UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

PREDICCIÓN DE SITIOS DE ENTRADA INTERNOS AL RIBOSOMA EN EUCARIONTES

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Doctor en Ciencias

PRESENTA:
ESTEBAN PEGUERO SÁNCHEZ

TUTOR PRINCIPAL
DR. ENRIQUE MERINO PÉREZ
Instituto de Biotecnología, UNAM

MIEMBROS DEL COMITÉ TUTOR
DRA. TZVETANKA DIMITROVA DINKOVA
Facultad de Química, UNAM
DR. ÁNGEL FERNANDO KURI MORALES
Departamento Académico de Computación, ITAM

Ciudad de México. Enero, 2017
CIUDAD UNIVERSITARIA

Índice

El presente trabajo de investigación se realizó en el Departamento de Microbiología Molecular del Instituto de Biotecnología de la UNAM bajo la dirección del Dr. Enrique Merino Pérez con financiamiento de la beca para doctorado CONACyT 384858.

Agradecimientos

Esta etapa de mi vida que ahora culmina ha sido una de las más satisfactorias, enriquecedoras e intensas. Muchas personas tanto en mi pasado como en el presente han contribuido de diversas maneras para lograr esta meta tan importante para mi.

Agradezco a **Luz María Sánchez** y **Javier Peguero** por todas las enseñanzas brindadas con el ejemplo y la convivencia diaria en el tiempo que estuvimos juntos.

A **Enrique Merino** por su apoyo incondicional, conocimiento y ejemplo de liderazgo. **Lucas**, tu laboratorio es uno de los lugares en donde más he crecido intelectualmente y eso es gracias a la gran labor que desempeñas como académico y docente. Muchas gracias por aceptarme en tu laboratorio y trabajar conmigo hombro con hombro, para mi fue una gran aventura llena de satisfacciones y desarrollo profesional y personal.

A **Carlos Camacho** por estar conmigo en cada peldaño a lo largo de los últimos 27 años. **Carlos**, muchas gracias por compartir conmigo tu experiencia de vida y mostrarme el camino hacia el autoconocimiento.

Agradezco a **Maricarmen Pintado** por compartir su vida conmigo. **Mary**, gracias por ser mi amiga, cómplice y compañera en tantos aspectos desde los más cotidianos hasta los más profundos. Hemos alimentado el yo, el tú y el nosotros para lograr la hermosa relación que tenemos.

Fabiola Pintado y **Guadalupe Martínez**, gracias por introducirme al mundo de los determinantes sociales y el empoderamiento. He aprendido mucho de ustedes y fue todo un placer el compartir un hogar.

A **Teresa Romero** por ser una excelente colega y amiga. He disfrutado mucho todas las actividades en las que hemos participado juntos, clubs, clases, proyectos, pelear por los certificados y tantas otras.

A **Rosa María Gutiérrez** por compartir conmigo su experiencia como investigadora de gran calidad.

Agradezco a **Ricardo Ciria** por ser el mejor profesor de programación que he conocido y siempre estar dispuesto a ayudarme. El aprendizaje que obtuve en tus clases me ha sido increíblemente útil, los conocimientos que me trasmitiste marcaron mi vida de manera muy positiva.

Agradezco a los miembros de mi comité tutoral, **Ángel Kuri** por sus valiosas enseñanzas que me permitieron entrar al mundo de las redes neuronales y la minería de datos y por sus valiosas aportaciones a mi trabajo y **Tzvetanka Dimitrova** por el valioso enfoque biológico que agregó a mi proyecto y sus comentarios críticos.

Estoy en deuda con los miembros de mi jurado **Alejandro Sánchez**, **Jorge Folch**, **Blanca Taboada**, **Martha Vázquez** y **Armando Mendoza** por su valiosas aportaciones a mi formación y a mi proyecto.

Muchas gracias **Mario Treviño** por compartir conmigo tu extensa experiencia en el ámbito académico, me ha sido muy útil como fuente de inspiración y guía.

Gracias a mis amigos **Emiliano Cantón** y **Jorge González** por tantos momentos compartidos y su amistad entrañable, **Rodrigo Arreola** por su incondicionalidad y **Daniel Vázquez** por mostrarme lo bueno que se puede llegar a ser para programar y por participar juntos en tantos proyectos.

Agradezco mis compañeros del Jiu-jitsu por tantas horas de entrenamiento y diversión, **Juan Pablo González** por todas las charlas entrañables, las roladas, los seminarios juntos y abrirme las puertas a tu mundo, **Luis Cueto** por todos los momentos desde que llegamos a Cuernavaca y comenzamos una bella amistad (y fuimos los únicos en presentar el examen de traje) y **Ernesto Gutiérrez** por todo el apoyo y mostrarme el estilo de vida del Jiu-jitsu.

Gracias a todos los integrantes del **laboratorio Merino** por compartir conmigo sus conocimientos y su valiosa retroalimentación en mis proyectos.

Un agradecimiento especial a **Claudia Treviño** por su invaluable ayuda para agilizar los trámites y en la escritura de proyectos. **Claudia**, además de ser una excelente investigadora tienes una calidad humana increíble y realmente haces una diferencia en el posgrado del IBT.

A **Antonio Bolaños**, por siempre hacer más allá de lo que te "toca" y ayudarme con incontables trámites.

A **Gloria Villa** y **Jalil Saab** por la gran labor que desempeñan para que el posgrado del IBT sea de los mejores.

Muchas gracias **Adrián Vergara** ya que gracias a tu importante labor y excelente servicio pude quedarme en incontables ocasiones a trabajar hasta altas horas de la noche y regresar seguro a casa.

Estoy agradecido con los excelentes profesores que tuve durante mi estancia en este instituto: **Gloria Saab**, **Tonatiuh Ramírez**, **Laura Palomares**, **Sergio Águila**, **Enrique Rudiño**, **Brenda Valderrama**, **Lourival Domingos**, **Marcela Ayala** y muchos otros.

Muchas gracias a **Lorenzo Segovia** por su impulsarme a lograr mis metas en tiempo y compartir conmigo su experiencia.

Al **CONACyT** por el financiamiento recibido para mis estudios y la asignación de la beca mixta –MZO 2017 Movilidad en el extranjero (291062).

Agradezco al Programa de Apoyos de Estudio de Posgrado (**PAEP**) por el financiamiento para asistir a los siguientes eventos:

- XXX Congreso Nacional de la Sociedad Mexicana de Bioquímica. Guadalajara, Jalisco. 2014.
- ECAS Course "Statistical Analysis of Network Data." Munich, Alemania. 2015.
- Estancia de investigación en el grupo del Prof. Rost en el departamento de informática de la Universidad Técnica de Munich. Munich, Alemania. 2016.

1 Resumen

Antecedentes: Los Sitios de Entrada Internos al Ribosoma (IRES) son elementos reguladores del inicio de la traducción en condiciones fisiológicas en donde el inicio canónico o cap-dependiente se encuentra comprometido. Algunos ejemplos de ello son la carencia de nutrientes, respuesta al calor, infecciones virales, mitosis, entre otras condiciones de estrés. Una característica notable de los IRES es que en su mayoría carecen de conservación de la secuencia primaria así como de la estructura secundaria lo cual ha dificultado su identificación mediante el uso de métodos computacionales. En la actualidad, únicamente un número limitado de estos elementos se encuentran verificados experimentalmente.

Resultados: En el presente trabajo se desarrolló un método bioinformático para la clasificación de IRES basado en genómica comparativa utilizando máquinas de soporte vectorial. Para este efecto se emplearon las regiones 5' no-codificantes (5'-UTR) de los mRNAs de 20 organismos no-redundantes pertenecientes al reino de los hongos, en donde este método predijo un total de 6,532 IRES. El análisis de las relaciones funcionales y evolutivas de dichas clasificaciones mostró enriquecimientos significativos en ontología de genes (GO), interacciones funcionales (STRING) y relaciones de homología. Algunos ejemplos sobresalientes de este estudio son la familia de las chaperonas HSP70 que notablemente tiene IRES verificados experimentalmente en otros organismos, como humano y mosca, y la familia de eIF4G, con IRES reportados en ratón y levadura. Estos resultados enfatizan la amplia conservación de estos elementos de regulación. El análisis de las interacciones funcionales de las predicciones en *S. cerevisiae* mostró una red altamente conectada y con topología modular enriquecida en términos GO lo que sugiere una co-regulación traduccional.

Conclusiones: Se desarrolló un método para la clasificación de IRES en hongos que podría ser utilizado para guiar posteriores análisis experimentales y bioinformáticos con el objetivo de incrementar nuestro conocimiento de la regulación traduccional. Este estudio sugiere que los IRES se encuentran conservados evolutivamente y que existen asociaciones funcionales en las proteínas traducidas por este mecanismo.

2 Introducción

2.1 Mecanismos de control traduccional

La regulación de la expresión génica es una de las muchas fuerzas motrices de los procesos evolutivos y de la diferenciación celular [1]. Las variaciones en dicha regulación permiten ampliar la gama de posibilidades con las que los organismos responden a estímulos tanto internos, como externos y por lo tanto facilitan el desarrollo de distintos fenotipos, incluso cuando mutaciones en las regiones codificantes no explican estas adaptaciones [2–4]. Diversos estudios que combinan proteómica y transcriptómica han revelado que el control post-transcripcional podría ser la fuente de hasta el 60% de la variación en las concentraciones de proteína y por lo tanto ser incluso más importante que el control transcripcional [5, 6], especialmente cuando se requieren cambios rápidos en las concentraciones de proteína [7]. Esto último podría explicar la prevalencia de este tipo de regulación en situaciones de estrés [8, 9]. Lo anterior pone de manifiesto que existe una gran cantidad de información biológica codificada en los mecanismos de regulación post-transcripcional.

El proceso de traducción consiste de tres etapas: iniciación, elongación y terminación. Sin embargo, se conoce que el punto donde existe mayor control es en su iniciación [10]. La iniciación de la traducción requiere la formación de una compleja maquinaria molecular conocida como complejo de pre-iniciación (CPI). El CPI está constituido por numerosas proteínas llamadas factores del inicio de la traducción en eucariontes (eIFs), la subunidad 40S del ribosoma y algunos factores auxiliares (esto se esquematiza en la **Figura 1A**). Hasta el momento se conocen trece eIFs y cinco factores auxiliares (revisado en [10]). En particular, se sabe que el eIF4E reconoce la modificación 5'-UTR en los RNA mensajeros (mRNAs) eucariontes. Este reconocimiento es necesario para la formación del CPI en el mecanismo "canónico" o cap-dependiente del inicio de la traducción. Una vez formado el CPI se lleva a cabo un escaneo (**Figura 1A**) del mRNA hasta que se encuentra el codón de inicio y comienza la síntesis de la proteína (revisado en [10]).

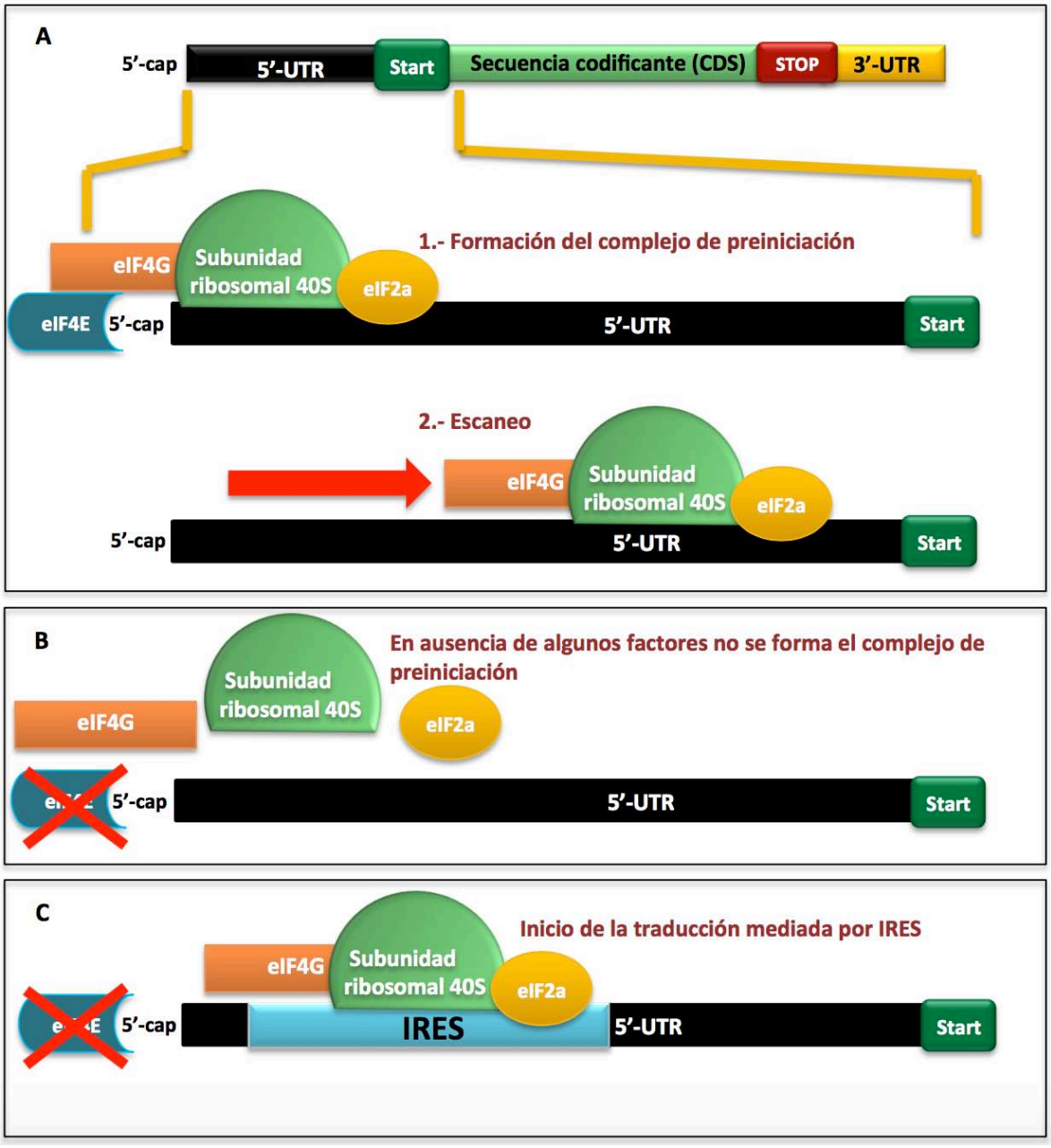

Figura 1. Inicio de la traducción.
A) Inicio de la traducción por el mecanismo canónico (sólo se muestran algunos de los factores necesarios para su formación). **B)** El complejo de preiniciación de la traducción no se forma en ausencia de algunos factores canónicos. **C)** Inicio de la traducción mediado por IRES.

En diversas condiciones de estrés, la célula responde disminuyendo la producción global de proteínas para ahorrar recursos energéticos [11] como por ejemplo durante las infecciones virales [12], las condiciones de hipoxia [13], la restricción de nutrientes (revisado en [8]). En estos escenarios algunos de los factores "canónicos" para el inicio cap-dependiente de la traducción se encuentran limitados y no se favorece la formación

CPI (**Figura 1B**). A pesar de esto, en ambos casos es necesario continuar con la producción de proteínas de respuesta a estrés, ya sea para promover la supervivencia de la célula, o para activar los mecanismos de muerte celular [14]. Los IRES son un tipo de elemento de regulación traduccional que permite la síntesis selectiva de proteínas durante estas circunstancias, aún cuando algunos eIFs canónicos no se encuentren presentes (**Figura 1C**). Una característica particular de los IRES es que permiten la formación del CPI aún sin el reconocimiento de la modificación 5'-cap [15]. Por este motivo el inicio de la traducción mediada por IRES también se conoce como inicio cap-independiente. Por lo tanto, la función reguladora de los IRES se lleva a cabo a nivel de la traducción [16, 17]. Tres excelentes revisiones sobre los mecanismos de control traduccional se pueden encontrar en [7, 10, 18].

El primer IRES se reportó en 1988 después de observar que las regiones 5'-UTR del poliovirus carecían de la modificación cap en su extremo 5' terminal como la mayoría de los mRNAs eucariontes. Posteriormente se demostró que su traducción procedía mediante un mecanismo alternativo de manera cap-independiente [19]. En los años siguientes, se describieron numerosos IRES en otros virus [20–22] donde este elemento de regulación permitía a las proteínas virales ser sintetizadas utilizando la maquinaria del hospedero aún cuando la traducción global se encontraba reprimida debido al proceso de infección [23]. Poco después el primer IRES eucarionte se descubrió en la región 5'-UTR de la chaperona humana BiP debido a que esta proteína seguía expresándose en células infectadas con poliovirus [24]. Se podría definir un IRES como la secuencia especializada en el mRNA que favorece el inicio de la traducción sin depender de la modificación 5'-cap y que permite a la subunidad 40S del ribosoma ser reclutada en la vecindad del codón de inicio sin requerir la etapa de escaneo del mecanismo canónico [15].

De manera general, los IRES se han clasificado en dos grupos: los que dependen de una alta estructuración del mRNA, y aquellos que carecen de ella [25–27]. Sin embargo, no se conocen motivos característicos ni en la estructura ni en la secuencia de los IRES que permitan su identificación a partir de la secuencia [28, 29]. Por este motivo la localización de estos elementos de regulación ha sido sujeta únicamente a métodos experimentales, requiriendo a su vez de diversas pruebas para su confirmación [30].

Aunado a lo anterior, la traducción de genes mediada por IRES es controlada por una amplia gama de diferentes mecanismos que recientemente comenzamos a comprender. Por ejemplo, se conocen diversas proteínas no-canónicas en el inicio de la traducción que interaccionan con los IRES y regulan su función. Estas proteínas son

conocidas como Factores que Actúan en *Trans* sobre IRES (ITAFs por sus siglas en inglés) [15]. Esto explicaría la observación de que algunos IRES sólo contribuyen al proceso de traducción en condiciones celulares particulares o tejidos específicos. Por ejemplo, durante la mitosis, los IRES de Unr, c-*myc* y PITSLREp58 favorecen la traducción 5'-cap dependiente en estas condiciones, mientras que otros IRES no [31]. Otro caso de esto se ha descrito en la apoptosis, donde el IRES de Apaf-1 es activo, mientras que el de XIAP se encuentra inhibido [32]. En condiciones de deficiencia de nutrientes en levadura, algunos elementos ricos en adenina en las 5'-UTRs funcionan como mediadores de la iniciación cap-independiente reclutando la proteína Pab1p que en este caso funciona como un ITAF [27]. Otro ejemplo de ITAF es la proteína de unión a polipirimidinas (PTB por sus siglas en inglés) que funciona como regulador de los IRES de los mensajeros que codifican a BAG-1 y p53 [33, 34]. Cabe señalar que la mayoría de los genes traducidos de manera de pendiente de IRES son traducidos de manera canónica en condiciones que no favorecen su expresión 5'-cap independiente [15]. Sin embargo, el conocimiento de los mecanismos moleculares de los IRES celulares es muy escaso en contraste con los IRES virales (**Figura 2**) en donde se han logrado progresos importantes de manera reciente [15].

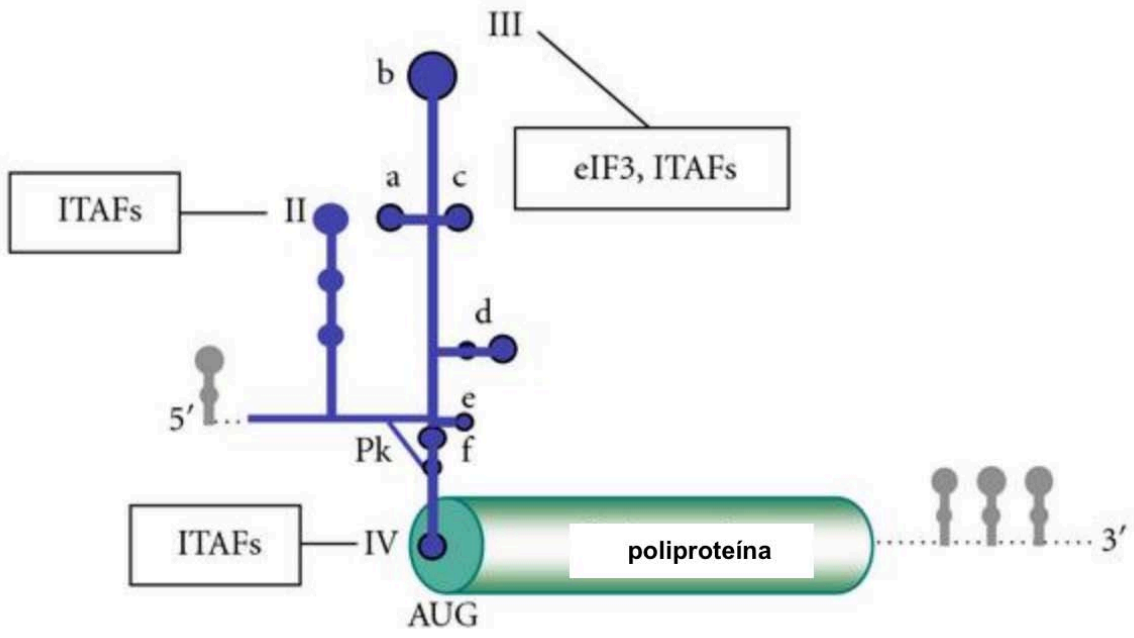

Figura 2. IRES del virus de la hepatitis C (HCV).
La estructura del IRES se muestra en azul y el genoma viral es representado por el cilindro verde (poliproteína). Se presentan las regiones de unión preferencial de eIFs e ITAFs. Figura adaptada de la referencia [35].

En la **Figura 2** se muestra un esquema del IRES del virus de la hepatitis C (HCV) que está conformado por 3 dominios llamados II, III y IV los cuales conforman su estructura

terciaria la cual es necesaria para la traducción eficiente de la proteínas virales [36]. En particular el dominio II es necesario para la formación del complejo ribosomal 80S ya que promueve la hidrólisis de GTP inducida por eIF5 y la salida de eIF2 del CPI [37, 38]. El dominio III consiste de 6 lazos (denominados IIIabcdef) [39]. En la ausencia de eIFs, el IRES del HCV puede establecer una interacción de alta afinidad con la subunidad ribosomal 40S mediante las regiones IIIabc, IIIef y IIId [40]. En condiciones de limitación de eIF2, el IRES del HCV es capaz de formar el complejo de preiniciación en presencia de eIF3 y eIF5B [41]. En el caso de este IRES, los factores no canónicos RACK1 y nucleolina funcionan como ITAFs. RACK1 regula el inicio de la traducción reclutando a la proteína cinasa C para que interaccione con la subunidad 40S [42, 43].

Tomando en cuenta la ausencia de conservación en estructura y secuencia así como la variedad de mecanismos potencialmente implicados en los IRES, resulta evidente que para la predicción *in silico* de IRES, es necesario el utilizar métodos computacionales diseñados para el análisis de datos multifactoriales y de complejidad elevada. A continuación se describen brevemente aquellos utilizados en nuestro estudio.

2.2 Aprendizaje de máquina

En los últimos años las técnicas experimentales utilizadas para la medición de una gran cantidad de variables a nivel molecular han avanzado de manera sustancial en su capacidad técnica, al mismo tiempo sus costos se han reducido de manera muy importante. Algunos ejemplos incluyen la generación de datos mediante microarreglos [44], espectrometría de masas [45] y secuenciación masiva de nueva generación [46]. Gracias al desarrollo de este tipo de metodologías y a la gran cantidad de datos que aportan, ha sido posible plantear nuevas preguntas sobre procesos biológicos. Por ejemplo: ¿Cuáles de los 10,000 genes analizados con microarreglos muestran expresión diferencial entre dos condiciones?, ¿qué variaciones genotípicas son predictivas de cierto fenotipo?, ¿qué proteínas, dentro de las 2,000 analizadas, se encuentran sobre-expresadas para las 8,000 muestras de pacientes con cáncer?, entre otras. Existen dos puntos en común en este tipo de preguntas: i) no existen modelos teóricos que describan este tipo de procesos debido a que son multifactoriales y además complejos ii) existe una gran cantidad de información que en muchas ocasiones es considerada ruidosa (existen falsos positivos y falsos negativos) [47]. En estos casos el reto es la generalización lo cual significa utilizar metodologías capaces de abstraer la complejidad detrás de los datos y ser capaces de predecir casos nuevos en base al entrenamiento con los casos conocidos [48]. Una de las

preguntas más importantes en este campo es el diseño de algoritmos que sean capaces de clasificar objetos con base en sus propiedades. Para responder esta pregunta se han desarrollado diversos algoritmos, como por ejemplo los árboles de clasificación, las redes neuronales y las máquinas de soporte vectorial [49]. Los árboles de clasificación se basan en el uso de reglas condicionales. Un ejemplo sencillo de un árbol de clasificación se presenta en la **Figura 3A**. En este caso el objetivo es desarrollar un método para la clasificación (o predicción) de los objetos de la "Clase 1" y de la "Clase 2" con base en el "Predictor A" y en el "Predictor B" (variables A y B) y por lo tanto establecer grupos (visualizados como regiones en la **Figura 3**) donde los elementos son homogéneos en sus propiedades (objetos similares). En este ejemplo, de antemano conocemos la clase a la que pertenecen cada uno de los elementos, los puntos rojos son de la "Clase 1" y los puntos azules son de la "Clase 2". La línea negra separa las regiones coloreadas y representa la frontera de decisión, es decir, un punto que cae en la zona roja será clasificado como perteneciente a la "Clase 1" por nuestro algoritmo y viceversa. Siguiendo el del árbol de clasificación, se establece la regla: si el Predictor B es mayor o igual a 0.197 (línea horizontal negra en la **Figura 3A**), entonces los objetos clasificados pertenecen a la clase 1, de otra manera pertenecen a la clase 2. Esto también se puede representar también como:

```
if Predictor B >= 0.197 then
 Clase = 1
else Clase = 2
```

En la **Figura 3A** podemos observar que existen algunos errores en la clasificación, puntos azules que caen en la zona roja y puntos rojos que caen en la zona azul (este es un ejemplo sencillo en donde el predictor B se mantiene constante y el predictor A varía en todo el rango de valores posibles). En este criterio se basan algunos de los métodos utilizados para medir el desempeño de diferentes modelos. La exactitud de la clasificación se calcula como el total de elementos fueron clasificados de manera correcta con relación al número total de elementos. En la **Figura 3B** se presenta el resultado de utilizar una máquina de soporte vectorial (SVM) con un kernel (función utilizada para llevar a cabo la separación de las clases) polinomial para llevar a cabo esta clasificación. En este caso el árbol de decisión alcanza una exactitud de clasificación de alrededor del 73%, mientras que este mismo parámetro para la máquina de soporte vectorial llega al 76% gracias a la curvatura de la función utilizada. En general, la exactitud de un algoritmo para una tarea de

clasificación dada depende de cada conjunto de datos en particular y por lo tanto no existe un método que sea superior a los demás en todos los escenarios [50, 51].

