martínez-Salas E., Ramos R., Lafuente E., López de Quinto S. (2001). Functional interactions in internal translation initiation directed by viral and cellular IRES elements. *Journal of General Virology*. **82**: 973-984.

38. Mathews M.B., Sonenberg N., Hershey J.W in *Translational Control of Gene Expression* (eds Mathews M.B., Sonenberg N., Hershey J.W) . (Cold Spring Harbor Laboratory, New York, 2000).

39. Merino E., Yanofsky C. (2005). Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends in Genetics.* **21**: 260- 264.

40. Mignone F., Pesole G. (2007). 5'-UTRs and Regulation. *Encyclopedia of Life Sciences.* 1-4.

41. Mokrejs M., Vopálensky V., Kolenaty O., Masek T., Feketová Z., Sekyrová P., Skaloudová B., Kriz V., Pospísek M. (2006) IRESite: the database of experimentally verified IRES structures (www.iresite.org). *Nucleic Acids Research.* **34**: 125- 130.

42. Morris M.J., Negishi Y, Pazsint C, Schonhoft J.D., Basu S. (2010). An RNA G-Quadruplex is essential for cap-independent translation initiation in human VEGF IRES. *Journal of Chemical American Society*. **132**: 17831-17839.

43. Pan T., Sosnick T. (2006). RNA folding during transcription. *Annual Review of Biophysics and Biomolecular Structure.* **35**: 161- 175.

44. Pelletier J, Sonenberg N. (1988) Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature.* **6180**: 320-325.

45. Pickering, B.M., Willis A.E. (2005). The implications of structured 5' untranslated regions on translation disease. *Seminars in Cell & Developmental Biology.* **16**: 39-47.

46. Pinkstaff J.K., Chapell S.A., Mauro V.P., Edelman G.M., Krushel L.A. (2001). Internal initiation of translation of five dendtritically localized neuronal mRNAs. *PNAS.* **98**: 2770–2775.

47. Qin X.L., Sarnow, P. (2004). Preferential translation of internal ribosome entry site-containing mRNAs during the mitotic cycle in mammalian cells. *Journal of Biological Chemistry*. **279**: 13721–13728.

48. Ringnér M., Krogh M. (2005). Folding free energies of 5'- UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Computational Biology.* **7**: 585-592.

49. Shore P., Sharrocks A.D. (1995). The MADS-box family of transcription factors. *European Journal of Biochemistry, FEMS.* **229**: 1-13.

50. Spriggs K.A., Bushell M., Mitchell S.A., Willis A.E. (2005). Internal ribosome entry segment-mediated translation during apoptosis: the role of IRES-trans-acting factors*. Cell Death and Differentiation*. **12**: 585–591.

51. Spriggs K.A., Cobbold L.C., Jopling C.L., Cooper R.E., Wilson L.A., Stoneley M., Coldwell M.J., Poncet D., Shen Y.C., Morley S.J., Bushell M., Willis A.E. (2009). Canonical initiation factor requirements of the Myc family of internal ribosome entry segments. *Molecular Cell Biology.* **29**: 1565-1574.

52. Stoneley M., Willis A.E. (2004) Cellular internal ribosome entry segments: structures, trans-acting factors and regulation of gene expression. *Oncogene*. **23**: 3200-3207.

53. Tatusov R.L., Koonin E.V., Lipman D.J. (1997). A genomic perspective on protein families. *Science.* **278**: 631-637.

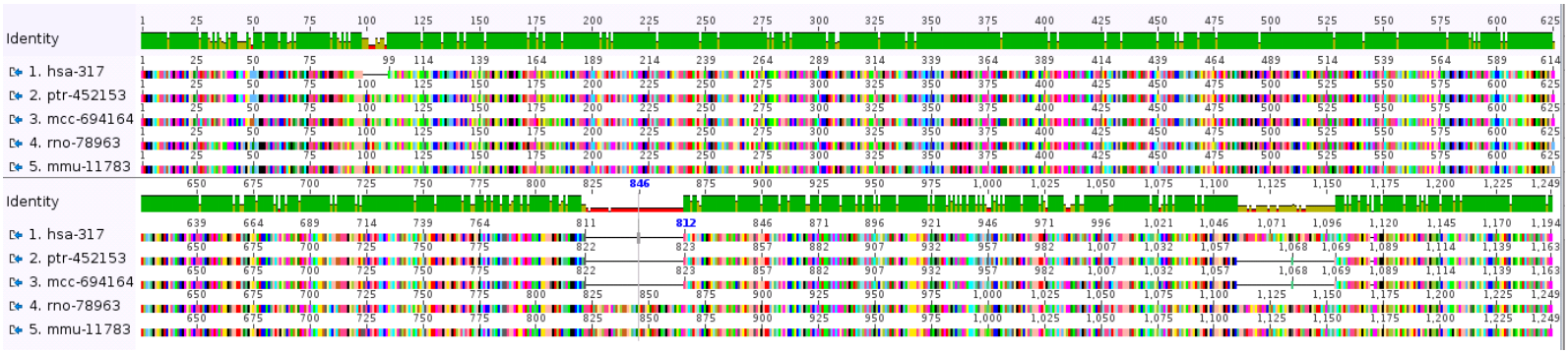
54. Warner J.R. (1999). The economics of ribosome biosynthesis in yeast. *Trends Sci.* **24:** 437-440.

