Facultad de Ciencias Bioquímicas y Farmacéuticas - Universidad Nacional de Rosario

Tesis de Doctorado

Estudios sobre la regulación de la expresión génica por microARNs en plantas mediante estrategias bioinformáticas

Presentada por: Uciel Pablo Chorostecki

Rosario, Argentina

2016

Estudios sobre la regulación de la expresión génica por microARNs en plantas mediante estrategias bioinformáticas

Uciel Pablo Chorostecki

Licenciado en Ciencias de la Computación

Universidad Nacional de Rosario

Esta Tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Ciencias Biológicas, de la Universidad de Rosario y no ha sido presentada previamente para la obtención de otro título en esta u otra Universidad. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el Instituto de Biología Molecular y Celular de Rosario (IBR-CONICET), dependiente de la Facultad de Cs. Bioquímicas y Farmacéuticas, durante el período comprendido entre el ?? y el ??, bajo la dirección del Dr. Javier Palatnik.

I would like to dedicate this thesis to my loving parents ...

Índice general

Índice de figuras

Índice de tablas

Resumen

Resumen: Deberá contener la siguiente información
1. Breve presentación del problema
2. Enfoque y planificación del problema
3. Datos significativos y hallazgos más importantes
4. Conclusiones

Chapter 1

Métodos

1.1 Predicción de genes regulados por miARNs en plantas

En la primer parte de esta tesis diseñamos una estrategia para la identificación de genes blanco regulados por miARNs basado en la conservación evolutiva del par miARN-gen blanco. La metodología aplicada es la siguiente.

1.1.1 MiARN consensos

Las 22 familias de miARNs conservadas en angiospermas fueron consideradas para esta parte del trabajo [4, 10]. MiR319 y miR159 que codifican para miARNs similares, fueron considerados como familias diferentes ya que regulan a genes blanco distintos [22]. Consideramos todos los miembros de estas familia, obtenidos de miRBASE[1], pertenecientes a *A. thaliana*, *Populus trichocarpa* y *Oryza Sativa*. Variaciones en las posiciones 1, 20 y 21 son muy comunes en las familias de miARNs [8]. Por esto, definimos como secuencia consenso, a las secuencias más comunes (posiciones 2-19) de distintos miembros de cada familia (tabla 2.1).

1.1.2 Predicción de genes regulados por miARNs

Conjunto de datos de plantas

Los datos de las secuencias pertenecen a librerías extraídas de "Gene Index Project"[2], que consiste en una base de datos de ESTs ensamblados. Seleccionamos un conjunto de datos pertenecientes a Angiospermas. Además utilizamos secuencias de ARNm completos de *A.*

[1] http://mirbase.org
[2] http://compbio.dfci.harvard.edu/tgi/

thaliana[3] y *Oryza Sativa*[4] (ver tabla A.1). La búsqueda la realizamos utilizando PatMatch[31], que es un programa de búsqueda de patrones de nucleótidos cortos o péptidos. El programa puede ser usado para encontrar coincidencias con un patrón de secuencia específico y permite el uso de códigos de secuencias ambiguas y expresiones regulares y por esto se puede utilizar la búsqueda con mismatches, inserciones y deleciones. Realizamos la búsqueda de potenciales genes blanco permitiendo tres mismatches con las secuencias consensos, mientras que las interacciones G:U y los bulges fueron considerados mismatches. Para realizar el alineamiento del par miARN-gen blanco, desarrollamos una versión modificada del algoritmo de programación dinámica Needleman-Wunsch[20], utilizando el lenguaje Perl[5]. Además, desarrollamos scripts para integrar los módulos de Blastx[3] utilizando el proteoma de Arabidopsis y el módulo RNAhybrid[12] que es una herramienta que permite encontrar la menor energía libre de hibridación (MFE) de dos secuencias de ARN.

Filtros

Las secuencias candidatas fueron etiquetadas con el identificador del locus (locus ID) con mejor puntuación (best hit) en *A. thaliana*, utilizando el módulo de Blastx (Corte del evalue de $10e^{-5}$). De este modo, genes blanco de distintas especies que tenían la misma etiqueta fueron agrupados juntos, ya que tendrían el mismo homólogo en *A. thaliana*. El filtro de conservación evolutiva hace referencia al número mínimo de especies donde la misma etiqueta estaba presente para un miARN particular. El filtro empírico está basado en trabajos previos[27] y hace referencia a la energía de interacción MFE (mínima energía libre de hibridación de al menos 72% del apareamiento perfecto). El otro filtro empírico requiere que entre el par miARN-gen blanco, solamente está permitido un mismatch entre la posición 2 y la 12 del miARN (1-11 de nuestra búsqueda modificada con las secuencias consenso).

Controles

Como control, realizamos las búsquedas del mismo modo que lo hicimos para los miARNs conservados, pero utilizando secuencia al azar. Para cada miARN conservado, generamos 20 secuencias al azar (scramble) dividiendo las secuencias originales de a di-nucleótidos y luego generando nuevas secuencias al azar conservando esa composición de los di-nucleótidos como fue descrito previamente [**?**]. De estas 20 secuencias al azar, elegimos las 10 que tenían el número más similar del total de genes blanco para el miARN real correspondiente. La relación señal/ruido fue calculada como el cociente entre el número de genes blanco para

[3] http://arabidopsis.org
[4] http://rice.plantbiology.msu.edu
[5] http://perl.org

los miARNs y el número de genes blanco del promedio obtenido para las secuencias al azar. Como un control adicional, seleccionamos dos miARNs que no están conservados durante la evolución, que son el miR158 y el miR173.

Ecotipos utilizados y condiciones de crecimiento

Las plantas de *A. thaliana* utilizadas para los experimentos en esta parte del trabajo corresponden a el ecotipo Columbia-0 Col-0. Las plantas fueron cultivadas en una cámara de crecimiento con un régimen de 16 h de luz (100 μE.m.$^{-2}s^{-1}$) y 8 h de oscuridad (condición día largo). La temperatura de crecimiento fue de 23°C durante el ciclo luz/oscuridad, mientras que la humedad fue mantenida en 65% de humedad relativa. Las plantas fueron regadas 2 veces por semana con agua. Para el crecimiento directo en tierra, las semillas fueron estratificadas a 4°C por 2 días en tubos de microcentrífuga con 1ml de 0,1% (p/v) agar, y luego sembradas en tierra. Las plantas de *Nicotiana tabacum* (cv Petit Havana) fueron crecidas en condición día largo durante 8 semanas y la segunda hoja fue utilizada para el análisis de ARN.

Cleavage site mapping of target mRNA and expression analysis

ARN Poly(A)+ fue extraído a partir de 50 mg de ARN total de plántulas de Col-0 utilizando el kit comercial PolyATract®(Promega) La ligación del Oligo Adaptador de ARN, transcripción reversa y 5' RACE fueron realizadas como se describió anteriormente [22] Two nested gene-specific reverse oligonucleotides were used for 5' RACE. Los productos de la PCR fueron resueltos en geles de agarosa al 2% y se detectaron por tinción con bromuro de etidio. La PCR en tiempo real cuantitativa (qPCR) para los genes blanco del miR396 y miR159 se realizó como se ha descrito anteriormente [22, 25] La lista de los cebadores para estos ensayos están descritos en las tablas A.3 y A.4. Las plantas que sobreexpresan el miR396 y miR159 se han descrito previamente [22, 25].

1.2 comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas

A partir de los resultados positivos obtenidos de la estrategia descrita anteriormente, decidimos desarrollar una herramienta web y dejarla disponible para la comunidad científica denominada comTAR que está disponible en un sub-dominio de la página web institucional del IBR: http://rnabiology.ibr-conicet.gov.ar/comtar.

1.2.1 MiARN y transcriptos

Como las secuencias del maduro del miARN puede variar en distintas especies, especialmente en la posición 1, 20 y 21 ([**?**], utilizamos secuencias del 2-19 (18nt) para realizar las búsquedas. Como además existen variaciones en las secuencias en los distintos miARNs de las mismas familias, utilizamos la más representativa teniendo en cuenta los genomas de Arabidopsis, álamo y arroz. De este modo comTAR contiene datos pre-calculados, de potenciales genes blanco para 22 miARNs conservados en plantas (ver tabla 2.1) donde el usuario puede navegar los resultados y cambiar los parámetros de entrada. Además, el usuario puede realizar la búsqueda de nuevos ARNs pequeños teniendo en cuenta esta consideración. El cálculo se hace en el cluster del CCT-Rosario y los datos se obtienen luego de unas horas. Como la herramienta web la realizamos tiempo después de haber hecho la estrategia para predicción de genes blanco, utilizamos una nueva base de datos más actualizada y completa denominada Phytozome[6] [13]. La misma corresponde a secuencias de transcriptos de plantas formado por archivos de nucleótidos en formato FASTA de transcriptos de ARNm (UTR, exones) con variantes de splicing.

1.2.2 Búsqueda de genes blanco

La búsqueda de genes blanco la realizamos de la misma manera que la descrita anteriormente con algunos cambios. Además de actualizar la base de datos y utilizar la de Phytozome, actualizamos la base de datos de *A. thaliana* por la del TAIR10. Las secuencias candidatas fueron etiquetadas con el mejor hit del locus ID del Arabidosis TAIR10, utilizando los archivos de anotación de Phytozome, y lo utilizamos como "TAG" (etiqueta). Por último, cada TAG de Arabidopsis fue indexado con una breve descripción funcional y computacional obtenida del TAIR10 y los genes blanco candidatos fueron agrupados por familias teniendo en cuanta la clasificación de familias del TAIR10.

1.2.3 Herramienta web y almacenamiento de datos

ComTAR fue diseñado como una aplicación web con un framework open-source en PHP denominado Codeigniter para la interfaz gráfica, pero el análisis está basado en un back-end escrito en Perl. Los datos que surgen de ese análisis fueron almacenados en una base de datos en MySQL[7]. El back-end es el encargado de realizar la búsqueda de secuencias y además ahí es donde se integraron las herramientas y scripts para aumentar la especificidad y sensibilidad

[6] http://phytozome.jgi.doe.gov
[7] http://mysql.com

de comTAR. También el back-end es el encargado de generar los resultados finales. Mientras el front-end es el responsable de mostrar los resultados (Figura 1.1). El TAG del mejor hit en Arabidopsis es el que determina el número de especies donde un gen blanco está presente, y el número mínimo de especie es un parámetro que es definido por el usuario.

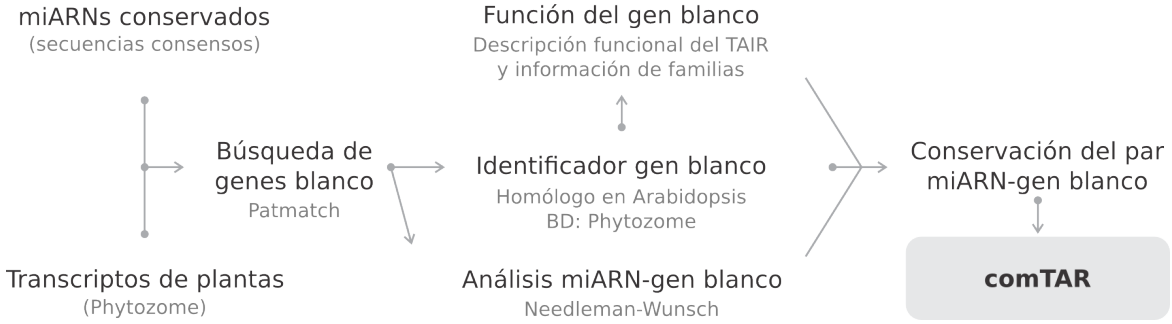


Fig. 1.1 comTAR. Diagrama de flujo que describe la herramienta

Capítulo 2

Aplicaciones bioinformáticas para el estudio de interacciones miARN-gen blanco

2.1 Introducción

2.2 Resultados

2.2.1 Predicción de genes regulados por miARNs.

Diseño de una estrategia para la identificación de genes blanco regulados por microARNs basado en la conservación evolutiva del par microARN-gen blanco.

Enfocamos nuestro análisis en 22 miARNs que están conservados en Angiospermas [6, 8]. En general estos miARNs están codificados por pequeñas familias hasta 32 miembros. En los genomas completos de Arabidopsis, poplar y arroz es común encontrar variaciones en la secuencia de los miARNs pertenecientes a una misma familia, especialmente en el primer nucleótido y los nucleótidos 20 y 21 [8].

Sin embargo, observamos que la región entre la posición 2 y 19 está bastante conservada y pudimos encontrar una secuencia consenso presente en la mayoría de los miembros de cada familia de miARNs en esas tres especies (tabla 2.1). Curiosamente, las bases variables fuera de esta región conservada son propensas a tener mismatches con genes blanco conocidos, lo que indica que podría existir una correlación entre la interacción miARN-gen blanco y la conservación de la secuencia del miARN.