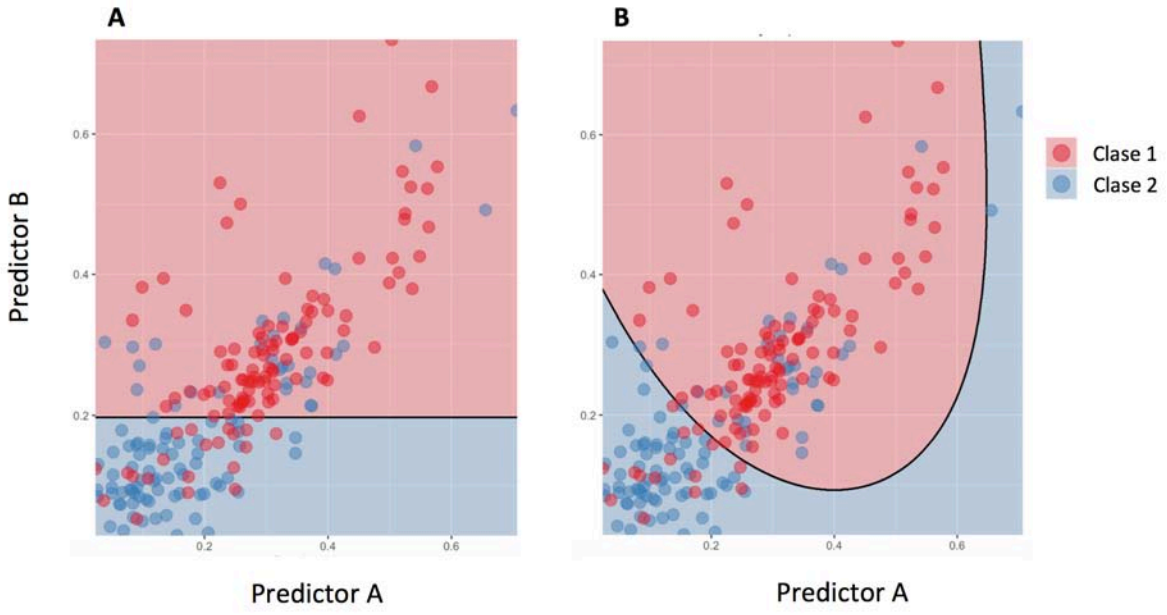

Figura 3. Clasificación de un conjunto de datos de dos dimensiones.
A.- Árbol de decisión utilizando el algoritmo C4.5 [52]. **B.-** El mismo conjunto de datos se clasificó utilizando una máquina de soporte vectorial con un kernel polinomial de segundo grado. (este conjunto de datos fue tomado de [53] y para la implementación de los algoritmos se utilizó el paquete "caret" [54]).

En caso descrito anteriormente se tienen únicamente dos variables (o predictores), sin embargo, en problemas de mayor complejidad es necesario utilizar un número mayor de éstas con el objetivo de incrementar la exactitud de la predicción. El caso bidimensional se puede visualizar fácilmente en una gráfica, pero los casos con más de tres variables no se pueden representar gráficamente, pero sí matemáticamente en un espacio n-dimensional.

2.3 Máquinas de soporte vectorial SVMs

Dentro del aprendizaje de máquina se utilizan diversos modelos para el análisis de datos, sin embargo, uno de los más utilizados debido a su gran flexibilidad y desempeño son las SVM [55]. Estos modelos fueron desarrollados a mediados de los años 60's por Vladimir Vapnik [56, 57]. Las SVM son utilizadas para llevar a cabo la discriminación de objetos que pertenecen a una de dos categorías (**Figura 3B**), es decir son clasificadores binarios. Se consideran modelos supervisados ya que requieren de un conjunto de datos de entrenamiento donde los ejemplos utilizados se encuentran etiquetados como pertenecientes a una de las dos categorías. El modelo se construye representando los ejemplos como puntos en el espacio de las variables y son mapeados de tal manera que

las categorías son separadas por el mayor margen posible. Los nuevos ejemplos son mapeados en el mismo espacio y asignados a la categoría correspondiente de acuerdo al modelo creado en el entrenamiento [49]. En este sentido, se considera que las distribuciones de las variables de los datos de entrenamiento y las de los datos nuevos son similares. Ejemplos de su aplicación en biología computacional son la discriminación de secuencias codificantes y no codificantes [58], la predicción de interacciones proteína-proteína (PPI), en donde el objetivo es determinar si un par dado de proteínas interaccionan físicamente [59] y de manera similar, la predicción de interacciones RNA-proteína [60], entre muchas otras. Para un tratamiento práctico del tema se puede consultar la obra de Kuhn [53] y para una discusión teórica más profunda se recomienda el trabajo de Vapnik [61].

2.4 Genómica comparativa

La genómica comparativa puede definirse como la disciplina que compara las características genómicas de diferentes organismos para generar conocimiento a partir de sus diferencias o similitudes. Los genomas contienen muchas características susceptibles a ser estudiadas por la genómica comparativa como son la secuencia genómica, la curvatura del DNA, la sintenia, las secuencias de regulación, las características fisicoquímicas como la energía mínima de plegamiento (MFE), la frecuencia nucleotídica, etc. De manera concisa, se puede decir que los principios fundamentales de la genómica comparativa son los siguientes: i) las características fenotípicas comunes a dos organismos se encuentran frecuentemente codificadas en el DNA que se encuentra conservado entre las especies, ii) las características del DNA/RNA que controlan la expresión de genes que son regulados de manera similar también tenderán a estar también conservadas y iii) las diferencias entre los fenotipos de organismos, pueden ser generalmente relacionadas con diferencias genotípicas (revisado en [62]).

La genómica comparativa ha sido muy exitosa para la predicción de sitios de regulación en diversos organismos mediante la comparación de sus secuencias no-codificantes [63–66] y también mediante la comparación de características diferentes a la secuencia, como la curvatura del DNA [67], entre otros.

3 Preguntas de investigación

¿Es posible llevar a cabo la predicción de IRES en hongos unicelulares mediante métodos bioinformáticos?

¿La presencia de IRES en genes ortólogos está conservada entre organismos pertenecientes al reino de los hongos?

¿Cuáles son las relaciones funcionales de las proteínas cuyos genes contienen IRES?

4 Hipótesis

La presencia de IRES se encuentra conservada en genes homólogos o funcionalmente relacionados en el reino de los hongos debido a que las fuerzas evolutivas que determinaron dicha presencia es común en organismos filogenéticamente relacionados.

5 Objetivo general

Desarrollar un método computacional para la predicción de presencia (casos positivos) o ausencia (casos negativos) de IRES en secuencias 5'-UTR y caracterizar las propiedades funcionales de los casos positivos *in silico*.

6 Objetivos específicos

a) Implementar y entrenar una máquina de soporte vectorial (SVM) para predecir IRES en secuencias 5'-UTR .
b) Utilizar la SVM ya entrenada para descubrir nuevos IRES mediante la clasificación de sus secuencias 5'-UTR.
c) Utilizar la genómica comparativa para estudiar la conservación de los casos positivos (presencia de IRES) en diferentes organismos .
d) Caracterizar la red de interacciones funcionales proteína-proteína de los productos correspondientes a los casos positivos (presencia de IRES).
e) Analizar el significado biológico de los resultados mediante un análisis de enriquecimiento funcional.

7 Resultados y discusión

7.1 Predicción de IRES en las regiones 5'-UTR

Con el objetivo de desarrollar un método para la predicción de IRES en las 5'-UTRs se desarrolló un método basado en aprendizaje de máquina utilizando genómica comparativa. Se probaron diferentes modelos y se seleccionó utilizar una SVM (ver la

sección “Selección del modelo a utilizar” dentro de la Metodología). Nuestro método se centra en los IRES no estructurados presentes en organismos unicelulares del reino de los hongos (consultar la lista de organismos en la metodología) como los presentes en *S. cerevisiae* [26, 27] y no incluye IRES altamente estructurados [28]. Esto con el objetivo de simplificar el análisis y además el número de IRES estructurados en *S. cerevisiae* al momento de llevar a cabo este estudio es muy limitado (sólo se ha reportado un IRES de este tipo en URE2 [68]). Se emplearon variables basadas en la composición de las secuencia, frecuencia de dinucleótidos, la energía mínima de plegamiento (MFE) de los RNAs y propiedades comparativas calculadas después de agrupar las 5’-UTRs en sus grupos de ortología correspondientes (tomando en cuenta los genes río abajo). El uso de estas características es utilizado en el campo de la biología de RNA ya que en esta molécula frecuentemente no presenta conservación de su secuencia [69], pero sí de su estructura (reflejado en la MFE [70]) y su composición (frecuencia de dinucleótidos). Se obtuvieron un total de 29 variables (consultar la lista de variables en la sección de metodología y la publicación (**Anexo 1**): “Peguero-Sanchez, Esteban, Liliana Pardo-Lopez, and Enrique Merino, ‘IRES-Dependent Translated Genes in Fungi: Computational Prediction, Phylogenetic Conservation and Functional Association’, *BMC Genomics*, 16 (2015), 1–15 <http://dx.doi.org/10.1186/s12864-015-2266-x>”).

Para evaluar el desempeño del modelo y en la optimización de parámetros de la SVM Se utilizó validación cruzada [53]. Se empleó el método SMOTE [71] para la generación de casos positivos sintéticos como se describe en la metodología. El conjunto de entrenamiento (formado por casos positivos y negativos), fue aleatoriamente dividido en diez partes. De estas partes, una fue utilizada para evaluación y el resto fue usado para ajustar el modelo. Las métricas determinadas fueron la exactitud y la kappa de Cohen [72]. El proceso continuó hasta que todas las partes fueran usadas individualmente para evaluar el modelo (validación cruzada de 10 dobleces). Posteriormente se tomó el promedio de la exactitud. Este proceso completo fue repetido 30 veces (un ciclo). En cada ciclo se evaluó el desempeño del modelo para una combinación de parámetros de la SVM (los parámetros de la SVM definen la frontera de decisión, ver la **Figura 3**). El mejor conjunto de parámetros se tomó como aquel que producía la exactitud promedio más alta en un ciclo completo dado. Esta exactitud fue del 94.3% y su kappa respectiva fue de 0.828, lo que indica que el modelo produce resultados superiores a los esperados de un proceso de predicción al azar [53, 72]. Adicionalmente se estimaron los valores de especificidad y selectividad que fueron 0.94 y 0.98 respectivamente. Posteriormente se utilizó el modelo para llevar a cabo la

predicción de las 99,759 secuencias 5'-UTR correspondientes a 20 organismos no redundantes. Dicha predicción es de naturaleza binaria, es decir con IRES (predicciones positivas) o sin IRES (predicciones negativas). Nuestro método predijo 6,532 secuencias con IRES y 93,227 como sin IRES. Las predicciones positivas representan el 6.8% del total de secuencias utilizadas. Este número es cercano a las estimaciones de la proporción de mRNAs que podrían ser traducidos de manera cap-independiente utilizando datos de microarreglos (10-15%) [73, 74]. Como control negativo se analizaron las secuencias de 60 nt río arriba del codón de paro de la traducción debido a que se espera que en esta región, la presencia de IRES sea mínima o inexistente. El número de secuencias usadas como control negativo fue de 99,759; de estas solamente 317 fueron clasificadas como con IRES por nuestro modelo. Esto corresponde a una frecuencia de falsos positivos del 5%.

7.2 En las secuencias 5'-UTR con predicciones de IRES en hongos se encuentran Patrones de conservación evolutivos.

De las 6,532 secuencias predichas positivamente (con IRES), 815 (344 si no se consideran las proteínas hipotéticas) mostraron características distintivas de conservación evolutiva analizadas mediante enriquecimiento de grupos de ortología (frecuencia de falsos positivos [FDR, por sus siglas en inglés, **F**alse **D**iscovery **R**ate] < 0.05%). Estas secuencias son encontradas en 86 grupos de ortología (Figura 1) de los 22,605 considerados [75]. Es decir, los genes correspondientes a estas secuencias se encuentran relacionados filogenéticamente y contienen predicciones positivas (IRES).

Varios de los grupos enriquecidos están implicados en la respuesta a estrés. Algunos de ellos tienen genes con IRES verificados experimentalmente en otros organismos. A continuación se discuten algunos de ellos.

El grupo más enriquecido tuvo un FDR de 1×10^{-26}, lo que corresponde a 78 predicciones positivas de un total de 265 genes en el grupo, comprendiendo los 20 organismos no redundantes usados en nuestro análisis (**Figura 4**). Las proteínas en este grupo son transportadores de azúcares y sensores de glucosa. Estos transportadores tienen una amplia gama de afinidades por lo que facilitan la adaptación a medios cambiantes desde el punto de vista nutricional [76]. Este tipo de variaciones ambientales produce condiciones de estrés en tanto las célula moviliza la maquinaria requerida para aprovechar fuentes de carbono alternativas (revisado en [77]). De manera importante, con nuestro método se predijo que *–HTX1, HTX5, HXT9, y GAL2–* tienen IRES y que se encuentran conservados en la mayoría de los organismos estudiados. Adicionalmente, los

seis transportadores –*HTX1, HTX2, HTX4, HTX5, HXT9, y GAL2*– se sobre expresan en condiciones de estrés osmótico [78]. La sobre expresión fue regulada de manera traduccional por una asociación incrementada de poli-ribosomas en la región 5'-UTR [78]. Este antecedente apoya la validez de nuestras predicciones debido a que la ocupación incrementada de ribosomas en el 5'-UTR en situaciones de estrés está ligada la traducción dependiente de IRES [79].

El segundo grupo más enriquecido (**Figura 4**) corresponde a la familia de HSP70 (36 predicciones de 107, FDR de 1.7×10^{-13}). Las proteínas de esta familia se encuentran conservadas virtualmente en todos los organismos y son importantes para contender con diversos tipos de estrés. Existe evidencia de control traduccional y un incremento en la ocupación de ribosomas en el mRNA de la chaperona *SSA4* (clasificada con IRES por nuestro modelo) en alta salinidad [80] y condiciones de deficiencia de nutrientes, en donde su eficiencia de traducción se incrementa 2.5 veces [81]. Dos miembros de esta familia tienen IRES verificados experimentalmente, uno en humanos y el otro en mosca, HSP70 en ambos casos [82]. Este resultado podría ser explicado por la hipótesis de que la traducción dependiente de IRES se encuentra conservada en mRNAs de proteínas relacionadas filogenéticamente.

El tercer grupo más enriquecido (**Figura 4**) incluye a la familia de proteínas inducidas por estrés Srp1p/Tip1p (16 predicciones de 40 miembros, FDR de 2.0×10^{-6}). Los miembros de esta familia son inducidos por diversas condiciones de estrés como lo son bajas temperaturas [83], hipoxia [84] y deficiencia de nitrógeno [85]. Algunos miembros de la familia *SRP/TIP1* son regulados por el factor transcripcional Mss11p [86]. Sobresale el hecho de que Mss11p, Msn1p y Flo8 son parte de la cascada de transducción de señales que regula la diferenciación de tipo pseudohifa y el crecimiento filamentoso [87], donde Mss11p y Flo8p se unen de manera cooperativa al promotor de *STA1*, desencadenando la respuesta de crecimiento invasivo [88]. Resulta interesante que los genes que codifican para Flo8p y Msn1p son traducidos de manera dependiente de IRES, así como otros siete genes involucrados en el crecimiento invasivo [27].

El cuarto grupo más enriquecido (36 predicciones de 185 miembros, FDR de 3.9×10^{-6}) representa un subconjunto de la superfamilia facilitadora principal (por sus siglas en inglés, MFS), más específicamente de genes que codifican para antiportadores de H^+. Los productos protéicos de estos genes son esenciales para la resistencia a multidrogas y la respuesta a estrés en levadura (revisado en [89]). En este contexto, se ha demostrado que *PDR15* es regulado de manera traduccional y su 5'-UTR muestra una ocupación

incrementada de ribosomas en su región 5'-UTR en respuesta a alta salinidad [80]. De manera análoga, *PDR5* y *PDR12*, muestran una correlación positiva entre la presencia de ribosomas en su 5'-UTR y su eficiencia de traducción en diversas etapas de desarrollo; esta tendencia ha sido observada en genes cuya expresión es regulada traduccionalmente. Esta misma correlación se observó para diversos IRES en levadura [79].

El quinto grupo más enriquecido (23 predicciones de 92, FDR de 1.5×10^{-5}) corresponde a la familia de las proteínas de unión al "cassette" de ATP (familia ABC por sus siglas en inglés). Sus miembros participan en muchos procesos biológicos que incluyen la detoxificación vacuolar, la resistencia pleiotrópica a drogas (PDR por sus siglas en inglés) y la adaptación al estrés (revisado en [90]). Yap1p participa en la regulación de la red PDR y su correspondiente gen (*YAP1*) tiene un IRES verificado experimentalmente [91]. Adicionalmente, la SVM predijo cuatro genes con IRES en esta misma red de regulación (*PDR5*, *PDR12*, *SNQ2* y *STP5*), dos de los cuales son regulados directamente por Yap1p (*SNQ2* y *STP5*) (revisado en [90]). Es importante resaltar que a pesar de que *YAP1* no fue utilizado para entrenar a la SVM, fue unos de los genes predichos con IRES. Los casos aquí descritos constituyen otro ejemplo consistente con la hipótesis de la existencia de múltiples genes regulados por IRES en la misma red de regulación o interacción [9].

Un segundo ejemplo de un gen que contiene un IRES verificado experimentalmente [92–94], que no fue utilizado para entrenar la SVM y que fue predicho de manera positiva es *HAP4*. Existe evidencia de que *YAP1* y *HAP4* provienen de la divergencia de un gen ancestral [95].

Considerando la evidencia estadística y los aspectos biológicos presentados, estos resultados validan el método de predicción de IRES aquí descrito y apoyan la hipótesis de conservación de la regulación dependiente de IRES en hongos unicelulares (**Figura 4**).

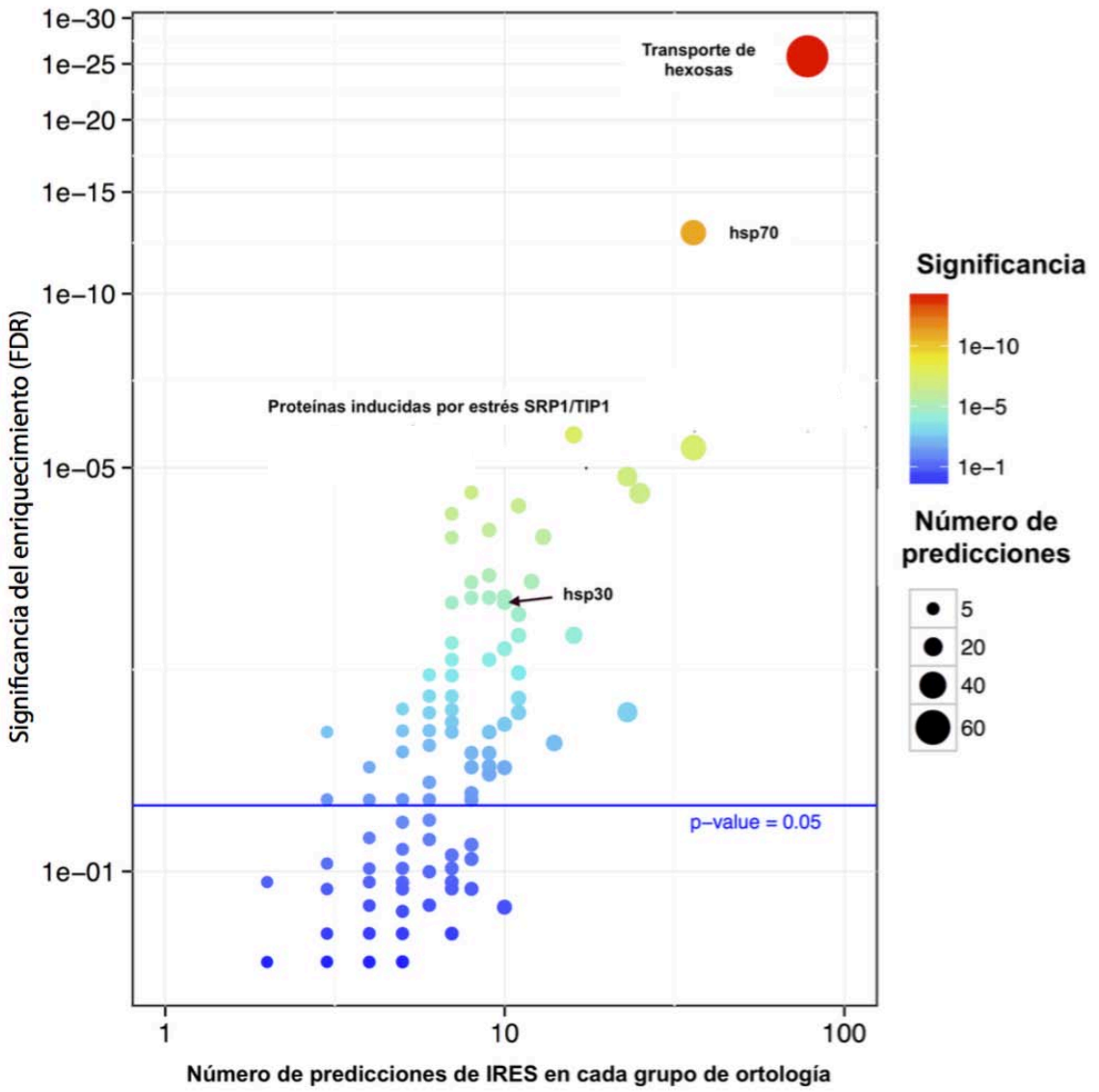

Figura 4. Grupos de ortología conteniendo genes clasificados como conteniendo IRES.
Algunos de los grupos más relevantes se encuentran etiquetados. La significancia estadística del enriquecimiento se presenta en el eje vertical y también codificada en color. El número de predicciones dentro de cada grupo es proporcional al tamaño de los puntos en la gráfica.

7.3 Selección de los mejores candidatos con IRES

El producto del enriquecimiento de cada grupo de ortología y el complemento de la probabilidad posterior de acuerdo a la clasificación de la SVM [96] (1 – probabilidad de tener IRES) como el criterio para asignar un puntaje. Se definió un corte de 0.05 para seleccionar las mejores predicciones (los valores más pequeños indican mejores predicciones). El grupo de las mejores predicciones está constituido por únicamente 801 genes del total de 6,532 (12%) y representa el 0.8% del total de genes considerados en el análisis (99,759 de 20 organismos). Cabe mencionar que en estas 801 proteínas no se

consideran proteínas hipotéticas. Para *S. cerevisiae*, 174 genes se incluyeron en esta categoría de mejores predicciones de un total de cerca de 6,000. La ventaja de este procedimiento de selección reside en que toma en cuenta la incertidumbre de la clasificación (estimada mediante la probabilidad posterior) y el enriquecimiento de los IRES en genes relacionados filogenéticamente en diferentes organismos. Esto último reflejaría la conservación de la regulación traduccional mediada por IRES. Todos los análisis subsecuentes se hicieron con el subconjunto de 174 genes para *S. cerevisiae*.

Al comparar el conjunto de genes descrito en la sección “En las secuencias 5’-UTR con predicciones de IRES en hongos se encuentran Patrones de conservación evolutivos” (344 genes sin considerar proteínas hipotéticas), con los 801 genes seleccionados como mejores predicciones se encontró que el primer conjunto está totalmente contenido en el segundo. Esto significa que dentro del conjunto de 801 genes encontramos 457 genes adicionales que no presentan conservación en grupos de homología, pero cuya probabilidad de ser IRES es alta.

7.4 Análisis de redes

Se llevó a cabo la construcción de una red de PPI funcionales con información de la base de datos STRING [97] mapeada sobre las mejores predicciones de IRES (**Figura 5A**). Esta base de datos brinda información sobre las interacciones físicas proteína-proteína y también considera un conjunto extendido de asociaciones funcionales como la participación de un par de proteínas en rutas metabólicas comunes, co-regulación y su participación en ensamblajes de proteínas (aún sin contacto físico directo) [97]. Para caracterizar la red de PPI desde el punto de vista de su conectividad, se seleccionó el parámetro de densidad ya que describe el nivel global de cohesión. La densidad de una red es calculada como la razón entre el número observado (grado de conectividad) y el número posible de conexiones (el número posible de conexiones se refiere al número de interacciones en una red completamente conectada) [98]. La densidad de una red de PPI tiene un significado biológico intuitivo en este contexto debido a que densidades altas se encuentran relacionadas a niveles elevados de asociación funcional.

Mediante simulación estadística llevamos a cabo pruebas para comparar la densidad de nuestra red contra redes aleatorias (ver métodos). El valor de densidad de nuestra red fue de 0.071, lo cuál es significativamente mayor al de una red aleatoria del mismo número de proteínas (0.0461; p-value de 1.3×10^{-4}). Este valor de densidad corresponde a un grado de conectividad promedio de 6.2, que es ligeramente superior al de

otras redes de PPI en *S. cerevisiae* [99]. Estos valores demuestran que la red de PPI formada por los productos proteicos de los mejores genes clasificados como conteniendo IRES es significativamente más cohesiva que una red aleatoria.

7.5 Las proteínas traducidas por IRES están funcionalmente asociadas en módulos que participan en procesos biológicos específicos

Se ha demostrado ampliamente que la mayoría de las funciones biológicas se llevan a cabo por conjuntos de proteínas en “módulos” y no por un solo tipo de molécula de manera aislada [100–102]. La definición más aceptada de modularidad considera un conjunto de proteínas conectadas de manera funcional (actuando en un proceso biológico en común) o físicamente (formando complejos). Esta definición implica que los proceso celulares ocurren a partir de la acción coordinada de un conjunto de moléculas. Por este motivo se espera que la densidad de conexiones entre las proteínas en un módulo sea mayor que si comparamos ese mismo módulo con el resto de la red. Esto se debe a que se espera que las proteínas en un módulo compartan un conjunto de funciones relativamente homogéneas [101]. Por este motivo, la descomposición de la red en subconjuntos relacionados funcionalmente puede ofrecer información valiosa sobre cómo funciona el sistema completo.