55. R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

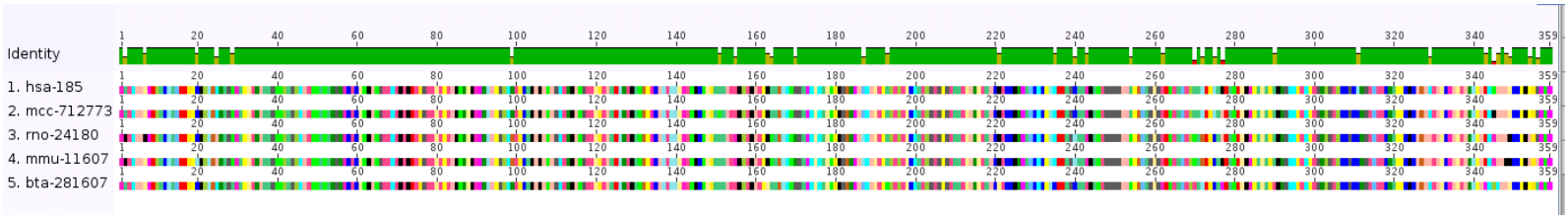
Anexo 1

Alineamientos de aminoácidos correspondientes a los genes ortólogos al gen de humano que presenta un IRES celular. Se presentan las comparaciones de secuencias, hechas por BLASTp, para cada gen de humano que presenta un IRES celular descrito. Las secuencias mostradas provienen de los genomas de *H. sapiens, P. troglodytes, M. mulatta, R. norvegicus* y *M. musculus.*

