CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

**UNIDAD IRAPUATO**

# Análisis evolutivo de la longitud de proteínas y desarrollo de un método independiente de homología para la clasificación funcional de proteínas ORFans

**Tesis que presenta**

**Obed Ramírez Sánchez**

**Para Obtener el Grado de**

**Doctor en Ciencias**

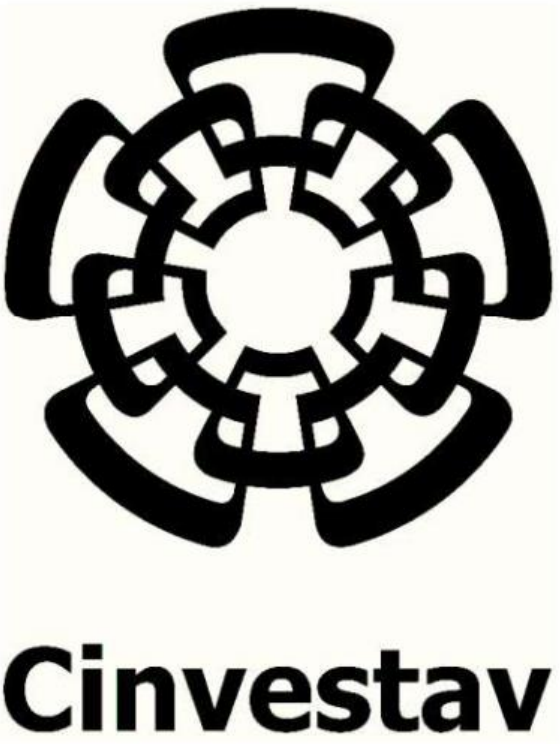
**En la Especialidad de**

**Biotecnología de plantas**

**Directores de Tesis: Axel Tiessen Favier / Luis Delaye Arredondo**

**Irapuato, Gto. México**

**Agosto, 2017**




El presente trabajo de investigación se desarrolló en los laboratorios de Genómica Evolutiva y Metabolómica y Fisiología Molecular, del departamento de Ingeniería Genética, ambos pertenecientes al CINVESTAV-IPN, Unidad de Biotecnología e Ingeniería Genética de Plantas. Bajo la dirección del Dr. Luis José Delaye Arredondo y del Dr. Axel Tiessen Favier.

**AGRADECIMIENTOS**

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico otorgado para la realización de este proyecto con la Beca No. 237183 otorgada durante el periodo comprendido de Septiembre del 2012 a Agosto del 2016.

Al Dr. Luis José Delaye Arredondo, Dr. Axel Tiessen Favier y Dr. Paulino Pérez Rodríguez, por su asesoría, impulso y paciencia, así como el espacio y materiales en sus respectivos laboratorios.

A la Dra. Gabriela Olmedo, Dra. Selene Fernandez Valverde, Dr. Cei Abreu-Goodger, Dr. Mauricio Carrillo por su asesoría y opinión crítico para la realización de este trabajo de investigación.

A todo el personal de CINVESTAV unidad Irapuato por su apoyo en las diferentes áreas, administrativas, secretariales, de mantenimiento y servicios.

A Areli, mi esposa y mejor amiga, gracias por tu apoyo, motivación y amor incondicionales.

A mis amig@s: Adrián, Alma, Cesi, Edgar, Issi, Josh, El Jules, Fabián, Mari, Maye, Medina, Pepe, Orlando, Sandris, Shey, Shini, Sofi, Vivis y Yisus. Por todo su apoyo, paciencia, cafés, consejos y (además) trabajo en equipo, durante esta aventura del doctorado.

**DEDICATORIA**

A mis padres, Moisés y Rosalba, que me inculcaron con su ejemplo
el valor del trabajo y la perseverancia.

*-La tierra es redonda como una naranja [les*
*reveló José Arcadio Buendía]*.
*Úrsula perdió la paciencia.*
*«Si has de volverte*
*loco, vuélvete tú solo -gritó-.*
*Pero no trates de*
*inculcar a los niños tus ideas de gitano.»*
Cien Años de Soledad. Gabriel García Márquez

# Índice de contenido

# Índice de tablas

## Índice de figuras

# Resumen

La presente tesis está dividida en dos proyectos independientes. 1) El primer proyecto se centró en el estudio de la evolución de la longitud de proteínas en eucariontes, en plástidos y en endosimbiontes. Algunas de las preguntas que guiaron este proyecto fueron: ¿Hay relación entre la estructura de exones y la longitud de las proteínas? ¿Cuál fue el impacto en la longitud de proteínas de plantas después de la migración de genes desde cloroplasto hacia el núcleo? En organismos endosimbiontes, ¿existe un proceso de reducción de longitud en sus proteínas, paralelo al de la reducción del genoma? Con respecto a organismos eucariontes, encontramos diferencias significativas entre el número de genes codificados en sus genomas y en el tamaño promedio de las proteínas. También encontramos que cada grupo filogenético utiliza combinaciones diferentes de número y longitud de exones para codificar sus genes. En particular, las plantas presentan menor longitud promedio de proteínas (363 aminoácidos (aa)) en comparación con animales (439 aa) y sus genes están codificados por un menor número de exones: 5.7 y 10.1. Con respecto a los plástidos, encontramos que la longitud promedio de proteínas en cientos de proteomas de plástidos se encuentran reducida (160 aa) en comparación a sus ancestros las cianobacterias (257 aa). Esto sugiere que la vida intracelular no solo causó una pérdida de ADN de los plástidos (reducción genómica) sino también una reducción en la longitud promedio de las proteínas. De manera opuesta, hubo un incremento de longitud en los genes de origen procarionte que migraron desde el cloroplasto hacia el núcleo eucarionte de las plantas primitivas después de la endosimbiosis. Otro hallazgo interesante fue el descubrir que el proceso de reducción de las proteínas de los endosimbiontes consiste en dos fenómenos opuestos: en una primera etapa, las proteínas grandes tienden a permanecer en el genoma, mientras que en una segunda etapa, dichas proteínas reducen su longitud conforme progresa la reducción del genoma. 2) El segundo proyecto tuvo como objetivo el desarrollo de un método independiente de homología para predecir diversas funciones de las proteínas. El método propuesto se basa en un ensamble de Máquinas de Soporte Vectorial (SVM), en el que cada clasificador SVM fue construido utilizando un alfabeto reducido diferente. En comparación a otros métodos recientes, el método propuesto logró

mejores resultados en 4 de los 6 conjuntos de datos en que fue evaluado. Con nuestro método se logró predecir la función de 1,786 proteínas huérfanas o ORFans (que no cuentan con homólogos en especies cercanas) de *A. thaliana*. El servidor web se encuentra implementado internamente en la institución en la siguiente dirección web: ira.cinvestav.svmensemble.com y está disponible para su uso por la comunidad.

# Abstract

This thesis is divided in two independent projects. 1) The first project focused on protein length evolution in eukaryotes, plastids and endosymbionts. Some of the following questions guided our work: Is there a relationship between protein length and exon structure? What was the impact on protein length of bacterial gene migration from the chloroplast to the nucleus? In endosymbiont organisms, is there a process of protein length reduction parallel to the genome reduction process? In eukaryotic organisms, we found significant differences between the number of genes by genome and their average protein length. We also found that each phylogenetic group uses different combination of number and length of exons to code their genes. On average, plant proteins are smaller (363 amino acids (aa)) than animal proteins (439 aa) and their genes are coded by less exons: 5.7 and 10.1. In plastids, we found that the average protein length in hundreds of plastid proteomes is reduced (160 aa) in comparison to cyanobacterial proteins (257 aa). This suggests that the intracellular lifestyle not only leads to DNA loss (genomic reduction) but is also accompanied by a reduction of the average protein size. In opposite way, there was an increment in the length of prokaryotic genes that migrated from the chloroplast to the eukaryotic nucleus in primitive plants after endosymbiosis. Another interesting finding was that the reduction of endosymbiont proteins is given by two opposed phenomena: In a first stage, large proteins tend to stay in the genome of the endosymbiont. In a second stage, the proteins that remain in the endosymbiont tend to reduce their size as genomic reduction progresses. 2) The second project consisted in the development of a homology independent method for protein function prediction. The proposed method is based on an ensemble of SVM classifiers; each SVM classifier was constructed using a different reduced amino acid alphabet. In comparison with other recent approaches, the optimized method achieved better predictions in four of six datasets where it was evaluated. Using our method we predicted the function of 1786 orphan proteins (without homologues in close species) from *A. thaliana*. The SVM classifier is available at: http://ira.cinvestav.svmensemble.com.

# Introducción

## Evolución de longitud de proteínas en eucariontes y en simbiontes

La longitud de las proteínas determina la manera en que éstas interactúan con su ambiente y en consecuencia tiene un profundo impacto en su función (Lin and Zewail, 2012; Xu and Nussinov, 1998). En términos generales, las proteínas con mayor longitud (>400 aa) pueden acomodar múltiples dominios (Ekman, et al., 2005; Petsko, 2004) lo que les permite interactuar en redes de regulación complejas (Brocchieri and Karlin, 2005; Zhang, 2000). Por ejemplo, el análisis de la red de interacción de proteína-proteína de *Saccharomyces cerevisiae* mostró que aquellas proteínas que participan en mayor número de interacciones, conocidas como "hubs", se caracterizan por ser multi dominio y tener una mayor longitud (Ekman, et al., 2006).

Los tres dominios de la vida muestran proteínas de distinta longitud. Las proteínas en eucariontes son más largas que las de las bacterias, y éstas a su vez son más largas que las de las arqueas (Brocchieri and Karlin, 2005; Tiessen, et al., 2012; Zhang, 2000). En eucariontes, el incremento en longitud se atribuye a un proceso gradual de adición de dominios mediante reacomodos en el genoma tales como la fusión de genes (Brocchieri and Karlin, 2005; Zhang, 2000). Dicho proceso ha sido interpretado como una estrategia evolutiva que permitió a los eucariontes desarrollar redes de regulación de mayor complejidad en comparación con las de las bacterias y las arqueas (Brocchieri and Karlin, 2005; Zhang, 2000).