Diseñamos una estrategia para identificar nuevos pares miARN-gen blanco principalmente basada en la conservación evolutiva de la secuencia del gen blanco (Figura 2.1). Las secuencias consenso de 18 nt de cada familia de miARN fueron usadas inicialmente para realizar la búsqueda de genes blanco en contigs de ESTs, de 41 especies de plantas, obtenidos de "Gene Index Project" un proyecto mantenido y administrado por la universidad de Harvard que contiene un catálogo completo de genes en una amplia gama de organismos incluyendo plantas. Además se utilizaron ARNm completos para *A. thaliana* y *Oryza Sativa* para ver la lista completa de especies, ver tabla A.1). Utilizando las secuencias consenso de 18nt y permitiendo 3 mismatches (errores), la búsqueda de genes blanco arrojó como resultado 38.597 genes distribuídos en las 43 especies (Figura 2.1, bin 1). Las interacciones G-U y los bulges fueron considerados como mismatches en esta primera búsqueda. Todos los genes blanco de *A. thaliana* conocidos hasta ese momento fueron identificados usando esta estrategia con la excepción de CSD2, un gen blanco del miR398 que contiene 4 mismatches (tabla A.1).

Teniendo en cuenta que la mayoría de los genes blanco arrojados presentan una escasa descripción del tipo genómica funcional, realizamos un BLASTx contra el proteoma de *A. thaliana*. El "locus ID" obtenido como "best hit" se utilizó como tag (etiqueta) para identificar al candidato en distintas especies (Figura 2.1). A pesar que esta estrategia no necesariamente identifica el gen ortólogo de Arabidopsis, sirve como propósito de clasificación de cada potencial gen blanco de miARN. Aunque la mayoría de los potenciales genes blanco pudieron ser fácilmente asignados con una etiqueta, algunos pocos casos, que incluye a los genes que representan ARNs no codificantes fueron perdidos en este paso.

Este enfoque permite la selección de los mejores candidatos basándose en la presencia de los genes blanco en un número distinto de especies. Utilizando 4 especies como el mínimo de especies requeridas (ya que tiene una buena especificidad), dio como resultado 3.781 genes que corresponden a 533 tags diferentes (Figura 2.1, bin 2).

La búsqueda también se puede hacer en combinación con filtros empíricos de interacción par miARN-gen blanco que tienen en cuenta la energía de interacción y la posición de los mismatches (ver Materiales y métodos). De los 38.597 candidatos iniciales, 9.375 pasan estos filtros (Figura 2.1, bin 4). Combinando filtros de energía y filtro de conservación evolutiva, la búsqueda arrojó como resultado 563 candidatos correspondientes a 146 tags (Figura 2.1, bin 5).

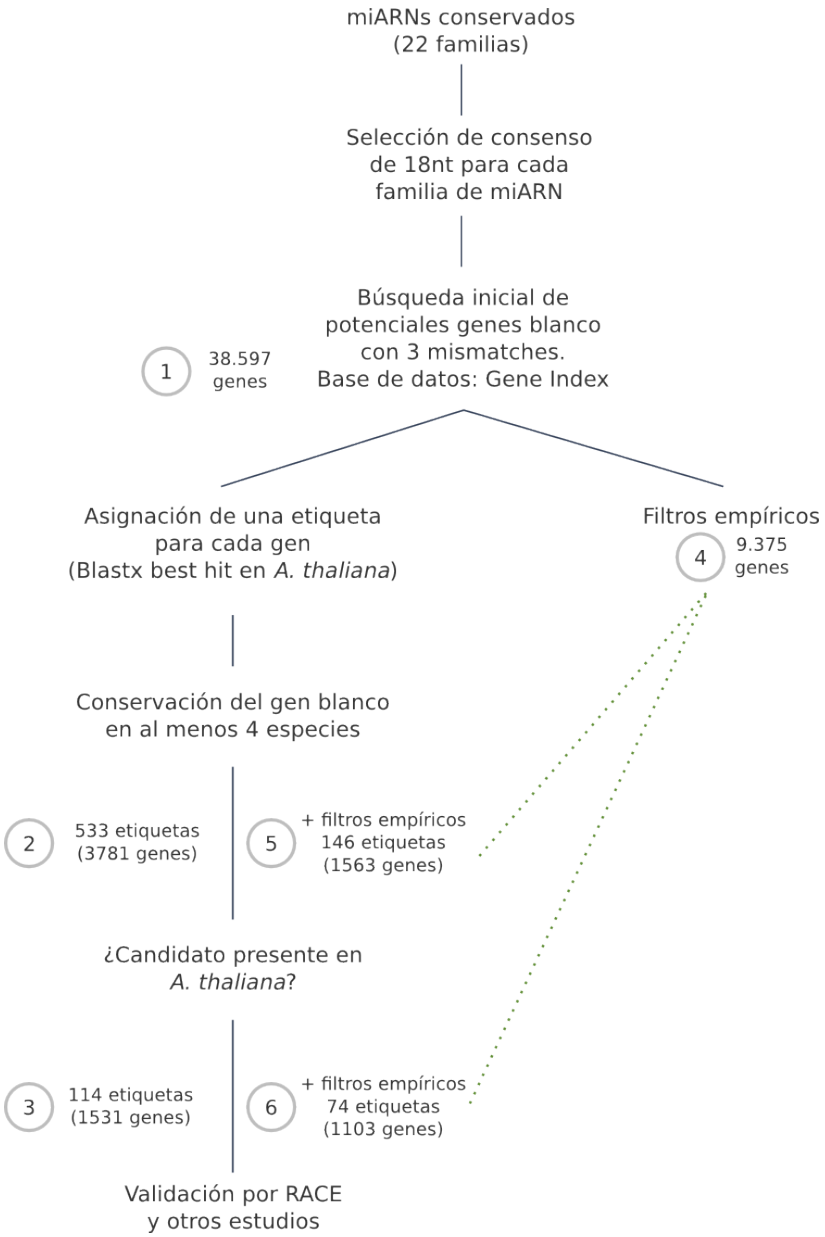


Figura 2.1 Esquema de la estrategia para la identificación de nuevos genes blanco. El número de genes blanco está identificado en cada paso. Luego de aplicar el análisis de conservación, todos los genes que tienen el mismo hit en Arabidopsis, fueron considerados como un solo gen blanco. El lado derecho muestra la búsqueda hecha con filtros empíricos: bin 5 y 6 incluyen genes blanco seleccionados con ambos filtros, empíricos y de conservación. Mientras que el bin 2 y 3 muestra los potenciales genes blanco seleccionados sólo con el filtro de conservación.

Table 2.1 miARNs y sus genes blanco en plantas

miARN	Consenso (18 nt)	Targets conocidos[a,b]
miR156	GACAGAAGAGAGTGAGCA	factores de transcripción SPL
miR159	TTGGATTGAAGGGAGCTC	factores de transcripción MYB, **NOZZLE (NZL)**
miR160	GCCTGGCTCCCTGTATGC	factores de transcripción ARF
miR162	CGATAAACCTCTGCATCC	DCL1
miR164	GGAGAAGCAGGGCACGTG	factores de transcripción NAC
miR166	CGGACCAGGCTTCATTCC	factores de transcripción HDZip
miR167	GAAGCTGCCAGCATGATC	factores de transcripción ARF, **IAA-ALANINE RESISTANT 3 (IAR3)**
miR168	CGCTTGGTGCAGGTCGGG	AGO1
mir169	AGCCAAGGATGACTTGCC	factores de transcripción CCAAT-HAP2
mir171	TTGAGCCGTGCCAATATC	factores de transcripción GRAS
miR172	GAATCTTGATGATGCTGC	factores de transcripción AP2
miR319	TGGACTGAAGGGAGCTCC	factores de transcripción TCP
miR390	AGCTCAGGAGGGATAGCG	TAS RNA
miR393	CCAAAGGGATCGCATTGA	TIR1 proteins, F-BOX proteins
miR394	TGGCATTCTGTCCACCTC	proteínas F-BOX
miR395	TGAAGTGTTTGGGGGAAC	ATP-sulfurilasas, transportadores de sulfato
miR396	TCCACAGCTTTCTTGAAC	factores de transcripción GRF, **MMG4.7, FLUORESCENT IN BLUE LIGHT (FLU)**
miR397	CATTGAGTGCAGCGTTGA	Laccases
miR398	GTGTTCTCAGGTCACCCC	Cu/Zn SODs, CytC oxidase protein subunit, Chaperona de cobre (CCS)
miR399	GCCAAAGGAGATTTGCCC	Enzima E2 de conjugación de ubiquitina
miR408	TGCACTGCCTCTTCCCTG	Blue copper proteins, Laccases, **P-TYPE ATPase (PAA2), PAC1 (Proteasome component)**
miR827	TAGATGACCATCAGCAAA	SPX proteins

a Los genes blanco fueron agrupados según sus funciones.
b Nuevos genes blanco validados experimentalmente en este estudio están indicados en negrita.

Parámetros empíricos y de conservación evolutiva pueden actuar de manera sinérgica para identificar genes blanco regulados por miARNs.

Potenciales genes blanco de miARNs fueron clasificados de acuerdo al mínimo número de especie en donde fueron detectados (Figura 2A-E). Como control para cada miARN generamos 10 secuencias "scramble" (al azar), dividiendo las secuencias originales de a di-nucleótidos y luego generando nuevas secuencias al azar conservando la composición de los di-nucleótidos. Estas secuencias al azar fueron utilizadas para realizar búsqueda de genes blanco del mismo modo que lo hicimos para las secuencias originales. La relación señal/ruido fue calculada como el cociente entre el número de genes blanco para los miARNs y el número promedio obtenido de las secuencias al azar. El radio fue de 1,2 para todos los miARNs juntos sin requerir conservación y esa relación incrementa con el número de especie en donde los genes blanco fueron detectados (Figura 2.2 A, recuadro). Los datos para todos los miARNs y sus potenciales genes blanco conservados en al menos 4 especies están incluidos en la tabla 2.2.

Luego estudiamos la selección de candidatos teniendo en cuenta los filtros empíricos. Para esto aplicamos una versión modificada de los filtros descritos anteriormente y requiriendo (i) una energía mínima de hibridación (MFE) de al menos 72% del apareamiento perfecto de cada secuencia consenso y (ii) que sólo un mismatch pudiera estar presente entre la posición 1 y la 11 de la secuencia consenso (2-12 del miARN). De la búsqueda inicial 9.375 genes pasaron estos filtros conteniendo el 97% de los genes validados anteriormente de Arabidopsis. (Figura 2.1, bin 4).

Al aplicar solamente este filtro empírico, dio como resultado una relación señal/ruido de 1,7, al agrupar todos los miARNs juntos (Figura 2.2 A). Observamos que aplicar si-

multáneamente los filtros empíricos y de conservación aumentaron significativamente la relación señal/ruido para todos los miARNs juntos (Figura 2.2 A recuadro) y también de cada miARN individualmente (Figura 2.2 B-E, recuadros y tabla 2.2). En varios casos, esta relación llega hasta 10 cuando se requiere de que el gen blanco este presente en más de 5 especies y que pase los filtros empíricos (Figure 2.2 A–D). Este efecto sinérgico indica que el filtro de conservación evolutiva y los parámetros empíricos pueden estar seleccionando aspectos diferentes de la interacción miARN-gen blanco.

Observamos que el número de genes blanco candidato y la relación señal/ruido es variable entre los distintos miARNs. El miR396 tiene la mayor cantidad de potenciales genes blanco, 92 de ellos presentes en al menos 4 especies y 26 de ellos pasan además los filtros empíricos (Tabla 2.2 y Figura 2.2 B). El miR408 y el miR398 también tienen un número alto de potenciales genes blanco y buenas relaciones de señal/ruido (Figura 2.2 C-D).

En contraste, ciertos miARNs como el miR162, miR168 y miR399 tienen un solo potencial gen blanco conservado en al menos 4 especies de acuerdo con nuestra búsqueda (Tabla 2.2 y Figura 2.2 E). Al menos en el caso del miR162 y del miR168 este resultado podría estar reflejando su rol específico en la regulación por retroalimentación de la biogénesis del miARN, ya que controlan los niveles de expresión DCL1 y AGO1 respectivamente [28],[29].

Como control adicional para nuestra estrategia hicimos la búsqueda de genes blanco del miR158 y miR173, que son miARNs presentes solamente en A. thaliana y especies bien cercanas (17). Como era esperado estos miARNs no generaron más candidatos que sus versiones al azar (Tabla 2.2 y Figura 2.2 F).

Luego chequeamos si los pares miARN-gen blanco altamente conservados tenían una interacción más fuerte que los que están presentes en pocas especies. Para esto calculamos la energía mínima de hibridación para cada interacción detectada en nuestro trabajo. Observamos que los pares miARN-gen blanco presentes en muchas especies tienden a tener energía de interacción mayores que los que están presentes en menos especies (Figure 2.3 A). De todos modos, la correlación no fue notoria y algunas interacciones miARN-gen blanco tuvieron una baja energía de hibridación (Figure 2.3 A). Estos resultados muestran que una alta conservación podría no ser necesariamente equivalente a una fuerte interacción, la misma podría proporcionar una explicación para los efectos sinérgicos causados por los filtros de evolución y empíricos sobre la relación señal/ruido.