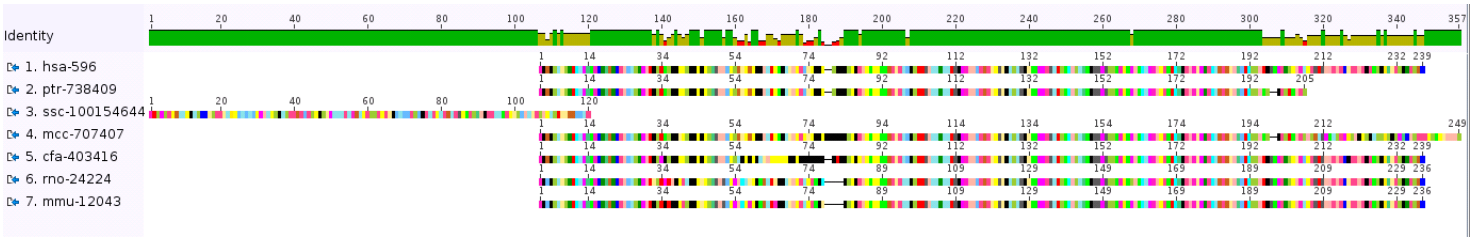
APAF1



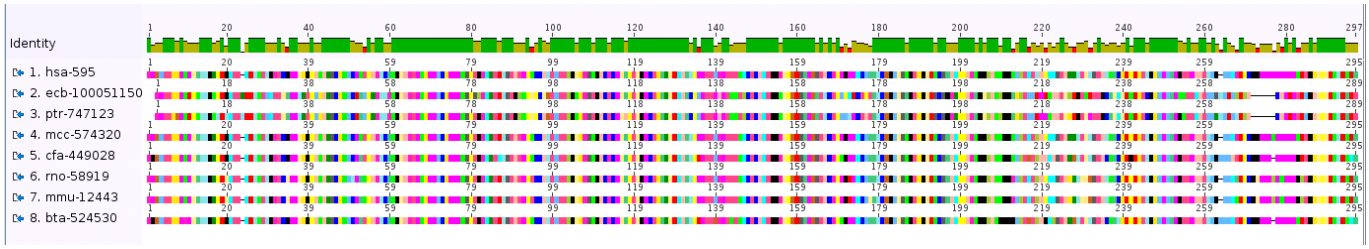
AGTR1



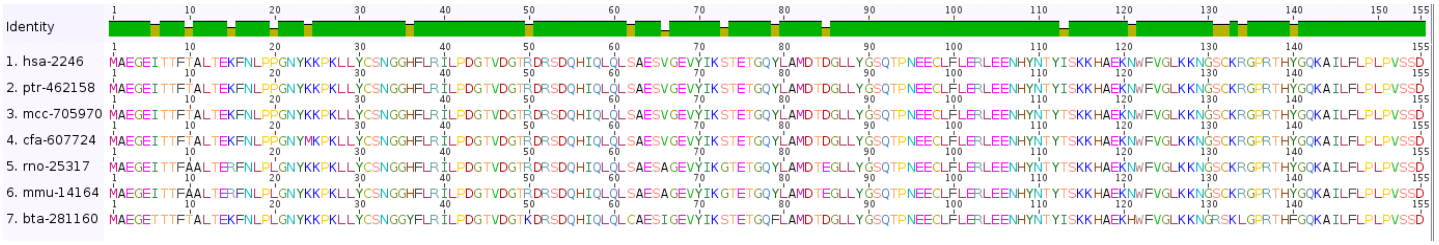
BCL2



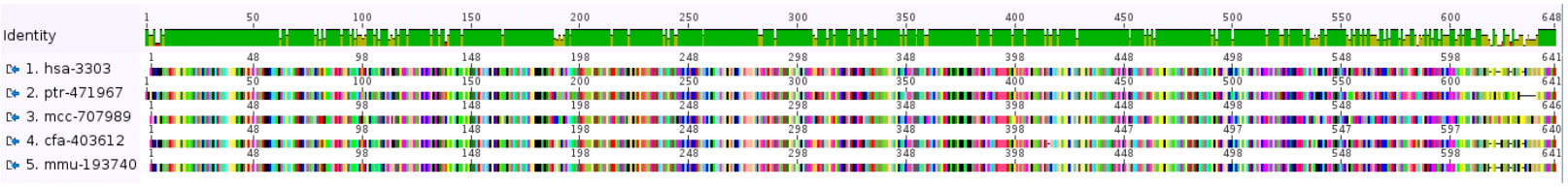
CCND1



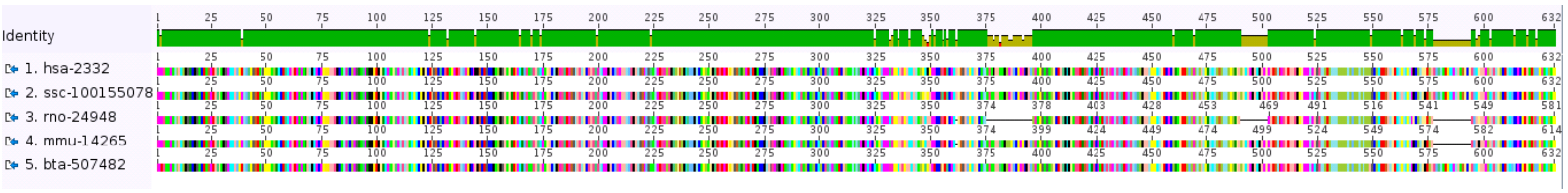
FGF2



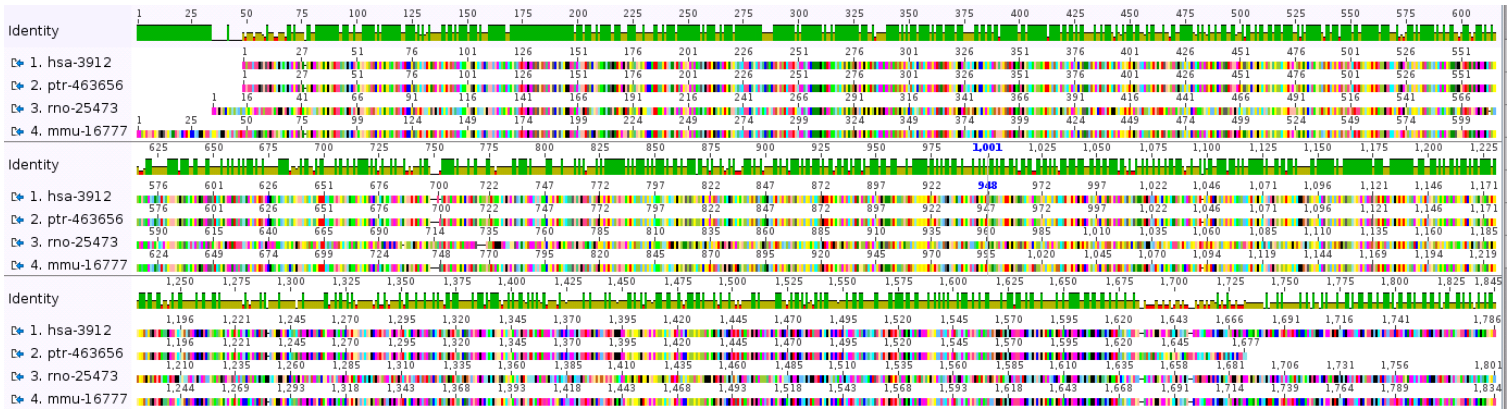
FMR1



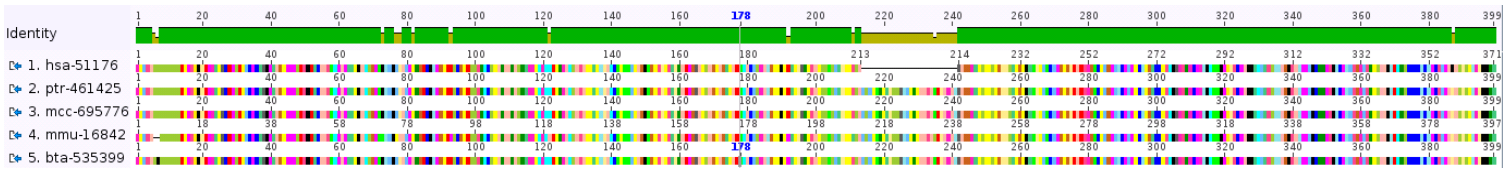
HSP70



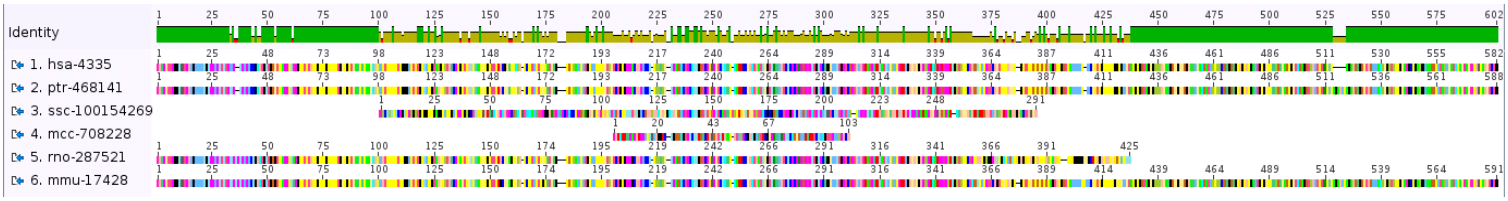
LAMB1



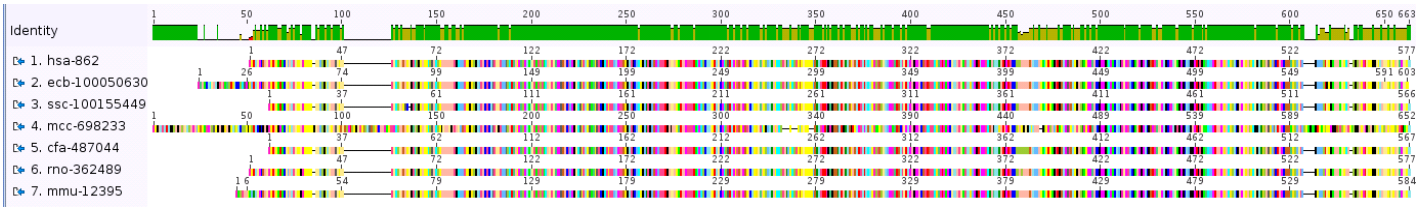
LEF1