Trabajos pioneros en el área, después de analizar un limitado conjunto proteomas, reportaron que la longitud de las proteínas estaba conservada entre distintas especies de eucariontes (Wang, et al., 2005; Xu, et al., 2006). Sin embargo, dichos análisis se basaron en comparar grupos de ortólogos, de los cuales se sabe que se encuentran altamente conservados en secuencia y longitud (Wang, et al., 2005). En un trabajo posterior, sin utilizar grupos de ortólogos, (Tiessen, et al., 2012) reportó diferencias significativas en la longitud de las proteínas de distintas especies tras analizar 140 proteomas de eucariontes y 1,302 proteomas de procariontes.

Distintos procesos evolutivos han contribuido a moldear la longitud de las proteínas en los organismos eucariontes: 1) endosimbiosis y la consecuente migración de genes bacterianos hacia el núcleo del hospedero eucarionte (Martin, et al., 2002); 2) duplicación de genes o de genomas (poliploidía) (Adams and Wendel, 2005); 3) reducción genómica y pérdida selectiva de genes (Kelkar and Ochman, 2013); 4) fusión de genes (Brocchieri and Karlin, 2005; Long, et al., 2003); 5) incorporación nuevos de genes por transferencia horizontal (Yue, et al., 2012); 6) ganancia o pérdida de exones (Coulombe-Huntington and Majewski, 2007 ); 7) genes surgidos *de-novo* a partir de secuencias previamente no codificantes (McCarrey and Thomas, 1987); 8) retrotrasposones que se integraron en genes (McCarrey and Thomas, 1987) y 9) genes que se dividieron en dos o más genes más pequeños por inserciones de trasposones (Moore, et al., 2008).

Por lo anterior, se entiende que el tamaño actual de las proteínas es el resultado de la interacción entre dichos mecanismos evolutivos. A continuación se discuten brevemente dos de ellos: a) *Duplicación de genes*. En eucariontes la duplicación de genes es considerado el principal mecanismo para generar nuevos genes. La duplicación de genes y del ADN en teoría no debería cambiar el tamaño promedio de las proteínas. Sin embargo, aunque cualquier gen puede sufrir un proceso de duplicación, se ha observado que las proteínas de mayor longitud se conservan en el genoma con mayor frecuencia que las proteínas con menor tamaño (He and Zhang, 2005). Presumiblemente esto se debe a que los múltiples dominios que contienen las hacen más propensos a procesos de subfuncionalización y neofuncionalización (He and Zhang, 2005). b) *Fusión de genes*. La fusión de genes adyacentes origina nuevas proteínas con mayor longitud (Brocchieri and Karlin, 2005). Si este proceso predomina, es posible que se generen genomas con proteínas de mayor longitud pero con menor número de genes. Recientemente, (Tiessen, et al., 2012) reportó una correlación negativa y significativa (R=-0.39) entre el número de genes y la longitud promedio de proteínas en eucariontes. En procariontes por contrario, la correlación entre número de genes y tamaño de proteínas fue positiva (Tiessen, et al., 2012). Esto sugiere que la evolución del número y tamaño de proteínas es diferente entre eucariontes y procariontes, posiblemente debido a la estructura de exones-intrones y por los mecanismos de splicing y edición de ARN, que son diferentes entre organismos con o sin núcleo.

Ciertos procesos evolutivos pueden actuar al mismo tiempo en direcciones opuestas. Por ejemplo, mientras que la división de genes (Moore, et al., 2008) incrementa el número de genes al mismo tiempo que los reduce de tamaño, la fusión de genes por su parte reduce el número de genes e incrementa su tamaño. El incremento o disminución promedio del tamaño de las proteínas dependerá de la frecuencia de uno u otro fenómeno.

Los genomas de procariontes son mucho más compactos en comparación a los de los eucariontes debido a presiones selectivas (Wang, et al., 2007). Los genomas compactos son propios de los organismos unicelulares, particularmente en estilos de vida intracelular: parasitarios (Wang, et al., 2007) y endosimbiontes (Wernegreen, 2012). Cuando dos organismos establecen una endosimbiosis obligada durante largo tiempo (millones de años), el tamaño del genoma del endosimbionte se reduce progresivamente, y la tasa de mutación aumenta considerablemente ocasionando que gran cantidad de sus genes pierdan funcionalidad y sean removidos del genoma durante el proceso (Wernegreen, 2012).

En comparación con proteínas homólogas de bacterias de vida libre, se ha observado una menor longitud de proteína en simbiontes con genomas extremadamente reducidos (como la bacteria *C. ruddii*, el Microsporidia *E. cuniculi*, y los nucleomorfos *G. theta* y *H. andersinii*) (Katinka, et al., 2001; Lane, et al., 2007; Nakabachi, et al., 2006).

**Predicción de función de proteínas huérfanas en *Arabidopsis thaliana* utilizando ensambles de SVMs**

En todos los genomas secuenciados a la fecha existe un porcentaje de genes que no tienen homólogos en otras especies, por ejemplo, el gen Ah24 restringido al género Amaranthus (Massange-Sanchez, et al., 2015). A estos genes, que se encuentran restringidos a una especie en particular (o a un clado de especies filogenéticamente cercanas) se les conoce como genes huérfanos u ORFans (Fischer and Eisenberg, 1999). Aunque inicialmente se les consideró como un artefacto producto de la limitada cantidad de genomas secuenciados en ese momento, actualmente se les considera un fenómeno real, se han encontrado genes ORFans en los tres dominios de la vida y en los virus (Cai, et al., 2006; Guo, et al., 2007; Schmid and Aquadro, 2001; Wilson, et al., 2007; Yin and Fischer, 2008). Se estima que los genes ORFans constituyen entre 5 y 15% de los genomas secuenciados a la fecha (Kuo and Kissinger, 2008; Ohm, et al., 2012; Wissler, et al., 2013).

Recientemente se ha encontrado que los genes ORFans se pueden originar de distintas formas: i) genes duplicados que han divergido a tal grado que sus parálogos son irreconocibles; ii) sobrelapamiento (overprinting); iii) domesticación (o exaptación) de retrotrasposones; iv) pseudogenes que han resucitado; v) y generación *de novo* a partir de secuencias no codificantes (Brosch, et al., 2011; Donoghue, et al., 2011; Palmieri, et al., 2014; Wissler, et al., 2013).

Funcionalmente, la mayoría de los genes ORFans que han sido estudiados participan tanto en respuestas a estrés abiótico (principalmente oxidativo y osmótico) como en interacciones bióticas como defensa y señalización (Arendsee, et al., 2014; Carvunis, et al., 2012; Colbourne, et al., 2011; Donoghue, et al., 2011; Guo, et al., 2007). Se ha propuesto que los genes ORFans constituyen una fuente de adaptación y generación de diversidad en los seres vivos (Wilson, et al., 2005).

Debido a que los genes ORFans no presentan plegamientos, motivos ni dominios conocidos, se desconoce la función de la gran mayoría de estos (Arendsee, et al., 2014). La falta de homología con otros genes implica que el uso de cualquier herramienta bioinformática basada en homología resulte inviable para su caracterización, por lo que la caracterización funcional de los genes ORFans representa un reto importante tanto para la bilogía experimental como para la bioinformática.

En años recientes, se han propuesto diversos métodos estadísticos independientes de homología provenientes del campo de aprendizaje de máquinas, también llamado aprendizaje estadístico (Hastie, et al., 2009). El aprendizaje estadístico es una disciplina madura con bases teóricas sólidas que ha sido aplicado en numerosas áreas de investigación incluyendo la bioinformática (Libbrecht and Noble, 2015). Algunos ejemplos de métodos de aprendizaje estadístico incluyen: regresión logística, máquinas de soporte vectorial (SVM por sus siglas en inglés), bosques aleatorios (random forest), redes neuronales, método del vecino más cercano (knn) y análisis discriminante (lda/qda), entre otros (Hastie, et al., 2009). Con respecto al método de SVM, algunas aplicaciones en la predicción funcional de proteínas incluyen: interacciones proteína-proteína (Zhang, et al., 2014), localización celular (Chou and Shen, 2010; Hasan, et al., 2017; Sperschneider, et al., 2016; Yu, et al., 2006) categorías funcionales (Li, et al., 2016; Saraç, et al., 2010), proteínas de

unión al ADN (Kumar, et al., 2009; Kumar, et al., 2007; Lin, et al., 2011; Liu, et al., 2015), predicción de canales de iones (Lin and Ding, 2011) y proteínas de unión a lípidos (Bakhtiarizadeh, et al., 2014), entre otros.

Los métodos de clasificación no basados en homología tienen en común la generación de covariables predictoras a partir de la estructura primaria de las proteínas. Estos predictores pueden ser simples: tales como usar la frecuencia de monómeros y pares de aminoácidos (aa) o sustituir el alfabeto de 20 aa por alfabetos reducidos (Peterson, et al., 2009). O más elaborados: construidos a partir de las propiedades físico-químicas de los aminoácidos (Li, et al., 2006). En general, a pesar de ser fáciles de obtener, dichos predictores han demostrado funcionar con niveles de exactitud (ACC) mayores al 80% (Chou, 2005; Ebrahimie, et al., 2011; Liu, et al., 2013; Qiu, et al., 2014; Yuan, et al., 2010).

**Panorama global de la tesis**

La presente tesis está dividida en tres capítulos. Los dos primeros se centran en el estudio de la evolución de la longitud de las proteínas en eucariontes y en procariontes (específicamente plantas, animales, hongos, plástidos y endosimbiontes). Mientras que el tercer capítulo se basó en el desarrollo de un método independiente de homología para predecir la función de las proteínas.