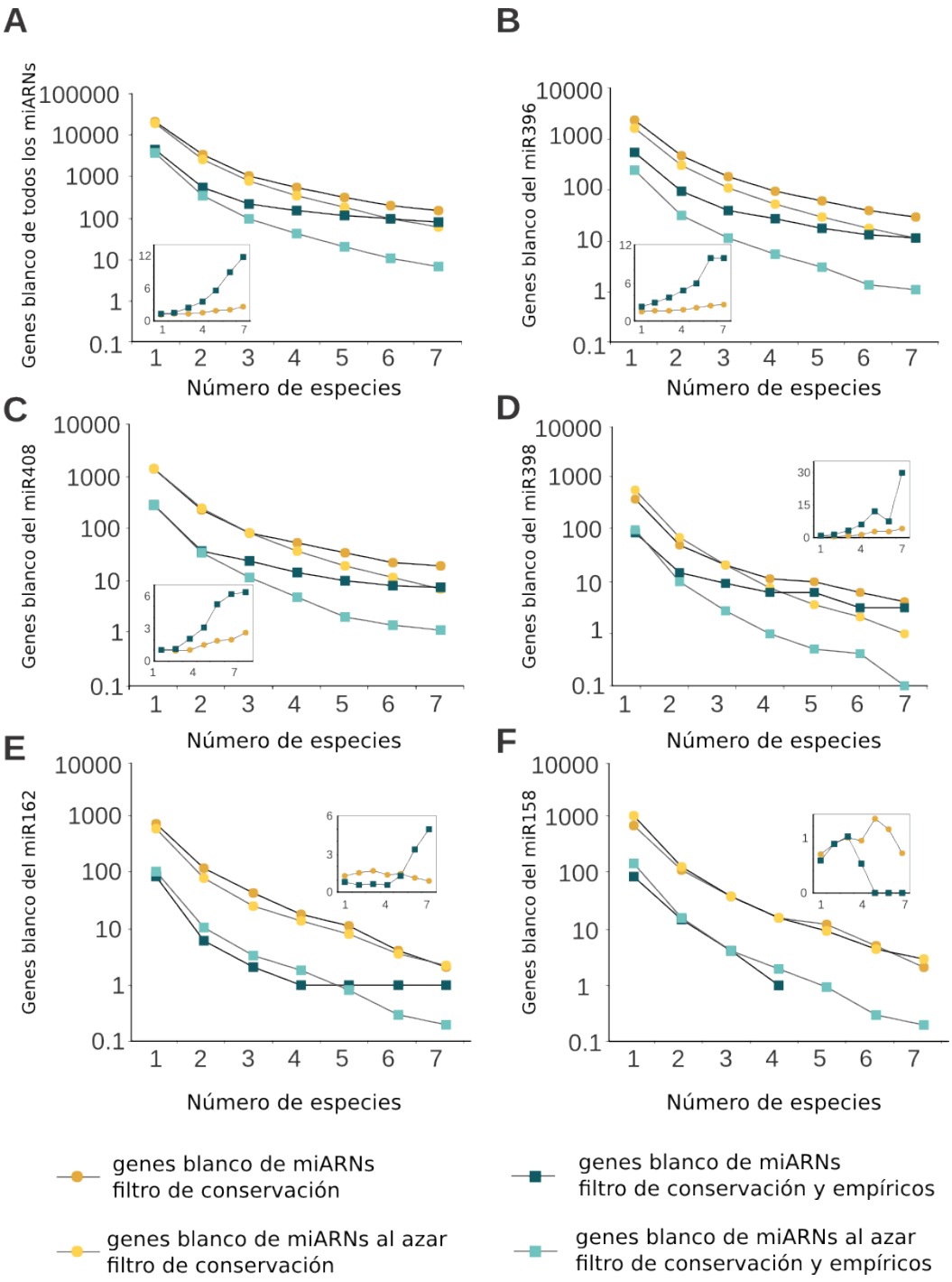


Figura 2.2 Conservación de potenciales genes blanco en distintas especies. Todos los miARNs (**A**), miR396 (**B**), miR408 (**C**), miR398 (**D**), miR162 (**E**), miR158 (**F**). Puntos naranja representan los genes blanco de miARNs usando filtro evolutivo. Puntos amarillos representan los genes blanco de las secuencias al azar usando filtro evolutivo. El cuadrado azul muestra los genes blanco de miARNs luego de aplicar filtros empíricos y evolutivos, mientras que el cuadrado celeste representa los genes blanco de las secuencias al azar en las mismas condiciones. Los recuadros muestran la relación señal/ruido.

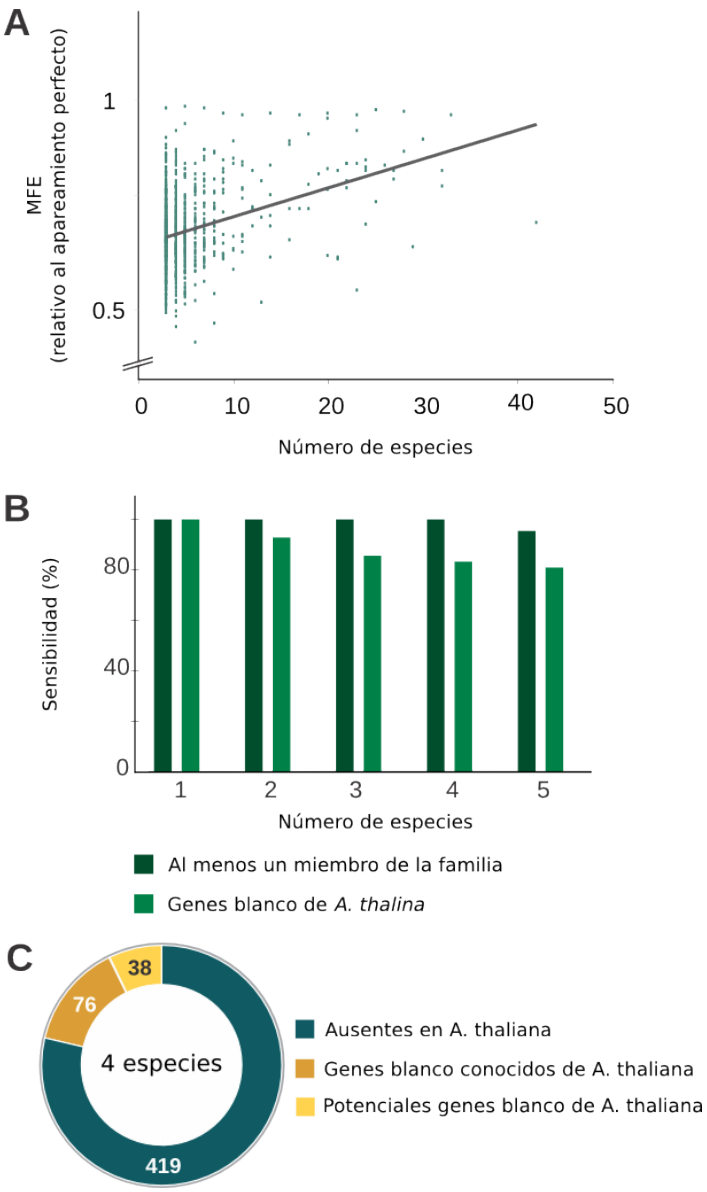


Figura 2.3 Selección de genes blanco por conservación evolutiva de la secuencia. **(A)** Relación entre MFE y el número de especies en donde cada gen blanco fue detectado. **(B)** Sensibilidad de la estrategia, analizado de dos modos distinto. Verde clarito: evaluando la presencia de genes validados en Arabidopsis y en verde oscuro teniendo en cuenta la presencia de por lo menos un gen blanco de cada familia regulada por miARNs. **(C)** Clasificación de los potenciales genes blanco presentes en al menos 4 especies.

Table 2.2 Detection of miRNA targets using different filters

	Sin filtros			Filtros empíricos			Conservación 4 especies			Todos los filtros		
	miARN	scramble	ratio	miARN	scramble	ratio	miARN	scramble	ratio	miARN	scramble	ratio
miR156	3915	3994.4 ± 149.9	1.0	890	704.7 ± 45.2	1.3	34	39.7 ± 3.1	0.9	10	5.4 ± 1.1	1.9
miR159	1663	1283.7 ± 47.8	1.3	472	254.9 ± 21.9	1.9	20	10.1 ± 1.1	2.0	6	1.5 ± 0.5	4.0
miR160	793	695.6 ± 30.5	1.1	277	157.5 ± 28.8	1.8	5	4.4 ± 0.9	1.1	4	0.5 ± 0.3	8.0
miR162	1191	930.2 ± 139.5	1.3	108	164.7 ± 24.1	0.7	18	13.5 ± 3.5	1.3	1	1.8 ± 0.5	0.6
miR164	2486	1480.2 ± 60.4	1.7	678	333.2 ± 32.2	2.0	39	12.4 ± 1.9	3.1	12	1.5 ± 0.5	8.0
miR166	879	815.5 ± 45.0	1.1	231	129 ± 14.5	1.8	16	10.6 ± 1.4	1.5	6	0.9 ± 0.4	6.7
miR167	1777	1364.2 ± 146.6	1.3	478	214.8 ± 27.5	2.2	22	20.2 ± 3.6	1.1	4	1.8 ± 0.5	2.2
miR168	962	797.5 ± 48.5	1.2	209	185 ± 14.2	1.1	6	4.4 ± 0.8	1.4	1	1.1 ± 0.5	0.9
miR169	1540	1047.2 ± 69.7	1.5	464	181.4 ± 15.6	2.6	26	11.1 ± 2.1	2.3	10	1.2 ± 0.2	8.3
miR171	884	723.4 ± 32.1	1.2	202	113.8 ± 13.4	1.8	7	6.6 ± 1.4	1.1	2	0.7 ± 0.3	2.9
miR172	3007	1693.7 ± 124.7	1.8	540	288.1 ± 40.3	1.9	34	17.7 ± 1.7	1.9	5	2.2 ± 0.6	2.3
miR319	1363	1274.2 ± 113.6	1.1	324	249.2 ± 22.3	1.3	18	15 ± 2.8	1.2	7	1.8 ± 0.5	3.9
miR390	873	814.4 ± 64.3	1.1	335	173 ± 22.5	1.9	8	4.7 ± 1.2	1.7	3	0.7 ± 0.5	4.3
miR393	986	844.6 ± 58.7	1.2	276	124.6 ± 11.1	2.2	14	7.1 ± 1.2	2.0	5	0.5 ± 0.2	10.0
miR394	1569	1531.4 ± 57.5	1.0	188	237.1 ± 25.0	0.8	26	21.4 ± 2.2	1.2	3	2.9 ± 0.5	1.0
miR395	1472	1226.7 ± 66.7	1.2	426	217.6 ± 16.5	2.0	11	8.8 ± 1.3	1.3	6	1.3 ± 0.3	4.6
miR396	4641	2979.3 ± 246.6	1.6	1246	390.5 ± 38.8	3.2	92	51.4 ± 5.9	1.8	26	5.4 ± 1.0	4.8
miR397	1426	1050.9 ± 27.9	1.4	368	236.5 ± 23.5	1.6	26	9.7 ± 0.8	2.7	10	1.6 ± 0.3	6.3
miR398	935	834 ± 34.5	1.1	376	144 ± 18.1	2.6	11	7.5 ± 1.6	1.5	6	1 ± 0.3	6.0
miR399	1192	1137.6 ± 72.0	1.0	275	207.8 ± 24.9	1.3	5	13.6 ± 1.7	0.4	1	1.5 ± 0.7	0.7
miR408	2782	2502.9 ± 103.6	1.1	695	468.7 ± 50.8	1.5	51	35.1 ± 3.0	1.5	14	4.6 ± 0.8	3.0
miR827	2261	2000.1 ± 119.8	1.1	317	297.1 ± 45.0	1.1	44	23.4 ± 3.9	1.9	4	2.3 ± 0.8	1.7
Total	38597	31021.7 ± 1859.8	1.2	9375	5473.2 ± 576.3	1.7	533	348.4 ± 47.0	1.5	146	42.2 ± 11.3	3.5
Control												
miR158	1364	1462.8 ± 69.1	0.9	170	208.7 ± 15.8	0.8	15	16 ± 1.7	0.9	1	1.9 ± 0.4	0.5
miR173	1386	1232.1 ± 101.7	1.1	243	215.6 ± 23.4	1.1	11	12 ± 2.4	0.9	1	1.5 ± 0.4	0.7

a Sin filtros, búsqueda inicial utilizando los miARN consenso de 18nt y 3 mismatches.
b Filtros empíricos, energía de al menos 72% del apareamiento perfecto y 1 mismatch en la posición 2-12 del par miARN-gen blanco.
c Conservación del ID tag en al menos cuatro especies.
d Todos los filtros, combinación de los filtros empíricos y de conservación en al menos cuatro especies.
e miARN, genes blanco para cada miARN específico.
f scramble, promedio de los genes blanco de 10 versiones al azar de cada miARN ± error estándar.