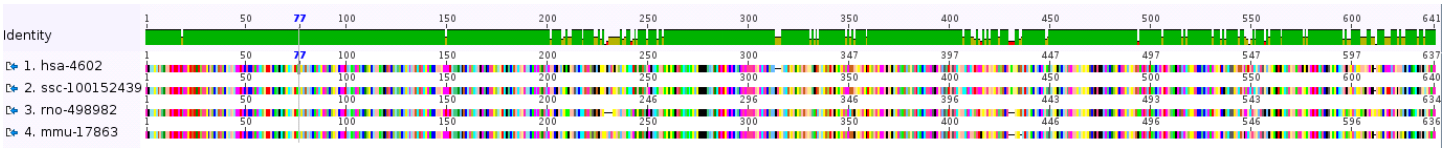
MNT



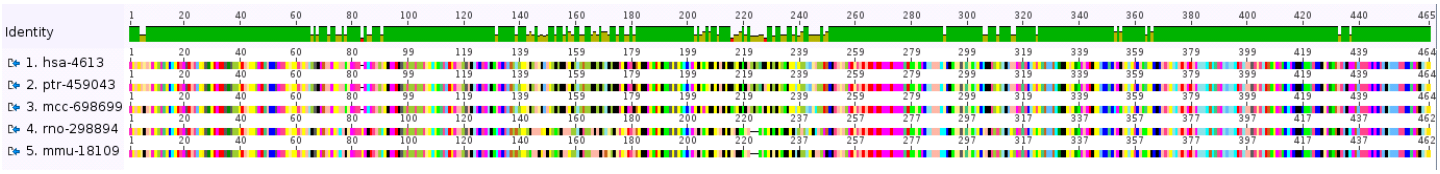
MTG8a



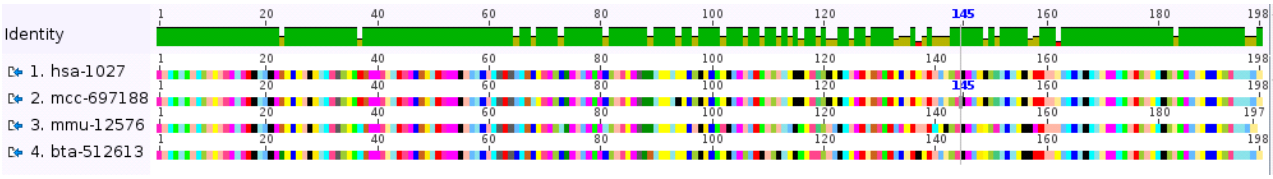
MYB



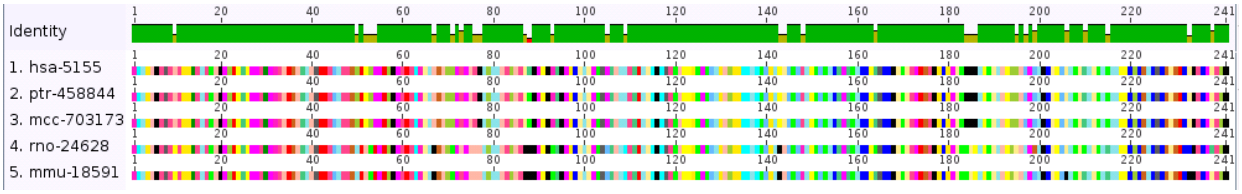
MYCN



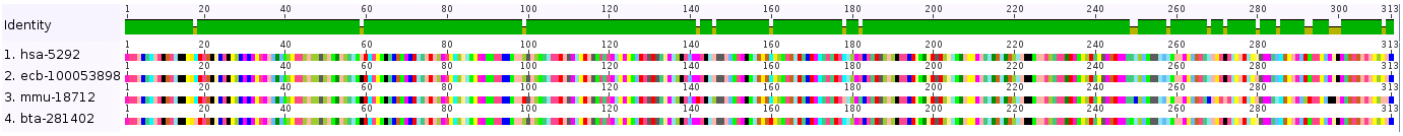
P27kip



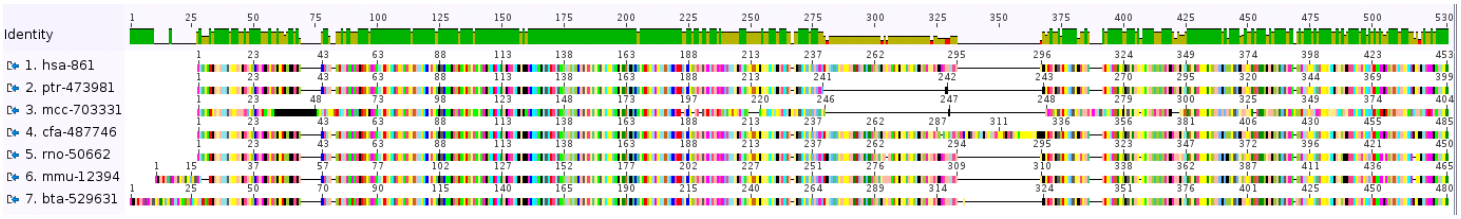
PDGF2/c-sis



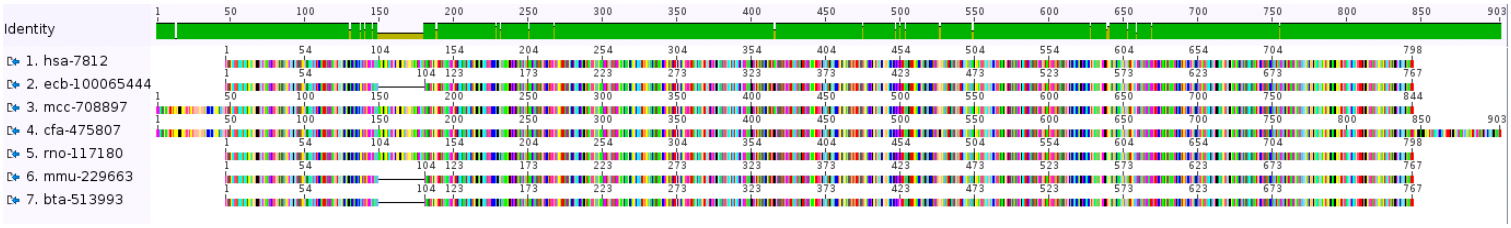
PIM1



RUNX1



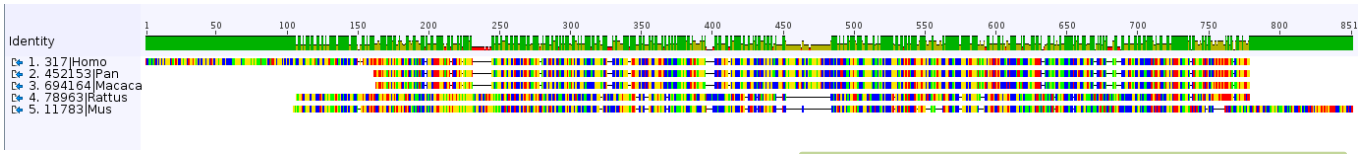
UNR



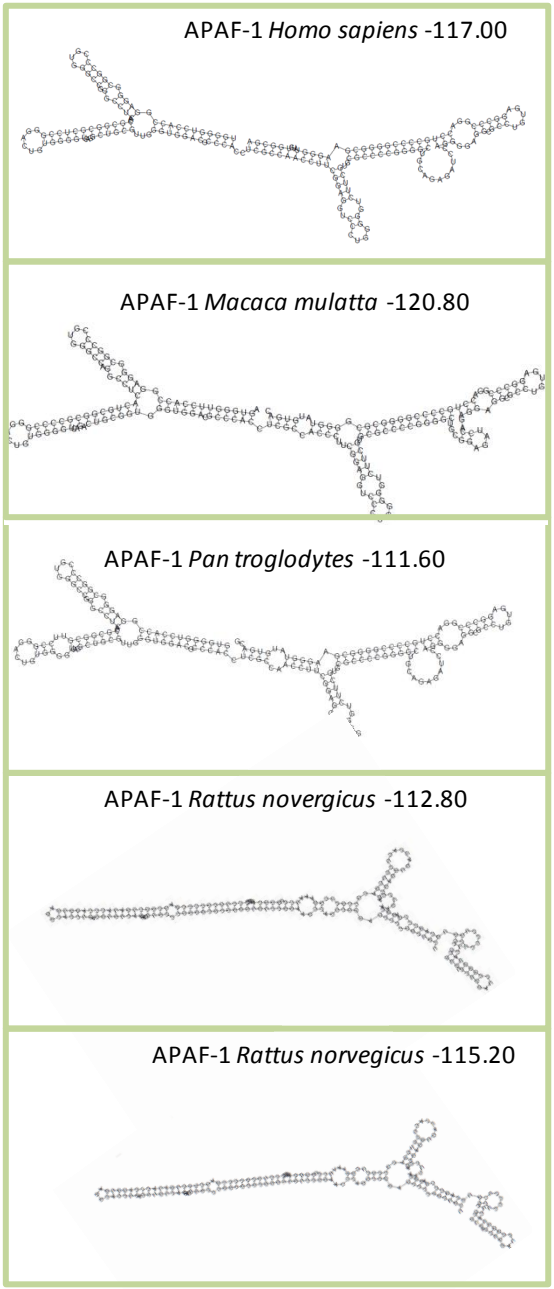
Anexo 2

Análisis de la conservación de secuencia primaria y estructura secundaria de las 5UTRs con IRES celulares descritos y sus 5UTRs ortólogas. Arriba) Alineamiento de las regiones 5'UTRs del gen que contiene un IRES y las 5'UTRs ortólogas correspondientes. Abajo izquierda) Alineamiento de la ventana de análisis de 100 nts. Abajo derecha) Comparación de estructura secundaria y valores de energía libre mínima de las regiones 5'UTRs analizadas.