Con el objetivo de estudiar la evolución de la longitud de proteínas en distintos linajes de eucariontes, en el **Capítulo 1** realizamos un análisis exhaustivo de tres conjuntos de datos independientes. Como resultado, encontramos diferencias significativas en la longitud de las proteínas entre los distintos linajes de eucariontes, mientas que el promedio general de longitud fue de 409 aa, nos pareció notable que las proteínas en Archaeplastida (363 aa) fueran más pequeñas en comparación a Opistokonta (clado que incluye a Metazoa y Fungi) (428 aa). Aunque ambos linajes de organismos eucariontes son pluricelulares tienen diferencias muy marcadas en la longitud promedio de sus proteínas. Por consiguiente, investigamos el papel de tres alternativas que pudieran explicar la longitud menor en Archaeplastica: 1) la estructura de exones (número y longitud exones), 2) endosimbiosis del cloroplasto, y 3) la localización en compartimentos celulares.

En el **capítulo 2**, investigamos los cambios de longitud de las proteínas en los proteomas de plástidos (primarios, secundarios y terciarios), nucleomorfos y endosimbiontes. Algunas de las preguntas que guiaron este trabajo fueron: 1) al igual que la reducción genómica, ¿la reducción de longitud en proteínas es una tendencia universal en endosimbiontes?; 2) ¿existe un límite en la reducción de longitud en proteínas?; 3) ¿de qué manera afecta a las proteínas del hospedero la migración de genes desde el simbionte hacia el núcleo de su hospedero?

El **capítulo 3** se basó en el desarrollo de un nuevo algoritmo de clasificación a partir de ensambles de SVMs para la predicción de diversas categorías funcionales: unión a ADN, actividad de transporte, actividad catalítica, localización celular, familias de transportadores de membrana y familias de enzimas.

En el algoritmo propuesto, combinamos un sistema de codificación (frecuencias de monómeros, dímeros-*n* y ventanas-*n*), el uso de alfabetos reducidos, selección de características y ensambles de SVMs. Como resultado, el desempeño final del método propuesto mejoró considerablemente con respecto a trabajos previos. Los ensambles obtenidos superan a otros métodos publicados con los mismos conjuntos de datos. El algoritmo que desarrollamos en este trabajo se diferencia de otros algoritmos por utilizar exclusivamente información derivada de la estructura primaria. En contraste otros métodos con los que se compara incorporan información de homología en forma de perfiles PSSM. Adicionalmente, desarrollamos una aplicación web disponible para la comunidad.

El método que desarrollamos fue utilizado para anotar 1786 proteínas ORFans de *Arabidopsis thaliana*, que representan el aproximadamente el 7% de su proteoma. La lista de anotaciones se encuentra disponible para que la comunidad dedicada al estudio de *A. thaliana* encamine sus experimentos en laboratorio y así facilitar la caracterización funcional de las proteínas huérfanas.

Los principales resultados del **Capitulo 1** fueron publicados en la revista *Genomics Proteomics and Bioinformatics* (Ramírez-Sánchez, et al., 2016), mientras que los resultados de los demás capítulos serán incluidos en dos manuscritos (Ramírez-Sánchez, et al., en preparación).

# Capítulo 1. Las proteínas en plantas son de menor tamaño en comparación con las de animales debido a que están codificadas por un menor número de exones

# Objetivos

### Objetivo general

Estudiar la evolución del tamaño de las proteínas en eucariontes y explorar distintos mecanismos que pudieran haber determinado su longitud actual en las plantas.

### Objetivos particulares

1. Determinar la relación filogenética entre la longitud de las proteínas eucariontes y su estructura de exones.
2. Corroborar si existe una relación negativa entre el número de proteínas codificadas en los genomas y la longitud promedio de sus proteínas.
3. Determinar si la migración de genes del cloroplasto hacia el núcleo eucarionte afectó la distribución de longitudes en proteínas nucleares.
4. Comparar las longitudes de las proteínas de diferentes componentes celulares entre plantas, hongos y animales.

# Métodos

### Construcción y curación de los conjuntos de datos

Con el fin de estudiar la evolución de la longitud de las proteínas entre distintos grupos de eucariontes utilizamos 3 conjuntos de datos. El conjunto 1 fue construido previamente por Tiessen, et al. (2012) a partir de los proteomas completos de 51especies de eucariontes, 24 especies de bacterias y 9 especies de arqueas (Febrero, 2010). Los datos contenidos en este conjunto de datos incluyen: longitud de proteína, número de exones y longitud de exones por secuencia. Además, en este conjunto de datos se eliminaron secuencias duplicadas, subsecuencias, variantes de splicing y transposones.

El conjunto 2 fue también construido previamente por Tiessen, et al. (2012) a partir de la base de datos de KEGG PATHWAY (Febrero, 2010) e incluye 140 proteomas de eucariontes. A diferencia de los conjuntos 1 y 3, el conjunto de datos 2 no contiene información sobre la estructura de exones.

Para construir el conjunto 3, utilizamos la base de datos NCBI RefSeq (Pruitt, et al., 2007) versión 70 (julio de 2015). Una vez procesados los archivos en formato GenBank, el conjunto de datos obtenido incluyó ~9.6 millones de secuencias pertenecientes a 5,837

especies. Sin embargo, una gran cantidad de especies estaban mínimamente representadas por número de secuencias. En consecuencia y para evitar sesgo en el análisis estadístico, seleccionamos solo aquellas especies representadas con más de 500 secuencias. De acuerdo con la distribución log-normal, un tamaño de muestra mayor que 500 permite obtener los intervalos de confianza (Tiessen, et al., 2012). Después de dicha selección, el conjunto de datos 3 quedó conformado por ~9.5 millones de secuencias de proteínas y ~74.4 millones de exones pertenecientes a 492 especies (**Figura 1.1**).

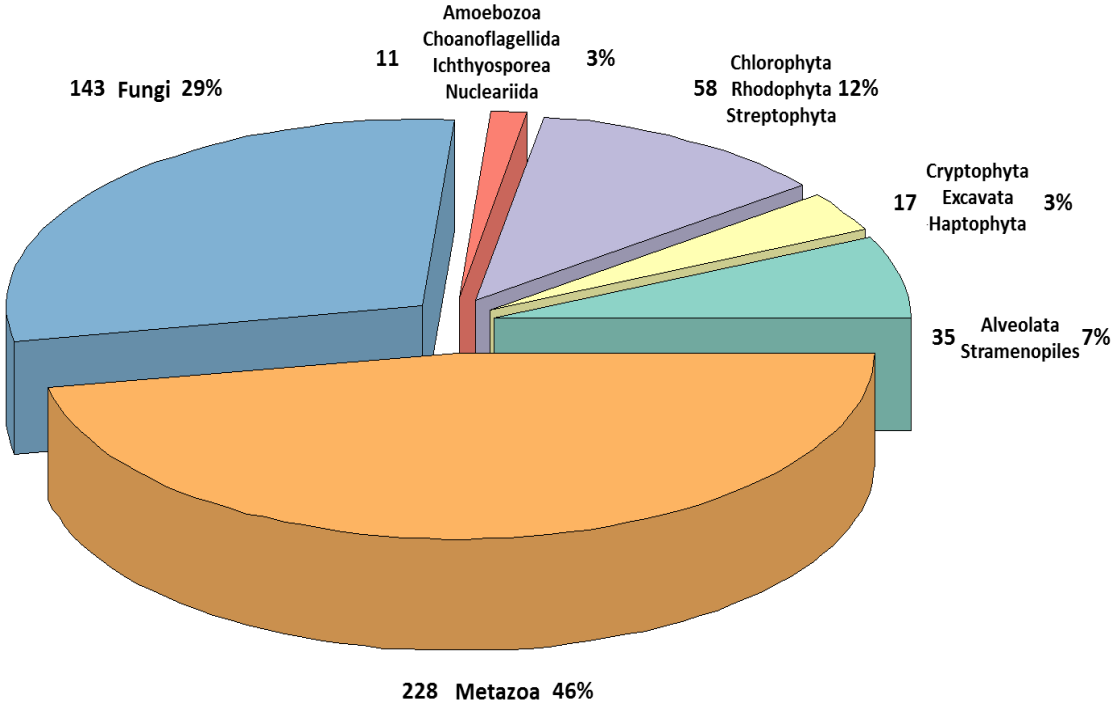



**Figura 1.1. Grupos filogenéticos incluidos en el conjunto de datos 3.** El conjunto de datos 3 fue construido a partir de la base de datos RefSeq (versión 70), contiene 9,522,269 de secuencias pertenecientes a 492 especies de eucariontes. En la figura se muestra el número de especies y el porcentaje de secuencias perteneciente a cada grupo filogenético. Los clados a los que pertenecen dichos grupos son: i) Opisthokonta (Ichthyospora, Choanoflagellida, Metazoa, Nucleariida y Fungi), ii) Amoebozoa, iii) SA (Stramenopila, Alveolata), iv) Archaeplastida (Chlorophyta, Streptophyta y Haptophyta), v) Excavata, vi) Cryptophyta y vii) Haptophyta.

La extracción de estructura de exones (número y longitud) la obtuvimos a partir de las líneas CDS (coding determining sequence) en los archivos GeneBank tal como fue descrito previamente (Kaplunovsky, et al., 2009). Por ejemplo, el CDS del gen XP_007325329.1 perteneciente a *Agaricus bisporus* es el siguiente:

"join(18372..18786,18829..19191,19447..19622)".

Éste CDS está compuesto por tres exones de longitud 415 pb (pares de bases), 363 pb y 176 pb, respectivamente. Adicionalmente, excluimos del conjunto 3 (3.6% de Protistas, 1.44% de Fungi, 0.96% de Streptophyta y 0.89% de Metazoa) aquellas secuencias con CDSs que contenían límites ambiguos (sin codones de inicio o paro explícitos). En la **Tabla 1.1** se muestra un resumen con las estadísticas descriptivas de secuencias de proteínas y secuencias de exones para cada especie.