Para llevar a cabo la detección de módulos dentro de la red de PPI de las mejores predicciones de IRES se utilizó el algoritmo Louvain [103] implementado en el paquete igraph [104]. Encontramos un total de nueve módulos en nuestro estudio (**Figura 5A**). Uno de esos módulos contenía una sola proteína y fue excluido de análisis posteriores. En la **Figura 5B** se muestra el número de proteínas en cada módulo. Con el objetivo de explorar las funciones biológicas de cada módulo, se utilizó el análisis de enriquecimiento de términos GO (FDR < 0.02). Este análisis reveló un enriquecimiento funcional significativo en siete de los ocho módulos considerados (**Figura 6**). Resulta interesante que los módulos enriquecidos muestran una alta especialización funcional ya que únicamente seis de los 117 términos GO sobre-representados fueron compartidos entre uno o más módulos. Una inspección más detallada de los módulos revela también una clara homogeneidad funcional en la mayoría de ellos. Estos resultados antes mencionados coinciden con reportes previos de enriquecimientos GO en módulos en redes biológicas [105].

Considero que el enriquecimiento de términos GO, su homogeneidad y especialización en los módulos de la red de PPI fortalecen la hipótesis de que los IRES facilitan la expresión de proteínas que funcionan de manera coordinada para llevar a cabo

procesos específicos. Ejemplo de IRES reportados en la literatura que funcionan en redes se encuentran en la infección por poliovirus, hipoxia y estrés del retículo endoplásmico [9], mitosis [9, 106], apoptosis [107], crecimiento invasivo en levadura [27], y el programa meiótico [79].

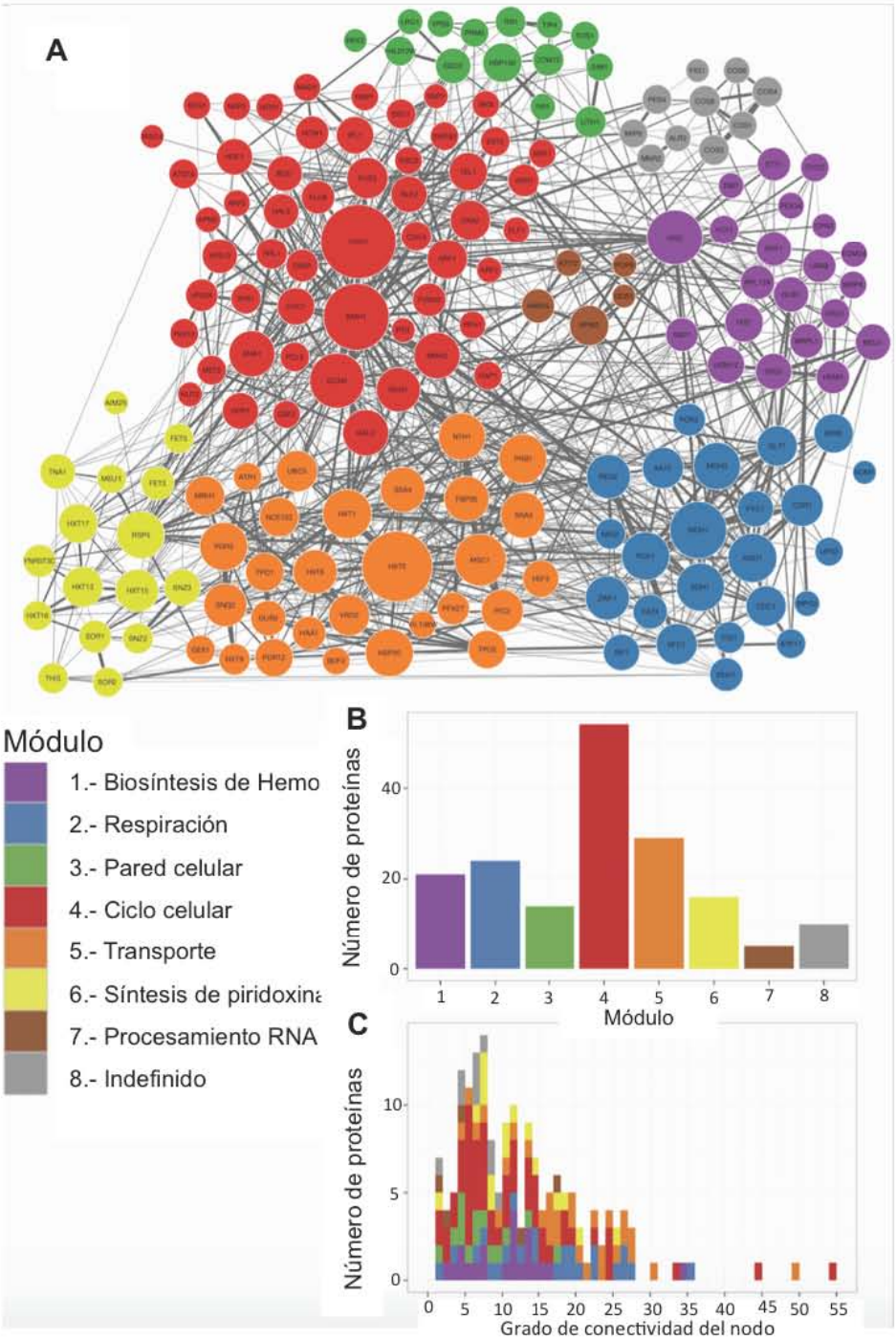

Figura 5. Red de PPI para las mejores predicciones. Cada módulo se representa por un color distinto. **A)** Red de PPI. El tamaño del nodo es proporcional al grado de conectividad **B)** Número de proteínas en cada módulo. **C)** Distribución del grado de conectividad por módulo.

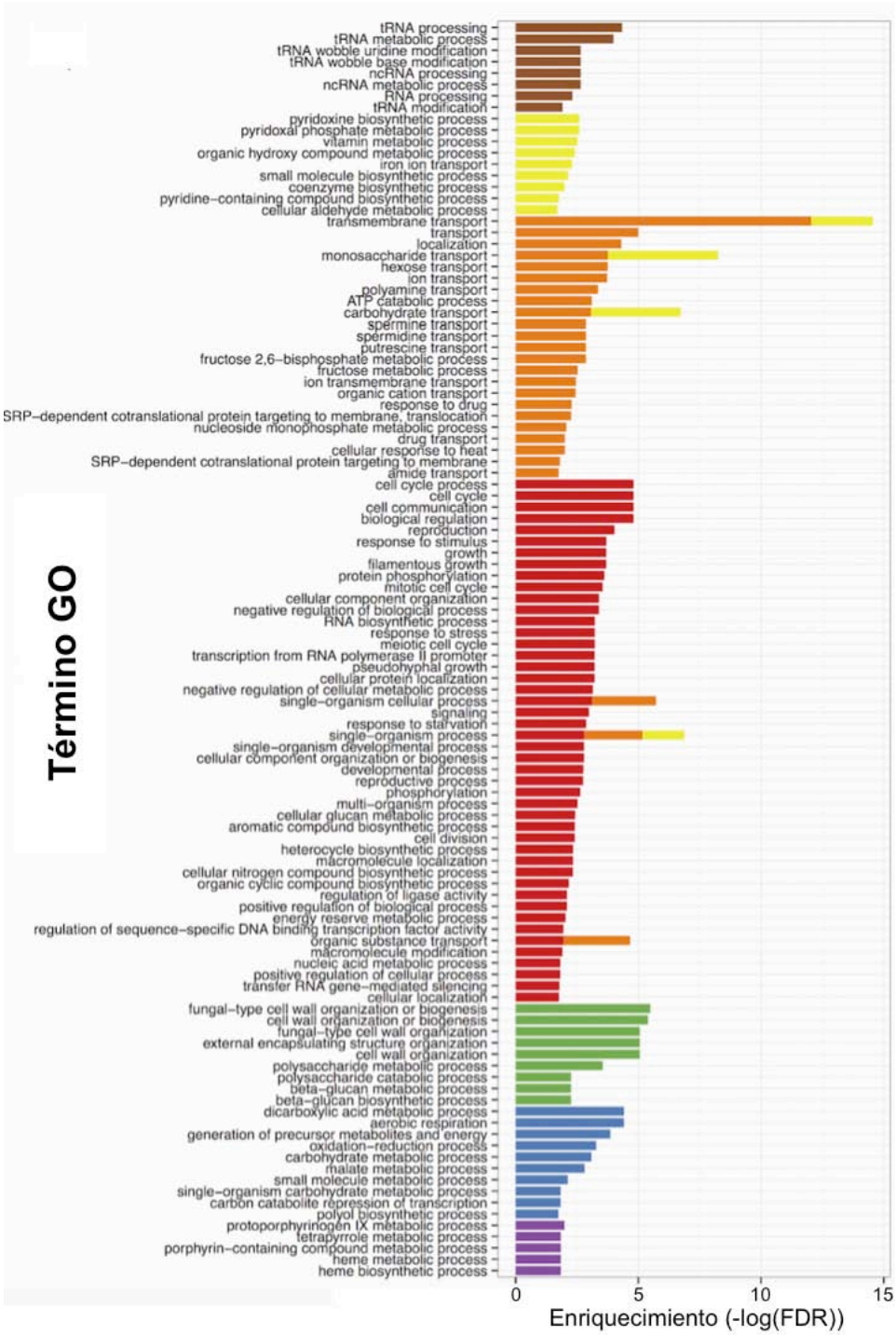

Figura 6. Enriquecimiento de términos GO Módulos para cada módulo de la red de PPI para las mejores predicciones.
Los colores corresponden a los módulos de la **Figura 5**.

7.6 Relevancia biológica de los IRES en sus respectivos módulos

En esta sección se describen algunos ejemplos relevantes de proteínas en el contexto de sus respectivos módulos en la red de IRES. Se utiliza el grado de conectividad de un nodo (o simplemente el grado de conectividad) como una medida de la importancia de la proteína en la red de PPI. El grado de conectividad representa el número de interacciones que ese nodo tiene en la red. El grado de conectividad ha sido asociado al significado biológico de las proteínas y a su importancia. Por ejemplo, se ha observado que el grado de conectividad se encuentra correlacionado con la letalidad en *S. cerevisiae* (no se reportan las características funcionales de las proteínas) [108] y genes de humano implicados en enfermedades (especialmente relacionados con la respuesta a estrés y la organogénesis) [109]. En la **Figura 5C** se muestra la distribución del grado de conectividad para toda la red y para cada módulo encontrado.

7.7 Módulo 1: Síntesis del grupo hemo

El grupo hemo es crucial para muchos procesos biológicos debido a que funciona tanto como un grupo prostético y como molécula de señalización. Por ejemplo, el grupo hemo es utilizado para controlar el crecimiento celular y la diferenciación, reduce el daño oxidativo, está implicado en la cadena respiratoria y actúa como cofactor enzimático (revisado en [110]).

En levadura, el grupo hemo controla la transcripción en respuesta a la variación en los niveles de oxígeno a través del activador Hap1p (revisado en [110]). De manera interesante, Hap1p está incluido en el módulo 4 (ciclo celular), lo que enfatiza la cercana relación que existe entre el grupo hemo y los procesos de desarrollo. Adicionalmente, esta asociación sugiere un caso en donde una molécula efectora (hemo) y el gen regulado (*HAP1*) son traducidos de manera IRES-dependiente, lo que resalta la asociación funcional de los IRES.

Otra molécula con una función regulatoria en este módulo es Gis2p. Esta proteína es la más conectada dentro de este módulo, con un grado de conectividad de 34. Gis2p es un activador traduccional de mRNAs con IRES [111, 112]. Adicionalmente, se han descrito funciones reguladoras de la actividad de algunos IRES en dos proteínas ortólogas a Gis2p, Znf9p en humano [113] y Cnbp en mosca [114], lo cual sugiere una función conservada. Gis2p está implicada en la respuesta a estrés mediada por su acumulación en cuerpos P y gránulos de estrés bajo condiciones de limitación de glucosa [115], y es parte de una red de regulación genética implicada en la inducción del crecimiento invasivo [116]. Resulta

relevante que siete genes adicionales, requeridos para el crecimiento invasivo se traducen de manera IRES-dependiente en *S. cerevisiae* (estos genes fueron utilizados para entrenar la SVM) [27].

La presencia de IRES en genes que codifican para proteínas regulatorias como las anteriormente mencionadas Hap1p y Gis2p implica la existencia de una red de regulación que responde a señales metabólicas o condiciones de estrés específicas en donde algunos de los componentes de la maquinaria traduccional podrían estar limitados.

Otro ejemplo que llama la atención dentro de este módulo es el caso de dos proteínas funcionalmente relacionadas, Sssq1p que es una chaperona miembro de la familia de HSP70 y su co-chaperona, Mdj1p [117]. Ssq1p tiene dos proteínas homólogas cuyos genes tienen IRES, Hsp70p en *D. melanogaster* y en humano [82].

Un ejemplo adicional en este módulo es el de Sbp1p que tiene dos homólogas, Pabp1 en levadura y Cirp en ratón en donde ambas son traducidas de manera dependiente de IRES.

7.8 Módulo 2: Respiración y generación de energía

La capacidad de responder a los cambios en los niveles de nutrientes es crucial para la supervivencia celular. *S. cerevisiae* utiliza preferencialmente glucosa como fuente de carbono, sin embargo, en condiciones de limitación de esta azúcar puede utilizar otras fuentes de carbono no fermentables. Tres proteínas en este módulo que pertenecen a la vía central de la gluconeogénesis son las malato deshidrogenasas (Mdh1p y Mdh2p) y la fosfoenolpiruvato carboxicinasa (Pck1p). Estas proteínas se degradan en presencia de glucosa [118]. Loss genes que las codifican se regulan por el factor transcripcional Znf1p en condiciones de limitación de glucosa [119]. Pck1p también participa en la respuesta a otros tipos de estrés y confiere tolerancia al frío en levadura [120].

7.9 Módulo 3: Organización de la pared celular

El espesor de la pared celular y su composición es modificado para contender con estímulos mecánicos, osmóticos y de calor en el medio ambiente. La pared celular de *S. cerevisiae* está compuesta por alrededor de un 85% de polisacáridos y el resto es proteína. Un gen que codifica para una de estas proteínas de pared celular se clasificó como conteniendo un IRES: Gsc2p, la cual es una β-1,3 glicano sintasa que se induce por diferentes estímulos como por ejemplo, calcio extracelular, tratamiento con el factor α [121], golpe de calor [122], exposición a agentes que dañan la pared celular [123], o tratamiento

con el agente reductor ditiotreitol [124], todos los cuales provocan una fuerte inducción de Gsc2p.

La pared celular sufre modificaciones en su forma durante diferentes etapas de desarrollo y crecimiento (revisado en [125]). Por ejemplo, los productos proteicos de HSP150/PIR2 y PIR1, cuyos mRNAs contienen IRES, se requieren para la estabilidad de la pared celular aunque su mecanismo se desconoce [126]. Sin embargo, las proteínas PIR tienen un impacto en la permeabilidad de la pared celular y este efecto es consistente con su papel en el entrecruzamiento de los β-1,3-glicanos [127, 128].

HSP150 y PIR1 se inducen por golpe de calor, tratamiento con calcofluoro blanco (CFW) o zimolasa y limitación de nitrógeno [127, 129, 130]. Estos genes también se regulan durante el ciclo celular y en respuesta a estrés [131]. Adicionalmente, existe evidencia de la regulación coordinada entre genes requeridos para la organización de la pared celular y los genes implicados en el ciclo celular [132]. Existe evidencia sobre el control traduccional mediado por la 5'-UTR en el caso de dos proteínas de este módulo, Uth1p y Sim1p [133].

7.10 Módulo 4: Ciclo celular

Este módulo es al más grande en términos del número de términos GO y del número de proteínas. Adicionalmente aquí se encuentran las proteínas con los grados de conectividad más altos. De manera importante, la relevancia de los IRES en la regulación traduccional del ciclo celular ha sido descrita con anterioridad. Aquí los IRES juegan papeles centrales en la regulación de la expresión de varias cinasas (esto se revisa en Kronja y Orr-Weaver [106]).

La proteína más conectada en el módulo4 es Hog1p (High Osmolarity Glycerol Response), con un grado de conectividad de 54. Resulta interesante que Hog1p tiene tres proteínas homólogas en humanos que se traducen de manera dependiente de IRES (PITSLREp, Pim1p y la subunidad tipo II de la cinasa dependiente de calcio/calmodulina) [82]. Esta observación apoya la hipótesis de que existe una conservación de IRES aún en organismos filogenéticamente distantes. Hog1p es una cinasa activada por mitógenos que tiene un papel importante en diferentes condiciones de estrés. Por ejemplo, la respuesta traduccional al choque hiperosmótico [78]. Además esta proteína está involucrada en el control del ciclo celular en respuesta al estrés [134]. Resulta digno de ser mencionado el que otras cinasas son parte de este módulo y tienen papeles importantes en el ciclo celular (Tel1p [135], Ipl1p [136], Vhs1p [137] y Mek1p [138]), crecimiento celular y proliferación

(Cka2p [139]), tolerancia a la sal (Hal5p [140]) y cruza (Fus3p [141]). Otra proteína relevante en este módulo es Gcn4p, que responde a la deficiencia de nitrógeno [142] y que además es controlada transcripcional y traduccionalmente en diversas situaciones de estrés [143]. Dentro de la red de las mejores predicciones, Gcn4p tiene un alto grado de conectividad (33), posiblemente reflejando su importancia. Además resalta el hecho de que una proteína paráloga de Gcn4p, Yap1p y su ortóloga Jun en pollo, tienen IRES verificados experimentalmente [82].

7.11 Módulo 5: Transporte

La relevancia biológica de algunos de los miembros de este grupo ya se ha discutido en la sección “En las secuencias 5’-UTR con predicciones de IRES en hongos se encuentran patrones de conservación evolutivos.” (HAA1, HXT1, HXT5, HXT6, HXT9 y SSA4). Las proteínas en este módulo se encuentran involucradas en diversos tipos de transporte y tienen funciones relevantes en el contexto de la respuesta a estrés . Por ejemplo, Snq2p, Tpo1p, Tpo2p, Pdr12p, etc., se encuentran relacionadas con la respuesta a estrés ácido [144]. Una proteína relevante dentro de este grupo es Yap1p que participa en la regulación PDR [90] y está codificada por un gen que tiene un IRES verificado experimentalmente [91]. Cabe señalar que *YAP1* no fue incluido en el entrenamiento de la SVM ya que se utilizó como control positivo. El otro control positivo clasificado exitosamente fue HAP4. HAP4 y YAP1 provienen de la divergencia de un ancestro común [95].

7.12 Módulo 6: síntesis de piridoxina

La función más estudiada de la vitamina 6 (Vit6) es como cofactor de reacciones enzimáticas. Sin embargo, ahora es claro que la Vit6 es un potente antioxidante que protege a las células del estrés oxidativo [145, 146]. Adicionalmente, Snz2p y Snz3p, que son parte de este módulo responden a la limitación de nutrientes [147].

7.13 Módulo 7: Procesamiento de RNA

Este módulo presentó enriquecimientos en términos GO relacionados al procesamiento y metabolismo de RNA, incluyendo ncRNA y tRNA. A pesar de que no se conocen IRES asociados a este tipo de procesos, esto no excluye que los genes presentados aquí sean regulados por IRES debido a que es un área relativamente inexplorada [148] y que tiene implicaciones importantes en la respuestas a estrés y en condiciones patológicas [149]. Resulta llamativo que un posible homólogo de la proteína

Hrr25p [150] (incluida en este módulo), la proteína Pim-1 tiene un IRES verificado experimentalmente [82].

7.14 Módulo 8: Funciones no definidas

A pesar de que no se encontraron términos GO enriquecidos en este módulo, cinco de sus 10 miembros son proteínas de Secuencia Conservada (COS), incluyendo Cos1p, Cos3p, Cos4p, y Cos8p. Estas proteínas presentan una alta conservación de su secuencia a pesar de que su función es desconocida [151].

Considero que el hecho de que los productos proteicos de genes clasificados con IRES se organicen en módulos cohesivos con términos GO enriquecidos resalta la asociación funcional de los genes traducidos de manera dependiente de IRES y complementa el análisis de redes y de genómica comparativa. El enfoque aquí propuesto, basado en módulos con funciones enriquecidas puede ser utilizado para analizar los resultados de estudios genómicos orientados a entender la regulación por IRES.

7.15 Comparación de las predicciones con genes regulados traduccionalmente

La técnica de perfil de ribosomas, desarrollada recientemente, ha contribuido al entendimiento de los procesos de traducción de proteínas al permitir determinar la posición de los ribosomas a lo largo del mRNA. Lo anterior permite también medir las tasas de traducción y la identificación de genes controlados traduccionalmente si se combina con RNA-seq [81]. Por este motivo, seleccionamos un estudio anteriormente publicado basado en el perfil de ribosomas [79] para encontrar la intersección entre genes controlados traduccionalmente y las predicciones de este estudio. Los genes que se seleccionaron fueron aquellos en donde se encontrara una correlación positiva entre la ocupación de ribosomas en el extremo 5'-UTR y su eficiencia de la traducción. El número de genes que presentaban esta característica fue de 110, mientras que nuestras mejores predicciones fueron 174. La intersección de estos conjuntos resultó ser de 14 genes. La probabilidad de tener esta intersección al azar considerando 6,000 genes en *S. cerevisiae* es de 5×10^{-7} calculada a partir de la prueba exacta de Fisher. Esto significa que los genes controlados de manera traduccional se encuentran sobre-representados en nuestras predicciones.

7.16 Las regiones 5'-UTR predichas con IRES presentan características distintivas con respecto al las 5'-UTR que no contienen IRES

Se compararon los promedios de las variables de aquellas 5'-UTRs con IRES con respecto al conjunto de entrenamiento negativo con el objetivo de definir aquellas que son

distintivas de los IRES (ver la sección “Selección de variables” dentro de métodos para una descripción de cada una de ellas). Se utilizó la prueba de suma de rangos de Wilcoxon para evaluar la diferencia de medias entre estos dos conjuntos. En total se encontró que 14 de las 25 presentan diferencias significativas (p-value < 0.001). De estas, 10 se encuentran por debajo de la media y sólo cuatro por arriba de ésta (**Figura 7**). Las dos variables que se desvían más de la media son los dinucleótidos “AA” (arriba de la media) y “GC” (debajo de la media), esto está de acuerdo con que el tipo de IRES en este caso son ricos en “A”.

La energía mínima de plegamiento también se encuentra debajo de lo esperado, cabe señalar que para este estudio se tomó el negativo de la energía mínima de plegamiento por lo que valores inferiores denotan menor estabilidad del RNA. Es interesante que existe una correlación inversa ya reportada entre la estabilidad estructural del mRNA y los niveles de expresión de los mRNAS IRES [26], lo que está de acuerdo con este resultado.

La variable “moda de la MFE” se encuentra considerablemente debajo de la media. Esto indicaría que la MFE es menor en los grupos de ortología que contienen IRES y coincide con los resultados reportados en la sección “En las secuencias 5’-UTR con predicciones de IRES en hongos se encuentran patrones de conservación evolutivos.” dentro de resultados.

Resulta relevante también que la longitud de la región intergénica en considerablemente mayor que el promedio, esto fue una de las hipótesis originales para seleccionar esta variable (ver la sección “Selección de variables” dentro de métodos). La **Figura 7** contrasta las diferencias que existen entre las propiedades de las regiones que contienen IRES y de aquellas que no las contienen y sin embargo, no necesariamente refleja la importancia de cada variable para el desempeño de la SVM. El análisis de la importancia de las variables para éste y otros tipos de modelos es un área de investigación en pleno desarrollo [152–156].

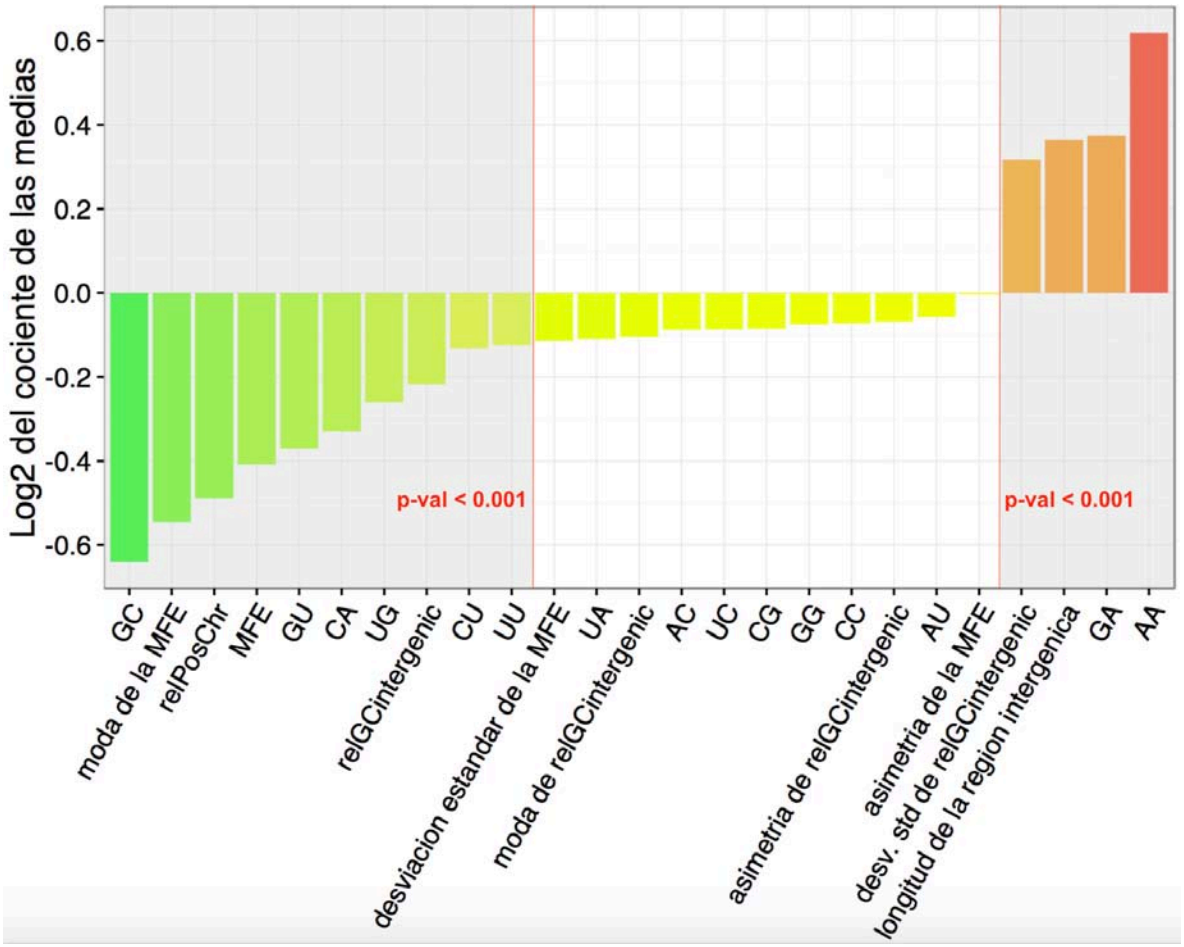

Figura 7. Diferencias promedio entre las 5'-UTR clasificadas como conteniendopredichas con IRES contra aquellas que no los tienen.
En el eje horizontal se muestra cada una de las 25 variables utilizadas para construir el modelo de clasificación. En el eje vertical se muestra el logaritmo base 2 de la relación (media de la variable en 5'-UTRs predichas como conteniendo IRES)/(media de la variable en 5'-UTRs sin IRES). Las zonas sombreadas se refieren a las variables que son significativamente diferentes entre ambos conjuntos.