Identificación de nuevos genes blanco en *A. thaliana* por conservación de la secuencia del gen blanco.

Para encontrar nuevos genes blanco nos enfocamos en los genes potenciales que fueron seleccionados de nuestra estrategia utilizando solamente conservación evolutiva, debido a que los parámetros empíricos ya fueron utilizados extensamente en trabajos anteriores. [2],[14],[27]. En primer lugar, analizamos la detección de genes blanco validados previamente en *A. thaliana* [basado en [10]] usando nuestra estrategia y encontramos que el 84% de ellos estaban presentes en al menos 4 especies (Figura 2.3 B). Consideramos esto como un buen resultado ya puede ser que no todos los genes blanco de Arabidopsis estén conservados evolutivamente.

Generalmente los miARNs en plantas regulan genes que codifican para proteínas de la misma familia, es por esto que evaluamos si por lo menos un miembro de cada familia era detectado en nuestro enfoque. Encontramos genes blanco pertenecientes a casi todas las familias de genes codificantes para proteínas presentes en cuatro especies (Figura 2.3 B), con la excepción de TAS3, que es regulado por el miR390, al ser un ARN no codificante no es detectado por Blastx.

Para encontrar nuevos genes blanco regulados por miARNs, nos enfocamos únicamente en los potenciales genes blanco conservados en 4 especies, donde una de ellas es *A. thaliana* (Figura 1, bin 1). Genes blanco de miARNs que no están presentes en *A. thaliana* podrían incluir genes que perdieron su regulación durante la evolución o genes que hayan adquirido control por un miARN conservado más reciente en otras especies. La conservación en cuatro especies fue elegida como un filtro evolutivo porque provee buena sensibilidad para genes blanco conocidos.

Identificamos 114 potenciales genes que satisfacen este criterio. Donde 76 de ellos son genes validados anteriormente o genes muy relacionados (Figura 2.3 C). Curiosamente encontramos 38 genes que no tienen relación con genes blanco conocidos de miARNs y decidimos estudiar este grupo con mayor detalle. Nos enfocamos primero en los genes que estaban presentes en un gran número de especies para tener mejor especificidad (Figura 2.2) e intentamos validarlos utilizando 5' RACE PCR modificada [17],[15].

Un potencial gen blanco del MiR408 era At5g21930 que codifica para P-TYPE ATPase OF ARABIDOPSIS 2 (PAA2) y estaba presente en 22 especies distintas incluido monocotiledóneas y dicotiledóneas. MiR408 es inusual debido a que tiene un 5'-A, sin embargo >30% de las secuencias maduro del miR408 corresponden a una variante corrida 1 nt que empieza con 5'-U [18] (Figura 2.4 A). La validación experimental reveló fragmentos de ARNm compatible con este último sitio de corte (Figura 2.4 A). PAA2 es necesaria para

el transporte de iones de cobre a plastocianina [21], y su regulación por el miR408 está relacionada con el rol de este miARN en la homeostasis de cobre [30].

Otro potencial candidato del miR408 era At3g22110 que codifica para PROTEASOME ALPHA SUBUNIT C1 (PAC1) y estaba presente en 20 especies. Por medio de 5' RACE PCR demostramos que este gen es gen blanco del miR408 (Figura 2.4 A). Curiosamente la interacción del par miARN-gen blanco tiene 3 mismatches en la región 5', y se hubiera perdido como potencial gen blanco si se aplicaban solamente los filtros empíricos.

Luego estudiamos los genes blanco del miR396, donde los genes SVP y SUI1 estaban presentes en 29 y 19 especies respectivamente. Pero en ambos casos fallamos al obtener producto de la PCR utilizando 5' RACE PCR modificada. La falta de regulación de este gen por el miR396 podría estar relacionado a la débil energía de hibridación del par miARN-gen blanco, aunque no podemos descartar que el miR396 esté controlando su traducción.

Otros dos potenciales genes blanco del miR396 eran At5g43060 y At3g14110 que codifican para la proteasa MMG4.7 y FLUORESCENT IN BLUE LIGTH (FLU), respectivamente. Y en ambos casos pudimos detectar el corte (Figura 2.4 C y D).

En contraste con el miR408 y miR396, donde tienen varios potenciales genes blanco, obtuvimos un solo potencial gen blanco para el miR159, un factor de transcripción MYB que regula desarrollo del estambre y polen. [19] El otro potencial gen blanco era At4g27330, conocido como NOZZLE/SPOROCYTLESS. Este factor de trascripción, que participa en desarrollo del estambre y óvulo [26, 32], fue también validado por 5' RACE PCR (Figura 2.4 E). Es interesante notar que al menos las funciones de NOZZLE y PAA2 pueden estar directamente relacionadas con el rol de genes blanco, ya descritos anteriormente, del miR159 y miR408 respectivamente.

PAA2, FLU y NOZZLE fueron detectados en mono y dicotiledóneas mientras que PAC1 y MMG4.7 fueron detectadas solamente en dicotiledóneas (Figura 2.4 A-E). Las posiciones del sitio de unión del miARN-gen blanco están altamente conservadas y muchas de las posiciones variables corresponden a mismatches con el miARN o variaciones del tipo G-C/G-U. Además este método no requiere que el sitio del gen blanco esté conservada, sino más bien que haya una interacción predicha con el miARN en distintas especies. De esta manera el sitio de NOZZLE, donde la secuencia cambia en diferentes especies (Figura 2.4 E), pudo ser detectado por este enfoque.

Identificación de nuevos genes blanco permitiendo interacciones G-U.

Los genes blanco identificados utilizando la estrategia descrita anteriormente, tienen varios mismatches y bulges con sus miARNs, lo que puede ayudar a explicar por que se perdieron en trabajos anteriores. También notamos que muchas de estas nuevas interacciones miARN-gen

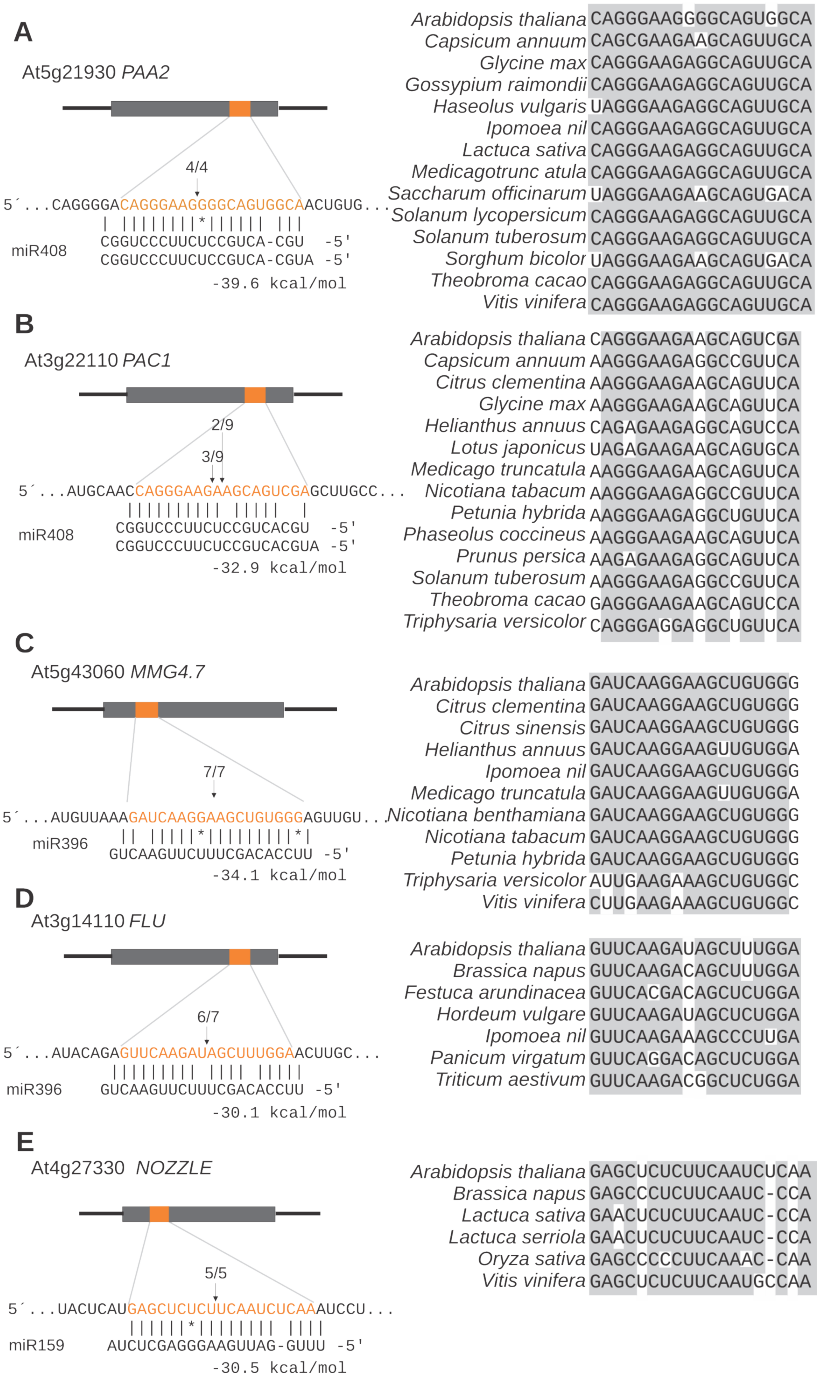


Figura 2.4 Nuevos genes blanco validados en *A. thaliana*. El alineamiento entre el miARN y los nuevos genes blanco identificados se muestran en la izquierda. La conservación evolutiva de la secuencia del sitio reconocido por el miARN en las especies seleccionadas se muestra a la derecha. La figura muestra las interacciones del miR408 con PAA2 (**A**), miR408 con PAC1 (**B**), miR396 con MMG4.7 (**C**), miR396 con FLU (**D**), miR159 con NOZZLE (**E**). Las flechas marcan el sitio de corte determinado por 5'RACE-PCR y los números indican la frecuencia de clonadao de cada fragmento.

blanco contenían posiciones que variaban alternadamente entre G-C y G-U en distintas especies. Como consideramos G-U como mismatch en nuestra búsqueda inicial, decidimos realizar nuevamente la búsqueda con los miARNs consenso de 18nt pero permitiendo ahora 4 mismatches, donde al menos uno de ellos tiene que ser del tipo G-U. Esta búsqueda permitiría interacciones miARN-gen blanco con sólo 14 bases apareadas perfectamente.

Para compensar el uso de estos parámetros relajados en términos de mismatches, requerimos que el gen blanco aparezca en al menos 10 especies distintas para aumentar la especificidad (Figura 2.5 A). Encontramos 125 potenciales genes blanco en *A. thaliana* teniendo en cuenta este criterio (Figura 2.5 A) y 34 de ellos no aparecían en las búsquedas anteriores. El gen blanco CSD2 regulado por el miR398, que no apareció anteriormente, fue detectado con estos parámetros.

Luego examinamos el último grupo de potenciales genes regulados por miARNs que estaban realizando funciones auxiliares a los genes blanco ya descritos para cada miARN. Y encontramos que el miR167 que regula factores de respuesta a auxina (ARFs), también regulaba potencialmente a un gen denominado IAA-ALANINE RESISTANT 3 (IAR3) (Figura 2.5 B y C), que está involucrado en el control de niveles libre de auxina [7, 24].

IAR3 en Arabidopsis tiene 3 mismatches con respecto al miR167, pero en la posición 12 de la interacción miARN-gen blanco, hay una interacción G-U en varias especies (Figura 2.5 B y C). La técnica de 5' RACE PCR confirmó que el gen realmente era gen blanco del miR167 (Figura 2.5 C).

Identificación de genes blanco específicos de *Solanaceae*.

Pensamos que la estrategia mostrada también se puede utilizar para encontrar genes blanco presentes específicamente en un grupo de especies relacionadas. Por lo tanto intentamos demostrar esto, encontrando potenciales genes blanco específicos de la familia de *Solanaceae*.

Elegimos esta familia en particular, ya que 6 especies estaban bien representadas en la biblioteca utilizada. La relación señal/ruido entre los genes blanco y las secuencias al azar era más de 2 cuando el filtro empírico o de conservación (en al menos 3 de las 6 especies *Solanaceae*) fueron aplicados (Figura 2.6 A). Curiosamente, al aplicar ambos filtros dio como resultado una relación señal/ruido por encima de 6 (Figura 2.6 A), confirmando nuestros previos hallazgos de que ambos filtros mejoran la detección de genes blanco de miARNs.