**Análisis estadístico**

Calculamos las medianas y los promedios de longitud de las proteínas, número de exones y longitud de exones para cada especie del conjunto 3. El promedio de exones por gen lo obtuvimos al dividir el total de exones entre el número de genes. Idénticos resultados se obtienen al calcular el número de exones por gen y después obtener un promedio general. La longitud (mediana y promedio) de exones la obtuvimos de manera similar al dividir la longitud total de los exones entre el número de genes. Una manera alternativa de calcular el número de exones y la longitud de exones por proteína es primero obtener el número/longitud de exones por proteína y posteriormente calcular la media y mediana por especie, lo que permite obtener la distribución de número/longitud de exones por gen. Utilizando ambas alternativas, se obtienen resultados similares.

Para realizar las comparaciones de longitud de proteínas entre especies o grupo filogenéticos utilizamos la prueba no paramétrica de Kruskal-Wallis (Kruskal and Wallis, 1952). Esta prueba es el equivalente al análisis de varianza (ANOVA), solo que no asume normalidad en los datos. Las suposiciones de esta prueba son:

- Las t muestras son muestras aleatorias de sus respectivas poblaciones, y además son independientes entre sí.
- La escala de medición es al menos ordinal.

La prueba de Kruskal-Wallis es una extención de la prueba de rangos de Wilcoxon-Mann-Whitney. El primer paso consiste en asignar los rangos a las $n_1 + n_2 + \cdots + n_t = N$ observaciones. Es juego de hipótesis de interés es

$$H_0 : \tau_1 = \cdots = \tau_t \text{ **vs** } H_1\text{: existe } i \neq j \text{ tal que } \tau_i \neq \tau_j$$

La estadística de prueba es:

$$T = \left(\frac{12}{N(N+1)}\sum_{i=n}^{t}\frac{R_i^2}{n_i}\right) - 3(N+1)$$

La regla de decisión es rechazar $H_0$ si $T > \chi^2_{\alpha,t-1}$. El P-valor es aproximado por $\Pr(\chi^2_{t-1} \geq T)$.

En donde,

$k =$ Cantidad de tratamientos,

$N =$ Número total de datos,

$n_i =$ Número total de repeticiones del tratamiento $i$,

$R_i =$ Suma de rangos del tratamiento $i$,

$$R_i = \sum_{j=1}^{n_i} R(Y_{ij}); i = 1, \ldots, t$$

$\chi^2_{\alpha,t-1}$ = valor de la distribución ji-cuadrada con $t-1$ grados de libertad, cuya área derecha sea igual a α.

Después de realizar dicha prueba realizamos las comparaciones por pares (Castellan, 1988), donde el juego de hipótesis de interés es:

$$H_0 : \tau_i = \tau_j \textbf{ vs } H_1: \tau_i \neq \tau_j$$

La regla de decisión es rechazar $H_0$ si

$$\left|\overline{R_i} - \overline{R_j}\right| > R_i - R_j = \pm Z_{\frac{\alpha}{t(t-1)}}\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

En donde,

$k, N$ y $n_i$ son iguales que en la prueba anterior,

$\overline{R_i} =$ Promedio de los rangos del tratamiento $i$,

$Z_{\frac{\alpha}{t(t-1)}} =$ Valor de la distribución normal estándar cuya área derecha es igual a $\frac{\alpha}{t(t-1)}$.

Adicionalmente, corregimos los P-valores para evitar el problema de falsos descubrimientos (Benjamini and Hochberg, 1995) y los grupos generados en las comparaciones múltiples fueron generados utilizando la librería de R "multcompView" (Spencer, 2012). Todos los análisis estadísticos los realizamos con el paquete estadístico R (R Core Team, 2016).

**Análisis de regresión filogenética**

Uno de los tópicos fundamentales en biología evolutiva es la manera en la cual la evolución de diferentes características biológicas se correlacionan a través de un linaje filogenético. Sin embargo, debido a que las especies descienden de manera jerárquica a partir de un ancestro común, las observaciones pertenecientes de cada especie no pueden ser tratadas como observaciones independientes desde el punto de vista del análisis estadístico. (Garland, et al., 1992). La no independencia filogenética reduce los grados de libertar para las pruebas de hipótesis, disminuye el poder estadístico y afecta la estimación de parámetros (Garland, et al., 1992).

El análisis de Contrastes Filogenéticos Independientes (PIC, por sus siglas en inglés) es un método comparativo para probar hipótesis de cómo los organismos se adaptan a su ambiente. El método PIC usa información filogenética para transformar las observaciones en valores que, en principio, son independientes e idénticamente distribuidos, de manera que puedan ser analizados con métodos estándar como la regresión lineal (Martins and Hansen, 1997).

El análisis PIC fue desarrollado por (Felsenstein, 1985) y se basa en que las diferencias o "contrastes" en los valores de un carácter entre especies hermanas (que comparten un ancestro común exclusivo) son independientes de las diferencias entre cualquier otro par de especies hermanas, y están, por tanto, libres del efecto filogenético. Los contrastes son calculados entre los valores de los caracteres de pares de especies hermanas, siguiendo la topología filogenética y recorriendo el árbol a partir de sus hojas. Este procedimiento resulta en $n$-1 contrastes a partir de las $n$ especies originales en las hojas. En cuanto los nodos ancestrales son determinados correctamente, cada uno de los contrastes resulta ser independiente en términos de los cambios evolutivos que han ocurrido para producir las diferencias entre dos miembros de un contraste. Debido a que los $n$-1 contrastes son estadísticamente independientes, estos pueden ser empleados en análisis estadísticos estándar.

Para realizar el análisis PIC, en primer lugar se realizó la reconstrucción filogenética de eucariontes a partir de la subunidad pequeña de ARN ribosomal. Para ello descargamos dichas secuencias de las bases de datos Protist Ribosomal Reference Database (PRRD)

(Guillou, et al., 2013) y SILVA (Quast, et al., 2013). El alineamiento de secuencias de ARN ribosomal lo realizamos con el programa SINA (Pruesse, et al., 2012) y la eliminación de "gaps" (huecos) con el programa trimAl (Capella-Gutiérrez, et al., 2009) con la opción "automated1". La estimación del modelo evolutivo con mejor ajuste la realizamos con el programa jModelTest (Darriba, et al., 2012), la reconstrucción filogenética con el programa PhyML y para colocar la raíz del árbol filogenético utilizamos la librería de R "phangorn" (Schliep, 2011) en base al criterio de "midpoint rooting".

Para realizar el análisis de PIC utilizamos la reconstrucción filogenética descrita en el párrafo anterior y la librería de R 'ape' (Paradis, et al., 2004). Para modelar la longitud de las proteínas utilizamos como variables explicativas el número de exones y su longitud. El modelo utilizado corresponde a un modelo de regresión lineal múltiple sin intercepto tal y como fue sugerido previamente (Garland, et al., 1992).

**Identificación de ortólogos entre cianobacterias y *Arabidopsis thaliana***

Obtuvimos los proteomas de 4 especies de arqueobacterias (*Pyrococcus furiosus, Methanobacterium AL, Methanococcus maripaludis, Archaeoglobus fulgidus*), 3 Gram positivas (*Mycoplasma genitalium, Bacillus subtilis, Mycobacterium sp. JDM601*), 3 cianobacterias (*Nostoc sp. PCC7107, Prochlorococcus marinus, Synechocystis sp. PCC6803*), 4 eubacterias (*Borrelia afzelii, Treponema azotonutricium, Chlamydia pecorum, Aquifex aeolicus*) y 4 proteobacterias (*Rickettsia akari, Helicobacter acinonychis, Haemophilus ducreyi, Escherichia coli*) a partir de la base de datos NCBI (agosto de 2015). Por otra parte, los proteomas nucleares de *Arabidopsis thaliana* y *Saccharomyces cerevisiae* los descargamos de la base de datos TAIR (https://www.arabidopsis.org/) y la base de datos de *Saccharomyces* (http://www.yeastgenome.org/).

Para obtener los genes de origen endosimbiótico en *A. thaliana*, primero construimos un conjunto de secuencias no redundantes utilizando el programa CD-Hit (Fu, et al., 2012; Li and Godzik, 2006) con los parámetros por defecto. Después comparamos cada secuencia de *A. thaliana* contra el resto de las secuencias (50,036) pertenecientes a los 20 proteomas mencionados en el párrafo anterior. Para cada secuencia de *A. thaliana* construimos una tabla a partir de los hits de BLAST que presentaron un *E-value* $< 10^{-10}$ y una identidad > 30%. Con dichas tablas realizamos alineamientos múltiples utilizando el programa

MUSCLE (Edgar, 2004) con los valores por defecto. Los gaps los eliminamos utilizando el programa trimAl (Capella-Gutiérrez, et al., 2009) con la opción "gappyout". Posteriormente, para inferir el modelo evolutivo utilizamos el programa ProtTest (Darriba, et al., 2011) y la reconstrucción de los árboles filogenéticos la realizamos con el programa PhyML (Guindon, et al., 2010). De acuerdo con una metodología descrita previamente (Martin, et al., 2002; Rujan and Martin, 2001), para inferir los genes de origen endosimbiótico seleccionamos aquellos clados en las filogenias donde ramificaron las secuencias de cianobacterias junto con las secuencias de *A. thaliana*.

**Comparación de longitud de proteínas en compartimentos celulares de *A. thaliana*, *H. sapiens* y *S. cerevisiae.***

Las anotaciones GO de componente celular de *A. thaliana*, *S. cerevisiae* y *H. sapiens* las descargamos de las bases de datos TAIR (https://www.arabidopsis.org/), Saccharomyces Genome (http://www.yeastgenome.org) y The Gene Ontology Consortium (http://geneontology.org/). Para el caso de *H. sapiens*, realizamos un mapeo a categorías GO Slim Generic utilizando el script Map2Slim (disponible a través del paquete go-perl en el repositorio CPAN). El análisis estadístico entre categorías GO Slim lo realizamos utilizando pruebas no paramétricas.