8 Conclusiones

Se desarrolló un método para la clasificación de IRES no estructurados en hongos. Utilizando este método, se predijeron IRES en las secuencias 5'-UTR de 20 organismos unicelulares no redundantes del reino de los hongos. Adicionalmente se llevó a cabo la caracterización de las relaciones filogenéticas y funcionales de dichas predicciones. Se encontró evidencia estadística de conservación de la traducción IRES dependiente en grupos de genes ortólogos que revela una presión selectiva subyacente, particularmente en genes relacionados con la respuesta a estrés. Además de esto, el análisis de la red de PPI de las predicciones permitió encontrar módulos biológicamente significativos que muestran especialización funcional. Este estudio es uno de los primeros en el análisis computacional de los IRES y sus asociaciones evolutivas y funcionales y representa un recurso útil para

guiar experimentos guiados por hipótesis y para la exploración de genes en el campo de la traducción independiente de 5'-cap.

9 Perspectivas

9.1 Ampliar el conjunto de positivos utilizando un subconjunto de IRES conservados

Una alternativa para incrementar el número de casos positivos y re-entrenar la SVM es el utilizar predicciones positivas (genes) que se encuentren conservadas en varios homólogos en diferentes organismos. De esta manera se podría disminuir el desbalance de los datos y mejorar el método de predicción.

9.2 Verificación experimental de las predicciones

Las pruebas estadísticas y la discusión sobre las características biológicas de las predicciones presentadas soportan la validez de este modelo, sin embargo es necesario llevar a cabo la validación experimental de al menos algunos de los mejores candidatos para confirmar su capacidad predictiva. Esta confirmación es especialmente necesaria dado que el modelo se entrenó con una cantidad pequeña de casos positivos (nueve casos) y esto representa una limitante potencial para su capacidad de generalización.

9.3 Re-entrenamiento de la SVM una vez se tengan más casos experimentales de IRES

Una vez confirmados experimentalmente un mayor número de IRES, es posible re-entrenar al modelo e hipotéticamente mejorar la predicción.

9.4 Selección de variables

Llevar a cabo un análisis de selección de variables para determinar su importancia en el desempeño del método y la robustez de los resultados.

10 Métodos

10.1 Secuencias de DNA

En este estudio se utilizaron las secuencias y anotaciones para 33 organismos unicelulares completamente secuenciados del reino de los hongos (columna izquierda **Tabla 1**). Para evitar la sobre-representación de los datos, se seleccionaron organismos no

redundantes (columna derecha **Tabla 1**) con base en su posición en un árbol filogenético de máxima verosimilitud (**Figura 8**) que fue construido utilizando el paquete PHANGORN [157]. Para cada par de organismos, aquel con el genoma más pequeño fue eliminado. Se tomaron en cuenta únicamente las proteínas ortólogas presentes en todos los organismos. Los ortólogos fueron seleccionados mediante el criterio del mejor hit bidireccional utilizando delta-BLAST [158] con un E-valor máximo de 1×10^{-6} y únicamente considerando alineamientos mayores o iguales al 50% de la menor de las proteínas. Estos 33 organismos fueron la totalidad de los contenidos en la base de datos del NCBI al momento de realizar este trabajo (ftp://ftp.ncbi.nlm.nih.gov/genomes/Fungi). Archivos adicionales al presente trabajo se pueden encontrar en la versión en línea del artículo anexo: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678720/, incluyendo la lista de **IRES utilizados en el entrenamiento, secuencias y descripción** (Additional file 5), lista de las **mejores predicciones, incluyendo su grupo de ortología, y secuencias** (Additional file 8) y la lista de las **mejores predicciones en *S. cerevisiae*, junto con su descripción** (Additional file 9).

Total de organismos	Organismos no redundantes
Aspergillus fumigatus	*Aspergillus fumigatus*
Aspergillus nidulans	*Aspergillus nidulans*
Aspergillus Niger	*Aspergillus Niger*
Aspergillus oryzae	*Candida glabrata*
Candida albicans	*Cryptococcus gattii*
Candida dubliniensis	*Debaryomyces hansenii*
Candida glabrata	*Encephalitozoon cuniculi*
Cryptococcus gattii	*Eremothecium cymbalariae*
Cryptococcus neoformans	*Eremothecium gossypii*
Debaryomyces hansenii	*Gibberella zeae*
Encephalitozoon cuniculi	*Myceliophthora thermophila*
Encephalitozoon intestinalis	*Naumovozyma castellii*
Eremothecium cymbalariae	*Naumovozyma dairenensis*
Eremothecium gossypii	*Pichia pastoris*
Gibberella zeae	*Pichia stipitis*
Kazachstania africana	*Saccharomyces cerevisiae*
Kluyveromyces lactis	*Schizosaccharomyces pombe*

Lachancea thermotolerans	*Tetrapisispora phaffii*
Magnaporthe grisea	*Torulaspora delbrueckii*
Myceliophthora thermophila	*Zygosaccharomyces rouxii*
Naumovozyma castellii	
Naumovozyma dairenensis	
Neurospora crassa	
Pichia pastoris	
Pichia stipitis	
Podospora anserina	
Saccharomyces cerevisiae	
Schizosaccharomyces pombe	
Tetrapisispora phaffii	
Thielavia terrestris	
Torulaspora delbrueckii	
Yarrowia lipolytica	
Zygosaccharomyces rouxii	

Tabla 1. Lista de organismos utilizados en este estudio La primera columna presenta el total de organismos y la segunda los organismos no redundantes.

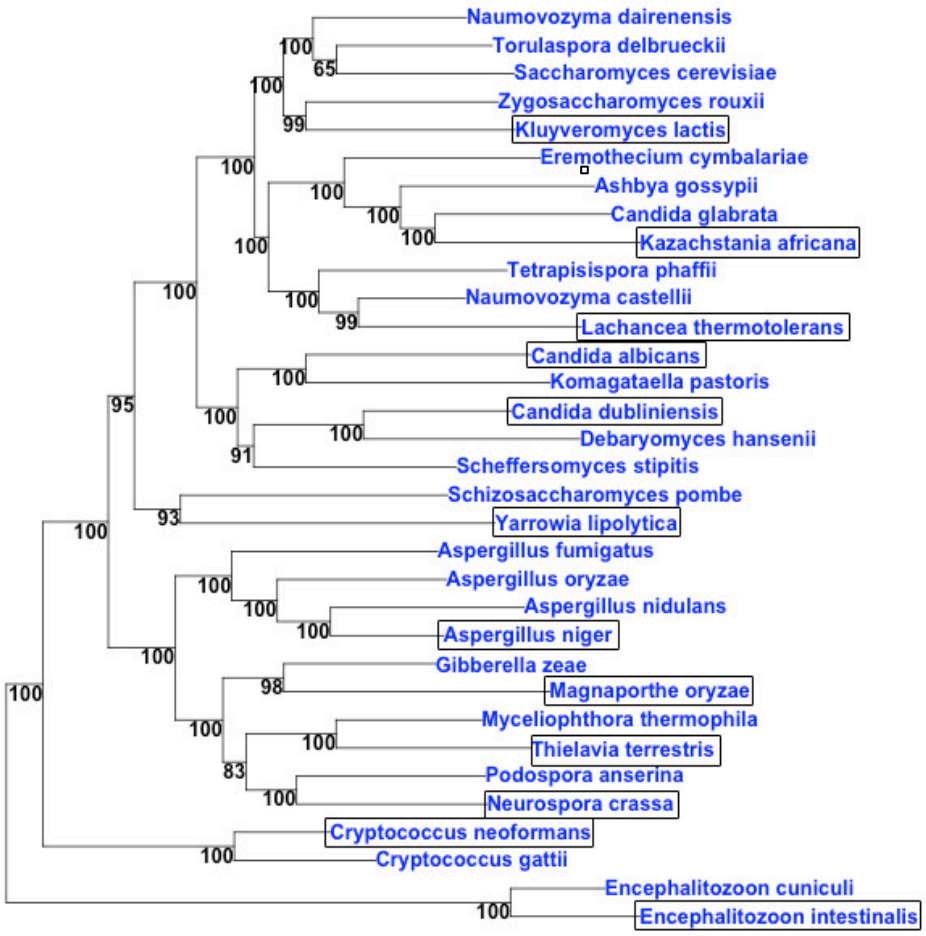

Figura 8. Árbol filogenético utilizado para seleccionar a los organismos no redundantes.
Los recuadros en negro muestran los organismos que fueron descartados. Los números indican el soporte de cada nodo calculado utilizando "bootstrapping" repetido 1000 veces.

10.2 Secuencias 5'-UTR

La mayoría de los IRES en hongos se encuentran en los primeros 60 nt río arriba del inicio de la traducción [26, 27]. Por este motivo se obtuvieron dichas secuencias para los organismos de estudio que son llamadas en este trabajo 60ntUTRs. Estas 60ntUTRs se asignaron a sus correspondientes grupos de ortología. En *S. cerevisiae,* hasta donde se sabe, existen 11 IRES no estructurados bien caracterizados (todos ellos en *S. cerevisiae*) [26, 27], nueve de ellos constituyen nuestros casos positivos, mientras que dos de ellos fueron usados como un control positivo. Para una revisión de los métodos experimentales adecuados para validar IRES se puede consultar [30]. Los casos negativos fueron obtenidos mediante un muestreo al azar de las 60ntUTRs de 12,500 genes de un conjunto total de cerca de 100,000 correspondientes a los 20 organismos no redundantes.

10.3 Selección de variables

La selección de variables de las 60ntUTRs para predecir los IRES fue basada en una revisión de la literatura, considerando aquellas que se han reportado como correlacionadas con la presencia de IRES o que han sido usadas para la clasificación de RNAs no codificantes (ncRNA). El conjunto de variables usadas en nuestro análisis fue agrupado de acuerdo al siguiente criterio: *i*) Energía mínima de plegamiento (MFE), que ha sido utilizada para clasificar ncRNAs [159], y se encuentra correlacionada con la expresión mediada por IRES en levadura [26]. *ii*) contenido de GC. Se ha propuesto que los IRES en *S. cerevisiae* relacionados con la deficiencia de nitrógeno tienen un contenido bajo de GC y particularmente son ricos en As [26]. Adicionalmente, se ha observado una correlación positiva entre el contenido bajo de GC y una actividad traduccional incrementada en condiciones de deficiencia de glucosa, que podría ser explicada, al menos de manera parcial debido a la presencia de IRES [160]. Por estas razones dos variables fueron incluidas: el contenido de GC de la región 60ntUTRs relativo al contenido de GC de su región intergénica correspondiente (relGCintergenic) y el contenido de GC de la 60ntUTRs relativo al contenido de GC del cromosoma (relGCchr). *iii*) Posición relativa del gen en el cromosoma (relPosChr). Esta variable se incluyó debido a que ha sido relacionada con la expresión selectiva de genes de respuesta a estrés [85, 161–163]. *iv*) Frecuencia de dinucleótidos (corresponde a 16 variables). Los enfoques basados en composición han sido exitosos para la clasificación de ncRNAs [164]. *v*) Medidas estadísticas de tendencia central y dispersión. Mientras mayor sea la conservación de IRES dentro de un grupo de ortología, se espera que ejerzan una mayor influencia en las características del grupo. Por esta razón se incluyeron la media, moda, desviación estándar y el sesgo. Estas propiedades fueron calculadas para la relGCintergenic y la MFE de todos los miembros de cada grupo de ortología. *vi*) Longitud de la región intergénica. Esta variable fue escogida debido a que regiones mayores podrían ser indicativas de la presencia de elementos de regulación como los IRES. Después de aplicar los criterios descritos se seleccionaron 29 variables.

10.4 Pre-procesamiento de las variables

Para lograr que los atributos inter-especie fueran comparables, se llevó a cabo la estandarización de las variables con base en sus medias y varianzas en relación a cada organismo. Por lo tanto, todas las variables tuvieron una media de cero y una varianza de 1. Para reducir la redundancia de las variables, se redujo su número, reteniendo únicamente aquellas que no estuvieran correlacionadas [165]. Un par de variables con un

coeficiente de correlación de Pearson mayor a 0.55 se consideraron como correlacionadas. Para cada combinación binaria de variables correlacionadas, una de ellas fue eliminada. Después de este paso se retuvieron 25 variables de las 29 originales. Este conjunto de variables fue el que se utilizó para entrenar y validar la SVM.

10.5 Generación de datos sintéticos mediante sobre-muestreo de la minoría

Los métodos de aprendizaje de máquina en general tienen un desempeño pobre cuando son aplicados a conjuntos de datos desbalanceados en donde los casos negativos son mucho más numerosos que los casos positivos o viceversa [71]. Sin embargo, en conjuntos de datos reales se han reportado desbalances que van desde 100:1 hasta 10,000:1 [166]. Este tipo de datos son comunes también en biología, por ejemplo, en la predicción de sitios del inicio de la traducción [167] y en la clasificación de pre-miRNA [168]. El conjunto de datos utilizado en este estudio presentó un importante desbalance debido a que el conjunto de casos positivos para entrenamiento consistió en nueve IRES bien caracterizados provenientes de *S. cerevisiae* [27]. Este conjunto de casos positivos es muy pequeño en comparación con todos los genes que podrían contener IRES (cerca de 100,000 para los 20 organismos usados). Para reducir este desbalance se utilizó la técnica de generación de datos sintéticos mediante sobre-muestreo de la minoría (SMOTE). Se seleccionó esta técnica debido a que ha demostrado incrementar de manera significativa el desempeño de las SVM utilizadas para clasificar nrRNAs [168], predecir interacciones RNA-proteína [60] y analizar otros conjuntos de datos en bioinformática [169]. Sin embargo, considerando que el desbalance es bastante pronunciado, esta es una limitación considerable de nuestro estudio que podría repercutir en su capacidad de generalización. Considerando esta limitante, se incluyeron dos casos positivos no usados en el entrenamiento como controles externos.

10.6 Selección del modelo a utilizar

Se llevó a cabo una prueba para estimar el mejor modelo para llevar a cabo la predicción de los IRES utilizando un subconjunto de los datos de 1000 5'-UTRs. Se evaluaron 11 diferentes modelos (**Figura 8**). Se escogió la SVM con un kernel polinomial (ver la siguiente sección).

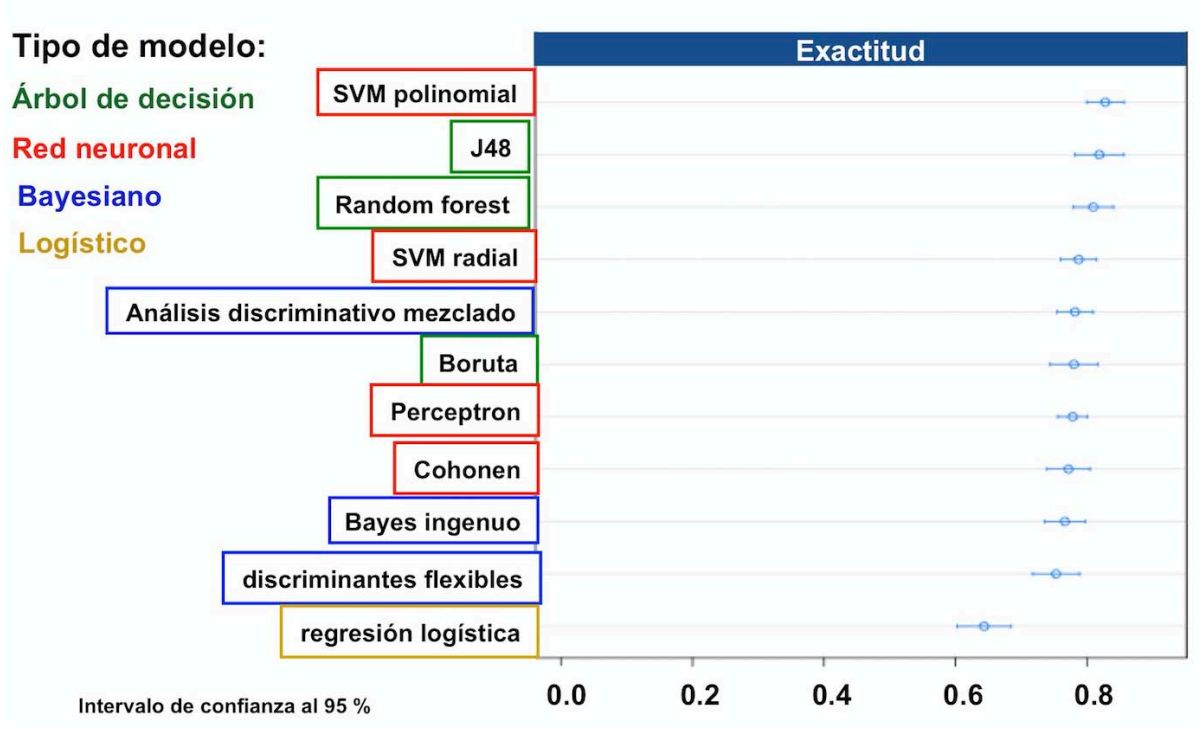

Figura 9. Comparación de modelos para llevar a cabo la predicción de IRES.
El modelo con la mejor exactitud fue la SVM con kernel polinomial. Los colores agrupan a los modelos según su tipo. Las barras horizontales corresponden a la desviación estándar de la exactitud de la validación cruzada repetida 10 veces.

10.7 Aprendizaje de máquina para la clasificación (predicción) de IRES

Una máquina de soporte vectorial con un kernel polinomial de segundo orden fue implementada usando el paquete caret [54] y el paquete kernlab [170]. Para incrementar la sensitividad el parámetro "cost" (revisado en [53]) fue ajustado a 2:1 (clasificación positiva:clasificación negativa). Para el entrenamiento y optimización de parámetros se utilizó validación cruzada (10-fold), la cual fue repetida 30 veces. La clasificación se llevó a cabo sobre el conjunto de 100,000 genes. Se estimaron las probabilidades posteriores de clase P(clase|probabilidad) utilizando el criterio de Platt [96].

10.8 Análisis de enriquecimiento para las predicciones

Los genes clasificados como conteniendo IRES fueron asignados a sus grupos de ortología correspondientes. Se utilizó la prueba exacta de Fisher para calcular la significancia del enriquecimiento [53]. Los p-valores resultantes fueron ajustados para comparaciones múltiples usando el procedimiento Benjamini-Hochberg [171] para controlar la tasa de falsos descubrimientos.

10.9 Comparación de las predicciones con datos experimentales de genes controlados traduccionalmente

Para comparar la probabilidad de una intersección al azar para las predicciones de IRES con el conjunto de datos experimental [79], se utilizó la prueba exacta de Fisher.

10.10 Análisis de ontología de genes

El análisis de GO [172] se llevó a cabo usando el paquete GOstats [173] corrigiendo para comparaciones múltiples usando el procedimiento de Benjamini-Hochberg [171].

10.11 Selección de los mejores candidatos

Para seleccionar los mejores candidatos a ser IRES, se utilizó un procedimiento simple que consistió en multiplicar la probabilidad posterior (1 - la probabilidad de que una clasificación es IRES de acuerdo a la distancia de la clasificación al hiperplano en la SVM) para cada clasificación por su enriquecimiento de ortología (p-valor). Se estableció un corte de 0.05 para este producto (valores más pequeños representan mejores candidatos).

10.12 Análisis de la red de PPI

Los datos de interacción fueron descargados de la base de datos de STRING versión 9.1 [97]. Las propiedades de la red fueron calculadas utilizando el paquete igraph [104]. Los grupos en la red se determinaron utilizando el método de Louvain [103]. Para comparar las propiedades de la red contra una red aleatoria se muestrearon 173 genes al azar sin reemplazo del genoma de *S. cerevisiae* y se calcularon las propiedades de esta red (i.e. densidad). Este proceso se repitió 100,000 veces , los datos se transformaron mediante el procedimiento box-cox [174] y se determinó el p-value mediante la función de distribución normal.

10.13 Determinación de relaciones de homología

Para determinar si un par de proteínas eran homólogas se llevó a cabo un alineamiento pareado mediante el software delta-blast [158] con un E-valor de 1×10^{-6}.

11 Referencias

1. Wittkopp PJ, Kalay G: **Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence**. *Nat Rev Genet* 2012, **13**:59–69.

2. Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424**:147–151.

3. McManus CJ, May GE, Spealman P, Shteyman A: **Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast.** *Genome Res* 2014, **24**:422–30.

4. Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H, Ricci E, Snyder MP: **Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans**. *Genome Res* 2015, **25**:1610–1621.

5. Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses.** *Nat Rev Genet* 2012, **13**:227–32.

6. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control.** *Nature* 2011, **473**:337–42.

7. Sonenberg N, Hinnebusch AG: **Regulation of translation initiation in eukaryotes: mechanisms and biological targets.** *Cell* 2009, **136**:731–45.

8. Liu B, Qian S-B: **Translational reprogramming in cellular stress response.** *Wiley Interdiscip Rev RNA* 2014, **5**:301–15.

9. Spriggs KA, Stoneley M, Bushell M, Willis AE: **Re-programming of translation following cell stress allows IRES-mediated translation to predominate.** *Biol Cell* 2008, **100**:27–38.

10. Jackson RJ, Hellen CUT, Pestova T V: **The mechanism of eukaryotic translation initiation and principles of its regulation.** *Nat Rev Mol Cell Biol* 2010, **11**:113–27.

11. Buttgereit F, Brand MD: **A hierarchy of ATP-consuming processes in mammalian cells.** *Biochem J* 1995, **312**(Pt 1):163–167.

12. He H, Liu M, Zheng Z, Liu Y, Xiao J, Su R, Hu C, Li J, Li S: **Synthesis and analgesic activity evaluation of some agmatine derivatives**. *Molecules* 2006, **11**:393–402.

13. Liu L, Cash TP, Jones RG, Keith B, Thompson CB, Simon MC: **Hypoxia-Induced Energy Stress Regulates mRNA Translation and Cell Growth**. *Mol Cell* 2006, **21**:521–531.

14. Fulda S, Gorman AM, Hori O, Samali A: **Cellular stress responses: Cell survival and cell death**. *Int J Cell Biol* 2010, **2010**:1–23.

15. Komar AA, Hatzoglou M: **Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states.** *Cell Cycle* 2011, **10**:229–40.

16. Bushell M, Stoneley M, Kong YW, Hamilton TL, Spriggs KA, Dobbyn HC, Qin X, Sarnow P, Willis AE: **Polypyrimidine tract binding protein regulates IRES-mediated gene expression during apoptosis.** *Mol Cell* 2006, **23**:401–12.

17. Sharathchandra A, Katoch A, Das S: **IRES mediated translational regulation of p53 isoforms.** *Wiley Interdiscip Rev RNA* , **5**:131–9.

18. Dever TE, Kinzy TG, Pavitt GD, Abastado JP, Miller PF, Jackson BM, Hinnebusch AG, Acker MG, Shin BS, Dever TE, Lorsch JR, Acker MG, Shin BS, Nanda JS, Saini AK, Dever TE, Advani VM, Belew AT, Dinman JD, Aitken CE, Lorsch JR, Albrecht G, Mosch HU, Hoffmann B, Reusser U, Braus GH, Algire MA, Maag D, Savio P, Acker MG, et al.: **Mechanism and Regulation of Protein Synthesis in Saccharomyces cerevisiae.** *Genetics* 2016, **203**:65–107.

19. Pelletier J, Sonenberg N: **Internal initiation of translation of eukaryotic mRNA directed**

by a sequence derived from poliovirus RNA. *Nature* 1988, **334**:320–5.

20. Jang SK, Kräusslich HG, Nicklin MJ, Duke GM, Palmenberg AC, Wimmer E: **A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation.** *J Virol* 1988, **62**:2636–43.

21. Tsukiyama-Kohara K, Iizuka N, Kohara M, Nomoto A: **Internal ribosome entry site within hepatitis C virus RNA.** *J Virol* 1992, **66**:1476–1483.

22. Ohlmann T, Mengardi C, López-Lastra M: **Translation initiation of the HIV-1 mRNA**. *Translation* 2014, **2**:e960242.

23. Hanson PJ, Zhang HM, Hemida MG, Ye X, Qiu Y, Yang D: **IRES-Dependent Translational Control during Virus-Induced Endoplasmic Reticulum Stress and Apoptosis**. *Front Microbiol* 2012, **3**:92.

24. Macejak DG, Sarnow P: **Internal initiation of translation mediated by the 5' leader of a cellular mRNA.** *Nature* 1991, **353**:90–4.

25. Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E: **Systematic discovery of cap-independent translation sequences in human and viral genomes**. *Science* 2016, **351**:240–240.

26. Xia X, Holcik M: **Strong eukaryotic IRESs have weak secondary structure.** *PLoS One* 2009, **4**:e4136.

27. Gilbert W V, Zhou K, Butler TK, Doudna JA: **Cap-independent translation is required for starvation-induced differentiation in yeast.** *Science* 2007, **317**:1224–7.

28. Baird SD, Turcotte M, Korneluk RG, Holcik M: **Searching for IRES.** *RNA* 2006, **12**:1755–85.

29. Baird SD, Lewis SM, Turcotte M, Holcik M: **A search for structurally similar cellular internal ribosome entry sites.** *Nucleic Acids Res* 2007, **35**:4664–77.

30. Thompson SR: **So you want to know if your message has an IRES?** *Wiley Interdiscip Rev RNA* 2012, **3**:697–705.

31. Schepens B, Tinton SA, Bruynooghe Y, Parthoens E, Haegman M, Beyaert R, Cornelis S: **A role for hnRNP C1/C2 and Unr in internal initiation of translation during mitosis.** *EMBO J* 2007, **26**:158–69.