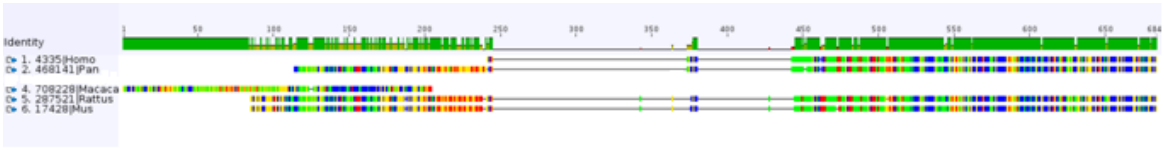
APAF1 y ortólogos



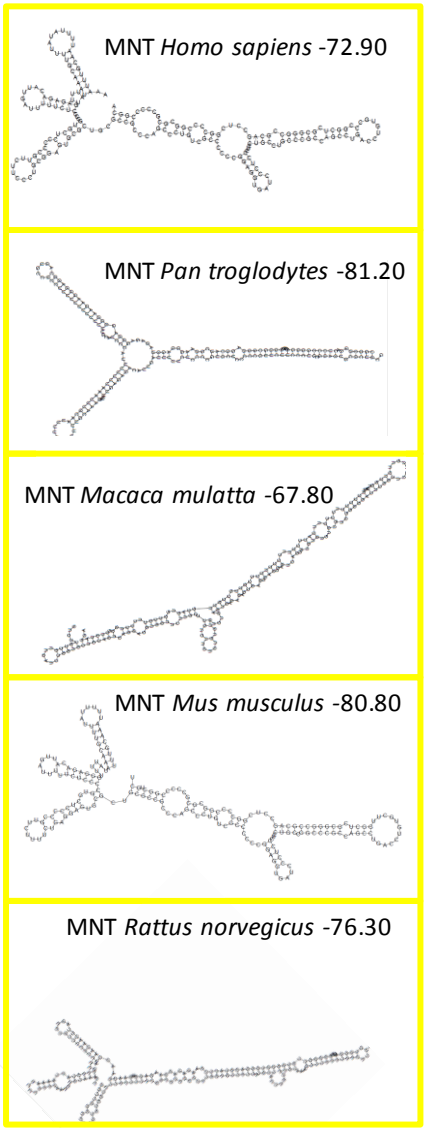

```
                    360        370        380        390        400
317|Homo      AGGCCGGGAA GACCTCCTCC CTTTGTGTCC AGTAGTGGGG TCCACCGGAG
452153|Pan    AGGCCGGGAA GACCTCCTCC CTTTGTGTCC AGGAGTGGGG TCCACCGGAG
694164|Maca   AGGCCGGGGA GCCCTCCTCC CTTTGTGTCC AGGAGTGGGT TCCACCGGAG
78963|Rattu   -GGCTGGAGT GGCC--GTGC TTTTGTGTCC -----TGGAT CCG-------
11783|Mus     -GACTGGAGT GGCC--GTGC TTTTGTGCCC -----TGGGT CCC-------
Clustal Con    * * **    * **   * *  ****** **      ***    *

                    410        420        430        440        450
317|Homo      GGCGGCCCGT GGGCCGGGCC TCACCGCGGC GCTCCGGGAC TGTGGGGTCA
452153|Pan    GGCGGCCCGT GGGCCGGGCC TCACCGCGGC GTTCCGGGAC TGTGGGGTCA
694164|Maca   GGCGGCCCGT GGGCCAGGCC TCACTGCGGC GCCCCGGGAC TGTGGGGTCA
78963|Rattu   GGTACCCTCC --------CT CCCCTGGTGC AGCCCGAGGC ---------A
11783|Mus     GGTACCCTCC --------C- ---CTGGTGC GGCCCGAGGC ---------A
Clustal Con   **   **             *     * *  **    *** * *          *

                    460        470        480        490        500
317|Homo      GGCTGCGTTG GGTGGACGCC CACCTCGCCA ACCTTCGGAG GTCCCTGGGG
452153|Pan    GGCTGCGTTG GGTGGACGCC CACCTCGCCA ACCTTCGGAG GTCCCTGGGG
694164|Maca   GACTGCGGTG GGTGGACGCC CACCTCGCCA CCCTTCGGAG GTCCCTGGGG
78963|Rattu   AGTCCATCGA GGTGATCACT C--CTCGACG CCGCTCGGAA ATCCTGGGCA
11783|Mus     AGCCCACCGA GGTGACCACC C--CTCGACG CCGCTTGGAG ATCCCGGGCA
Clustal Con              ****  * *  *  **** *   *  * ***   ***  **

                    510        520        530        540        550
317|Homo      GTCTTCGTGC GCCCCGGGGC TGCAGAGATC CAGGGGAGGC GCCTGTGAGG
452153|Pan    GTCTTCGTGC GCCCCGGGGC TGCAGAGATC CAGGGGAGGC GCCTGTGAGG
694164|Maca   GTCTTCGTGC GCCCCGGGGC TGCGGAGATC CAGAGGAGGC GCCTGTGAGG
78963|Rattu   TCCCCCCTGC GCCCCGAG-C GGCTGATACC CAGGTGAGGC ACCTGAGGTG
11783|Mus     TCCACCCTGC GCCCCGAG-C AGCTGATACC CAGG------ ----GAGGTG
Clustal Con     *  * *** ****** * *  ** ** * * ***           * *  *

                    560        570        580        590        600
317|Homo      CCCGGACCTG CCCCGGGGCG AAGGGTATGT GGCGAGACAG AGCCCTGCAC
452153|Pan    CCCGGACCTG CCCCGGGGCG AAGGGTATGT GACGAGACAG AGCCCTGCAC
694164|Maca   CCCGGACCTG CCCCGGGGCG CGGGGTATGT GACGGGACAG AGCCCCACAC
78963|Rattu   TCAGGACCTG CCC-GGGGCG CGGGGT---G TCCGGAAGCC AGGCGGGAGC
11783|Mus     TCAGGACCTG CCC-GGGGCG CGGGGT---C GCCGGAAGCC AGGCGGGAGC
Clustal Con    * ******* *** ******   ****       **  *    ** *     *

                    610        620        630        640        650
317|Homo      CCCTAATTCC CGGTGGAAAA CTCCTGTTGC CGTTTCCCTC CACCGGCCTG
452153|Pan    CCCTAATTCC CGGTGGAAAA CTCCTGTTGC CGTTTCCCTC CACCGACCTG
694164|Maca   CCCTGATACC CGGTGGAAAA CCTCTGTTGC CGTTTCCCTC CACCGGCCTG
78963|Rattu   CCCGGCTGCT TTTTGGCAAT CGATTCTCAT CTGTGACCCT CCCCGAGCTG
11783|Mus     CCCGGCTGCT TTCTGGCAAT CTAGTCTCAT AAGTGACCCT CCCTGGGCTG
Clustal Con   ***   * *     *** **  *   * *       *  **   * * *  ***
```