## Resultados y discusión

**Los eucariontes muestran una gran diversidad de longitudes de proteínas**

En este trabajo, determinamos las longitudes de las proteínas en distintos grupos de eucariontes utilizando 3 conjuntos de datos independientes. En la **Tabla S1.1** se muestran para cada especie del conjunto 1 las estadísticas detalladas (promedios y medianas) de las longitudes de las proteínas, el número de exones y la longitud de exones. En la (**Figura 1.2**)¡Error! No se encuentra el origen de la referencia. se puede observar que las proteínas procariontes son en general más pequeñas que aquellas de eucariontes, lo que concuerda con observaciones previas (Brocchieri and Karlin, 2005; Tiessen, et al., 2012; Zhang, 2000). Adicionalmente, dicha figura también muestra que las proteínas de plantas son de menor longitud en comparación con las de stramenopiles, animales y hongos. El número de especies, el total de proteínas, la cantidad de proteínas por especie, el total de exones y el

número de exones por especie del conjunto 3 se resumen en la **Tabla 1.1**. Adicionalmente, en la **Tabla S1.2** se pueden encontrar las estadísticas detalladas para cada especie del conjunto 3.

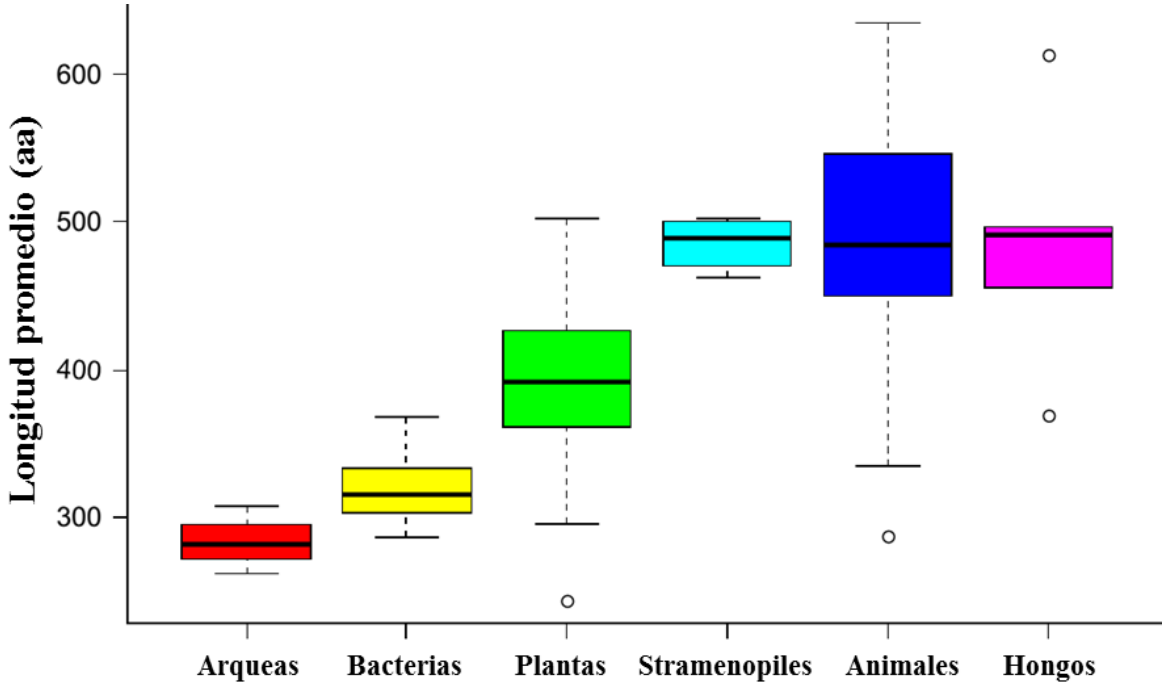



**Figura 1.2. Distribución de la longitud de las proteínas en distintos grupos filogenéticos del conjunto de datos 1.** Primero se calculó la longitud promedio para cada especie, posteriormente se obtuvo la distribución de longitudes promedio para cada grupo filogenético. Las líneas sólidas en los diagramas de cajas y bigotes indican medianas. aa = aminoácidos.

El análisis estadístico en el conjunto 3 reveló diferencias significativas entre distintos grupos filogenéticos de eucariontes al considerar las variables de longitud de proteína, número de exones o longitud de exones (**Tabla 1.2**). Al analizar los conjuntos 1 y 2 obtuvimos resultados similares, lo que nos permitió generalizar los resultados obtenidos a partir del conjunto 3 que contiene mayor número y diversidad de especies.

En la **Figura 1.3** se pueden observar los promedios y medianas de longitud de proteína para los 14 grupos filogenéticos del conjunto 3, dichos grupos fueron organizados de acuerdo a la versión moderna de la filogenia de eucariontes (Adl, et al., 2012; Burki, 2014; Pawlowski, et al., 2012). En dicha figura se puede observar que las proteínas de Opisthokonta (Ichthyosporea, Choanoflagellida, Metazoa, Nucleariida y Fungi) fueron las más largas. Entre estos, las proteínas en los grupos de Ichthyosporea, Nucleariida y

Choanoflagellida fueron de mayor longitud con respecto a Metazoa, y éstas a su vez más largas que las de Fungi. La longitud de las proteínas en Archaeplastida (Chlorophyta, Rodophyta y Streptophyta) fue similar al grupo de SA (Stramenopiles y Alveolata), al de Amoebozoa y al de Excavata. Finalmente, los grupos de Cryptophyta y Haptophyta presentaron las proteínas de menor tamaño en los eucariontes.

**Tabla 1.1. Estadísticas resumen del conjunto de datos 3.**

| Grupo | Total de especies | % de especies | Total de proteínas | Proteínas por especie | Total de exones | Exones por especie |
|---|---|---|---|---|---|---|
| Alveolata | 24 | 4.9 | 166,806 | 6,950 | 596,916 | 24,872 |
| Amoebozoa | 7 | 1.4 | 70,337 | 10,048 | 218,096 | 31,157 |
| Chlorophyta | 9 | 1.8 | 78,172 | 8,686 | 424,697 | 47,189 |
| Choanoflagellida | 2 | 0.4 | 19,817 | 9,909 | 164,858 | 82,429 |
| Cryptophyta | 4 | 0.8 | 22,383 | 5,596 | 146,349 | 36,587 |
| Excavata | 12 | 2.4 | 160,927 | 13,411 | 171,251 | 14,271 |
| Fungi | 143 | 29.1 | 1,303,212 | 9,113 | 4,285,632 | 29,969 |
| Haptophyta | 1 | 0.2 | 31,735 | 31,735 | 118,186 | 118,186 |
| Ichthyosporea | 1 | 0.2 | 8510 | 8,510 | 40,799 | 40,799 |
| Metazoa | 228 | 46.3 | 5,743,160 | 25,189 | 58,003,499 | 254,401 |
| Nucleariida | 1 | 0.2 | 6,115 | 6,115 | 29,416 | 29,416 |
| Rhodophyta | 3 | 0.6 | 21,255 | 7,085 | 39,453 | 13,151 |
| Stramenopila | 11 | 2.2 | 175,058 | 15,914 | 614,663 | 55,878 |
| Streptophyta | 46 | 9.3 | 1,692,582 | 36,795 | 9,571,726 | 208,081 |
| ***Total*** | **492** | **100** | **9,522,269** | **19,354** | **74,425,541** | **151,271** |

El conjunto de datos 3 fue construido a partir de la base de datos RefSeq (versión 70) con las secuencias de 492 especies. Los clados a los que pertenecen los 14 grupos filogenéticos son: i) Opisthokonta (Ichthyospora, Choanoflagellida, Metazoa, Nucleariida y Fungi), ii) Amoebozoa, iii) SA (Stramenopila, Alveolata), iv) Archaeplastida (Chlorophyta, Streptophyta y Haptophyta), v) Excavata, vi) Cryptophyta y vii) Haptophyta.