32. Ungureanu NH, Cloutier M, Lewis SM, De Silva N, Blais JD, Bell JC, Holcik M: **Internal ribosome entry site-mediated translation of Apaf-1, but not XIAP, is regulated during UV-induced cell death**. *J Biol Chem* 2006, **281**:15155–15163.

33. Pickering BM, Mitchell SA, Spriggs KA, Stoneley M, Willis AE: **Bag-1 internal ribosome entry segment activity is promoted by structural changes mediated by poly(rC) binding protein 1 and recruitment of polypyrimidine tract binding protein 1.** *Mol Cell Biol* 2004, **24**:5595–605.

34. Grover R, Ray PS, Das S: **Polypyrimidine tract binding protein regulates IRES-mediated translation of p53 isoforms**. *Cell Cycle* 2008, **7**:2189–2198.

35. Pacheco A, Martinez-Salas E: **Insights into the biology of IRES elements through riboproteomic approaches.** *J Biomed Biotechnol* 2010, **2010**:458927.

36. Wang C, Le SY, Ali N, Siddiqui A: **An RNA pseudoknot is an essential structural element of the internal ribosome entry site located within the hepatitis C virus 5' noncoding region.** *RNA* 1995, **1**:526–37.

37. Otto GA, Puglisi JD, Blobel G, Sabatini D, Brown EA, Zhang H, Ping LH, Lemon SM, Buratti E, Tisminetzky S, Zotti M, Baralle FE, Dmitriev SE, Pisarev AV, Rubtsova MP, Dunaevsky YE, Shatsky IN, Fletcher SR, Ali IK, Kaminski A, Digard P, Jackson RJ, Fukushi S, Okada M, Stahl J, Kageyama T, Hoshino FB, Katayama K, He Y, Katze MG, et al.: **The pathway of HCV IRES-mediated translation initiation.** *Cell* 2004, **119**:369–80.

38. Locker N, Easton LE, Lukavsky PJ: **HCV and CSFV IRES domain II mediate eIF2 release during 80S ribosome assembly**. *EMBO J* 2007, **26**:795–805.

39. Lukavsky PJ: **Structure and function of HCV IRES domains**. *Virus Res* 2009, **139**:166–171.

40. Kieft JS, Zhou K, Jubin R, Doudna JA: **Mechanism of ribosome recruitment by hepatitis C IRES RNA.** *RNA* 2001, **7**:194–206.

41. Terenin IM, Dmitriev SE, Andreev DE, Shatsky IN: **Eukaryotic translation initiation machinery can operate in a bacterial-like mode without eIF2**. *Nat Struct Mol Biol* 2008, **15**:836–841.

42. Yu Y, Ji H, Doudna JA, Leary JA: **Mass spectrometric analysis of the human 40S ribosomal subunit: Native and HCV IRES-bound complexes**. *Protein Sci* 2009, **14**:1438–1446.

43. Sengupta J, Nilsson J, Gursky R, Spahn CMT, Nissen P, Frank J: **Identification of the versatile scaffold protein RACK1 on the eukaryotic ribosome by cryo-EM**. *Nat Struct Mol Biol* 2004, **11**:957–962.

44. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2**:418–27.

45. Gstaiger M, Aebersold R: **Applying mass spectrometry-based proteomics to genetics, genomics and network biology.** *Nat Rev Genet* 2009, **10**:617–27.

46. Schuster SC: **Next-generation sequencing transforms today's biology.** *Nat Methods* 2008, **5**:16–8.

47. de Ridder D, de Ridder J, Reinders MJT: **Pattern recognition in bioinformatics.** *Brief Bioinform* 2013, **14**:633–47.

48. James G, Witten D, Hastie T, Tibshirani R: *An Introduction to Statistical Learning*. New York: Springer; 2013.

49. Cortes C, Vapnik V: **Support Vector Networks**. *Mach Learn* 1995, **20**:273~-~297.

50. Wolpert DH, H. D: **The Lack of A Priori Distinctions Between Learning Algorithms**. *Neural Comput* 1996, **8**:1341–1390.

51. Wolpert DH, Macready WG: **No free lunch theorems for optimization**. *IEEE Trans Evol*

Comput 1997, **1**:67–82.

52. Quinlan JR: *C4.5: Programs for Machine Learning*. Morgan Kaufmann; 1993.

53. Kuhn M, Johnson K: *Applied Predictive Modeling*. New York: Springer; 2013.

54. Kuhn M: **Building Predictive Models in R Using the caret Package**. *J Stat Softw* 2008, **28**:1–26.

55. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G: **Support vector machines and kernels for computational biology.** *PLoS Comput Biol* 2008, **4**:e1000173.

56. Vapnik V, Lerner A: **Pattern Recognition using Generalized Portrait Method**. *Autom Remote Control* 1963, **24**.

57. Vapnik V, Chervonenkis A: **A note on one class of perceptrons**. *Autom Remote Control* 1964, **25**.

58. Liu J, Gough J, Rost B: **Distinguishing protein-coding from non-coding RNAs through support vector machines.** *PLoS Genet* 2006, **2**:e29.

59. Hamp T, Rost B: **Evolutionary profiles improve protein-protein interaction prediction from sequence.** *Bioinformatics* 2015, **31**:1945-50.

60. Livi CM, Blanzieri E: **Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures.** *BMC Bioinformatics* 2014, **15**:123.

61. Vapnik V: *The Nature of Statistical Learning Theory*. New York, NY: Springer New York; 2000.

62. Hardison RC: **Comparative genomics.** *PLoS Biol* 2003, **1**:E58.

63. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in Saccharomyces genomes by phylogenetic footprinting.** *Science* 2003, **301**:71–6.

64. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241–54.

65. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5**:276–87.

66. Martinsen L, Johnsen A, Venanzetti F, Bachmann L: **Phylogenetic footprinting of non-coding RNA: hammerhead ribozyme sequences in a satellite DNA family of Dolichopoda cave crickets (Orthoptera, Rhaphidophoridae).** *BMC Evol Biol* 2010, **10**:3.

67. Jáuregui R, Abreu-Goodger C, Moreno-Hagelsieb G, Collado-Vides J, Merino E: **Conservation of DNA curvature signals in regulatory regions of prokaryotic genes.** *Nucleic Acids Res* 2003, **31**:6770–7.

68. Komar AA, Lesnik T, Cullin C, Merrick WC, Trachsel H, Altmann M, Altmann M, Trachsel H, Altmann M, Handschin C, Trachsel H, Altmann M, Sonenberg N, Trachsel H, Altmann M, Blum S, Pelletier J, Sonenberg N, Wilson T, Trachsel H, Altmann M, Muller P, Wittmer B, Ruchti F, Lanker S, Trachsel H, Alwine J, Kemp D, Stark G, Ashe M, et al.: **Internal initiation drives the synthesis of**

Ure2 protein lacking the prion domain and affects [URE3] propagation in yeast cells. *EMBO J* 2003, **22**:1199–209.

69. Pang KC, Frith MC, Mattick JS, Frith MC, al. et, Carninci P, al. et, Cavaille J, al. et, Huttenhofer A, al. et, Lau NC, al. et, Lagos-Quintana M, al. et, Lee RC, Ambros V, Okazaki Y, al. et, Ota T, al. et, Kapranov P, al. et, Cawley S, al. et, Kampa D, al. et, Stolc V, al. et, Bertone P, et al.: **Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function.** *Trends Genet* 2006, **22**:1–5.

70. Gruber AR, Bernhart SH, Hofacker IL, Washietl S: **Strategies for measuring evolutionary conservation of RNA secondary structures**. *BMC Bioinformatics* 2008, **9**:122.

71. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP: **SMOTE: synthetic minority over-sampling technique**. *J Artif Intell Res* 2002, **16**:321–357.

72. Cohen J: **A Coefficient of Agreement for Nominal Scales**. *Educ Psychol Meas* 1960, **20**:37–46.

73. Spriggs KA, Bushell M, Mitchell SA, Willis AE: **Internal ribosome entry segment-mediated translation during apoptosis: the role of IRES-trans-acting factors.** *Cell Death Differ* 2005, **12**:585–91.

74. Johannes G, Carter MS, Eisen MB, Brown PO, Sarnow P: **Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray**. *Proc Natl Acad Sci* 1999, **96**:13118–13123.

75. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva E V: **OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs.** *Nucleic Acids Res* 2013, **41**(Database issue):D358-65.

76. Ozcan S, Johnston M: **Function and Regulation of Yeast Hexose Transporters**. *Microbiol Mol Biol Rev* 1999, **63**:554–569.

77. Johnston M, Magasanik B, Ullmann A, Cereghino GP, Scheffler IE, Vallari RC, al. et, Jiang H, al. et, Lagunas R, Samuel D, Yin Z, al. et, Ostling J, al. et, Hardie DG, al. et, Yang X, al. et, Tu J, Carlson M, Lutfiyya LL, Johnston M, Treitel MA, Carlson M, Devit M, al. et, Woods A, al. et, Wilson WA, et al.: **Feasting, fasting and fermenting. Glucose sensing in yeast and other cells.** *Trends Genet* 1999, **15**:29–33.

78. Warringer J, Hult M, Regot S, Posas F, Sunnerhagen P: **The HOG Pathway Dictates the Short-Term Translational Response after Hyperosmotic Shock**. *Mol Biol Cell* 2010, **21**:3080–3092.

79. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS: **High-resolution view of the yeast meiotic program revealed by ribosome profiling.** *Science* 2012, **335**:552–7.

80. Melamed D, Pnueli L, Arava Y: **Yeast translational response to high salinity: global analysis reveals regulation at multiple levels.** *RNA* 2008, **14**:1337–51.

81. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.** *Science* 2009, **324**:218–

23.

82. Mokrejs M, Masek T, Vopálensky V, Hlubucek P, Delbos P, Pospísek M: **IRESite--a tool for the examination of viral and cellular internal ribosome entry sites.** *Nucleic Acids Res* 2010, **38**(Database issue):D131-6.

83. Kowalski LRZ, Kondo K, Inouye M: **Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae**. *Mol Microbiol* 1995, **15**:341–353.

84. Donzeau M, Bourdineaud J-P, Lauquin GJ-M: **Regulation by low temperatures and anaerobiosis of a yeast gene specifying a putative GPI-anchored plasma membrane**. *Mol Microbiol* 1996, **20**:449–459.

85. Ai W, Bertram PG, Tsang CK, Chan T-FF, Zheng XFSFS: **Regulation of Subtelomeric Silencing during Stress Response**. *Mol Cell* 2002, **10**:1295–305.

86. Bester MC, Jacobson D, Bauer FF: **Many Saccharomyces cerevisiae Cell Wall Protein Encoding Genes Are Coregulated by Mss11, but Cellular Adhesion Phenotypes Appear Only Flo Protein Dependent.** *G3 (Bethesda)* 2012, **2**:131–41.

87. Gagiano M, van Dyk D, Bauer FF, Lambrechts MG, Pretorius IS: **Msn1p/Mss10p, Mss11p and Muc1p/Flo11p are part of a signal transduction pathway downstream of Mep2p regulating invasive growth and pseudohyphal differentiation in Saccharomyces cerevisiae**. *Mol Microbiol* 1999, **31**:103–116.

88. Kim TS, Kim HY, Yoon JH, Kang HS: **Recruitment of the Swi/Snf complex by Ste12-Tec1 promotes Flo8-Mss11-mediated activation of STA1 expression.** *Mol Cell Biol* 2004, **24**:9542–56.

89. Sá-Correia I, dos Santos SC, Teixeira MC, Cabrito TR, Mira NP: **Drug:H+ antiporters in chemical stress response in yeast.** *Trends Microbiol* 2009, **17**:22–31.

90. Jungwirth H, Kuchler K: **Yeast ABC transporters-- a tale of sex, stress, drugs and aging.** *FEBS Lett* 2006, **580**:1131–8.

91. Zhou W, Edelman GM, Mauro VP: **Transcript leader regions of two Saccharomyces cerevisiae mRNAs contain internal ribosome entry sites that function in living cells.** *Proc Natl Acad Sci U S A* 2001, **98**:1531–6.

92. Seino A, Yanagida Y, Aizawa M, Kobatake E: **Translational control by internal ribosome entry site in Saccharomyces cerevisiae.** *Biochim Biophys Acta* 2005, **1681**:166–74.

93. Hecht K, Bailey JE, Minas W: **Polycistronic gene expression in yeast versus cryptic promoter elements.** *FEMS Yeast Res* 2002, **2**:215–24.

94. Iizuka N, Najita L, Franzusoff A, Sarnow P: **Cap-dependent and cap-independent translation by internal initiation of mRNAs in cell extracts prepared from Saccharomyces cerevisiae.** *Mol Cell Biol* 1994, **14**:7322–30.

95. Petryk N, Zhou Y-F, Sybirna K, Mucchielli M-H, Guiard B, Bao W-G, Stasyk O V., Stasyk OG, Krasovska OS, Budin K, Reymond N, Imbeaud S, Coudouel S, Delacroix H, Sibirny A, Bolotin-

Fukuhara M: **Functional Study of the Hap4-Like Genes Suggests That the Key Regulators of Carbon Metabolism HAP4 and Oxidative Stress Response YAP1 in Yeast Diverged from a Common Ancestor**. *PLoS One* 2014, **9**:e112263.

96. Platt JC: **Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods**. In *Adv Large Margin Classif*. Edited by Smola AJ, Bartlett P, Schoelkopf B, Schuurmans D. MIT press; 2000:61--74.

97. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Res* 2013, **41**(Database issue):D808-15.

98. Kolaczyk ED, Csárdi G: *Statistical Analysis of Network Data with R*. *Volume 65*. New York, NY: Springer New York; 2014. [*Use R!*]

99. Peterson GJ, Pressé S, Peterson KS, Dill KA, Hughes D, Cirz R, Chin J, Andes D, Crécy-Lagard V de, Craig W, Taylor I, Linding R, Warde-Farley D, Liu Y, Pesquita C, Lynch M, O'Hely M, Walsh B, Force A, Ting C, Tsaur S, Sun S, Browne W, Chen Y, Dutkowski J, Tiuryn J, Ito T, Tashiro K, Muta S, Ozawa R, et al.: **Simulated Evolution of Protein-Protein Interaction Networks with Realistic Topology**. *PLoS One* 2012, **7**:e39052.

100. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**(6761 Suppl):C47-52.

101. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks.** *Proc Natl Acad Sci U S A* 2003, **100**:12123–8.

102. Kaltenbach H-M, Stelling J: **Modular analysis of biological networks.** *Adv Exp Med Biol* 2012, **736**:3–17.

103. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E: **Fast unfolding of communities in large networks**. *J Stat Mech Theory Exp* 2008, **2008**:P10008.

104. Csardi G, Nepusz T: **The igraph software package for complex network research**. *InterJournal* 2006, **Complex Sy**:1965.

105. Dong J, Horvath S: **Understanding network concepts in modules.** *BMC Syst Biol* 2007, **1**:24.

106. Kronja I, Orr-Weaver TL: **Translational regulation of the cell cycle: when, where, how and why?** *Philos Trans R Soc Lond B Biol Sci* 2011, **366**:3638–52.

107. Holcik M, Sonenberg N: **Translational control in stress and apoptosis.** *Nat Rev Mol Cell Biol* 2005, **6**:318–27.

108. Jeong H, Mason SP, Barabási AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411**:41–2.

109. Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F: **Further understanding human disease genes by comparing with housekeeping genes and other genes.** *BMC Genomics* 2006, **7**:31.

110. Mense SM, Zhang L: **Heme: a versatile signaling molecule controlling the activities of diverse regulators ranging from transcription factors to MAP kinases.** *Cell Res* 2006,

16:681–92.

111. Sammons MA, Samir P, Link AJ: **Saccharomyces cerevisiae Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9.** *Biochem Biophys Res Commun* 2011, **406**:13–9.

112. Tsvetanova NG, Klass DM, Salzman J, Brown PO: **Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae.** *PLoS One* 2010, **5**:e12671.

113. Gerbasi VR, Link AJ: **The myotonic dystrophy type 2 protein ZNF9 is part of an ITAF complex that promotes cap-independent translation.** *Mol Cell Proteomics* 2007, **6**:1049–58.

114. Antonucci L, D'Amico D, Di Magno L, Coni S, Di Marcotullio L, Cardinali B, Gulino A, Ciapponi L, Canettieri G: **CNBP regulates wing development in Drosophila melanogaster by promoting IRES-dependent translation of dMyc.** *Cell Cycle* 2014, **13**:434–9.

115. Rojas M, Farr GW, Fernandez CF, Lauden L, McCormack JC, Wolin SL: **Yeast Gis2 and its human ortholog CNBP are novel components of stress-induced RNP granules.** *PLoS One* 2012, **7**:e52824.

116. Shively CA, Eckwahl MJ, Dobry CJ, Mellacheruvu D, Nesvizhskii A, Kumar A: **Genetic networks inducing invasive growth in Saccharomyces cerevisiae identified through systematic genome-wide overexpression.** *Genetics* 2013, **193**:1297–310.

117. Westermann B, Gaume B, Herrmann JM, Neupert W, Schwarz E: **Role of the mitochondrial DnaJ homolog Mdj1p as a chaperone for mitochondrially synthesized and imported proteins.** *Mol Cell Biol* 1996, **16**:7063–71.

118. Hung G-C, Brown CR, Wolfe AB, Liu J, Chiang H-L: **Degradation of the gluconeogenic enzymes fructose-1,6-bisphosphatase and malate dehydrogenase is mediated by distinct proteolytic pathways and signaling events.** *J Biol Chem* 2004, **279**:49138–50.

119. Tangsombatvichit P, Semkiv M V, Sibirny AA, Jensen LT, Ratanakhanokchai K, Soontorngun N: **Zinc cluster protein Znf1, a novel transcription factor of non-fermentative metabolism in Saccharomyces cerevisiae.** *FEMS Yeast Res* 2015, **15**.

120. Vicent I, Navarro A, Mulet JM, Sharma S, Serrano R: **Uptake of inorganic phosphate is a limiting factor for Saccharomyces cerevisiae during growth at low temperatures.** *FEMS Yeast Res* 2015, **15**.

121. Mazur P, Morin N, Baginsky W, el-Sherbeini M, Clemas JA, Nielsen JB, Foor F: **Differential expression and function of two homologous subunits of yeast 1,3-beta-D-glucan synthase.** *Mol Cell Biol* 1995, **15**:5671–81.

122. Zhao C, Jung US, Garrett-Engele P, Roe T, Cyert MS, Levin DE: **Temperature-induced expression of yeast FKS2 is under the dual control of protein kinase C and calcineurin.** *Mol Cell Biol* 1998, **18**:1013–22.

123. Agarwal AK, Rogers PD, Baerson SR, Jacob MR, Barker KS, Cleary JD, Walker LA, Nagle DG, Clark AM: **Genome-wide expression profiling of the response to polyene, pyrimidine, azole, and echinocandin antifungal agents in Saccharomyces cerevisiae.** *J Biol*

Chem 2003, **278**:34998–5015.

124. Travers KJ, Patil CK, Wodicka L, Lockhart DJ, Weissman JS, Walter P: **Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation.** *Cell* 2000, **101**:249–58.

125. Orlean P: **Architecture and biosynthesis of the Saccharomyces cerevisiae cell wall.** *Genetics* 2012, **192**:775–818.

126. Kapteyn JC, Van Egmond P, Sievi E, Van Den Ende H, Makarow M, Klis FM: **The contribution of the O-glycosylated protein Pir2p/Hsp150 to the construction of the yeast cell wall in wild-type cells and beta 1,6-glucan-deficient mutants.** *Mol Microbiol* 1999, **31**:1835–44.

127. Yun DJ, Zhao Y, Pardo JM, Narasimhan ML, Damsz B, Lee H, Abad LR, D'Urzo MP, Hasegawa PM, Bressan RA: **Stress proteins on the yeast cell surface determine resistance to osmotin, a plant antifungal protein.** *Proc Natl Acad Sci U S A* 1997, **94**:7082–7.

128. Klis FM, Mol P, Hellingwerf K, Brul S: **Dynamics of cell wall structure in Saccharomyces cerevisiae.** *FEMS Microbiol Rev* 2002, **26**:239–56.

129. Russo P, Simonen M, Uimari A, Teesalu T, Makarow M: **Dual regulation by heat and nutrient stress of the yeast HSP150 gene encoding a secretory glycoprotein.** *Mol Gen Genet* 1993, **239**:273–80.

130. Toh-e A, Yasunaga S, Nisogi H, Tanaka K, Oguchi T, Matsui Y: **Three yeast genes, PIR1, PIR2 and PIR3, containing internal tandem repeats, are related to each other, and PIR1 and PIR2 are required for tolerance to heat shock.** *Yeast* 1993, **9**:481–94.

131. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273–97.

132. Negishi T, Ohya Y: **The cell wall integrity checkpoint: coordination between cell wall synthesis and the cell cycle.** *Yeast* 2010, **27**:513–9.

133. Wanless AG, Lin Y, Weiss EL: **Cell morphogenesis proteins are translationally controlled through UTRs by the Ndr/LATS target Ssd1.** *PLoS One* 2014, **9**:e85212.

134. Duch A, de Nadal E, Posas F: **The p38 and Hog1 SAPKs control cell cycle progression in response to environmental stresses.** *FEBS Lett* 2012, **586**:2925–31.

135. Giannattasio M, Sommariva E, Vercillo R, Lippi-Boncambi F, Liberi G, Foiani M, Plevani P, Muzi-Falconi M: **A dominant-negative MEC3 mutant uncovers new functions for the Rad17 complex and Tel1.** *Proc Natl Acad Sci U S A* 2002, **99**:12997–3002.

136. Francisco L, Wang W, Chan CS: **Type 1 protein phosphatase acts in opposition to IpL1 protein kinase in regulating yeast chromosome segregation.** *Mol Cell Biol* 1994, **14**:4731–40.

137. Muñoz I, Simón E, Casals N, Clotet J, Ariño J: **Identification of multicopy suppressors of cell cycle arrest at the G1-S transition in Saccharomyces cerevisiae.** *Yeast* 2003, **20**:157–69.

138. Rockmill B, Roeder GS: **A meiosis-specific protein kinase homolog required for**

chromosome synapsis and recombination. *Genes Dev* 1991, **5**:2392–404.

139. Ahmed K: **Joining the cell survival squad: an emerging role for protein kinase CK2**. *Trends Cell Biol* 2002, **12**:226–230.

140. Mulet JM, Leube MP, Kron SJ, Rios G, Fink GR, Serrano R: **A novel mechanism of ion homeostasis and salt tolerance in yeast: the Hal4 and Hal5 protein kinases modulate the Trk1-Trk2 potassium transporter.** *Mol Cell Biol* 1999, **19**:3328–37.

141. Elion EA, Grisafi PL, Fink GR: **FUS3 encodes a cdc2+/CDC28-related kinase required for the transition from mitosis into conjugation.** *Cell* 1990, **60**:649–64.

142. Hinnebusch AG: **Mechanisms of gene regulation in the general control of amino acid biosynthesis in Saccharomyces cerevisiae.** *Microbiol Rev* 1988, **52**:248–73.

143. Hinnebusch AG, Natarajan K: **Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress.** *Eukaryot Cell* 2002, **1**:22–32.

144. Mira NP, Teixeira MC, Sá-Correia I: **Adaptive response and tolerance to weak acids in Saccharomyces cerevisiae: a genome-wide view.** *OMICS* 2010, **14**:525–40.

145. Bilski P, Li MY, Ehrenshaft M, Daub ME, Chignell CF: **Vitamin B6 (pyridoxine) and its derivatives are efficient singlet oxygen quenchers and potential fungal antioxidants.** *Photochem Photobiol* 2000, **71**:129–34.

146. Chumnantana R, Yokochi N, Yagi T: **Vitamin B6 compounds prevent the death of yeast cells due to menadione, a reactive oxygen generator.** *Biochim Biophys Acta* 2005, **1722**:84–91.

147. Padilla PA, Fuge EK, Crawford ME, Errett A, Werner-Washburne M: **The Highly Conserved, Coregulated SNO and SNZ Gene Families in Saccharomyces cerevisiae Respond to Nutrient Limitation**. *J Bacteriol* 1998, **180**:5718–5726.

148. Barrett LW, Fletcher S, Wilton SD: *Untranslated Gene Regions and Other Non-Coding Elements*. Basel: Springer Basel; 2013. [*SpringerBriefs in Biochemistry and Molecular Biology*]

149. Su W-Y, Xiong H, Fang J-Y: **Natural antisense transcripts regulate gene expression in an epigenetic manner.** *Biochem Biophys Res Commun* 2010, **396**:177–81.

150. Huang B, Lu J, Byström AS: **A genome-wide screen identifies genes required for formation of the wobble nucleoside 5-methoxycarbonylmethyl-2-thiouridine in Saccharomyces cerevisiae.** *RNA* 2008, **14**:2183–94.

151. Despons L, Wirth B, Louis VL, Potier S, Souciet J-L: **An evolutionary scenario for one of the largest yeast gene families.** *Trends Genet* 2006, **22**:10–5.

152. Chen Y-W, Lin C-J: **Combining SVMs with Various Feature Selection Strategies**. In *Featur Extr Found Appl*. Edited by Guyon I, Nikravesh M, Gunn S, Zadeh LA. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006:315–324.

153. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**:2507–17.

154. Sugumaran V, Muralidharan V, Ramachandran KI: **Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing**. *Mech Syst Signal Process* 2007, **21**:930–942.

155. Becker N, Werft W, Toedt G, Lichter P, Benner A: **penalizedSVM: a R-package for feature selection SVM classification.** *Bioinformatics* 2009, **25**:1711–2.

156. Lertampaiporn S, Thammarongtham C, Nukoolkit C, Kaewkamnerdpong B, Ruengjitchatchawalya M: **Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification.** *Nucleic Acids Res* 2013, **41**:e21.

157. Schliep KP: **phangorn: phylogenetic analysis in R.** *Bioinformatics* 2011, **27**:592–3.

158. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL: **Domain enhanced lookup time accelerated BLAST**. *Biol Direct* 2012, **7**:12.

159. Bompfünewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, Will S: **Variations on RNA folding and alignment: lessons from Benasque.** *J Math Biol* 2008, **56**:129–44.

160. Castelli LM, Lui J, Campbell SG, Rowe W, Zeef LAH, Holmes LEA, Hoyle NP, Bone J, Selley JN, Sims PFG, Ashe MP: **Glucose depletion inhibits translation initiation via eIF4A loss and subsequent 48S preinitiation complex accumulation, while the pentose phosphate pathway is coordinately up-regulated.** *Mol Biol Cell* 2011, **22**:3379–93.

161. De Las Peñas A, Pan S-J, Castaño I, Alder J, Cregg R, Cormack BP: **Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing.** *Genes Dev* 2003, **17**:2245–58.