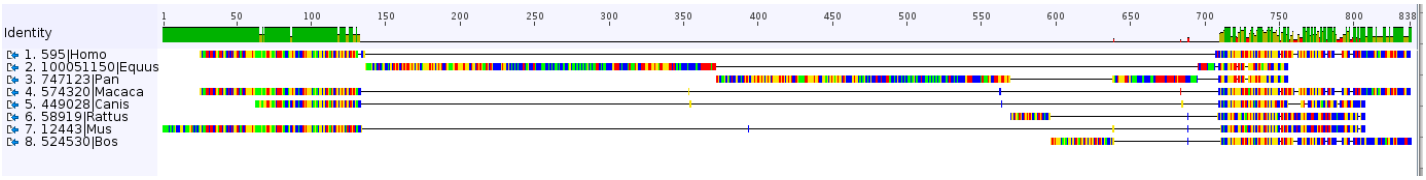
MNT y ortólogos




```
....|....| ....|....| ....|....| ....|....| ....|....|
                10         20         30         40         50
4335|Homo    ---------- ---------- ---------- ---------- ----------
468141|Pan   ---------- ----------C CCTTGTTGTG TGTGTCGGTG
287521|Rat   GGCGGCGGCC GGGGGCAGCG CGGCTCCTTC CCTTGTTGTG TGTGTCGGTG
17428|Mus    GGCGGCGGCC GGGGGCAGCG CGGCTCCTTC CCTTGTTGTG TGTGTCGGTG
708228|Mac   ---------- ---------- ---------- ---------- ----------
Clustal Co

                ....|....| ....|....| ....|....| ....|....| ....|....|
                       60         70         80         90        100
4335|Homo    ---------- ---------- ---------- ---------- ----------
468141|Pan   CCTCCTCGCC ATCTTGTTGC AAAGCCCTTT CTTGTCGGCG GGACTCCCGG
287521|Rat   CCTCCTCGCC ATCTTGTTGC AAAGCCCTTT CTTGTCGGCG GGACTCC-GG
17428|Mus    CCTCCTCGCC ATCTTGTTGC AAAGCCCTTT CTTGTCGGCG GGACTCC-GG
708228|Mac   ---------- ---------- ---------- ---------- ----------
Clustal Co

                ....|....| ....|....| ....|....| ....|....| ....|....|
              110        120        130        140        150
4335|Homo    ---------- ---------- ---------- ---------- ----------
468141|Pan   GGGCCGCGGG GCGGGAGGCA TCGGAAGGGA GGTAGAGAGG GAGGGGAAGA
287521|Rat   GGGCCGCGGG GCGGGAGGCA TCGGAAGAGA GGAAGAGGAA GAAGG-AAGG
17428|Mus    GGGCCGCGGG GCGGGAGGCA TCGGAAGGGA GGAAGAGGAA GAAGG-AAGG
708228|Mac   ---------- ---------- ---------- ---------- ----------
Clustal Co

                ....|....| ....|....| ....|....| ....|....| ....|....|
              160        170        180        190        200
4335|Homo    ------GCAG TGCCGCCTTT TTTTTTTTTT TGCATCCCA- TTTTTTTAAA
468141|Pan   AGGGAGGCAG TGCCGCCTTT TTTTTTTTTT -GCATCCCA- TTTTTTTAAA
287521|Rat   AAGAAGGCAG TGC-GCCTTT TTTATTTTTT -GCATCAAAA TTTTTTTTAA
17428|Mus    AAGAAGGCAG TGCCGCCTTT TTTATTTTTT -GCATCAAAA TTTTTTT-AA
708228|Mac   ---------- ---------- ---------- ---------- ----------
Clustal Co

                ....|....| ....|....| ....|....| ....|....| ....|....|
              210        220        230        240        250
4335|Homo    TTTGCAATTT TATATTTTGC AAATATTTTG AGAGACATTG ATTTTTCTCC
468141|Pan   TTTGCAATTT TATATTTTGC AAATATTTTG AGAGACATTG ATTTTTCTCC
287521|Rat   TTTGCAAATT TTTATTTTGC AAATATTTTG G-AGACATTG ATTTTTCTCC
17428|Mus    TTTGCAAATT TTTATTTTGC AAATATTTTG GGAGACATTG ATTTTTCTCC
708228|Mac   -GTGCCGCCG TTGATC---C GGTGCTGCAG TGAGGAGGTG GTTCTTGCCC
Clustal Co                      *          *   **                *

                ....|....| ....|....| ....|....| ....|....| ....|....|
              260        270        280        290        300
4335|Homo    CCGTGCTCCC CCGTTCTTCC CTGCGGAGTG CGCTGCGCCG CCCAG----C
468141|Pan   CCGTGCTCCC CCGTTCTTCC CTGCGGAGTG CGCTGCGCCG CCCAG----C
287521|Rat   CCGTGCTCCC CCGTTCTTTC CTGAGGAGTG CGCTGCTCCG CCCAG----C
17428|Mus    CCGTGCTCCC CCGTTCTTTT CTGAGGAGTG CGCTGCGCCG CCCAG----C
708228|Mac   GTGT-TGTGT GTGTGTGTGA GTGAGAGAGC GAGTGAGTGA GTGAG----T
```