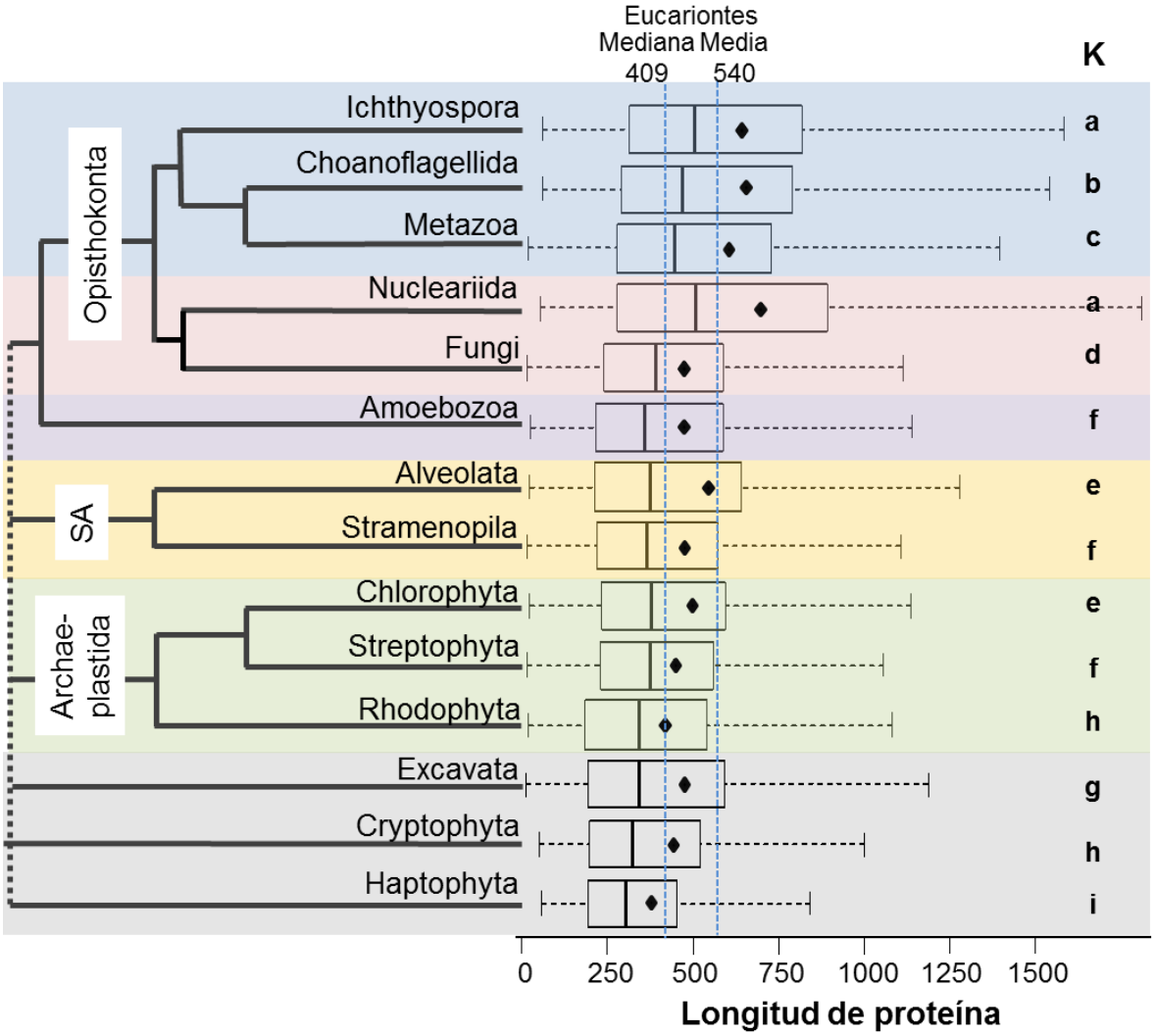



**Figura 1.3. Longitud de proteínas a través del árbol de la vida de eucariontes (conjunto 3).** Los 14 diferentes grupos filogenéticos fueron ordenados de acuerdo a su origen evolutivo. Las líneas sólidas en los diagramas de cajas y bigotes indican medianas de cada grupo, mientras que los puntos en forma de diamante indican promedios de cada grupo. Las líneas azules muestran la mediana y el promedio generales para los eucariontes. Las distintas letras en la columna K indican diferencias significativas de longitud de proteínas entre grupos filogenéticos (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05).

**Tabla 1.2. Estadísticas y análisis estadístico de longitud de proteína, número y longitud de exones de 14 grupos filogenéticos de eucariontes (conjunto 3).**

| Grupo | Promedio longitud de proteína (aa) | Mediana longitud de proteína | K | Promedio número de exones | Mediana número de exones | K | Promedio longitud de exones (nc) | Mediana longitud de exones | K |
|---|---|---|---|---|---|---|---|---|---|
| Alveolata | 535 | 364 | e | 3.6 | 2 | i | 449 | 161 | h |
| Amoebozoa | 463 | 351 | f | 3.1 | 2 | j | 448 | 192 | f |
| Chlorophyta | 490 | 369 | e | 5.4 | 3 | g | 270 | 139 | j |
| Choanoflagellida | 648 | 459 | b | 8.2 | 6 | b | 237 | 113 | m |
| Cryptophyta | 435 | 316 | h | 6.5 | 5 | c | 199 | 82 | n |
| Excavata | 471 | 334 | g | 1.1 | 1 | m | 1,330 | 935 | a |
| Fungi | 462 | 381 | d | 3.3 | 2 | k | 421 | 195 | e |
| Haptophyta | 367 | 294 | i | 3.7 | 3 | h | 295 | 164 | g |
| Ichthyosporea | 637 | 494 | a | 4.8 | 4 | d | 398 | 152 | i |
| Metazoa | 595 | 439 | c | 10.1 | 7 | a | 176 | 126 | l |
| Nucleariida | 690 | 499 | a | 4.8 | 4 | e | 429 | 216 | c |
| Rhodophyta | 411 | 334 | g | 1.9 | 1 | n | 665 | 330 | b |
| Stramenopila | 467 | 356 | f | 3.5 | 2 | l | 399 | 196 | d |
| Streptophyta | 436 | 363 | f | 5.7 | 4 | f | 230 | 128 | k |

Las letras diferentes en la columna K indican diferencias significativas entre grupos filogenéticos. Grupos que comparten la misma letra no mostraron diferencias significativas (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05). aa = aminoácidos; nc = nucleótidos.

### Los protistas muestran la mayor diversidad de longitudes de proteínas

Los protistas mostraron la mayor diversidad en longitudes de proteínas (**Figura 1.3**). Por ejemplo, los grupos de Ichthyosporea y Choanoflagellida tuvieron las proteínas más largas entre todos los grupos de eucariontes. En contraste, el grupo de Cryptophyta y el grupo de Haptophyta tuvieron las proteínas de menor longitud. Esta situación concuerda con la filogenia del ARN ribosomal, en donde se observa que los protistas son el grupo más diverso entre los eucariontes (Schlegela, 1994), reflejando su distinto origen evolutivo. Históricamente, el grupo de protistas fue construido en base a características microscópicas y morfológicas. En consecuencia, dichos organismos forman un grupo polifilético. En realidad, los protistas representan un agrupamiento artificial que no refleja una historia evolutiva sino tal vez solo un tipo de vida unicelular. A su vez, distintos linajes de protistas están filogenéticamente más relacionados con animales, plantas u hongos, respectivamente.

Algunos autores han propuesto que la mayor longitud de proteínas de eucariontes, en comparación a procariontes, refleja "una tendencia evolutiva hacia organismos más complejos" (Brocchieri and Karlin, 2005) en términos de sus redes de interacción de proteínas. Sin embargo, a pesar de que las especies multicelulares pueden ser vistas como "más complejas" que las unicelulares, de acuerdo con nuestros resultados muchos organismos unicelulares (Alveolata, Amoebozoa, Choanoflagellida e Ichthyosporea) poseen una longitud promedio similar o aún mayor que la de los multicelulares (Metazoa y Streptophyta). De esta manera, en el contexto de la evolución de eucariontes, otros factores como el tamaño del proteoma y la tasa de splicing alternativo deben ser considerados en conjunto con la longitud de proteínas como un indicativo de la "complejidad" en los organismos. Una hipótesis que abordaremos más adelante, es si la complejidad biológica está relacionada con el número de exones y de dominios, más que con la longitud *per se* de las proteínas.

**Longitud de proteínas en el clado Archaeplastida**

Una comparación entre organismos fotosintéticos reveló diferencias entre grupos del clado de Archaeplastida (P-valor < 0.01). Por ejemplo, las proteínas de las algas verdes (Chlorophyta) fueron 12% más largas que otros grupos de plantas. El alga roja *Cyanidioschyzon merolae* también tuvo proteínas de mayor longitud (~504 aa) en comparación con las de Streptophytas (~436 aa). Por otra parte, las especies de monocotiledoneas (*Oryza sativa* (379-448 aa), *Zea mays* (345-402 aa), *Sorghum bicolor* (361-418 aa) y *Brachipodium distachyon* (428-457 aa)) tuvieron proteínas ligeramente de mayor longitud que las especies de dicotiledóneas (*Carica papaya* (~296 aa), *Medicago truncatula* (245-295 aa) y *Populus trichocarpa* (375-390 aa)). Es interesante recalcar que, a pesar de tener un genoma compacto, el promedio de longitud en *Arabidopsis thaliana* (403-410 aa) no fue particularmente menor en comparación con otras plantas.

**Los genomas de plantas codifican para una mayor cantidad de proteínas pero tienen menor longitud en comparación con animales y hongos**

De acuerdo a la anotación de genomas actual, los hongos (Fungi) y las plantas (Streptophyta) tuvieron en promedio más proteínas pero de menor longitud en comparación con animales (Metazoa) (**Figura 1.4**). El número de proteínas en las especies de plantas

(36,795 en promedio por genoma) fue mayor que en las especies de animales (25,189) y hongos (9,113). Mientras que, el 90% de las proteínas en plantas estuvo en el rango de longitud de 649-877 aa, en animales dicho rango fue de 909-1125 aa. Existe pues una diferencia promedio de 250 aa entre proteínas de plantas y animales, lo que representa un incremento del 30%. Por otro lado, las plantas tienen ~46% más genes codificados en su genoma comparados con los animales (Tabla 1.1), posiblemente debido a un mayor número de duplicaciones genómicas (Adams and Wendel, 2005).

Además, encontramos una correlación negativa de -0.39 y -0.42 (Prueba de correlación de Pearson, P-valor = 0.005) entre el promedio de la longitud y la cantidad de proteínas de los proteomas completos en los conjuntos 1 y 2. La correlación negativa entre la longitud de proteína y el número de genes por especie sugiere la existencia de un proceso de fusión de genes y/o duplicación de dominios en eucariontes, particularmente en animales. El mayor número de proteínas en plantas (particularmente en las angiospermas) podría deberse a eventos de duplicaciones genómicas (Adams and Wendel, 2005).

Recientemente, Buljan et al. (2010) reportó que uno de los procesos mediante los cuales las proteínas de animales incrementan su longitud es a través de un proceso de duplicación de genes y la posterior fusión de genes adyacentes. Mientras que otros mecanismos, como la transcripción reversa y la inserción de genes, han tenido únicamente una contribución menor en el proceso de adquisición de nuevos dominios en proteínas de animales (Buljan, et al., 2010).

La fusión de genes tiene profundas consecuencias en la maquinaria de splicing alternativo. Koralewski et. al. (2011) observó que, en comparación con plantas, los genes de animales tienen mayores tasas de splicing alternativo (¡Error! No se encuentra el origen de la referencia.). Adicionalmente, también encontró una correlación positiva entre el número de exones y la tasa de splicing alternativo.