162. Hansen KR, Burns G, Mata J, Volpe TA, Martienssen RA, Bähler J, Thon G: **Global effects on gene expression in fission yeast by silencing and RNA interference machineries.** *Mol Cell Biol* 2005, **25**:590–601.

163. Halme A, Bumgarner S, Styles C, Fink GR: **Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast.** *Cell* 2004, **116**:405–15.

164. Panwar B, Arora A, Raghava GPS: **Prediction and classification of ncRNAs using structural information.** *BMC Genomics* 2014, **15**:127.

165. Kuri-Morales A: **An automated search space reduction methodology for large databases**. In *Adv Data Mining Appl Theor Asp*. *Volume 7987*. Edited by Perner P. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:11–24. [*Lecture Notes in Computer Science*]

166. Provost F, Fawcett T: **Robust Classification for Imprecise Environments**. *Mach Learn* , **42**:203–231.

167. Silva LM, Teixeira FC de S, Ortega JM, Zárate LE, Nobre CN: **Improvement in the prediction of the translation initiation site through balancing methods, inclusion of acquired knowledge and addition of features to sequences of mRNA.** *BMC Genomics* 2011, **12 Suppl 4**(Suppl 4):S9.

168. Batuwita R, Palade V: **microPred: effective classification of pre-miRNAs for human miRNA gene prediction.** *Bioinformatics* 2009, **25**:989–95.

169. Batuwita R, Palade V: **A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems**. In *2009 Int Conf Mach Learn Appl*. IEEE; 2009:545–550.

170. Karatzoglou A, Smola A, Hornik K, Zeileis A: **kernlab - An *S4* Package for Kernel Methods in *R***. *J Stat Softw* 2004, **11**:1–20.

171. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *J R Stat Soc Ser B* 1995, **57**:289–300.

172. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25–9.

173. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23**:257–8.

174. Box GEP, Cox DR: **An analysis of transformations**. *J R Stat Soc Ser B* 1964:211–252.

12 Anexo 1: Artículo original derivado de este proyecto

Peguero-Sanchez et al. BMC Genomics (2015) 16:1059
DOI 10.1186/s12864-015-2266-x

BMC Genomics

RESEARCH ARTICLE Open Access

IRES-dependent translated genes in fungi: computational prediction, phylogenetic conservation and functional association

Esteban Peguero-Sanchez, Liliana Pardo-Lopez and Enrique Merino*

Abstract

Background: The initiation of translation via cellular internal ribosome entry sites plays an important role in the stress response and certain physiological conditions in which canonical cap-dependent translation initiation is compromised. Currently, only a limited number of these regulatory elements have been experimentally identified. Notably, cellular internal ribosome entry sites lack conservation of both the primary sequence and mRNA secondary structure, rendering their identification difficult. Despite their biological importance, the currently available computational strategies to predict them have had limited success. We developed a bioinformatic method based on a support vector machine for the prediction of internal ribosome entry sites in fungi using the 5′-UTR sequences of 20 non-redundant fungal organisms. Additionally, we performed a comparative analysis and characterization of the functional relationships among the gene products predicted to be translated by this cap-independent mechanism.

Results: Using our method, we predicted 6,532 internal ribosome entry sites in 20 non-redundant fungal organisms. Some orthologous groups were enriched with our positive predictions. This is the case of the HSP70 chaperone family, which remarkably has two verified internal ribosome entry sites, one in humans and the other in flies. A second example is the orthologous group of the eIF4G repression protein Sbp1p, which has two homologous genes known to be translated by this cap-independent mechanism, one in mice and the other in yeast. These examples emphasize the wide conservation of these regulatory elements as a result of selective pressure. In addition, we performed a protein-protein interaction network characterization of the gene products of our positive predictions using *Saccharomyces cerevisiae* as a model, which revealed a highly connected and modular topology, suggesting a functional association. A remarkable example of this functional association is our prediction of internal ribosome entry sites elements in three components of the RNA polymerase II mediator complex.

Conclusions: We developed a method for the prediction of cellular internal ribosome entry sites that may guide experimental and bioinformatic analyses to increase our understanding of protein translation regulation. Our analysis suggests that fungi show evolutionary conservation and functional association of proteins translated by this cap-independent mechanism.

Keywords: IRES, mRNA, Translation, Comparative genomics, Fungi, SVM, Protein-protein interaction networks, Stress response

* Correspondence: merino@ibt.unam.mx
Departamento de Microbiología Molecular, Instituto de Biotecnología, UNAM, Av. Universidad 2001, Cuernavaca, Morelos CP 62210, Mexico

Background

Eukaryotic cells regulate the synthesis of proteins using various mechanisms. Among them, protein translation control provides faster changes in protein levels when compared, for example, to transcriptional responses [1]. Under stress and other physiological and physiopathological conditions, translation is heavily repressed to conserve cellular resources. Nevertheless, a set of proteins, mostly related to stress responses that mediate cell adaptation to diverse stimuli or that are necessary for the regulation of developmental processes, are selectively synthesized. The prevalence of translational control has been assessed in yeast and other fungal organisms [2–5]. One of the mechanisms that allows such selective protein expression under these conditions is internal ribosome entry site (IRES)-dependent translation [1, 6–9].

Importantly, translation initiation is widely considered to be the most regulated step in protein translation [1]. Under normal conditions, translation initiation proceeds via the canonical or 5'-cap-dependent mechanism. In this process, the translation machinery recognizes the 5'-m^7G-cap modification of the mRNA, paving the way for translation initiation. However, under certain circumstances, some components of the translation machinery are depleted, and 5'-cap recognition is suppressed. These conditions hinder the canonical translation initiation mechanism. IRESs allow the binding of the translation machinery to mRNA independently of 5'-cap recognition, enabling translational initiation to proceed [6, 7, 10]. The first IRES was reported in the 5'-UTR of picornaviruses [11]. Subsequently, a number of IRESs were described in multiple viral transcripts. They enable viral protein production using the host translational machinery when the global synthesis is repressed due to the infection process. Shortly afterwards, the first cellular IRES was identified in the 5'-UTR of the BiP chaperone, allowing its translation in poliovirus-infected cells [12]. Currently, there are more than 100 reported cellular IRESs [13].

Until now, a high-throughput method for the detection of IRESs is not available; each candidate has to be tested individually in a procedure that involves different stringent controls to verify its activity [14]. Thus, we believed that a bioinformatic approach to discover new potential IRESs would vastly reduce the number of candidates to be tested. Nevertheless, the prediction of cellular IRESs presents a considerable challenge due to their lack of sequence and structure conservation, even in homologous genes [15]. For this reason, and to the best of our knowledge, current predictive strategies have had very limited success [16]. To develop a computational methodology to identify IRES-specific patterns and accurately predict these regulatory elements in fungal species, we implemented a support vector machine (SVM) method using 5'-UTR sequence characteristics and comparative genomic features. Subsequently, an enrichment analysis of our initial IRES predictions in clusters of orthologous yeast genes allowed us to identify the most likely IRES candidates. In this article, using *S. cerevisiae* as a model, we present a detailed analysis of the protein-protein interaction (PPI) network of genes translated by these top IRES predictions. The notable enrichment in orthologous groups, the PPI analysis and the functional evaluation of our predictions enabled us to formulate biological hypotheses concerning the evolutionary conservation and genome-wide associations of IRESs.

Results and discussion

Prediction of IRESs in 5'-UTR regions

To identify 5'-UTR regions bearing IRESs, we developed a method based on machine learning and comparative genomics. Our method is focused on the unstructured A-rich IRESs found in fungal organisms, such as those identified in *S. cerevisiae* [17, 18], and does not include highly structured IRESs found in higher organisms [16]. We employed sequence composition-based features, the minimum folding energies (MFEs) of the RNAs, and certain orthologous group comparative properties to generate a total of 29 features, which are listed in Additional file 1 (see Methods).

Cross-validation was performed to evaluate the performance of our method and for parameter optimization. The SMOTE procedure allowed the generation of synthetic positive cases that were used for training and testing, as described in the Methodology section. This set (consisting of positive and negative cases) was randomly split into ten parts; of these, one was used for testing, and the rest were used for training. The process continued until all the parts were individually used for training (10 steps). The performance measures were determined (accuracy and Cohen's kappa) for each of the steps, and the mean values were calculated. This entire process was repeated 30 times. From a total of 32 different combinations of SVM parameters, the optimized parameters were sigma and cost. We selected the model with the best performance measures achieving an accuracy of 94.3 % and a Cohen's kappa of 0.828 [19, 20]. The estimated values of sensitivity and specificity using a confusion matrix for our model were 0.94 and 0.98, respectively. Thereafter, we used our model to make predictions for 99,759 sequences from 20 independent fungal organisms. Our method classified 6,532 sequences as containing IRESs (positive predictions) and 93,227 sequences as not containing IRESs (negative predictions). The positive predictions represent 6.8 % of the total sequences used. This number is in close agreement with the estimate of the proportion of cellular mRNAs that could be translated using cap-independent mechanisms, according to cDNA microarray data (10-15 %) [21, 22]. In order to have a negative control with the exactly the same

number of sequences as in our IRES analysis, for every gene initially considered in our study, we analyzed the 60 nt immediately upstream of the translation termination codon since it is expected that in this coding region, the presence of IRES would be minimal or nonexistent. The number of sequences analyzed as negative control was 99,759 and of these only 317 were predicted as containing IRESs (false positives). This corresponds to 0.3 % of the total negative control set. Considering that our IRES analysis included the same number of UTR sequences (99,759) and that 6,532 of them were predicted as containing IRESs, the false discovery rate of our predictions was evaluated to be 5 %.

Evolutionarily conserved patterns related to IRES-dependent translation are found in the 5′-UTRs of fungi

Of the aforementioned 6,532 positive predictions, 815 showed distinctive features of evolutionary conservation as analyzed by orthologous group enrichment (False discovery rate (FDR) < 0.05). These predictions included 86 orthologous groups (OG) out of the 22,605 considered (Fig. 1).

Several of the enriched groups contain genes implicated in the stress response. Some of these groups have homologous genes experimentally verified as genes containing IRESs in other organisms. We discuss the biological relevance of the most enriched groups below.

The most significant group yielded an enrichment *p*-value of 1 x 10^{-26}, corresponding to 78 positive predictions out of the 265 gene members of this group encoded in the 20 non-redundant fungal genomes used in our analysis. The proteins in this group are transmembrane sugar transporters and glucose sensors (hexose transporters group). These transporters have a wide array of affinities and are regulated by glucose concentration, allowing adaptation to changing conditions in nutrient levels; their function and regulation are reviewed in [23]. Furthermore, experiments using ribosome profiling analysis have shown that six hexose transporters genes—*HXT1*, *HXT2*, *HXT4*, *HXT5*, *HXT9*, and *GAL2*—are translationally up-regulated in response to osmotic stress [24]. Importantly, the genes encoding four of these proteins—Hxt1p, Hxt5p, Hxt9p and Gal2p—were predicted to contain IRESs using our method in most of our studied organisms. Translational up-regulation was preferentially mediated by strengthened polysome association in the 5′-UTR after osmotic stress and not only by increased mRNA levels. Additionally, an increase in polysomal mRNA led to incremental protein production [24]. This finding is in good agreement with our predictions because increased ribosome occupancy

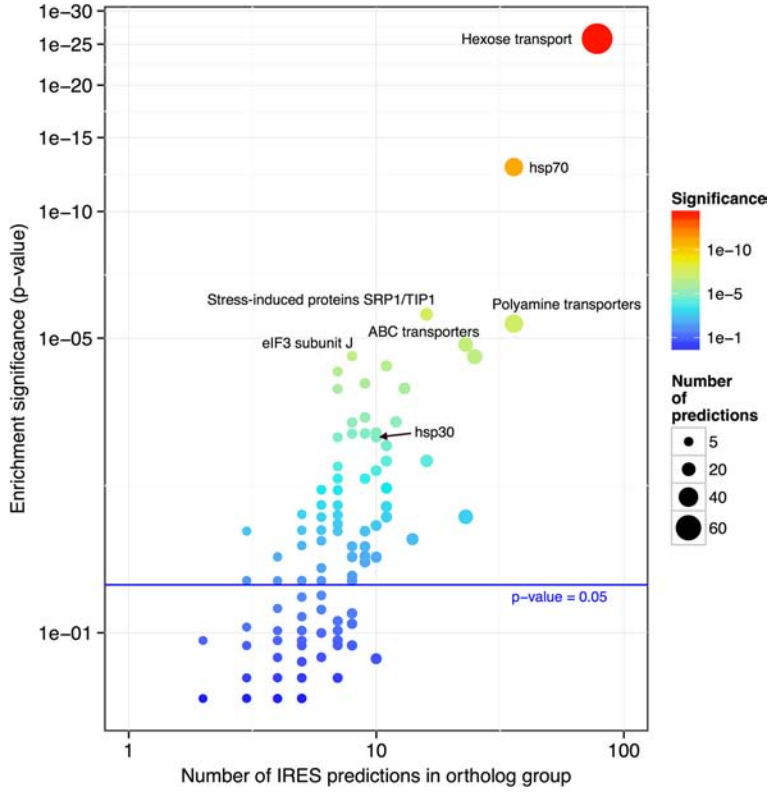

Fig. 1 Predictions of enriched orthologous groups. Some of the most relevant groups are labeled. The statistical significance of the enriched orthologous groups is highlighted. The number of predictions within each group is proportional to the size of the points

in the 5'-UTR has been linked to IRES-dependent translation [3].

The second-most enriched group corresponds to the HSP70 family (36 predictions out of 107, p-value of 1.7 x 10^{-13}). Acting as chaperones, the proteins in this family are conserved in virtually all organisms and are used by cells to contend with several types of stress, including heat. There is evidence of translational control and increased ribosome occupancy in the mRNA of *SSA4* (which was predicted to contain an IRES by our model) in response to different stress conditions, such as high salinity [25] and starvation conditions, in which its translation efficiency increased 2.5-fold [2]. There are two members of this family that have experimentally verified IRESs, one in humans and the other in flies [13]. This result may be explained by the hypothesis that IRES-dependent translation initiation is conserved across species in phylogenetically related proteins.

The third-most enriched group includes the stress-induced Srp1p/Tip1p family (16 predictions out of 40, p-value of 2.0 x 10^{-6}). Several members of this family are known to be induced by various stress conditions, including low temperatures [26], hypoxia [27] and nitrogen starvation [28]. A number of members of the SRP/TIP1 family are regulated by the transcriptional factor Mss11p [29]. Remarkably, Mss11p, Msn1p and Flo8p are part of the signal transduction pathway that regulates pseudohyphal differentiation and filamentous growth [30]. Furthermore, Mss11p and Flo8p bind cooperatively to the *STA1* promoter, leading to the filamentous and invasive growth response [31]. Significantly, the genes coding for Flo8p and Msn1p are translated in an IRES-dependent manner, as are 7 additional genes involved in invasive growth [17]. One hypothesis that could explain these observations is that IRES-dependent translation is required when the selective co-expression of proteins under stress conditions is needed, for example, in some regulatory or interaction networks.

The fourth-most enriched group (36 predictions out of 185, p-value of 3.9 x 10^{-6}) represents a subset of the major facilitator superfamily, more specifically genes that code for H^+ antiporters. These enzymes are crucial for multidrug resistance and chemical stress responses in yeast [32]. In this regard, it has been demonstrated that *PDR15* is translationally regulated and that its 5'-UTR shows increased levels of ribosome occupancy in response to high salinity [25]. Similarly, *PDR5* and *PDR12* (which were positive IRES predictions according to our model) showed a positive correlation between ribosome 5'-UTR occupancy and translational efficiency during different developmental stages; this trend has been linked to translationally regulated genes. Importantly, a similar correlation was observed for yeast IRESs [3].

The fifth-most significant group (23 predictions out of 92, p-value of 1.5 x 10^{-5}) is the ATP-binding cassette (ABC) family. Its members participate in many biological processes that include vacuolar detoxification, pleiotropic drug resistance (PDR) and stress adaptation (reviewed in [33]). Yap1p participates in the PDR regulation network, and its encoding gene has a verified IRES [34]. Additionally, we predicted four genes containing IRESs in this regulatory network (*PDR5*, *PDR12*, *SNQ2* and *STP5*), two of which have been shown to be directly regulated by Yap1p (*SNQ2* and *PDR5*) [33]. It is important to note that although *YAP1* was not used to train our SVM, it was one of the genes predicted as having an IRES. The above-described case constitutes another example consistent with the hypothesis of multiple IRES-dependent genes in the same regulatory or interaction network [7]. A second example of a gene with an experimentally verified IRES [35–37], that was not used in the training procedure of our SVM, but successfully identified as having an IRES by our method is HAP4. Remarkably, there is functional evidence that HAP4 and YAP1 diverged from a common ancestor [38]. Considering the statistical and biological aspects of the aforementioned predictions, we believe that our results validate our method and support IRES-dependent translation conservation in fungi (Fig. 1).

Selection of top IRES predictions

The product of the orthologous group enrichment and the complement of the SVM class posterior probability [39] (1-probability that a prediction is an IRES based on SVM output) was used as a criterion for ranking our predictions. We defined a threshold of 0.05 (lower values indicate better predictions) for selecting the best predictions. Our top IRES predictions included only 801 out of the 6,532 total positive predictions (12 %) and represented 0.8 % of the entire set of genes considered in our analysis (99,759). For *S. cerevisiae*, 174 genes out of its nearly 6,000 coding genes were included in the top-predictions category. The advantage of our selection procedure (see Methods) is that it takes into account the similarity of features found in each sequence compared with those of the experimentally verified IRESs used in this study (given by the posterior class probability) and the enrichment of IRES predictions in phylogenetically related genes across organisms, which could indicate a selective pressure to conserve IRES-dependent translational control. All further analyses in our study were performed using this sub-set of 174 top IRES predictions in *S. cerevisiae*.

Gene ontology enrichment analysis of the predicted IRESs in *S. cerevisiae*

We performed a gene ontology (GO) [40, 41] enrichment analysis corresponding to "Biological Process (BP)" terms

for the top IRES predictions of 174 genes. We found 28 significantly enriched GO terms using FDR adjustment (FDR < 0.1) after summarizing them using REVIGO [42] (Fig. 2). It is worth noting that a number of the enriched GO terms presented here have been associated with 5'-cap-dependent translation suppression and selective protein production through 5'-cap-independent translation, and in several cases, a detailed study of those genes translated in a selective manner led to the discovery of new IRESs. Some of the aforementioned conditions include developmental processes, transport, cell communication [43], filamentous growth [17] and response to stress (reviewed in [7, 44]). As such, these results indicate that our predictions are clearly different from those produced randomly, not only at the phylogenetic level (as shown by orthologous group enrichment) but also at the functional level of proteins participating in specific biological processes.

Network analysis of the predictions

A functional PPI network comprising the 174 top IRES predictions in *S. cerevisiae* was constructed using data from the STRING database [45]. This database provides information not only for direct physical protein-protein interactions but also considers a broader set of "functional protein-protein associations" comprising participation in common metabolic pathways, co-regulation, and participation in larger structural assemblies [45]. To characterize the PPI network from the perspective of its connectivity, we selected the parameter of network density because it describes the global level of cohesion. Network density is calculated as the ratio of observed connections to possible connections (possible connections refers to the number of links in a fully connected network) [46]. Network density has an intuitive biological meaning in this context because larger network densities are related to higher levels of functional association.

Statistical simulation was used to test how the network density of the PPI network of the top IRES predictions compares with a random network (see Methods). Remarkably, the network density had a value of 0.071, higher than that of the expected value of a random network (0.0461; *p*-value of 1.3×10^{-4}). These values show that the PPI network built from the top predicted IRESs was more cohesive and significantly different from a random network of the same size.

IRES-dependent translated proteins are functionally associated into biologically significant modules that participate in specific processes

It has been widely demonstrated that cells perform most of their functions in a modular fashion [47–51]. The prevailing definition of modularity considers a set of proteins connected working together physically or functionally to perform related functions [47, 52]. This definition implies that cellular processes occur via the coordinated action of a number of molecules. Therefore, it is expected that the density of connections in each module will be higher than the density of connections in the entire network because the proteins within each module share a common, relatively homogenous set of functions [48]. In addition, the decomposition of a network into functionally related sub-parts can offer valuable information on how the complete system works, thus facilitating its analysis.

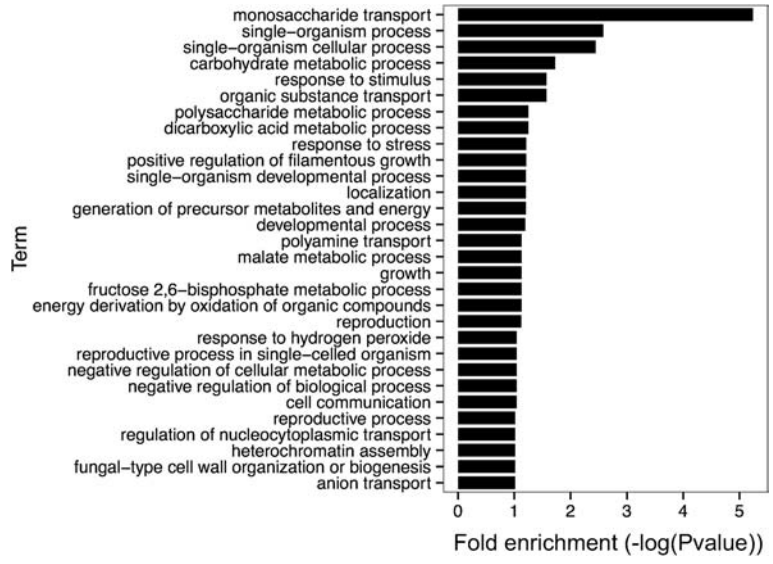

Fig. 2 GO-term enrichment analysis for the top IRES predictions. The GO-term enrichment analysis was performed using the GOstats package for R. Fold enrichment values are represented as the minus base 10 log of their corresponding p-values

To characterize the PPI network of the top IRES predictions in terms of its modular structure, we used the Louvain method multi-level unsupervised clustering (module-finding) algorithm [53] implemented in the igraph package [54]. In our study, a total of 9 modules were obtained. One of these modules contained only one protein and was therefore excluded from further evaluations. To explore the biological function of each module, we conducted GO-term enrichment analysis using the same procedure as that applied to the complete set of predictions. This analysis revealed a significant GO-term enrichment (FDR < 0.02) in 7 out of the 8 modules (Fig. 3a). Interestingly, each module exhibited a unique functional specialization because only 6 of the total 117 GO terms were shared (represented by multiple colored bars in Fig. 3a). A closer inspection inside each module also revealed substantial functional homogeneity (Fig. 3a).

The aforementioned results are in good agreement with previous findings of comparable functional module enrichments [55] and support the validity of the clustering procedure used. To facilitate discussion, representative names were assigned to each module based on GO terms: Module 1, heme biosynthesis; Module 2, respiration and energy generation; Module 3, cell wall organization; Module 4, cell cycle; Module 5, transport; Module 6, pyridoxine biosynthesis; Module 7, RNA processing; and Module 8, undefined function, because no GO term enrichment was found.

In general, higher density values are found when proteins are properly classified into biological modules [55]. For this reason, we compared the modules with the entire network of top IRES predictions in terms of density. Additionally, we compared the density of each module with that of a randomly generated network of a corresponding size (Fig. 4). All modules showed statistically significant higher density values compared with both random networks and the entire network of top IRES predictions. These results are in good agreement with previous studies in which a comparable trend in the density of clusters was observed [55, 56].