Encontramos 132 potenciales genes blanco presentes en al menos 3 especies *Solanaceae*. De este grupo, 41 genes no fueron detectados en otras especies (Figura 6B). El gen blanco más común fue la metalotioneína MT2A, presente en las 6 *Solanaceae*, como potencial gen

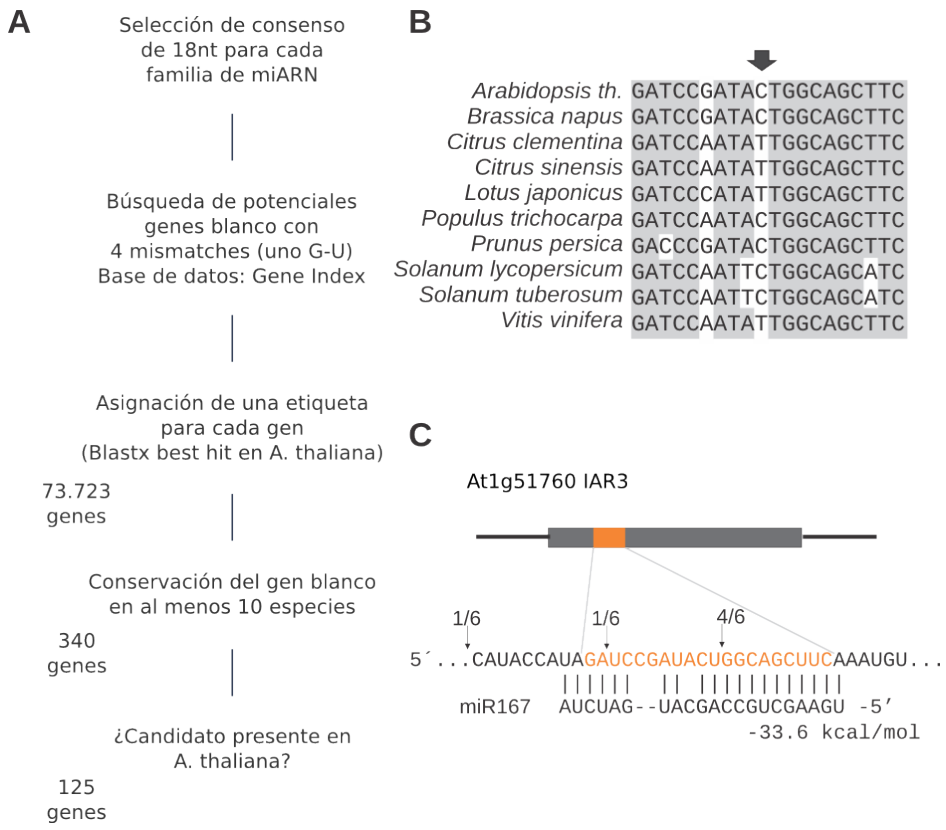


Figura 2.5 Identificación de un nuevo gen blanco de miARN, relajando los parámetros de interacción pero incrementando el parámetro de conservación evolutiva. **(A)** Esquema de la estrategia modificada para identificar genes blanco de miARNs. **(B)** Conservación del sitio blanco reconocido por el miARN en distintas especies. La flecha indica una variación de G-C o G-U con el miARN dependiendo de la especie. **(C)** Alineamiento en Arabidopsis thaliana del gen blanco IAR3 con el miR167. La flecha indica la posición del corte indicada por 5'RACE-PCR y el número indica la frecuencia de clonado de cada fragmento.

blanco del miR398, mientras que MT2B, homólogo de este gen, fue detectado en 5 especies (Figura 2.6 B-D).

Luego, aprovechamos las plantas transgénicas de tabaco que contienen un transgén 35S.mir398 (A.F. Lodeyro, N. Carrillo y J.F. Palatnik resultados no publicados) y chequeamos la expresión de estos genes. Encontramos que CSD2, un gen blanco conservado del miR398, disminuía su expresión > 10 veces en las plantas transgénicas 35S:miR398 comparadas con la planta salvaje (Figura 2.6 E). Curiosamente, observamos que tanto MT2A como MT2B disminuyeron sus niveles de transcripción > 5 veces en estas plantas (Figura 2.6 E). Estos resultados concuerdan con la regulación de MT2A y MT2B por el miR398, aunque no necesariamente demuestra una interacción directa. Además, estos resultados demuestran que los genes blanco presentes en un grupo específico de especies pueden ser encontrados utilizando esta estrategia.

2.2.2 comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas.

A partir de la estrategia descrita en el capítulo anterior, que fue utilizada para encontrar y validar experimentalmente genes blanco regulados por miARNs en Arabidopsis thaliana, desarrollamos una herramienta web denominada comTAR[1] (Conserved plant miRNA target prediction tool) [5]. La misma se puede utilizar para predecir potenciales genes blanco regulados por miARNs en plantas y está basada en la conservación evolutiva del par miARN-gen blanco con un número relajado de mismatches. ComTAR permite distintas opciones/parámetros de búsqueda que pueden ser modificados por el usuario:

- Filtro de mismatch: Solamente un mismatch está permitido entre la posición 1 y la 11 de la secuencia del miARN consenso. (Sí/No).
- Corte por energía de hibridación: Se define que un gen blanco es predicho si la mínima energía de hibridación está por debajo del corte elegido.
- El número mínimo de especies donde un mismo TAG está presente para un miARN particular.

Buscar potenciales genes blanco de miARN

Esta es la búsqueda por defecto. El usuario puede realizar la búsqueda de genes blanco de miARNs conservados. En la primer pantalla se muestra los potenciales genes blanco para un

[1] http://rnabiology.ibr-conicet.gov.ar/comtar

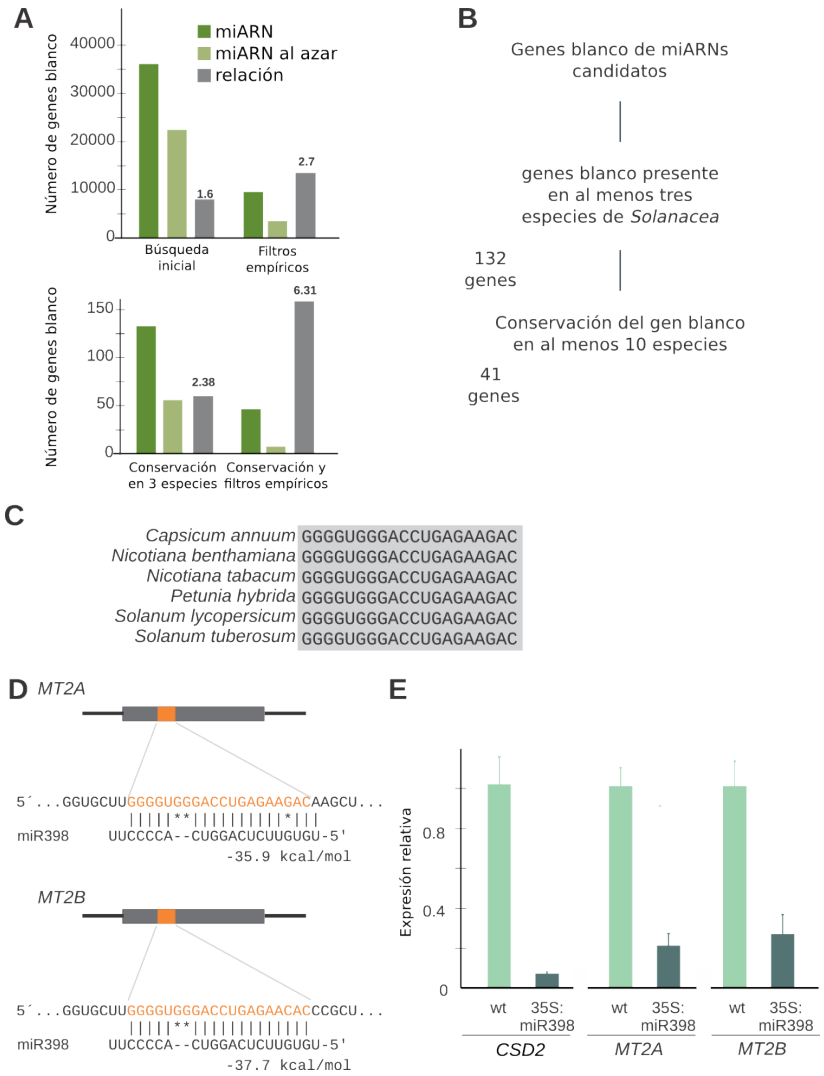


Figura 2.6 Identificación de genes blanco de miARN, específicos de *Solanaceae*. **(A)** Predicción de genes blanco de miARN en cinco especies de *Solanaceae*. El número de genes blanco de todos los miARNs conservados juntos se muestra luego de aplicar distintos filtros. También se muestran los genes blanco obtenidos a partir de las secuencias al azar. **(B)** Esquema que muestra la estrategia para identificar genes blanco específicos de *Solanaceae*. **(C)** Conservación del sitio reconocido por el mir398 con MT2A específico de *Solanaceae*. **(D)** Esquema que muestra el sitio de unión entre el miR398 y MT2A y MT2B. **(E)** Niveles de transcriptos de CSD2, MT2A y MT2B en plantas salvajes y plantas transgénicas de tabaco (cv Petit havana) que sobreexpresan el miR398.

miARN dado (Figura 2.7), con una breve descripción del gen, la familia a la que pertenece y además en cuantas y cuáles especies está presente. También, para cada especie que está presente, se tiene acceso por pantalla al alineamiento del miARN-gen blanco, la energía de hibridación y los filtros empíricos de interacciones conocidas del par miARN-gen blanco (Figura 2.8).

Buscar familias de potenciales genes blanco de miARN

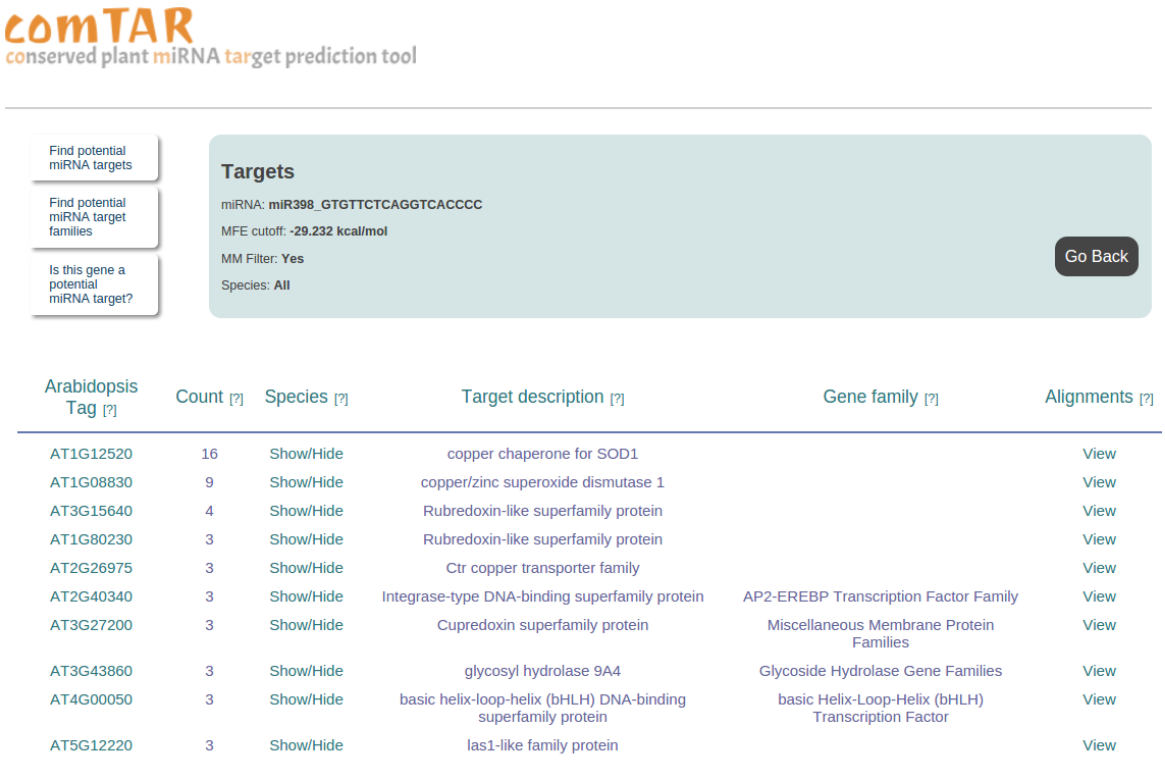
Debido a que los miARNs en plantas en general regulan genes que codifican a proteínas de las misma familias, la herramienta tiene otra funcionalidad donde permite la búsqueda de genes agrupados por familias en vez de agruparlos por TAG. De este modo genes en distintas especies con diferentes TAG, pero que pertenecen a la misma familia pueden ser detectados como familias de potenciales genes blanco.

¿Es este gen un potencial gen blanco de algun miARN conservado?

El usuario puede introducir un locus TAG en particular (tanto de Arabidopsis como el 'gene ID' del Phytozome) y se identifica si este gen en particular puede ser un potencial gen blanco de algun miARN y en cuantas especies aparece. En Arabidopsis se utiliza el LocusID como identificador, mientrás que en Phytozome este identificador varía según la especie y se puede ver la precedencia de cada especie en el sitio de Phytozome.