CCND1 y ortólogos




```
                  10         20         30         40         50
595|Homo    ---------- ---------- ---------- ---------- --GCTTAACA
574320|Mac  ---------- ---------- ---------- ---------- --GCTTAACA
12443|Mus   ---------- -------TTT TCTCTGCCCG GCTTTGATCT CTGCTTAACA
100051150|  -GTTTGCGTC ACTGCCTCGC AGAGCAGAGA AGAGCGAGCT GGGGAGAGCG
747123|Pan  AGCTTGCGTC ACCGCCT--C AGAGCGGAGA AGAGCGAGCA GGGGAGAGCG
Clustal Co

                  60         70         80         90        100
595|Homo    ACAGTAACGT CACACGGACT ACAGGGGAGT TTTGTTG-AA GTTGCAAAGT
574320|Mac  ACAGTAACGT CACACGGACT ACAGGGGAGT TTTGTTG-AA GTTGCAAAGT
12443|Mus   ACAGTAACGT CACACGGACT ACAGGGGAGT TTTGTTG-AA GTTGCAAAGT
100051150|  AGACCAGTTT GAAGGGGAGG GCAGGCA-GA GTTAGCAGCC CCCGAGGCGC
747123|Pan  AGACCAGTTT TAAGGGGAGG ACCGGTGCGA GTGAGGCAGC CCCGAGGCTC
Clustal Co

                 110        120        130        140        150
595|Homo    CCTGGAGCCT CCAGAGGGCT GTCGGCGCAG TAGCAG-CGA GCAGCAGAGT
574320|Mac  CCTGGAGCCT CCAGAGGGCT GTCGGCGCAG TAGCAG-CGA GCAGCAGAGT
12443|Mus   CCTGCAGCCT CCAGAGGGCT GTCGGCGCAG TAGCAG-AGA GCTACAGACT
100051150|  T-CTCTCCCA CTGCCCATCC CTTTCCTCTC TTCTCCTCCT GCTTCTCTCT
747123|Pan  TGCTCGCCCA CCACCCAATC CTCGCCTCTC TTCTGCTCCA CCTTCTCTCT
Clustal Co                          * *   *          *  *    *

                 160        170        180        190        200
595|Homo    CCGCACG--- -----CTCCG GCGAGGGGCA GAAGAGCGCG AGGGAGCGCG
574320|Mac  CCGCACG--- -----CTCCG GCGAGGGGCA GAATAGCGCG AGGGAGCGCG
12443|Mus   CCGCGCG--- -----CTCCG GAGACCGGCA GTACAGCGCG AGGCAGCGCG
100051150|  CTGCTCTTAA CTCTTCCCCG AAAACTCCCT ATTTAGCCAA AGGAAGGAGG
747123|Pan  CTGCCCTCAC CTCTCCCCCG AAAACCCCCT ATTTAGCCAA AGGAAGGAGG
Clustal Co * ** *          * ***    *           ***    *** **   *

                 210        220        230        240        250
595|Homo    GGGCAG--CA GAAGCGAGAG CCGAGCGCGG ACCCAGCCA- GGACCCACAG
574320|Mac  GGGCAG--CA GAAGCGAGAG CCGAGCGCGG ACCCAGCCC- GGACCCACAG
12443|Mus   CGTCAG---- -CAGCCGCCA CCG-----GA GCCCAACCG- AGACC-ACAG
100051150|  TAAGGG---- -GAACACTCT CCCCTCCCCC TCCAAAAAAA TAAAACAAAA
747123|Pan  TCAGGG---- -GAACGCTTT CCCCTCCCCT TCCAAAAAAC AAAAACAGAA
Clustal Co               * *      **          **            *

                 260        270        280        290        300
595|Homo    ----CCCTCC CCAGCTGCCC AGGAAGAGCC CCAGCC---- ----------
574320|Mac  ----CCCTCC CTAGCTGCCC AGGCAGAGCC CCAGCC---- ----------
12443|Mus   ----CCCTCC CCAGACGGCC GCGCC----- ---------- ----------
100051150|  -----CTTTT CCAGTCCGGA GAAAGCAGGG GAACGAGTGG GCACCGGGCT
747123|Pan  AAACCCTTTT CCAGGCCGGG GAAAGCAGGA GGGAGAGGGG CCGCCGGGCT
```