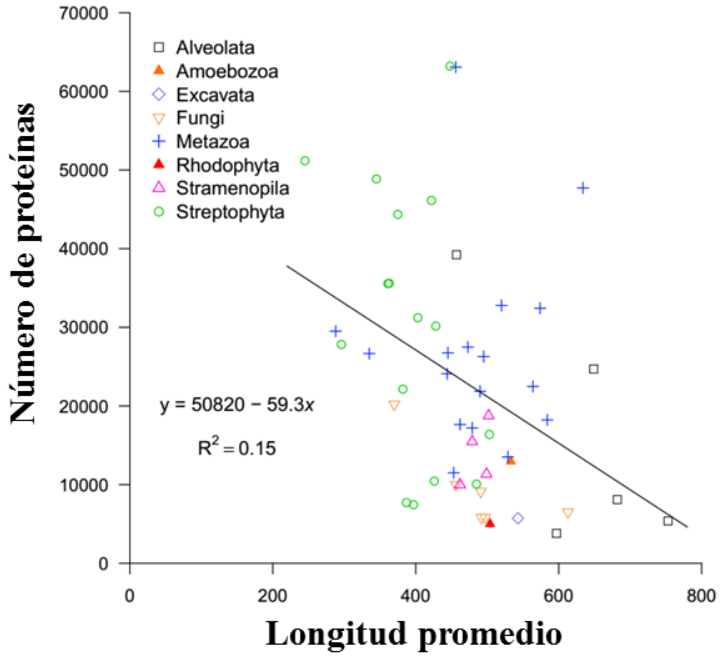



**Figura 1.4. Relación entre el número de proteínas por genoma y la longitud promedio (en número de aminoácidos) de proteínas en 51 especies de eucariontes (conjunto 1).** El valor de correlación obtenido fue de -0.39 (Prueba de correlación de Pearson, P-valor = 0.005). Cada punto representa una especie diferente. La línea sólida fue obtenida utilizando el modelo de regresión lineal.

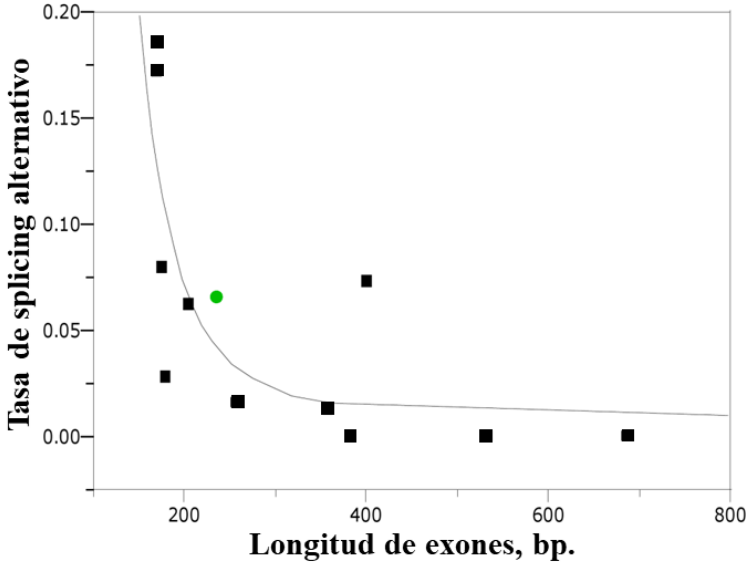



**Figura 1.5. Relación entre la tasa de splicing alternativo y el promedio de longitud de exones de 12 especies de eucariontes (Koralewski and Krutovsky, 2011).** Los proteomas de eucariontes en este análisis fueron obtenidos de la base de datos NCBI GenBank. Los promedios del número/longitud de exones primero fueron calculados para cada gen y con ellos posteriormente se obtuvieron los promedios para cada especie. La tasa de splicing alternativo fue calculada como el cociente entre la suma del número de splicing alternativo por gen y el total de genes de cada especie. El círculo verde corresponde a *A. thaliana*.

**Las proteínas específicas de plantas son de menor longitud en comparación con animales y hongos**

Al conjunto de proteínas que se encuentran conservadas entre diferentes especies de un mismo grupo de organismos se les denomina como proteínas específicas de dicho grupo. De manera que, las proteínas específicas de plantas se encuentran conservadas sólo entre plantas pero no en otros eucariontes como animales u hongos.

Para obtener los conjuntos de proteínas específicas de plantas y proteínas conservadas entre eucariontes, consultamos la base de datos Plant Specific Database (http://genomics.msu.edu/plant_specific/) (Gutiérrez, et al., 2004). Por una parte, las proteínas específicas de plantas fueron significativamente de menor tamaño (Prueba de Wilcox, P-valor < 0.001) en comparación con el proteoma completo de *A. thaliana* (346 aa y 402 aa, media y mediana, respectivamente) y que el panproteoma formado con todos los proteomas del grupo Streptophyta (363 aa y 436 aa, media y mediana). Por otra parte, las proteínas conservadas entre eucariontes que *A. thaliana* comparte con Metazoa y Fungi (mediana de 392aa y promedio de 458aa) fueron significativamente de mayor longitud que las proteínas en *A. thaliana* y Streptophyta (Prueba de Wilcox, valor-p < 0.001).

Es notable que las proteínas en plantas fueran significativamente de menor tamaño que las de animales y hongos (P-valor = 1e-16). Por tanto, la siguiente parte de este trabajo la enfocamos en explorar tres escenarios que pudieran explicar este fenómeno: 1) estructura de exones, 2) endosimbiosis, y 3) compartimentos celulares. Cabe señalar éstos no son los únicos escenarios posibles, otras alternativas como la duplicación de dominios y fusión de dominios en animales merecen también ser abordados en futuros trabajos. Adicionalmente, existe una hipótesis nula de que las diferencias en longitudes de proteínas en distintos grupos de eucariontes se deban a diferencias aleatorias que poseían los ancestros en cada grupo, también es necesario contrastar esta hipótesis en un trabajo futuro.

**El número más que el tamaño de los exones determina la longitud de las proteínas**

La longitud de una proteína cualquiera está determinada por una combinación entre el número de exones y su longitud (**Figura 1.6 A**). Por ejemplo, las proteínas de mayor longitud pueden estar compuestas por pocos exones de gran longitud, o por un gran número

de exones pero de longitud reducida. En cambio, las proteínas de menor longitud se componen de un número reducido de exones o bien por tamaños de exones más pequeños.

Diversos linajes eucariontes utilizan distintas estrategias generales para determinar el tamaño de sus proteínas (**Figura 1.6 B,** ¡Error! No se encuentra el origen de la referencia.). Por ejemplo, el grupo Metazoa tuvo numerosos exones de longitud muy corta; en cambio el grupo Fungi tuvo pocos exones pero de longitud mucho mayor; el grupo de Streptophyta presentó una configuración intermedia entre Metazoa y Fungi. Es decir, las plantas tienen menos exones que los animales, pero más exones que hongos. Además, tienen exones más cortos que hongos pero más largos que los animales. En resumen, cada linaje filogenético utiliza diferentes estrategias para codificar el tamaño de sus proteínas: exones largos (Excavata y Rodophyta), gran número exones (Metazoa y Choanoflagellida), intermedio entre número y longitud (Fungi y Amoebozoa).

Para determinar cuál de ambas características (número o longitud de exones) ha influido más en la longitud de las proteínas durante la evolución, realizamos un análisis de correlación entre dichas características y la longitud de proteína. Como resultado de dicho análisis, encontramos una correlación positiva de 0.55 entre el número de exones y la longitud de proteína (Prueba de Spearman, P-valor < 0.001), mientras que la longitud de exones no mostró correlación significativa con la longitud de proteínas (P-valor = 0.06). Al repetir este análisis con cada uno de los grupos filogenéticos encontramos que el grado de asociación entre número de exones y longitud de proteína es distinto para cada grupo (**Tabla 1.3**): el grupo Metazoa presentó el mayor valor de correlación, el grupo Fungi presentó el menor, mientras que el grupo de Streptophyta tuvo un valor intermedio entre ambos. Adicionalmente, obtuvimos los coeficientes de regresión y los parámetros de regresión lineal para cada una de las especies del conjunto 3 (**Tabla S1.3**).

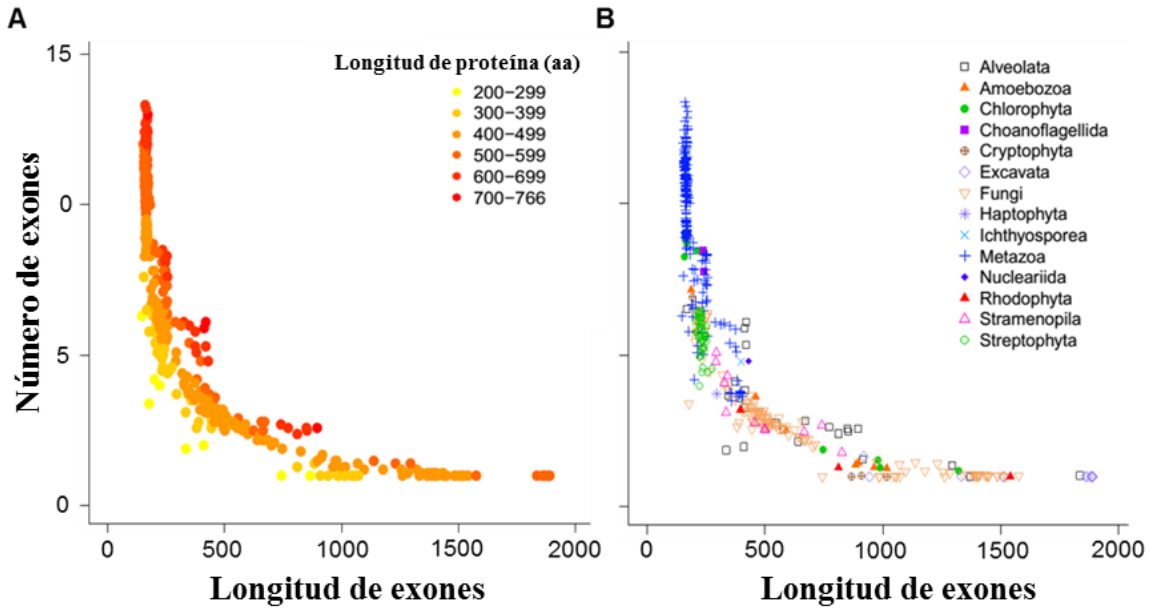



**Figura 1.6. Relación entre la longitud de proteína y la estructura de exones en el conjunto 3.** En la gráfica se muestra el promedio del número de exones contra el promedio de longitud de exones. Cada punto representa una especie. A) La longitud de proteína es presentada en un gradiente de color. B) Cada símbolo representa a un grupo filogenético distinto.