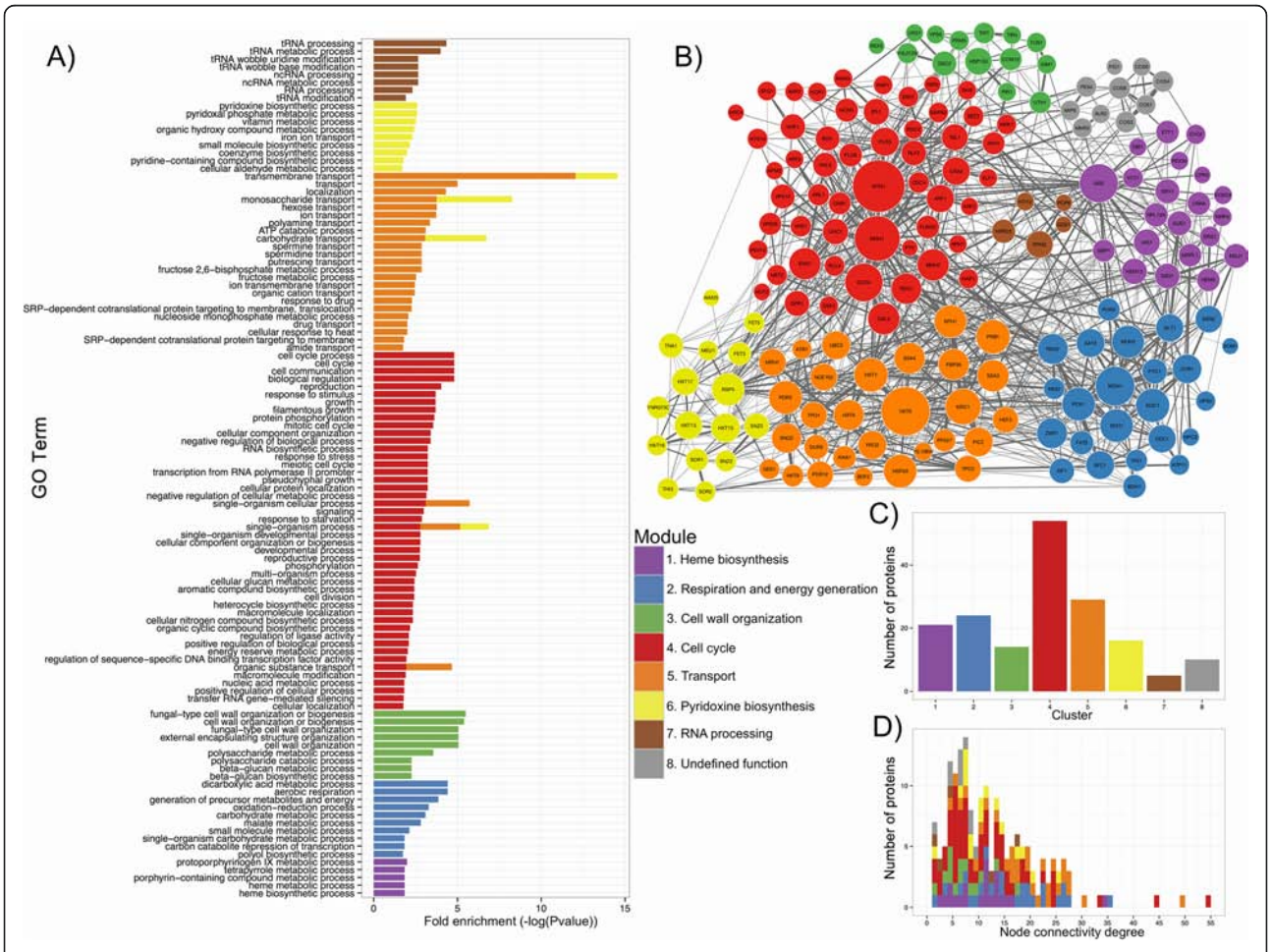

Fig. 3 Network clustering of the proteins translated by the set of top IRES predictions in *S. cerevisiae*. Each group is represented by a different color (**a**) GO-term enrichment analysis. The p-values of the GO-term enrichments are represented in logarithmic terms. **b** PPI network. The size of each vertex is proportional to its connectivity degree value. **c** Number of proteins in each module. **d** Node connectivity distribution by module

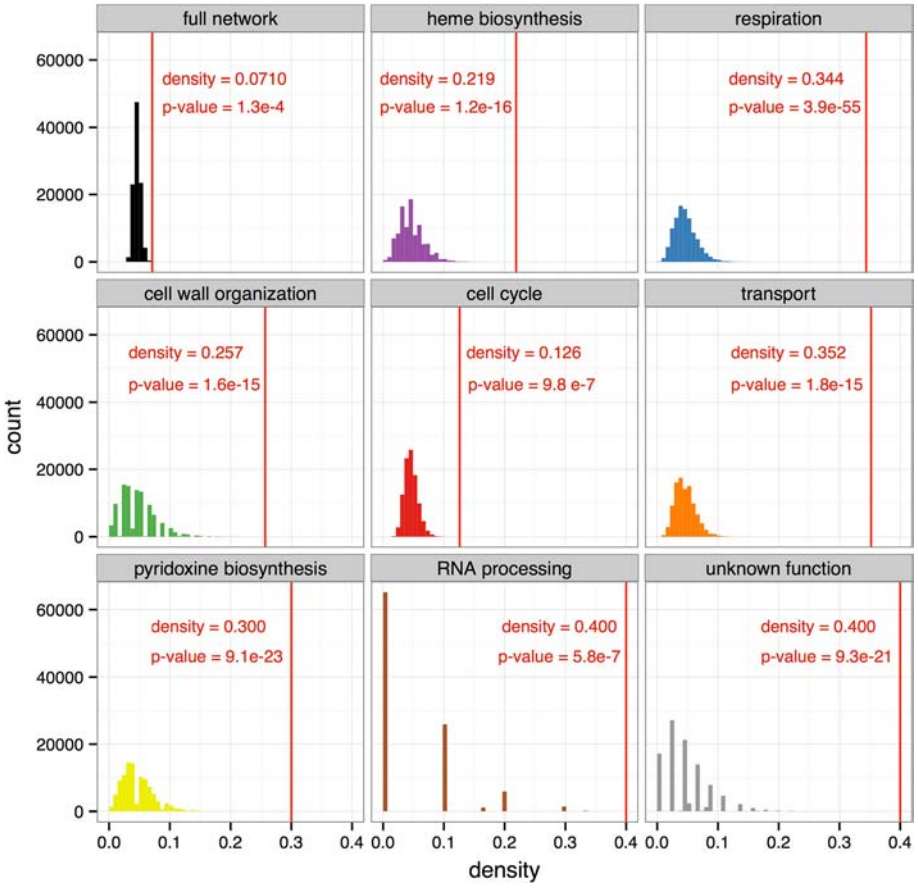

Fig. 4 PPI network density of the predicted proteins compared with simulated random networks. The full network and the modules are more cohesive than random networks

We believe the higher density of the modules when compared with either size-equivalent random networks or with the entire network, the GO-term enrichments found in each module, their functional specialization (few shared GO terms between modules), and their functional homogeneity (similar or related functions within a module) strengthen the biological relevance of our IRES predictions. These results support the hypothesis that IRES-dependent translation facilitates the expression of proteins working in a coordinated manner in specific biological processes, such as those previously reported in poliovirus infection, hypoxia and ER stress [7], mitosis [7, 9], apoptosis [10], invasive growth in yeast [17], and the meiotic program [3].

Biological relevance of the top IRES predictions in the context of their respective modules

In this section, we describe some relevant examples of the top IRES predictions in the biological contexts of their respective modules. We use node connectivity degree (or simply connectivity degree) as a measure of protein importance within the PPI network. The connectivity degree of a given node (protein) represents the number of interactions that this node has in a particular network. The connectivity degree in PPI networks has been linked to the biological significance of proteins. For example, it has been observed that protein connectivity degree is positively correlated with lethality [57] and disease-related genes [58]. The complete list of top IRESs predictions, their module classifications, connectivity degrees and descriptions are presented in Additional file 2.

Module 1: Heme biosynthesis. Heme is crucial in many fundamental biological processes and serves as a prosthetic group and a signaling molecule. For example, heme is used in controlling cell growth and differentiation, reducing oxidative damage, generating energy by respiration, and as an enzyme cofactor [59]. In yeast, heme controls transcription in response to oxygen levels through the activator Hap1p [59]. Interestingly, Hap1p is indeed included in the cell-cycle module (module 4), emphasizing the close relationship between the heme group and developmental processes. Additionally, this association suggests a case in which the effector molecule (heme) and the regulated gene (*HAP1*) are translated in an

IRES-dependent manner, highlighting the functional association of IRESs.

Another protein with a regulatory function in this module is Gis2p. This protein is the most connected element in this module, with a connectivity degree value of 34. Gis2p is a translational activator of mRNAs with IRESs [60, 61]. Additionally, similar regulatory functions in IRES translation were found for the Gis2p orthologs Znf9p in humans [62] and Cnbp in flies [63], indicating a potentially conserved function. Furthermore, Gis2p is implicated in stress response by its accumulation in P-bodies and stress granules under glucose deprivation conditions [64], and it is part of the genetic network implicated in the induction of invasive growth [65]. Remarkably, seven other genes required for invasive growth are also known to be translated in an IRES-dependent manner in *S. cerevisiae* [17].

The presence of IRES elements in regulatory genes, such as the abovementioned *HAP1* and *GIS2*, implies the existence of a wider hierarchical regulatory network that responds to specific metabolic or stress conditions in which some components of the translation machinery may be depleted.

Another remarkable example of two closely related proteins in this module is Ssq1p, a mitochondrial chaperone of the HSP70 family, and its co-chaperone, Mdj1p [66]. It is worth noting that Ssq1p has two homologs encoded by genes with IRESs, Hsp70p in *D. melanogaster* and in humans [13]. Ssq1p is required for the assembly of iron-sulfur clusters into proteins [67].

An additional example worth noting within this module is Sbp1p. It has two homologous proteins, Pab1p in yeast and Cirp in mice, which are known to be translated in an IRES-dependent manner.

Module 2: Respiration and energy generation. The ability to respond to nutrient changes is a crucial requirement for cell survival. *S. cerevisiae* preferentially uses glucose as a carbon source, although in glucose starvation conditions, alternative non-fermentable carbon sources can be used. Two proteins included in this module that belong to the central pathway of gluconeogenesis are malate dehydrogenase (Mdh1p and Mdh2p) and phosphoenolpyruvate carboxykinase (Pck1p). These proteins are degraded in the presence of glucose [68], and their corresponding genes are transcriptionally regulated by the zinc-finger transcription factor Znf1p following glucose starvation [69]. Pck1p also participates in other stress conditions, and it was found to confer cold tolerance in yeast [70].

Module 3: Cell wall organization. The cell wall adjusts its thickness and composition to contend with environmental stimuli, such as mechanical, osmotic, and heat shock stresses. The *S. cerevisiae* cell wall is composed largely of polysaccharides (85 %) and proteins (15 %), one of which is predicted to have an IRES: Gsc2p, a β-1,3-glucan synthase that can be induced by different environmental stimuli. For example, extracellular calcium, treatment with α-factor [71], heat shock [72], exposure to cell wall-damaging agents [73], or treatment with the reducing agent dithiothreitol [74] leads to strong Gsc2p induction.

The cell wall adapts its shape during different developmental and growth stages (reviewed in [75]). For example, the products of *HSP150*/*PIR2* and *PIR1*, which are predicted to contain IRES sequences, are required for cell wall stability [76], but how these proteins contribute is still unclear. However, PIR proteins are known to impact the permeability of the cell wall, and this effect is consistent with the role of these proteins in cross-linking β-1,3-glucans [77, 78].

Hsp150p and Pir1p are induced by heat shock, treatment with CFW or Zymolyase, and nitrogen limitation [77, 79, 80]. They are also regulated during cell cycle progression and in response to stress [81]. Additionally, there is evidence of coordinated regulation between genes required for cell wall organization or biogenesis and cell cycle genes [82]. Importantly, evidence has been found for translational control through 5'-UTRs in the case of two of the proteins in this module, Uth1p and Sim1p [83].

Module 4: Cell cycle. Module 4 is the largest module in terms of the number of enriched GO terms (Fig. 3a) and the number of genes (Fig. 3c). Additionally, this module includes the nodes displaying the highest connectivity degrees.

Importantly, the translational regulation of cell cycle processes has been described in several studies, and a number of IRESs play central roles regulating the expression of different kinases (reviewed in [9]).

The most connected protein in module 4 is Hog1p (High Osmolarity Glycerol response), which has a connectivity degree of 54 (Fig. 3c). Remarkably, Hog1p has 3 homologous proteins in humans that are translated in an IRES-dependent manner (PITSLREp, Pim1p and calcium/calmodulin-dependent protein kinase type II subunit alpha) [13]. This finding supports IRES conservation even in distant organisms. Hog1p is a mitogen-activated protein kinase that has important roles in different stress conditions. For example, the translational response to hyperosmotic shock is strongly dependent on Hog1p [24], and this protein has been shown to control cell cycle progression in response to stress [84]. It is worth noting that other kinases are part of this module and have important regulatory roles in the cell cycle (Tel1p [85], Ipl1p [86], Vhs1p [87], and Mek1p [88]), cell growth and proliferation (Cka2p [89]), salt tolerance (Hal5p [90]), and mating (Fus3p [91]).

Another relevant protein in this module is Gcn4p, which is a master translation factor that activates the

response to amino acid starvation [92] and is controlled at both the transcriptional and translational levels by diverse signals of stress [93]. Within the network of top IRES predictions, Gcn4p has a high connectivity degree (33), possibly reflecting its importance. Significantly, the Gcn4p paralog Yap1p and the ortholog Jun protein in chicken are encoded by genes with experimentally verified IRESs [13].

A final example of a remarkably significant protein in this module is Med10p (Nut2p), a subunit of the RNA II mediator complex that is required for transcriptional activation because its concentration is elevated in response to DNA replication stress [94, 95]. In addition, the mRNAs of two other proteins that are part of this complex, Med7p and Med18p (Srb5p), were predicted to be encoded by genes with IRESs. This finding has biological relevance because several components of the same complex are translated in an IRES-dependent manner, providing selective co-regulation under stress conditions. Additionally, there is evidence of the co-regulation of some proteins in these complexes at both the transcriptional and translational levels [96–98].

Module 5: Transport. The biological relevance of some members of this module has already been discussed because they are part of the most enriched orthologous groups (*HAA1*, *HXT1*, *HXT5*, *HXT6*, *HXT9*, *PDR5*, *PDR12*, *PDR15*, *SSA4*, and *SNQ2*; see *Enrichment analysis of IRES predictions in Orthologous Groups*). The proteins in this module are involved in the transport of a wide range of molecules. For example, Ssa3p and Ssa4p participate in SRP-dependent co-translational protein-membrane targeting and translocation [99], HXT members sense and transport glucose [23], Snq2p is an ABC transporter that confers multidrug resistance [100], Tpo1p and Tpo2p function as polyamine transporters [101], Yro2p is a plasma membrane protein involved in resistance to weak acid stress [102, 103], and Haa1p is a transcriptional activator that regulates *TPO2* and *YRO2* [104]. In the context of stress response, transport mechanisms play fundamental roles. For example, multiple transporters are involved in the response to weak acid stress, including the aforementioned Snq2p, Tpo1p, Tpo2p, Pdr12p, etc. [105]. Other stress conditions relevant to the members of this module are osmotic stress, oxidative stress, heat shock and detoxification [33] (see Additional file 2).

Module 6: Pyridoxine biosynthesis (VitB6). The most studied role of VitB6 is as a cofactor of enzymatic reactions. However, it is now clear that VitB6 is a potent antioxidant that protects cells from oxidative stress [106, 107]. Moreover, Snz2p and Snz3p, which are part of this module, are known to respond to nutrient limitation [108].

Module 7: RNA processing. This module exhibited an enrichment of terms related to RNA processing, including ncRNA and tRNA. Although there are no IRESs known to be associated with RNA processing, we believe that the predicted genes presented in this study are plausible because RNA-based regulation is an area that we are just beginning to understand [109], and our findings may have direct implications in stress response and pathological processes [110]. It is worth noting that the protein Hrr25p, which is a part of this module and is involved in tRNA wobble uridine modification [111], is a likely homolog of the Pim-1 protein, which has an experimentally verified IRES [13].

Module 8: Undefined function. Although module 8 was not enriched with specific GO terms, 5 out of the 10 members of this module were COnserved Sequence (COS) proteins (Cos1p, Cos3p, Cos4p, Cos6p and Cos8p), which are highly conserved in sequence, although their functions are unknown [112].

Interestingly, two RNA-binding paralogs, Pes4p and Mip6p, are included in this module. Both proteins are also homologous to Pap1p, the translation of which is IRES-dependent and part of the translation initiation complex [17], providing evidence of selective IRES conservation in homologous genes.

We believe that the fact that our IRES predictions clustered in cohesive GO-enriched modules highlights the functional association of IRES-translated genes and is complementary with our comparative genomics and network analyses. Furthermore, our modular organization-based approach could be used to analyze the results from genome-wide studies addressing IRES-dependent translation.

Comparison of predictions with translationally regulated genes

Ribosome profiling is a recently developed technique that has contributed to the understanding of the translation process by enabling the determination of the positions and dynamics of active ribosomes along the message, allowing the identification of translationally controlled genes. For this reason, we selected a previously published study [3] based on ribosome profiling to determine the intersection of our predictions and those genes that were found to be translationally controlled. Selected genes had the additional characteristic that their 5'-UTR ribosome occupancy rates were positively correlated with their translation efficiencies, indicating that augmented protein production is a consequence of increased ribosome occupancy. The number of translationally regulated *S. cerevisiae* genes in the aforementioned study was 110, whereas the number of top *S. cerevisiae* IRES predictions was 174; the intersection between these two sets of genes is 14 genes. The probability of having an intersection of this size at random considering 6,000 coding genes in *S. cerevisiae* is 5×10^{-7} (Fisher's exact test), which is a good indication of the accuracy of our predictions. Representative examples of ribosome footprints for genes with experimentally verified IRESs, genes

predicted as having IRES and genes predicted as not having IRESs, are presented in Additional file 3. Data obtained from reference [2].

Conclusions

We developed an accurate computational method based on a SVM for the identification of unstructured A-rich IRESs in fungal organisms. Using this method, we predicted IRES elements in the 5'-UTR sequences of 20 non-redundant fungal genomes and performed a comparative analysis and characterization of the functional relationships among the proteins encoded by the genes predicted to have IRES elements. We found statistically significant conservation of IRES-dependent translation in some groups of orthologous genes that revealed an underlying selective pressure, particularly in stress-related genes. In addition, our network analyses allowed us to identify biologically meaningful modules exhibiting specialized functions, providing evidence of a strong functional association between IRES-dependent translated proteins. Our study represents a useful resource for hypothesis-driven experiments and gene function exploration in the field of cap-independent translational regulation.

Methods

DNA sequence data

In this study we used sequence and annotation data from 33 completely sequenced fungal genomes. To avoid data overrepresentation, non-redundant genomes were selected based on their position in a maximum likelihood phylogenetic tree that was constructed using the PHANGORN package [113] available in the R software [114]. For each pair of phylogenetically close organisms, the one with the smaller genome was eliminated, leaving the organism with the larger genome [115]. The final set of non-redundant organisms used in our analysis consisted of 20 organisms. The complete list of organisms used, the list of non-redundant organisms and the phylogenetic tree are provided in Additional files 4, 5 and 6, respectively.

Obtaining the 5'-UTR sequences in this study

It has been shown that several yeast IRESs are located within the region corresponding to the first 60 nt immediately upstream of the translation initiation codon [17, 18]. Consequently, using a Perl script, the aforementioned region was obtained for each gene in each of our non-redundant yeast genomes. We termed these sequences 60ntUTRs, and they were used in our subsequent analyses. These 60ntUTRs were sorted in accordance with the orthologous groups of their corresponding genes. In *S. cerevisiae*, as far as we know, there are 11 experimentally confirmed and well-characterized IRESs with the common characteristic of being A-rich sequences [17, 18], 9 of these constitute the positive cases for the training of our SVM, whilst 2 of them were used as our internal positive control. The list of these genes and their characteristics are given in Additional file 7. Negative cases were obtained by the random sampling of 12,500 sequences from the complete pool of 60ntUTRs. This number of negative cases was selected because when compared with the number of positive cases generated by SMOTE (5,000) gives a ratio of 2.5:1. It should be noted that it was not easy to select cases that represent a truly 100 % negative control supported by experimental studies. Nevertheless, based on microarray analyses it has been estimated that only 10-15 % of mRNAs remain attached to polyribosomes under different stress conditions and considering that only 4 % of them might exhibit cap-independent translation [43]; we estimated that only 4-6 % of the genes used in the negative set for the training of our SVM might contain an IRES.

Feature selection for the prediction of IRES elements

Feature selection of the 60ntUTRs to predict IRES elements was based on a literature review, considering those variables that have been reported as correlated with the presence of IRESs or those that have been used to classify non-coding RNA (ncRNA). The set of features used in our analysis was grouped according to the following criteria: i) Minimum folding energy (MFE), which has been used to classify non-coding RNAs (ncRNA) [116] and is correlated with IRES expression strength in yeast [18]. The MFE of each of the 60ntUTRs was calculated using the RNAfold program of the Vienna RNA package, version 1.8.5 [117]. ii) GC content. It has been proposed that IRES-possessing *S. cerevisiae* genes related to nitrogen starvation tend to be A-rich [17]. Additionally, a positive correlation has been observed between the low GC content of UTRs and increased translational activity in glucose starvation conditions, which might be explained, at least partially, by IRES elements promoting protein expression [118]. For these reasons, two features were included: the GC content of the 60ntUTRs relative to the GC content of their corresponding intergenic regions (relGCintergenic) and the GC content of the 60ntURs relative to the chromosomal GC content (relGCchr). iii) Relative gene position in the chromosome (relPosChr). This feature was used because it has been shown to be related to the selective expression of stress-response genes [28, 119–121]. iv) Di-nucleotide frequencies. Composition-based approaches have been successfully used to develop ncRNA classification methods [122]. Therefore, we calculated di-nucleotide frequencies (16 variables) as input features. v) Measures of statistical dispersion. The higher the conservation of IRESs in an orthologous group, the more influence these elements will have on the properties of their group. For this reason, the

mean, mode, standard deviation, and skewness of each of the ortholog groups were calculated for the relGCintergenic and the MFE of its sequence members. vi) Length of the intergenic region. This feature was selected for IRES identification because longer regions could be indicative of the presence of regulatory elements, such as IRESs. After applying the criteria described above, a total of 29 features were used as inputs in our SVM. The lists of features before and after selection are provided in Additional file 1. In order to have a relative estimation of the likely contribution of these features in our SVM, we compared the average values of the features of the positive predictions *versus* the average values of the negative set of sequences used to train our SVM. The result of these comparisons is show in the figure of Additional file 8. As it was expected, in this figure, the *AA* and the *GC dinucleotides* presented the most significant values. Other features with important coefficient values were the *intergenic region length* and those features related with the minimum folding energy of the 60ntUTRs.

Feature pre-processing

To render the inter-species attributes comparable, feature standardization was performed to rescale the variables by their means and variances relative to their distributions in each organism. Subsequently, all the features had a mean of 0 and a variance of 1. To avoid data redundancy, we reduced the number of features to retain only uncorrelated variables [123]. A pair of features exhibiting a Pearson correlation factor greater than 0.55 were considered to be correlated. For each binary combination of correlated features, one was eliminated. After this step, 25 features from the original set of 29 were retained (see Additional file 1).

Synthetic minority oversampling

In general, machine learning methods perform poorly when they are applied to imbalanced datasets in which negative cases heavily outnumber positive cases [124]. However, in real data sets imbalances ranging from 100:1 up to 10,000:1 have been reported [125]. This type of datasets are common in biological studies as well, for example, in the prediction of translation initiation sites [126] and pre-miRNA classification [127]. The dataset used in this work is imbalanced because the number of A-rich IRESs in *S. cerevisiae* used in the training of our SVM (9 sequences) is very small compared to all possible genes containing IRESs (nearly 100,000 for the 20 selected organisms). To address this imbalance, synthetic minority oversampling technique, SMOTE [124] implemented in the DMwR package [128], was used. SMOTE is an oversampling technique that generates synthetic minority class samples by randomly choosing elements along the line segments joining some of the k minority class nearest neighbors [124]. This technique was selected because it has been shown to significantly improve the performance of SVMs used to classify non-coding RNAs (ncRNAs) [127], to predict RNA-protein interactions [129], and to analyze other imbalanced bioinformatic datasets [130] when this imbalance is superior to 100:1 [129, 131, 132]. It is worth mentioning that in our analysis, this imbalance was more significant than those previously reported (nearly 10,000:1). Additionally, considering the limited number of positive cases, this represents a potential constraint for the generalization of our predictions outside the training set. Considering this concern, our study includes two external positive cases of experimentally confirmed IRES not used in the training procedure (see Methods section). Furthermore, we performed an extensive and detailed statistic analysis of the enrichment of our IRES prediction in specific orthologous and functional groups (PPI network analysis) that supports the validity and generalization capacity of our IRES identification method (see Results and discussion section). For the SMOTE procedure, the parameter k was set to 300, the over-sampling to 900, and the under-sampling to 500, resulting in a subset of 17,500 genes.

Machine learning for IRES prediction

A support vector machine with a second-order polynomial kernel implemented in the caret package [133] was used for training on the selected features for IRES prediction. To increase the sensitivity, a cost of 2:1 (positive prediction: negative prediction) was set for the SVM [134]. Cross-validation (10-fold) repeated 30 times was used to measure the performance of the SVM. Predictions were evaluated using the set of 100,000 genes. Posterior class probabilities P(class|input) were calculated for each prediction according to Platt's methodology [39]. The complete list of predictions is provided in Additional file 9.

Enrichment analysis for the predictions

Genes predicted to contain IRESs were assigned to their corresponding orthologous groups. Fisher's exact test was performed to determine enrichment significance [19], and the resulting p-values were corrected for multiple testing using the Benjamini-Hochberg procedure [135].

Comparison of predictions with experimental data

To compare the probability of a random intersection of the IRES predictions with sets of translationally-controlled genes determined experimentally [3], Fisher's exact test was used [19].

Gene ontology analysis

The GO enrichment analysis [40, 41] was performed using the GOstats package [136] in the R software [114], correcting for multiple comparisons via the Benjamini-Hochberg method [135].

Selection of top IRES predictions

To select the best IRES predictions, we used a simple procedure that consisted of multiplying the posterior class probability [39] (the probability that a prediction is an IRES based on SVM output) of each prediction by its corresponding orthologous group enrichment *p*-value. We ranked the predictions according to this product (lower values indicate better predictions) and established a cutoff of 0.05. We used only genes classified as having IRESs by the SVM to avoid increasing the misclassification. The list of top predictions can be found in Additional file 2.

Protein-protein network analysis

Interaction data were obtained from the STRING database version 9.1 [45] using the STRINGdb package in the R software [45]. Graph properties were calculated using the Louvain method [53] implemented in the igraph package [54]. The corresponding number of genes in the complete network or in the particular module (Fig. 3c) to be evaluated was sampled at random from the entire list of protein-coding genes in *S. cerevisiae.* Afterwards, a network was constructed using the sampled genes, and its graph properties were calculated. This process was repeated 100,000 times for each case. The data were Box-Cox transformed to approximate normal distributions [137]. The normal distribution function was applied to calculate the *p*-values using the R software [114].

Protein homology determination

In order to determine if two proteins are homologous, pairwise comparison was performed using delta-blast [138], with an e-value threshold of 1×10^{-6} .

Availability of supporting data

DNA sequences and annotations were obtained from the Entrez Genome Database (ftp://ftp.ncbi.nlm.nih.gov/genomes/) [139]. The complete list of organisms and their corresponding accession numbers are in Additional file 4. The R software is available from https://www.r-project.org/. Groups of orthologous genes were downloaded from ftp://cegg.unige.ch/OrthoDB7/OrthoDB7_ALL_FUNGI_tabtext.gz [140]. The data of genes having IRESs was obtained from http://iresite.org/ [13] and from [17], and its respective supplement: http://www.sciencemag.org/content/317/5842/1224/suppl/DC1.

The list of translationally controlled genes determined by ribosome profiling was downloaded from the materials and methods supplement of [3]: http://www.sciencemag.org/content/335/6068/552/suppl/DC1. All the programs used in our analysis are available at our web page http://www.ibt.unam.mx/biocomputo/IRES_programs.html or at the figshare website http://dx.doi.org/10.6084/m9.figshare.1598203.