Buscar tu secuencia particular

En esta parte del programa el usuario puede realizar la búsqueda de nuevos ARNs pequeños teniendo en cuenta que la secuencia introducida tiene que ser de 18nt de largo (posiciones 2-19). Luego de la búsqueda, se da un link al usuario y después de unas horas, cuando haya sido procesado el cálculo, el usuario puede entrar a ese link y navegar los resultados por pantalla.



Arabidopsis Tag [?]	Count [?]	Species [?]	Target description [?]	Gene family [?]	Alignments [?]
AT1G12520	16	Show/Hide	copper chaperone for SOD1		View
AT1G08830	9	Show/Hide	copper/zinc superoxide dismutase 1		View
AT3G15640	4	Show/Hide	Rubredoxin-like superfamily protein		View
AT1G80230	3	Show/Hide	Rubredoxin-like superfamily protein		View
AT2G26975	3	Show/Hide	Ctr copper transporter family		View
AT2G40340	3	Show/Hide	Integrase-type DNA-binding superfamily protein	AP2-EREBP Transcription Factor Family	View
AT3G27200	3	Show/Hide	Cupredoxin superfamily protein	Miscellaneous Membrane Protein Families	View
AT3G43860	3	Show/Hide	glycosyl hydrolase 9A4	Glycoside Hydrolase Gene Families	View
AT4G00050	3	Show/Hide	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	basic Helix-Loop-Helix (bHLH) Transcription Factor	View
AT5G12220	3	Show/Hide	las1-like family protein		View

Figura 2.7 Resultados de la búsqueda con parámetros por defecto para el miR398

Sequence ID	Species	5'-target-3' Alignment 3'-miRNA-5' [?]	MFE [?]
Aquca_013_00504.1	Aquilegia coerulea	TGGGCGACCTGGGAACAT *|||*||||||*|||||* CCCCACTGGACTCTTGTG	-31.7
471402	Arabidopsis lyrata	TGGGAGACCTGGGAACAC *|||*||||||*|||||| CCCCACTGGACTCTTGTG	-32.1
AT1G12520.1	Arabidopsis thaliana	TGGGAGACCTGGGAACAC *|||*||||||*|||||| CCCCACTGGACTCTTGTG	-32.1
Bradi5g18900.3	Brachypodium distachyon	TTGGTGACCTGGGAACGC **|||||||||*||||*| CCCCACTGGACTCTTGTG	-33.5
Bra026968	Brassica rapa	TGGGCGACCTGGGAACAC *|||*||||||*|||||| CCCCACTGGACTCTTGTG	-32.5
Carubv10011816m	Capsella rubella	TGGGAGACCTGGGAACAC *|||*||||||*|||||| CCCCACTGGACTCTTGTG	-32.1
evm.model.supercontig_29.47	Carica papaya	TAGGTGACCTGAGAACAT **||||||||||||||* CCCCACTGGACTCTTGTG	-34.2
Ciclev10021134m	Citrus clementina	TTGGTGACCTGGGAACAC **|||||||||*|||||| CCCCACTGGACTCTTGTG	-33.9
orange1.1g020436m	Citrus sinensis	TTGGTGACCTGGGAACAg **|||||||||*|||||* CCCCACTGGACTCTTGTG	-32.7

Figura 2.8 Parte de la salida de comTAR mostrando el par miR398/SOD1 (At1g12520) en diferentes especies

Capítulo 3

My second chapter

3.1 Enfoque bioinformático para el estudio de la evolución y biogénesis de microARNs en plantas.

Conclusiones 4

Conclusiones

4.1 Primera parte

En cuanto a la primera parte, a través de diferentes estrategias y estudios, hemos alcanzado las siguientes conclusiones:

- Diseñamos una estrategia para identificar genes blanco regulados por miARNs en plantas, basado en la conservación evolutiva del par microARN-gen blanco.

- El enfoque requiere que la interacción miARN-gen blanco, pueda ocurrir en el contexto de un conjunto mínimo de parámetros que interactúan en diferentes especies. Pero la secuencia del gen blanco en sí, no necesariamente tiene que estar conservada.

- Además, nuestro enfoque permite ajustar el número de especies requeridas como un filtro para realizar la búsqueda con diferentes sensibilidades y relaciones señal/ruido.

- Utilizando esta estrategia identificamos y validamos experimentalmente nuevos genes blanco en *A. thaliana*, a pesar de que este sistema ya había sido estudiado en detalles en distintos enfoques genómicos a gran escala ([1, 2, 11, 14, 23, 27]).

- Tres de los nuevos genes blanco validados tienen bulges. Parámetros empíricos usualmente le otorgan una gran penalidad a ellos, que puede llegar a ser el doble que un mismatch regular [14], sin embargo es probable que genes blancos con bulges asimétricos sean más frecuente de lo que se pensaba previamente en plantas.

- El enfoque ofrece una estrategia alternativa a otras predicciones que se basan en parámetros empíricos del par miARN-gen blanco [2, 6, 9, 14].

- Una ventaja de la estrategia presentada es que la interacciones miARN-gen blanco conservadas probablemente participen en procesos biológicos relevantes.
- Además, esta estrategia puede ser fácilmente modificada para incorporar datos de otras bibliotecas, y/o para realizar la búsqueda de genes blanco presentes en un grupo específico de especies de plantas.

En la segunda parte de esta Tesis,

References

[1] Addo-quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2009). Endogenous siRNA and microRNA targets identified by sequencing of the Arabidopsis degradome. *NIH Public Access*, 18(10):758–762.

[2] Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005). microrna-directed phasing during trans-acting sirna biogenesis in plants. *Cell*, 121(2):207 – 221.

[3] Altschup, S. F., Gish, W., Pennsylvania, T., and Park, U. (1990). Basic Local Alignment Search Tool 2Department of Computer Science. pages 403–410.

[4] Axtell, M. J. and Bowman, J. L. (2008). Evolution of plant microRNAs and their targets. *Trends in Plant Science*, 13(7):343–349.

[5] Chorostecki, U. and Palatnik, J. F. (2014). comTAR: a web tool for the prediction and characterization of conserved microRNA targets in plants. *Bioinformatics (Oxford, England)*, 30(14):2066–7.

[6] Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of mirna genes. *The Plant cell*, 23(2):431–442.

[7] Davies, R. T., Goetz, D. H., Lasswell, J., Anderson, M. N., and Bartel, B. (1999). IAR3 Encodes an Auxin Conjugate Hydrolase from Arabidopsis. 11(March):365–376.

[8] Debernardi, J. M., Rodriguez, R. E., Mecchia, M. A., and Palatnik, J. F. (2012). Functional Specialization of the Plant miR396 Regulatory Network through Distinct MicroRNA–Target Interactions. *PLoS Genet*, 8(1):e1002419.

[9] Fahlgren, N. and Carrington, J. (2010). mirna target prediction in plants. In Meyers, B. C. and Green, P. J., editors, *Plant MicroRNAs*, volume 592 of *Methods in Molecular Biology*, pages 51–57. Humana Press.

[10] Fahlgren, N., Jogdeo, S., Kasschau, K. D., Sullivan, C. M., Chapman, E. J., Laubinger, S., Smith, L. M., Dasenko, M., Givan, S. a., Weigel, D., and Carrington, J. C. (2010). MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. *The Plant cell*, 22(4):1074–89.

[11] German, M. a., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. a., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nature biotechnology*, 26(8):941–6.

[12] Giegerich, R., Rehmsmeier, M., Steffen, P., and Ho, M. (2004). Fast and effective prediction of microRNA / target duplexes. (2003):1507–1517.

[13] Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(Database issue):D1178–86.

[14] Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of plant micrornas and their targets, including a stress-induced mirna. *Molecular Cell*, 14(6):787 – 799.

[15] Kasschau, K. D., Xie, Z., Allen, E., Llave, C., Chapman, E. J., Krizan, K. a., and Carrington, J. C. (2003). P1/HC-Pro, a Viral Suppressor of RNA Silencing, Interferes with Arabidopsis Development and miRNA Function. *Developmental Cell*, 4(2):205–217.

[16] Krüger, J. and Rehmsmeier, M. (2006). Rnahybrid: microrna target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(suppl 2):W451–W454.

[17] Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. 297(September):2053–2056.

[18] Maunoury, N. (2011). AGO1 and AGO2 Act Redundantly in miR408-Mediated Plantacyanin Regulation. 6(12).

[19] Millar, A. a. and Gubler, F. (2005). The Arabidopsis GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. *The Plant cell*, 17(3):705–21.

[20] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.

[21] Niyogi, K. K., Pilon, M., Shikanai, T., Abdel-ghany, S. E., and Mu, P. (2005). Two P-Type ATPases Are Required for Copper Delivery in Arabidopsis thaliana Chloroplasts. 17(April):1233–1251.

[22] Palatnik, J. F., Wollmann, H., Schommer, C., Schwab, R., Boisbouvier, J., Rodriguez, R., Warthmann, N., Allen, E., Dezulian, T., Huson, D., Carrington, J. C., and Weigel, D. (2007). Sequence and expression differences underlie functional specialization of Arabidopsis microRNAs miR159 and miR319. *Developmental cell*, 13(1):115–25.

[23] Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes & development*, 20(24):3407–25.

[24] Rampey, R. A., Leclere, S., Kowalczyk, M., Ljung, K., Bartel, B., Biology, C., and Texas, R. A. R. (2004). A Family of Auxin-Conjugate Hydrolases That Contributes to Free Indole-3-Acetic Acid Levels during Arabidopsis Germination 1. 135(June):978–988.

[25] Rodriguez, R. E., Mecchia, M. A., Debernardi, J. M., Schommer, C., Weigel, D., and Palatnik, J. F. (2010). Control of cell proliferation in Arabidopsis thaliana by microRNA miR396. 112:103–112.

[26] Schiefthaler, Balasubramanian, Sieber, Chevalier, Wisman, and Schneitz (1999). Molecular analysis of NOZZLE , a gene involved in pattern formation and early sporogenesis during sex organ development in Arabidopsis thaliana. *Proc. Natl Acad. Sci.*, 96(September):11664–11669.

[27] Schwab, R., Palatnik, J. F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of micrornas on the plant transcriptome. *Developmental Cell*, 8(4):517 – 527.

[28] Vazquez, F., Gasciolli, V., Crété, P., and Vaucheret, H. (2004). The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. *Current biology : CB*, 14(4):346–51.

[29] Xie, Z., Kasschau, K. D., and Carrington, J. C. (2003). Negative Feedback Regulation of Dicer-Like1 in Arabidopsis by microRNA-Guided mRNA Degradation. *Current Biology*, 13(9):784–789.

[30] Yamasaki, H., Abdel-Ghany, S. E., Cohu, C. M., Kobayashi, Y., Shikanai, T., and Pilon, M. (2007). Regulation of copper homeostasis by micro-RNA in Arabidopsis. *The Journal of biological chemistry*, 282(22):16369–78.

[31] Yan, T., Yoo, D., Berardini, T. Z., Mueller, L. A., Weems, D. C., Weng, S., Cherry, J. M., and Rhee, S. Y. (2005). Patmatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Research*, 33(suppl 2):W262–W266.

[32] Yang, W.-c., Ye, D., Xu, J., and Sundaresan, V. (1999). The SPOROCYTELESS gene of Arabidopsis is required for initiation of sporogenesis and encodes a novel nuclear protein. pages 2108–2117.

Appendix A

Anexo

Predicción de genes regulados por microARNs.