Additional files

Additional file 1: Features used to train our SVM. The initial 29 features to train our SVM are listed. After the selection of non-correlated features, 25 features from the original set of 29 were retained. A pair of features exhibiting a Pearson correlation factor greater than 0.55 were considered to be correlated. (XLS 21 kb)

Additional file 2: Top predictions. List of top genes predicted as having IRESs, including the names of the genes, their GI identifiers, their corresponding modules (clusters) in the PPI network, short descriptions/aliases, and long descriptions. (XLS 98 kb)

Additional file 3: Comparison of ribosome profiles. Ribosome profiles for *Saccharomyces cerevisiae* grown in rich media and under starvation conditions are presented. A) and B) are two genes that have verified IRESs. C) and D) are positive predictions and E), and F) are negative predictions. The regions presented comprise the 60 nt upstream and the 100 nt downstream the start codon. The translation initiation site is shown (blue vertical lines). The genes having IRESs present an increment ribosomes occupancy within the 60 nt upstream the translation initiation in starvation conditions compared to rich media. The two positive predictions depicted similar increments. These increments are not observed in the two negative predictions. The profiles were obtained using the web page http://gwips.ucc.ie/. (PNG 980 kb)

Additional file 4: List of organisms initially considered for the analysis. (XLS 29 kb)

Additional file 5: List of non-redundant organisms in the analysis. Non-redundant genomes were selected based on their position in a maximum likelihood phylogenetic tree that was constructed using the PHANGORN package [113] available in the R software (http://www.r-project.org) [114]. (XLS 21 kb)

Additional file 6: Phylogenetic tree. Phylogenetic tree used to select the non-redundant organisms. Bootstrap supporting values are indicated. (PNG 79 kb)

Additional file 7: Description of experimentally verified IRESs in *Saccharomyces cerevisiae*. A table containing their sequences, description and references. (XLS 34 kb)

Additional file 8: Comparison of the feature averages of the IRES predictions to the negative set used for training. The y-axis represents the log2 ratio of the feature averages for the positive predictions to the negative set. Higher values for the positive predictions are depicted in red and lower values in green. The Wilcoxon rank sum test was used to test the differences between means comparing the positive and negative sets. (PNG 88 kb)

Additional file 9: Complete predictions. List of genes predicted to have IRESs, including the GI identifier, gene name, orthologous group, posterior probability of being an IRES, orthologous group enrichment, product of the posterior probability and group enrichment, and the mRNA sequence used. (XLS 1571 kb)

Abbreviations

60ntURs: First 60 nt immediately upstream of the translation initiation codon; ABC: ATP-binding cassette; BP: Biological Process; COS: Conserved sequence; FDR: False discovery rate; GO: Gene ontology; IRES: Internal ribosome entry sites; MFE: Minimum folding energy; ncRNA: non-coding RNAs; nt: Nucleotide; OG: Orthologous group; PDR: Pleiotropic drug resistance; PPI: Protein-protein interaction; relGCchr: Relative to the chromosomal GC content; relGCintergenic: Relative to the GC content of their corresponding intergenic regions; relPosChr: Relative gene position in the chromosome; SVM: Support vector machine.

Competing interest

The authors have declared that no competing interests exist.

Authors' contributions

EPS co-developed the project idea, designed and performed the analysis, interpreted the biological significance of the results, and wrote the manuscript. LPL assisted in the biological interpretation of the results and helped in drafting

the manuscript. EM co-developed the project idea, coordinated the study, participated in its design, helped in its interpretation, and refined the manuscript. All authors participated in discussions and read and approved the final manuscript.

Acknowledgements
We wish to thank Ricardo Ciria and Juan-Manuel Hurtado for computer support and Shirley Ainsworth for bibliographical assistance. We acknowledge the Programa de Maestría y Doctorado en Ciencias Bioquímicas at the Instituto de Biotecnología-UNAM. EPS thanks the Consejo Nacional de Ciencia y Tecnologia (CONACyT) for its support and for the PhD scholarship (number 384858).

Computational requirements
The most computer-intensive steps in our method are the cross-validation and optimization processes. These steps were repeated 30 times on 32 different combinations of the SVM parameters in order to select the best parameter pair. This required 38 h of CPU time using a SUPERMICRO SC748TQ-R1400B with a 64- core computer. We used the R software for all the machine learning and statistical analyses.

Received: 20 July 2015 Accepted: 1 December 2015
Published online: 15 December 2015

References

1. Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009;136:731–45.
2. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324:218–23.
3. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science. 2012;335:552–7.
4. Lackner DH, Schmidt MW, Wu S, Wolf DA, Bähler J. Regulation of transcriptome, translation, and proteome in response to environmental stress in fission yeast. Genome Biol. 2012;13:R25.
5. Duncan CDS, Mata J. The translational landscape of fission-yeast meiosis and sporulation. Nat Struct Mol Biol. 2014;21:641–7.
6. Liu B, Qian S-B. Translational reprogramming in cellular stress response. Wiley Interdiscip Rev RNA. 2014;5:301–15.
7. Spriggs KA, Stoneley M, Bushell M, Willis AE. Re-programming of translation following cell stress allows IRES-mediated translation to predominate. Biol Cell. 2008;100:27–38.
8. Cornelis S, Bruynooghe Y, Denecker G, Van Huffel S, Tinton S, Beyaert R. Identification and Characterization of a Novel Cell Cycle–Regulated Internal Ribosome Entry Site. Mol Cell. 2000;5:597–605.
9. Kronja I, Orr-Weaver TL. Translational regulation of the cell cycle: when, where, how and why? Philos Trans R Soc Lond B Biol Sci. 2011;366:3638–52.
10. Holcik M, Sonenberg N. Translational control in stress and apoptosis. Nat Rev Mol Cell Biol. 2005;6:318–27.
11. Pelletier J, Sonenberg N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature. 1988;334:320–5.
12. Macejak DG, Sarnow P. Internal initiation of translation mediated by the 5′ leader of a cellular mRNA. Nature. 1991;353:90–4.
13. Mokrejs M, Masek T, Vopálensky V, Hlubucek P, Delbos P, Pospísek M. IRESite–a tool for the examination of viral and cellular internal ribosome entry sites. Nucleic Acids Res. 2010;38(Database issue):D131–6.
14. Thompson SR. So you want to know if your message has an IRES? Wiley Interdiscip Rev RNA. 2012;3:697–705.
15. Baird SD, Lewis SM, Turcotte M, Holcik M. A search for structurally similar cellular internal ribosome entry sites. Nucleic Acids Res. 2007;35:4664–77.
16. Baird SD, Turcotte M, Korneluk RG, Holcik M. Searching for IRES. RNA. 2006;12:1755–85.
17. Gilbert WV, Zhou K, Butler TK, Doudna JA. Cap-independent translation is required for starvation-induced differentiation in yeast. Science. 2007;317:1224–7.
18. Xia X, Holcik M. Strong eukaryotic IRESs have weak secondary structure. PLoS One. 2009;4, e4136.
19. Kuhn M, Johnson K. Applied Predictive Modeling. New York: Springer; 2013.
20. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960;20:37–46.
21. Spriggs KA, Bushell M, Mitchell SA, Willis AE. Internal ribosome entry segment-mediated translation during apoptosis: the role of IRES-trans-acting factors. Cell Death Differ. 2005;12:585–91.
22. Johannes G, Carter MS, Eisen MB, Brown PO, Sarnow P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. Proc Natl Acad Sci. 1999;96:13118–23.
23. Ozcan S, Johnston M. Function and Regulation of Yeast Hexose Transporters. Microbiol Mol Biol Rev. 1999;63:554–69.
24. Warringer J, Hult M, Regot S, Posas F, Sunnerhagen P. The HOG Pathway Dictates the Short-Term Translational Response after Hyperosmotic Shock. Mol Biol Cell. 2010;21:3080–92.
25. Melamed D, Pnueli L, Arava Y. Yeast translational response to high salinity: global analysis reveals regulation at multiple levels. RNA. 2008;14:1337–51.
26. Kowalski LRZ, Kondo K, Inouye M. Cold-shock induction of a family of TIP1-related proteins associated with the membrane in Saccharomyces cerevisiae. Mol Microbiol. 1995;15:341–53.
27. Donzeau M, Bourdineaud J-P, Lauquin GJ-M. Regulation by low temperatures and anaerobiosis of a yeast gene specifying a putative GPI-anchored plasma membrane. Mol Microbiol. 1996;20:449–59.
28. Ai W, Bertram PG, Tsang CK, Chan T-FF, Zheng XFSFS. Regulation of Subtelomeric Silencing during Stress Response. Mol Cell. 2002;10:1295–305.
29. Bester MC, Jacobson D, Bauer FF. Many Saccharomyces cerevisiae Cell Wall Protein Encoding Genes Are Coregulated by Mss11, but Cellular Adhesion Phenotypes Appear Only Flo Protein Dependent. G3 (Bethesda). 2012;2:131–41.
30. Gagiano M, van Dyk D, Bauer FF, Lambrechts MG, Pretorius IS. Msn1p/Mss10p, Mss11p and Muc1p/Flo11p are part of a signal transduction pathway downstream of Mep2p regulating invasive growth and pseudohyphal differentiation in Saccharomyces cerevisiae. Mol Microbiol. 1999;31:103–16.
31. Kim TS, Kim HY, Yoon JH, Kang HS. Recruitment of the Swi/Snf complex by Ste12-Tec1 promotes Flo8-Mss11-mediated activation of STA1 expression. Mol Cell Biol. 2004;24:9542–56.
32. Sá-Correia I, dos Santos SC, Teixeira MC, Cabrito TR, Mira NP. Drug:H+ antiporters in chemical stress response in yeast. Trends Microbiol. 2009;17:22–31.
33. Jungwirth H, Kuchler K. Yeast ABC transporters– a tale of sex, stress, drugs and aging. FEBS Lett. 2006;580:1131–8.
34. Zhou W, Edelman GM, Mauro VP. Transcript leader regions of two Saccharomyces cerevisiae mRNAs contain internal ribosome entry sites that function in living cells. Proc Natl Acad Sci U S A. 2001;98:1531–6.
35. Seino A, Yanagida Y, Aizawa M, Kobatake E. Translational control by internal ribosome entry site in Saccharomyces cerevisiae. Biochim Biophys Acta. 1681;2005:166–74.
36. Hecht K, Bailey JE, Minas W. Polycistronic gene expression in yeast versus cryptic promoter elements. FEMS Yeast Res. 2002;2:215–24.
37. Iizuka N, Najita L, Franzusoff A, Sarnow P. Cap-dependent and cap-independent translation by internal initiation of mRNAs in cell extracts prepared from Saccharomyces cerevisiae. Mol Cell Biol. 1994;14:7322–30.
38. Petryk N, Zhou Y-F, Sybirna K, Mucchielli M-H, Guiard B, Bao W-G, et al. Functional Study of the Hap4-Like Genes Suggests That the Key Regulators of Carbon Metabolism HAP4 and Oxidative Stress Response YAP1 in Yeast Diverged from a Common Ancestor. PLoS One. 2014;9, e112263.
39. Platt JC. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Smola AJ, Bartlett P, Schoelkopf B, Schuurmans D, editors. Adv Large Margin Classif. Cambridge, MA, USA: MIT press; 2000. p. 61–74.
40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.
41. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2014;43(Database issue):D1049–56.
42. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6, e21800.
43. Wellensiek BP, Larsen AC, Stephens B, Kukurba K, Waern K, Briones N, et al. Genome-wide profiling of human cap-independent translation-enhancing elements. Nat Methods. 2013;10:747–50.
44. Komar AA, Hatzoglou M. Cellular IRES-mediated translation: the war of ITAFs in pathophysiological states. Cell Cycle. 2011;10:229–40.
45. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41(Database issue):D808–15.
46. Kolaczyk ED, Csárdi G. Statistical Analysis of Network Data with R, vol. 65. New York, NY: Springer New York; 2014 [*Use R!*].

47. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761 Suppl):C47–52.
48. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A. 2003;100:12123–8.
49. Kaltenbach H-M, Stelling J. Modular analysis of biological networks. Adv Exp Med Biol. 2012;736:3–17.
50. Chung SS, Pandini A, Annibale A, Coolen ACC, Thomas NSB, Fraternali F. Bridging topological and functional information in protein interaction networks by short loops profiling. Sci Rep. 2015;5:8540.
51. Pereira-Leal JB, Enright AJ, Ouzounis CA. Detection of functional modules from protein interaction networks. Proteins. 2004;54:49–57.
52. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5:101–13.
53. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008, P10008.
54. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006; Complex Sy:1965.
55. Dong J, Horvath S. Understanding network concepts in modules. BMC Syst Biol. 2007;1:24.
56. Lancichinetti A, Kivelä M, Saramäki J, Fortunato S. Characterizing the community structure of complex networks. PLoS One. 2010;5, e11976.
57. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411:41–2.
58. Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics. 2006;7:31.
59. Mense SM, Zhang L. Heme: a versatile signaling molecule controlling the activities of diverse regulators ranging from transcription factors to MAP kinases. Cell Res. 2006;16:681–92.
60. Sammons MA, Samir P, Link AJ. Saccharomyces cerevisiae Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9. Biochem Biophys Res Commun. 2011;406:13–9.
61. Tsvetanova NG, Klass DM, Salzman J, Brown PO. Proteome-wide search reveals unexpected RNA-binding proteins in Saccharomyces cerevisiae. PLoS One. 2010;5, e12671.
62. Gerbasi VR, Link AJ. The myotonic dystrophy type 2 protein ZNF9 is part of an ITAF complex that promotes cap-independent translation. Mol Cell Proteomics. 2007;6:1049–58.
63. Antonucci L, D'Amico D, Di Magno L, Coni S, Di Marcotullio L, Cardinali B, et al. CNBP regulates wing development in Drosophila melanogaster by promoting IRES-dependent translation of dMyc. Cell Cycle. 2014;13:434–9.
64. Rojas M, Farr GW, Fernandez CF, Lauden L, McCormack JC, Wolin SL. Yeast Gis2 and its human ortholog CNBP are novel components of stress-induced RNP granules. PLoS One. 2012;7, e52824.
65. Shively CA, Eckwahl MJ, Dobry CJ, Mellacheruvu D, Nesvizhskii A, Kumar A. Genetic networks inducing invasive growth in Saccharomyces cerevisiae identified through systematic genome-wide overexpression. Genetics. 2013;193:1297–310.
66. Westermann B, Gaume B, Herrmann JM, Neupert W, Schwarz E. Role of the mitochondrial DnaJ homolog Mdj1p as a chaperone for mitochondrially synthesized and imported proteins. Mol Cell Biol. 1996;16:7063–71.
67. Mühlenhoff U, Gerber J, Richhardt N, Lill R. Components involved in assembly and dislocation of iron-sulfur clusters on the scaffold protein Isu1p. EMBO J. 2003;22:4815–25.
68. Hung G-C, Brown CR, Wolfe AB, Liu J, Chiang H-L. Degradation of the gluconeogenic enzymes fructose-1,6-bisphosphatase and malate dehydrogenase is mediated by distinct proteolytic pathways and signaling events. J Biol Chem. 2004;279:49138–50.
69. Tangsombatvichit P, Semkiv MV, Sibirny AA, Jensen LT, Ratanakhanokchai K, Soontorngun N. Zinc cluster protein Znf1, a novel transcription factor of non-fermentative metabolism in Saccharomyces cerevisiae. FEMS Yeast Res. 2015;15.
70. Vicent I, Navarro A, Mulet JM, Sharma S, Serrano R. Uptake of inorganic phosphate is a limiting factor for Saccharomyces cerevisiae during growth at low temperatures. FEMS Yeast Res. 2015;15.
71. Mazur P, Morin N, Baginsky W, el-Sherbeini M, Clemas JA, Nielsen JB, et al. Differential expression and function of two homologous subunits of yeast 1,3-beta-D-glucan synthase. Mol Cell Biol. 1995;15:5671–81.
72. Zhao C, Jung US, Garrett-Engele P, Roe T, Cyert MS, Levin DE. Temperature-induced expression of yeast FKS2 is under the dual control of protein kinase C and calcineurin. Mol Cell Biol. 1998;18:1013–22.
73. Agarwal AK, Rogers PD, Baerson SR, Jacob MR, Barker KS, Cleary JD, et al. Genome-wide expression profiling of the response to polyene, pyrimidine, azole, and echinocandin antifungal agents in Saccharomyces cerevisiae. J Biol Chem. 2003;278:34998–5015.
74. Travers KJ, Patil CK, Wodicka L, Lockhart DJ, Weissman JS, Walter P. Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. Cell. 2000;101:249–58.
75. Orlean P. Architecture and biosynthesis of the Saccharomyces cerevisiae cell wall. Genetics. 2012;192:775–818.
76. Kapteyn JC, Van Egmond P, Sievi E, Van Den Ende H, Makarow M, Klis FM. The contribution of the O-glycosylated protein Pir2p/Hsp150 to the construction of the yeast cell wall in wild-type cells and beta 1,6-glucan-deficient mutants. Mol Microbiol. 1999;31:1835–44.
77. Yun DJ, Zhao Y, Pardo JM, Narasimhan ML, Damsz B, Lee H, et al. Stress proteins on the yeast cell surface determine resistance to osmotin, a plant antifungal protein. Proc Natl Acad Sci U S A. 1997;94:7082–7.
78. Klis FM, Mol P, Hellingwerf K, Brul S. Dynamics of cell wall structure in Saccharomyces cerevisiae. FEMS Microbiol Rev. 2002;26:239–56.
79. Russo P, Simonen M, Uimari A, Teesalu T, Makarow M. Dual regulation by heat and nutrient stress of the yeast HSP150 gene encoding a secretory glycoprotein. Mol Gen Genet. 1993;239:273–80.
80. Toh-e A, Yasunaga S, Nisogi H, Tanaka K, Oguchi T, Matsui Y. Three yeast genes, PIR1, PIR2 and PIR3, containing internal tandem repeats, are related to each other, and PIR1 and PIR2 are required for tolerance to heat shock. Yeast. 1993;9:481–94.
81. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9:3273–97.
82. Negishi T, Ohya Y. The cell wall integrity checkpoint: coordination between cell wall synthesis and the cell cycle. Yeast. 2010;27:513–9.
83. Wanless AG, Lin Y, Weiss EL. Cell morphogenesis proteins are translationally controlled through UTRs by the Ndr/LATS target Ssd1. PLoS One. 2014;9, e85212.
84. Duch A, de Nadal E, Posas F. The p38 and Hog1 SAPKs control cell cycle progression in response to environmental stresses. FEBS Lett. 2012;586:2925–31.
85. Giannattasio M, Sommariva E, Vercillo R, Lippi-Boncambi F, Liberi G, Foiani M, et al. A dominant-negative MEC3 mutant uncovers new functions for the Rad17 complex and Tel1. Proc Natl Acad Sci U S A. 2002;99:12997–3002.
86. Francisco L, Wang W, Chan CS. Type 1 protein phosphatase acts in opposition to IpL1 protein kinase in regulating yeast chromosome segregation. Mol Cell Biol. 1994;14:4731–40.
87. Muñoz I, Simón E, Casals N, Clotet J, Ariño J. Identification of multicopy suppressors of cell cycle arrest at the G1-S transition in Saccharomyces cerevisiae. Yeast. 2003;20:157–69.
88. Rockmill B, Roeder GS. A meiosis-specific protein kinase homolog required for chromosome synapsis and recombination. Genes Dev. 1991;5:2392–404.
89. Ahmed K. Joining the cell survival squad: an emerging role for protein kinase CK2. Trends Cell Biol. 2002;12:226–30.
90. Mulet JM, Leube MP, Kron SJ, Rios G, Fink GR, Serrano R. A novel mechanism of ion homeostasis and salt tolerance in yeast: the Hal4 and Hal5 protein kinases modulate the Trk1-Trk2 potassium transporter. Mol Cell Biol. 1999;19:3328–37.
91. Elion EA, Grisafi PL, Fink GR. FUS3 encodes a cdc2+/CDC28-related kinase required for the transition from mitosis into conjugation. Cell. 1990;60:649–64.
92. Hinnebusch AG. Mechanisms of gene regulation in the general control of amino acid biosynthesis in Saccharomyces cerevisiae. Microbiol Rev. 1988;52:248–73.
93. Hinnebusch AG, Natarajan K. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. Eukaryot Cell. 2002;1:22–32.
94. Gustafsson CM, Myers LC, Beve J, Spåhr H, Lui M, Erdjument-Bromage H, et al. Identification of new mediator subunits in the RNA polymerase II holoenzyme from Saccharomyces cerevisiae. J Biol Chem. 1998;273:30851–4.
95. Tkach JM, Yimit A, Lee AY, Riffle M, Costanzo M, Jaschob D, et al. Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat Cell Biol. 2012;14:966–76.
96. Siwiak M, Zielenkiewicz P. Co-regulation of translation in protein complexes. Biol Direct. 2015;10:18.
97. Jin J, Iakova P, Jiang Y, Lewis K, Sullivan E, Jawanmardi N, et al. Transcriptional and translational regulation of C/EBPβ-HDAC1 protein complexes controls different levels of p53, SIRT1, and PGC1α proteins at the early and late stages of liver cancer. J Biol Chem. 2013;288:14451–62.

98. Webb EC, Westhead DR. The transcriptional regulation of protein complexes; a cross-species perspective. Genomics. 2009;94:369–76.
99. Becker J, Walter W, Yan W, Craig EA. Functional interaction of cytosolic hsp70 and a DnaJ-related protein, Ydj1p, in protein translocation in vivo. Mol Cell Biol. 1996;16:4378–86.
100. Servos J, Haase E, Brendel M. Gene SNQ2 of Saccharomyces cerevisiae, which confers resistance to 4-nitroquinoline-N-oxide and other chemicals, encodes a 169 kDa protein homologous to ATP-dependent permeases. Mol Gen Genet. 1993;236:214–8.
101. Tomitori H, Kashiwagi K, Sakata K, Kakinuma Y, Igarashi K. Identification of a gene for a polyamine transport protein in yeast. J Biol Chem. 1999;274:3265–7.
102. De Hertogh B, Carvajal E, Talla E, Dujon B, Baret P, Goffeau A. Phylogenetic classification of transporters and other membrane proteins from Saccharomyces cerevisiae. Funct Integr Genomics. 2002;2:154–70.
103. Takabatake A, Kawazoe N, Izawa S. Plasma membrane proteins Yro2 and Mrh1 are required for acetic acid tolerance in Saccharomyces cerevisiae. Appl Microbiol Biotechnol. 2015;99:2805–14.
104. Fernandes AR, Mira NP, Vargas RC, Canelhas I, Sá-Correia I. Saccharomyces cerevisiae adaptation to weak acids involves the transcription factor Haa1p and Haa1p-regulated genes. Biochem Biophys Res Commun. 2005;337:95–103.
105. Mira NP, Teixeira MC, Sá-Correia I. Adaptive response and tolerance to weak acids in Saccharomyces cerevisiae: a genome-wide view. OMICS. 2010;14:525–40.
106. Bilski P, Li MY, Ehrenshaft M, Daub ME, Chignell CF. Vitamin B6 (pyridoxine) and its derivatives are efficient singlet oxygen quenchers and potential fungal antioxidants. Photochem Photobiol. 2000;71:129–34.
107. Chumnantana R, Yokochi N, Yagi T. Vitamin B6 compounds prevent the death of yeast cells due to menadione, a reactive oxygen generator. Biochim Biophys Acta. 1722;2005:84–91.
108. Padilla PA, Fuge EK, Crawford ME, Errett A, Werner-Washburne M. The Highly Conserved, Coregulated SNO and SNZ Gene Families in Saccharomyces cerevisiae Respond to Nutrient Limitation. J Bacteriol. 1998;180:5718–26.
109. Barrett LW, Fletcher S, Wilton SD. Untranslated Gene Regions and Other Non-Coding Elements. Basel: Springer Basel; 2013 [*SpringerBriefs in Biochemistry and Molecular Biology*].
110. Su W-Y, Xiong H, Fang J-Y. Natural antisense transcripts regulate gene expression in an epigenetic manner. Biochem Biophys Res Commun. 2010;396:177–81.
111. Huang B, Lu J, Byström AS. A genome-wide screen identifies genes required for formation of the wobble nucleoside 5-methoxycarbonylmethyl-2-thiouridine in Saccharomyces cerevisiae. RNA. 2008;14:2183–94.
112. Despons L, Wirth B, Louis VL, Potier S, Souciet J-L. An evolutionary scenario for one of the largest yeast gene families. Trends Genet. 2006;22:10–5.
113. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27:592–3.
114. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Graph Stat. 1996;5:299–314.
115. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res. 2010;38, e130.
116. Bompfünewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, et al. Variations on RNA folding and alignment: lessons from Benasque. J Math Biol. 2008;56:129–44.
117. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie Chem Mon. 1994;125:167–88.
118. Castelli LM, Lui J, Campbell SG, Rowe W, Zeef LAH, Holmes LEA, et al. Glucose depletion inhibits translation initiation via eIF4A loss and subsequent 48S preinitiation complex accumulation, while the pentose phosphate pathway is coordinately up-regulated. Mol Biol Cell. 2011;22:3379–93.
119. De Las PA, Pan S-J, Castaño I, Alder J, Cregg R, Cormack BP. Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. Genes Dev. 2003;17:2245–58.
120. Hansen KR, Burns G, Mata J, Volpe TA, Martienssen RA, Bähler J, et al. Global effects on gene expression in fission yeast by silencing and RNA interference machineries. Mol Cell Biol. 2005;25:590–601.
121. Halme A, Bumgarner S, Styles C, Fink GR. Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast. Cell. 2004;116:405–15.
122. Panwar B, Arora A, Raghava GPS. Prediction and classification of ncRNAs using structural information. BMC Genomics. 2014;15:127.
123. Kuri-Morales A. An automated search space reduction methodology for large databases. In: Perner P, editor. *Adv Data Mining Appl Theor Asp. Volume 7987*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 11–24 [*Lecture Notes in Computer Science*].
124. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
125. Provost F, Fawcett T. Robust Classification for Imprecise Environments. Mach Learn. 2001;42:203–231.
126. Silva LM, Teixeira FC, Ortega JM, Zárate LE, Nobre CN. Improvement in the prediction of the translation initiation site through balancing methods, inclusion of acquired knowledge and addition of features to sequences of mRNA. BMC Genomics. 2011;12 Suppl 4:S9.
127. Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics. 2009;25:989–95.
128. Torgo L. Data Mining with R: Learning with Case Studies. Boca Raton, FL, USA: Chapman & Hall/CRC; 2010.
129. Livi CM, Blanzieri E. Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. BMC Bioinformatics. 2014;15:123.
130. Batuwita R, Palade V. A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems. In: 2009 Int Conf Mach Learn Appl. Washington, DC, USA: IEEE; 2009. p. 545–550.
131. Johnson RA, Chawla NV, Hellmann JJ. Species distribution modeling and prediction: A class imbalance problem. In: 2012 Conf Intell Data Underst. Washington, DC, USA: IEEE; 2012. p. 9–16.
132. Macisaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, Gifford DK, et al. A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. Bioinformatics. 2006;22:423–9.
133. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Softw. 2008;28:1–26.
134. Veropoulos K, Campbell C, Cristianini N. Controlling the Sensitivity of Support Vector Machines. In: Proc Int Jt Conf AI. 1999. p. 55–60.
135. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57:289–300.
136. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23:257–8.
137. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Ser B. 1964;211–252.
138. Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. Biol Direct. 2012;7:12.
139. Robbertse B, Tatusova T. Fungal genome resources at NCBI. Mycology. 2011;2:142–60.
140. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 2013;41(Database issue):D358–65.