For 64bit OS

edit $~/.bashrc file and add following lines

```
PATH=/usr/local/texlive/2011/bin/x86_64-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

```
1  #!/usr/bin/perl -w
2  use DBI;
3  use FindBin;
4  use strict;
5
6  # Constants
7  package Constants;
8  use constant MM => 4;
9  use constant MM_TYPE => 'is';
10 use constant DB => 'patmatch_2013';
11 use constant PLANTDB => 'phytozome';
12
13 # MYSql
14 my $host   = "localhost";
15 my $userid;
16 my $passwd;
17
18 # Arguments
19 my $pattern = $ARGV[0] || die "Must give pattern";
20 my $table_name = $ARGV[1] || die "Must give table name";
21
22 $pattern  = uc($pattern);
23 $pattern  =~ tr/uU/tT/;
24
25 my $mismatches = MM;
26 my $mismatch_types = MM_TYPE;
```

Table A.1 Especies y base de datos utilizadas para la búsquedas de genes blanco de miARNs conservados

Specie	Database
Allium_cepa	http://compbio.dfci.harvard.edu/tgi/
Aquilegia	http://compbio.dfci.harvard.edu/tgi/
Arabidopsis_thaliana	http://arabidopsis.org/
Beta_vulgaris	http://compbio.dfci.harvard.edu/tgi/
Brassica napus	http://compbio.dfci.harvard.edu/tgi/
Capsicum_annuum	http://compbio.dfci.harvard.edu/tgi/
Citrus_clementina	http://compbio.dfci.harvard.edu/tgi/
Citrus_sinensis	http://compbio.dfci.harvard.edu/tgi/
Coffea_canephora	http://compbio.dfci.harvard.edu/tgi/
Euphorbia_esula	http://compbio.dfci.harvard.edu/tgi/
Festuca_arundinacea	http://compbio.dfci.harvard.edu/tgi/
Glycine_max	http://compbio.dfci.harvard.edu/tgi/
Gossypium	http://compbio.dfci.harvard.edu/tgi/
Gossypium_raimondii	http://compbio.dfci.harvard.edu/tgi/
Haseolus_vulgaris	http://compbio.dfci.harvard.edu/tgi/
Helianthus_annuus	http://compbio.dfci.harvard.edu/tgi/
Hordeum_vulgare	http://compbio.dfci.harvard.edu/tgi/
Ipomoea_nil	http://compbio.dfci.harvard.edu/tgi/
Lactuca_sativa	http://compbio.dfci.harvard.edu/tgi/
Lactuca_serriola	http://compbio.dfci.harvard.edu/tgi/
Lotus_japonicus	http://compbio.dfci.harvard.edu/tgi/
Malus_x_domestica	http://compbio.dfci.harvard.edu/tgi/
Medicago_truncatula	http://compbio.dfci.harvard.edu/tgi/
Mesembryanthemum_crystallinum	http://compbio.dfci.harvard.edu/tgi/
Nicotiana_benthamiana	http://compbio.dfci.harvard.edu/tgi/
Nicotiana_tabacum	http://compbio.dfci.harvard.edu/tgi/
Oryza_sativa	http://www.jcvi.org/
Panicum_virgatum	http://compbio.dfci.harvard.edu/tgi/
Petunia_hybrida	http://compbio.dfci.harvard.edu/tgi/
Phaseolus_coccineus	http://compbio.dfci.harvard.edu/tgi/
Populus	http://compbio.dfci.harvard.edu/tgi/
Prunus_persica	http://compbio.dfci.harvard.edu/tgi/
Saccharum_officinarum	http://compbio.dfci.harvard.edu/tgi/
Secale_cereale	http://compbio.dfci.harvard.edu/tgi/
Solanum_lycopersicum	http://compbio.dfci.harvard.edu/tgi/
Solanum_tuberosum	http://compbio.dfci.harvard.edu/tgi/
Sorghum_bicolor	http://compbio.dfci.harvard.edu/tgi/
Theobroma_cacao	http://compbio.dfci.harvard.edu/tgi/
triphysaria	http://compbio.dfci.harvard.edu/tgi/
Triphysaria_versicolor	http://compbio.dfci.harvard.edu/tgi/
Triticum_aestivum	http://compbio.dfci.harvard.edu/tgi/
Vitis_vinifera	http://compbio.dfci.harvard.edu/tgi/
Zea_mays	http://compbio.dfci.harvard.edu/tgi/

Table A.2 My caption

microRNA	Target	ID
miR156/miR157	SPL	At1g27370
miR156/miR157	SPL	At1g53160
miR156/miR157	SPL	At2g33810
miR156/miR157	SPL	At3g15270
miR156/miR157	SPL	At5g43270
miR156/miR157	SPL	At1g69170
miR156/miR157	SPL	At2g42200
miR156/miR157	SPL	At3g57920
miR156/miR157	SPL	At5g50670
miR159/miR319	TCP	At1g30210
miR159/miR319	TCP	At1g53230
miR159/miR319	TCP	At2g31070
miR159/miR319	MYB	At3g11440
miR159/miR319	TCP	At3g15030
miR159/miR319	TCP	At4g18390
miR159/miR319	MYB	At5g06100
miR159/miR319	MYB	At2g26950
miR159/miR319	MYB	At2g32460
miR159/miR319	MYB	At5g55020
miR160	ARF	At1g77850
miR160	ARF	At2g28350
miR160	ARF	At4g30080
miR162	DCL	At1g01040
miR164	NAC	At1g56010
miR164	NAC	At3g15170
miR164	NAC	At5g07680
miR164	NAC	At5g53950
miR164	NAC	At5g61430
miR164	NAC	At3g12977
miR164	NAC	At5g39610
miR166/miR165	HD-ZIPIII	At1g30490
miR166/miR165	HD-ZIPIII	At1g52150
miR166/miR165	HD-ZIPIII	At2g34710
miR166/miR165	HD-ZIPIII	At5g60690
miR166/miR165	HD-ZIPIII	At4g32880
miR167	ARF	At1g30330
miR167	ARF	At5g37020
miR168	AGO	At1g48410
miR169	HAP2	At1g17590
miR169	HAP2	At1g54160
miR169	HAP2	At1g72830
miR169	HAP2	At3g05690
miR169	HAP2	At3g20910
miR169	HAP2	At5g06510
miR170/miR171	SCL	At2g45160
miR170/miR171	SCL	At3g60630
miR170/miR171	SCL	At4g00150
miR172	AP2	At2g28550
miR172	AP2	At4g36920
miR172	AP2	At5g60120
miR172	AP2	At5g67180
miR172	AP2	At2g39250
miR172	AP2	At3g54990
miR390/miR391	TAS3	At3g17185
miR390/miR391	TAS3	At5g49615
miR390/miR391	TAS3	At5g57735
miR393	TIR1/AFB	At1g12820
miR393	bHLH	At3g23690
miR393	TIR1/AFB	At3g26810
miR393	TIR1/AFB	At3g62980
miR393	TIR1/AFB	At4g03190
miR394	F-Box	At1g27340
miR395	APS	At3g22890
miR395	AST	At5g10180
miR395	APS	At5g43780
miR395	APS	At4g14680
miR396	GRF	At2g22840
miR396	GRF	At2g36400
miR396	GRF	At2g45480
miR396	GRF	At4g24150
miR396	GRF	At4g37740
miR396	GRF	At5g53660
miR396	GRF	At3g52910
miR397	LAC	At2g29130
miR397	LAC	At2g38080
miR397	LAC	At5g60020
miR398	CSD	At1g08830
miR398	CSD	At2g28190
miR398	CytC oxidase	At3g15640
miR399	E2-UBC	At2g33770
miR399	E2-UBC	At2g33770
miR408	LAC	At2g30210
mir408	PLC	At2g02850
miR827	SPX	At1g02860

Table A.3 Oligonucleotide primers used for RT-qPCR

Gene	Locus ID	Forward primer	Reverse Primer
PAA2	At5g21930	GTCCTCTTATCAGGGGACAGG	CATAGTTGCTTGTGCAAGACTCAG
MYB33	At5g06100	CTATGGAAACCGACATTCACCTG	CTTGGCTTCCAGAAGCAACATATCG
NZZ	At4g27330	TCGGGTCAGGTTATGATCGA	AGGGTTTCCTTCCATGTAGCTCC
PP2A	At1g13320	CCTGCGGTAATAACTGCATCT	CTTCACTTAGCTCCACCAAGCA
tMT2A	tobacco	TACCCAGATTTGAGCTACAACGAG	GCAGGAGATTCACCCATTTCCATA
tMT2B	tobacco	TACCCAGATTTGAGCTACAACGAA	AGGGGATTCACCCATTTCCATT

Table A.4 Oligonucleotide primers used for 5' RACE

Gen	Locus ID	5' RACE	5' RACE nested
General		CGACTGGAGCACGAGGACACTGA	GGACACTGACATGGACTGAAGGAGTA
PAA2	At5g21930	GACTTATGGAGCTGCAGAAGTAATG	CATAGTTGCTTGTGCAAGACTCAG
IAR3	At1g51760	ATCTTCTGATCCCATTAATGGTTGCATCTCG	CATATTCACGCTCGCTTGCCTTGTGATAACC
NZZ	At4g27330	CATTTAAAGCTTCAAGGACAAATCAATGGTATTAGG	AGGGTTTCCTTCCATGTAGCTCC
MMG4.7	At5g43060	ATGGTAACAACCTTAGCATTTTTCC	CTTCGGTATCAATACCWCCATT
UDP	At2g47650	AATGGGCCGACATGTTCTCC	CCTCGGTGATAGTCCATGGT
SVP	At2g22540	GCAACTTTCCTTCATTCATC	TTTCATCTGCCTCAGCTCAC
loricrin-related	AT5g64550	ACCATGAGCTTTGCAGTAGT	CCTCAGCACTTCGTGTACAG
	At3g14110	CGGAAGGATCAGTCAGTCTC	CCCAGCTCGGTATAACAGTC
	At3g22110	GTTTCATCGCCAAAGGTAAC	CCAGGCGAATAAGACTAGAG
AVA-P2	At1g19910	CTCTAGACTGACCAGCTCGA	GGATGATACCAACAATGAGA

```
27
28  # Check pattern
29  my $SYNTAX_CHECKER_BIN = "perl " . $FindBin::Bin . '/patmatchPatternChecker.pl';
30  my $patStatus = `$SYNTAX_CHECKER_BIN 'dna' $pattern `;
31  chomp($patStatus);
32
33  # Variables
34  my ($gen_name,$hit_start,$hit_end,$target,$mirna) ='';
35  my $deltaG = 0;
36
37  # Files RNAhybrid target and microrna 5'3' y blast
38  my $target_file = $FindBin::Bin . "/extra_files/target_rnahybrid.txt";
39  my $mirna_file = $FindBin::Bin . "/extra_files/mirna_rnahybrid.txt";
40  my $blast_file_sequence = $FindBin::Bin . "/extra_files/" . "seq_blast.txt";
41  my $blast_database  = $FindBin::Bin . "/extra_files/blast/TAIR10_pep_20101214_updated";
42
43  my $sequence_file = '';
44  my $tab = $table_name . "_" .  $pattern;
45
46  hybrid_mirna_file($pattern,$mirna_file);
47  fill_table_mirnas($pattern,$table_name,$tab);
48
49  my $tab_db = create_table($tab);
50  my @specie_db = species(PLANTDB);
51
52  foreach my $file (@specie_db){
53
54    my $fasta_file = $file->{'fasta'};
55    my $specie = $file->{'specie'};
56
57    $sequence_file = $FindBin::Bin . '/databases/' . PLANTDB . "/" . $fasta_file;
58    print $fasta_file . " - ";
59    print $specie . "\n";
60
61    my ($gen_sv,$target_sv,$align_sv,$miR_sv,$deltaG_sv,$nro_mm_sv,$ins_sv) = '';
62    my ($del_sv,$sust_mm_sv,$gu_mm_sv) = '';
63    my ($filtro_mm,$family,$sub_family,$alias) = '';
64
65    my %res_blast;
66    my @res_ned,@res_mm;
67    my @res_family;
68
69    if ($patStatus eq "OK") { # syntax OK, run PatMatch
70      my $SCAN_PIPELINE = "perl " . $FindBin::Bin . '/scan_pipeline.pl';
71      open(OUTPUT, "$SCAN_PIPELINE -c '$pattern' '$sequence_file' '$mismatches' '$mismatch_types' |");
72      open(INFO,$sequence_file);
73      while (my $line = <OUTPUT>) {
74        if ( $line =~ />(.*?):\[(\d*?),(\d*?)\]/ ) {
75          #keep output parameters
```

```
76            $gen_name  = $1;
77            $hit_start = $2;
78            $hit_end   = $3;
79            $gen_sv    = $gen_name;
80          }
81          elsif ($line =~ /(.*)(\d*?)/ ) {
82            $target = trim($1);
83            $mirna = $pattern;
84            $mirna = reverse complement($mirna);
85            # Needleman
86            @res_ned = Needleman($mirna,$target);
87            $target_sv = $res_ned[0];
88            $align_sv  = $res_ned[1];
89            $miR_sv  = $res_ned[2];
90            # Mismatches
91            @res_mm = mismatches($target_sv,$align_sv,$miR_sv);
92            $nro_mm_sv  = trim($res_mm[0]);
93            $ins_sv  = trim($res_mm[1]);
94            $del_sv  = trim($res_mm[2]);
95            $sust_mm_sv = trim($res_mm[3]);
96            $gu_mm_sv   = trim($res_mm[4]);
97
98            hybrid_target_file($target,$target_file);
99
100           #RNAhybrid
101           my $RNAhybrid_BIN = 'RNAhybrid -d 0 -m 12000';
102           my $RNAhybrid_Status = `$RNAhybrid_BIN '-t' $target_file '-q' $mirna_file`;
103           chomp($RNAhybrid_Status);
104           if ( $RNAhybrid_Status =~ m/mfe: (.*)kcal\/mol/ ){
105             $deltaG = $1;
106             $deltaG =~ s/ //g;
107             $deltaG_sv = $deltaG;
108           }
109
110           $filtro_mm = mm_position($align_sv);
111           insert($tab_db,$specie,$gen_sv,$target_sv,$align_sv,$miR_sv,$nro_mm_sv,$ins_sv,$del_sv,$sust_mm_sv,$gu_mm_sv,$filtro_mm,$deltaG_sv);
112         }
113       }
114       close(OUTPUT);
115       close(INFO);
116       blast_from_db_annotation($fasta_file,$specie,$tab_db);
117       family_from_db($fasta_file,$tab_db);
118     }
119
120     else
121     {
122       print "Invalid pattern syntax";
123     }
124   }
125
126   sub fill_table_mirnas{
127     my ($sequence,$name,$table_reference) = @_;
128     $target = reverse(complement($sequence));
129     hybrid_target_file($target,$target_file);
130     my $hyb_perf;
131
132     #RNAhybrid
133     my $RNAhybrid_BIN = 'RNAhybrid -d 0 -m 12000';
134     my $RNAhybrid_Status = `$RNAhybrid_BIN '-t' $target_file '-q' $mirna_file`;
135     chomp($RNAhybrid_Status);
136     if ( $RNAhybrid_Status =~ m/mfe: (.*)kcal\/mol/ ){
137       $deltaG = $1;
138       $deltaG =~ s/ //g;
139       $hyb_perf = $deltaG;
140     }
141
142     my $sth;
143     my $db = DB ;
144     my $connectionInfo = "dbi:mysql:$db;$host";
145     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
146
147     my $query_insert = "INSERT INTO mirnas
148       (name,sequence,table_reference,hyb_perf)
149       values
150       ('$name',
```

```
151         '$sequence' ,
152         '$table_reference' ,
153         '$hyb_perf'
154        );";
155
156     $sth = $dbh->prepare($query_insert);
157     $sth->execute();
158
159   }
160
161   sub species{
162     (my $db_plants) = @_;
163
164     my $db = DB;
165     my $connectionInfo="dbi:mysql:$db;$host";
166     my $sth;
167     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
168
169     my $specie;
170     my $fasta_file;
171     my $specie_data = {};
172
173     my $ref;
174     my @res;
175
176     my $query = "SELECT fasta,specie from plants where db = '$db_plants'";
177
178     $sth = $dbh->prepare($query);
179     $sth->execute();
180     $sth->bind_columns(\$fasta_file,\$specie);
181
182     if ($sth->rows > 0 ){
183       while($ref = $sth->fetchrow_hashref() ) {
184         push(@res, $ref);
185       }
186     }
187
188     return @res;
189   }
190
191   sub insert{
192     my ($table_db,$especie,$gen,$target,$align,$mirna,$mm,$ins,$del,$sust, $gu,$filtro_mm,$deltag) = @_;
193     my $sth;
194     my $db = DB ;
195     my $connectionInfo = "dbi:mysql:$db;$host";
196     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
197
198     my $query_insert = "INSERT INTO $table_db
199           (file,gen,target,align,mirna,mm,ins,del,sust,gu,filtro_mm,deltag)
200           values
201           ('$especie','$gen','$target','$align','$mirna','$mm','$ins','$del','$sust','$gu','$filtro_mm','$deltag');";
202
203     $sth = $dbh->prepare($query_insert);
204     $sth->execute();
205
206   }
207   sub create_table{
208
209     my ($tabmiRNA) = @_;
210     my $sth;
211     my $db = DB;
212     my $connectionInfo = "dbi:mysql:$db;$host";
213     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
214
215     my $query_create = "CREATE TABLE $tabmiRNA (id INT AUTO_INCREMENT PRIMARY KEY,
216               file VARCHAR(60) NOT NULL,
217               gen VARCHAR(30) NOT NULL,
218               target  VARCHAR(30) NOT NULL,
219               align  VARCHAR(30) NOT NULL,
220               mirna  VARCHAR(30) NOT NULL,
221              mm INT NOT NULL,
222               ins INT NOT NULL,
223               del INT NOT NULL,
224               sust INT NOT NULL,
225               gu INT NOT NULL,
```

```
226                 similar_ath VARCHAR(20) NOT NULL,
227                 similar_osa VARCHAR(20) NOT NULL,
228                 filtro_mm INT NOT NULL,
229                 family TEXT(1000) NOT NULL,
230                 sub_family VARCHAR(20) NOT NULL,
231                 alias VARCHAR(20) NOT NULL,
232                 deltag FLOAT NOT NULL
233               );";
234
235     $sth = $dbh->prepare($query_create);
236     $sth->execute();
237     $sth->finish();
238     $dbh->disconnect;
239     return $tabmiRNA;
240   }
241
242
243   sub blast_from_db_annotation{
244     my ($file, $specie, $mirna) = @_;
245
246     my $db= DB;
247     my $connectionInfo="dbi:mysql:$db;$host";
248     my $sth;
249     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
250     my $annotation_table = 'annotation_' . $file;
251     my $annotation;
252     my $query = "SELECT annotation from plants where fasta = '$file'";
253
254     $sth = $dbh->prepare($query);
255     $sth->execute();
256     $sth->bind_columns(\$annotation);
257
258
259     $sth->fetch();
260
261     unless($annotation eq 'empty') {
262       my $query_update = "UPDATE $mirna miR
263                 LEFT JOIN $annotation_table a
264                 ON  miR.gen = a.gen
265                 SET miR.similar_ath = SUBSTRING(a.similar_ath,1,9) ,
266                   miR.similar_osa = SUBSTRING(a.similar_osa,1,14)
267                 WHERE file = '$specie'";
268
269       $sth = $dbh->prepare($query_update);
270       $sth->execute();
271     }
272   }
273
274
275   sub family{
276     (my $gen) = @_;
277
278     my $db = DB;
279     my $connectionInfo="dbi:mysql:$db;$host";
280     my $sth;
281     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
282     my $family;
283     my $sub_family;
284     my $alias;
285     my $query = "SELECT family,sub_family,gene_name from gene_families where locus_tag = '$gen';";
286
287     $sth = $dbh->prepare($query);
288     $sth->execute();
289     $sth->bind_columns(\$family,\$sub_family,\$alias);
290
291     if ($sth->rows > 0 ){
292       while($sth->fetch()) {
293         my @results = ($family,$sub_family,$alias);
294         return @results;
295       }
296     }
297     else{
298       my @results = ('','','');
299       return @results;
300     }
```

```
301 }
302
303 sub family_from_db{
304   (my $file , $mirna) = @_;
305
306   my $db = DB;
307   my $connectionInfo="dbi:mysql:$db;$host";
308   my $sth;
309   my $dbh = DBI->connect($connectionInfo ,$userid ,$passwd);
310   my $gen;
311
312   my $query_update = "UPDATE $mirna miR
313             LEFT JOIN gene_families f
314             ON  miR.similar_ath = f.locus_tag
315             SET miR.family     = f.family ,
316               miR.sub_family  = f.sub_family ,
317               miR.alias      = f.gene_name" ;
318
319   $sth = $dbh->prepare($query_update);
320   $sth->execute();
321 }
322
323 # Save the miRNA for RNAhybird
324 sub hybrid_mirna_file{
325   my ($mirna ,$file_mirna) = @_;
326   open(MIRNA, ">$file_mirna") || die;
327   print MIRNA ">mirna\n$mirna\n";
328   close(MIRNA);
329 }
330 # Save the target for RNAhybird
331 sub hybrid_target_file {
332   my ($target ,$file_target) = @_;
333   open(OUT, ">$file_target") || die;
334   print OUT ">target\n$target\n";
335   close(OUT);
336 }
337
338 sub mm_position{
339   (my $align) = @_;
340   my @align = split (//,$align);
341   my $i = 0;
342   my $j = 0;
343   @align = reverse(@align);
344   foreach (@align) {
345     if ($align[$i] eq "*"){
346       if ($i+2 < 13 ){
347         $j++;
348       }
349     }
350     $i++;
351   }
352   if ($j > 1 ){
353     return 0;
354   }
355   else{
356     return 1;
357   }
358
359 }
360
361 sub mismatches{
362   my ($target_mm ,$align_mm ,$mirna_mm) = @_;
363   my $pos  = 0;
364   my $nro_mm = 0;
365   my $del_mm = 0;
366   my $ins_mm = 0;
367   my $sust_mm = 0;
368   my $gu_mm = 0;
369   my @align_mm = split(//, $align_mm);
370   my @target_mm = split(//, $target_mm);
371   my @mirna_mm = split(//, $mirna_mm);
372   # mismatches
373   foreach (@align_mm) {
374     if  ($_ =~ /\*/) {
375       if (($target_mm[$pos] eq 'G' and $mirna_mm[$pos] eq 'T') or ($target_mm[$pos] eq 'T' and $mirna_mm[$pos] eq 'G')){
```

```
376             $gu_mm++;
377           }
378           else{}
379           $nro_mm ++;
380         }
381         $pos++;
382       }
383       # deletions
384       foreach (@target_mm) {
385         if  ($_ =~ /-/) {
386             $del_mm ++;
387         }
388       }
389       # insertions
390       foreach (@mirna_mm) {
391         if  ($_ =~ /-/) {
392             $ins_mm ++;
393         }
394
395       }
396       # sustitutions
397       $sust_mm = $nro_mm - ($del_mm + $ins_mm + $gu_mm);
398
399       my @results = ($nro_mm,$ins_mm,$del_mm,$sust_mm,$gu_mm);
400       return @results;
401
402   }
403
404   sub Needleman {
405     my ($seq1,$seq2) = @_;
406     # scoring scheme
407     my $MATCH =  1; # +1 for letters that match
408     my $MISMATCH = -1; # -1 for letters that mismatch
409     my $GAP   = -1; # -1 for any gap
410
411     # initialization
412     my @matrix;
413     $matrix[0][0]{score}   = 0;
414     $matrix[0][0]{pointer} = "none";
415
416     for(my $j = 1; $j <= length($seq1); $j++) {
417       $matrix[0][$j]{score}   = $GAP * $j ;
418       $matrix[0][$j]{pointer} = "left";
419     }
420     for (my $i = 1; $i <= length($seq2); $i++) {
421       $matrix[$i][0]{score}   = $GAP * $i ;
422       $matrix[$i][0]{pointer} = "up";
423     }
424
425     # fill
426     for(my $i = 1; $i <= length($seq2); $i++) {
427       for(my $j = 1; $j <= length($seq1); $j++) {
428         my ($diagonal_score, $left_score, $up_score);
429
430         # calculate match score
431         my $letter1 = substr($seq1, $j-1, 1);
432         my $letter2 = substr($seq2, $i-1, 1);
433         if ($letter1 eq $letter2) {
434           $diagonal_score = $matrix[$i-1][$j-1]{score} + $MATCH;
435         }
436         else {
437           $diagonal_score = $matrix[$i-1][$j-1]{score} + $MISMATCH;
438         }
439
440         # calculate gap scores
441         $up_score   = $matrix[$i-1][$j]{score} + $GAP;
442         $left_score = $matrix[$i][$j-1]{score} + ($GAP*10);
443
444         # choose best score
445         if ($diagonal_score >= $up_score) {
446           if ($diagonal_score >= $left_score) {
447             $matrix[$i][$j]{score}   = $diagonal_score;
448             $matrix[$i][$j]{pointer} = "diagonal";
449           }
450           else {
```

```
451             $matrix[$i][$j]{score}   = $left_score;
452             $matrix[$i][$j]{pointer} = "left";
453                   }
454       }
455       else {
456           if ($up_score >= $left_score) {
457                       $matrix[$i][$j]{score}   = $up_score;
458                       $matrix[$i][$j]{pointer} = "up";
459                   }
460                   else {
461                       $matrix[$i][$j]{score}   = $left_score;
462                       $matrix[$i][$j]{pointer} = "left";
463                   }
464       }
465     }
466   }
467
468   # trace-back
469
470   my $align1 = "";
471   my $align2 = "";
472   my $align3;
473
474   # start at last cell of matrix
475   my $j = length($seq1);
476   my $i = length($seq2);
477
478   while (1) {
479     last if $matrix[$i][$j]{pointer} eq "none"; # ends at first cell of matrix
480     my $letter1 = substr($seq1, $j-1, 1);
481       my $letter2 = substr($seq2, $i-1, 1);
482
483     if ($matrix[$i][$j]{pointer} eq "diagonal") {
484             if ($letter1 eq $letter2){
485                 $align3 .= "|";
486             }
487             else
488             {
489                 $align3 .= "*";
490             }
491       $align1 .= substr($seq1, $j-1, 1);
492       $align2 .= substr($seq2, $i-1, 1);
493       $i--;
494       $j--;
495
496
497     }
498     #no tendria que entrar aca, salvo si esta al final
499     elsif ($matrix[$i][$j]{pointer} eq "left") {
500       $align1 .= substr($seq1, $j-1, 1);
501       $align2 .= "-";
502             $align3 .= "*";
503       $j--;
504
505     }
506     elsif ($matrix[$i][$j]{pointer} eq "up") {
507       $align1 .= "-";
508       $align2 .= substr($seq2, $i-1, 1);
509             $align3 .= "*";
510       $i--;
511
512     }
513   }
514
515   $align1 = reverse $align1;
516   $align2 = reverse $align2;
517   $align3 = reverse $align3;
518   $align1 = complement($align1);
519
520   my @results = ($align2,$align3,$align1);
521   return @results;
522 }
523 sub complement
524 {
525   my $sequence = shift;
```

```
526     $sequence =~ tr/atcgATCG/tagcTAGC/;
527     return $sequence;
528 }
529
530 sub trim
531 {
532     (my $string) = @_;
533     $string =~ s/^\s+//;
534     $string =~ s/\s+$//;
535     $string =~ s/\n//g;
536     return $string;
537 }
```