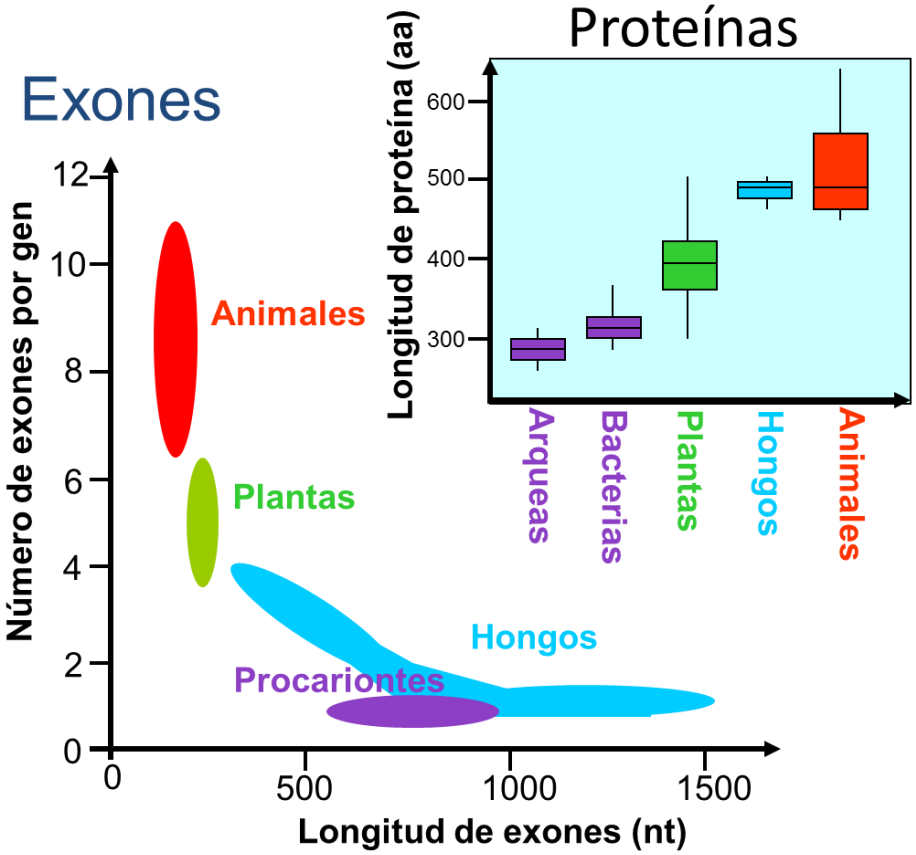



**Figura 1.7. Modelo simplificado de la relación entre la longitud de proteína y la estructura de exones usando los resultados de los conjuntos de datos 1-3.** Esta versión simplificada fue generada manualmente e incluye los grupos filogenéticos de Arquea, Bacteria, Streptopyta, Metazoa y Fungi. Se muestra la estrategia utilizada por hongos y animales para lograr proteínas de mayor longitud que las plantas (verde): más exones (animales, rojo) o exones largos (hongos, azul). Los organismos procariontes (morado) utilizan la "estrategia de un exón" y por lo tienen proteínas de menor longitud en comparación de eucariontes, las proteínas de las plantas tienen una longitud intermedia.

**Tabla 1.3. Correlación y parámetros de regresión lineal entre la longitud de proteína y el número de exones para cada uno de los 14 linajes de eucariontes del conjunto 3.**

| Grupo | Coeficiente Spearman | P-valor | $\beta_1$ | $R^2$ |
|---|---|---|---|---|
| Alveolata | 0.2 | < 0.001 | 89.39 | 0.34 |
| Amoebozoa | 0.24 | < 0.001 | 84.77 | 0.44 |
| Chlorophyta | 0.34 | < 0.001 | 62.59 | 0.56 |
| Choanoflagellida | 0.49 | < 0.001 | 62.95 | 0.52 |
| Cryptophyta | 0.46 | < 0.001 | 47.76 | 0.43 |
| Excavata | 0.1 | < 0.001 | 395.69 | 0.45 |
| Fungi | 0.2 | < 0.001 | 88.94 | 0.46 |
| Haptophyta | 0.43 | < 0.001 | 77.18 | 0.55 |
| Ichthyosporea | 0.23 | < 0.001 | 92.1 | 0.45 |
| Metazoa | 0.71 | < 0.001 | 53.94 | 0.79 |
| Nucleariida | 0.51 | < 0.001 | 130.8 | 0.64 |
| Rhodophyta | 0.11 | < 0.001 | 123.92 | 0.37 |
| Stramenopila | 0.35 | < 0.001 | 90.55 | 0.49 |
| Streptophyta | 0.45 | < 0.001 | 57.22 | 0.66 |

$B_1$ corresponde al coeficiente de regresión para el número de exones y $R^2$ al coeficiente de determinación en el modelo de regresión lineal sin intercepto.

Tal como Felsenstein señaló, el obtener una correlación al analizar características biológicas podría carecer de sentido si se ignoran durante el análisis las relaciones filogenéticas (Martins and Hansen, 1997). Por tanto, para tomar en cuenta la estructura filogenética en el conjunto 3, realizamos una reconstrucción filogenética de eucariontes utilizando las secuencias del ARN ribosomal de 233 especies (ver métodos; **Tabla S1.4**). Con la reconstrucción filogenética obtuvimos las distancias filogenéticas necesarias para realizar el análisis PIC. Una vez que se tomaron en cuenta las distancias filogenéticas, realizamos el análisis estadístico y obtuvimos una correlación de 0.89 (Prueba de Spearman P-valor < 0.001) entre número de exones y longitud de proteína. En el caso de longitud de exones y longitud de proteína el valor de correlación fue bajo (0.19) y no significativo. Este análisis final nos permitió confirmar que existe una relación entre el promedio de exones y el tamaño final de las proteínas. Sin embargo, con estos resultados no es claro si dicha relación es en efecto la causa en proteínas más largas en animales.

**La longitud promedio de las proteínas en plantas no se debe a la migración de genes desde el cloroplasto hacia el núcleo**

La segunda explicación al porqué las plantas tienen proteínas de menor tamaño en comparación a otros eucariontes puede ser la migración de una gran cantidad de genes provenientes del cloroplasto hacia el núcleo después de la endosimbiosis. Tres hechos soportan esta hipótesis: 1) las cianobacterias son los ancestros de los cloroplastos en plantas (Margulis, 1981), 2) las cianobacterias tienen proteínas de menor longitud que los eucariontes unicelulares (Tiessen, et al., 2012), 3) los animales y hongos solo tuvieron una endosimbiosis (mitocondria) pero las células vegetales han tenido dos eventos de endosimbiosis (mitocondria y plástido). Por lo tanto, el tamaño promedio actual de las proteínas en plantas puede deberse a la migración masiva de genes de menor longitud, provenientes del cloroplasto, hacia el núcleo de las plantas primitivas.

Una estimación sencilla se obtiene siguiendo el siguiente razonamiento: después de la endosimibiósis, la migración de un estimado de 2,000 proteínas provenientes del cloroplasto con una longitud mediana de 257 aa hacía el núcleo de un eucarionte ancestral de plantas con ~22,900 proteínas con longitud mediana de 400 aa resulta en un nuevo proteoma con aproximadamente 24,900 proteínas (un incremento de 8.7%). Sin embargo, la longitud mediana resultante (ponderando la contribución de cada una) es de 391 aa, la cual se encuentra aún en el rango de hongos y animales.

Una forma alternativa de poner a prueba esta hipótesis es comparar la longitud de proteínas nucleares de *A. thaliana* con la longitud de aquellas que probablemente tuvieron su origen en cianobacterias. De acuerdo con trabajos previos (Dagan, et al., 2013; Martin, et al., 2002; Rujan and Martin), los genes transferidos del cloroplasto hacia el núcleo pueden ser identificados al construir árboles filogenéticos que contengan homólogos eucariontes y procariontes, y posteriormente seleccionar aquellas filogenias donde plantas y cianobacterias ramifiquen juntas. Siguiendo esta estrategia, identificamos 1,339 genes en *A. thaliana* cuyos ancestros muy probablemente se encuentran en cianobacterias. Dicha cantidad coincide con reportes previos (Martin, et al., 2002; Rujan and Martin 2001).

Contrario a nuestras expectativas, la longitud mediana de las 1,339 proteínas que migraron del cloroplasto al núcleo fue de 396 aa, lo cual es una longitud significativamente mayor

que la mediana del proteoma completo en *A. thaliana* (349 aa), mayor que el proteoma de su cloroplasto (160 aa), mayor que el panproteoma de cianobacterias (257 aa) (compuesto por *Nostoc PCC*, *Prochlorococcus marinus* y *Synechocystis PCC*) y mayor que sus ortólogos en cianobacterias (334 aa). Por lo tanto, estos resultados claramente permiten excluir la hipótesis de la migración de genes de origen cianobacteriano hacia el núcleo como una explicación al porqué las plantas tienen proteínas de longitud menor en comparación con otros eucariontes.

**Las proteínas de las plantas poseen un sesgo hacia menor tamaño en varios compartimentos celulares**

Una posible tercera explicación del porqué las proteínas en plantas son de menor tamaño con respecto a otros eucariontes como animales y hongos podría ser debido a diferencias específicas en los diferentes compartimentos celulares. La célula vegetal tiene una estructura diferente comparada con la célula animal. No está claro si los compartimentos celulares afectan significativamente la distribución de la longitud de proteínas en estos grupos de eucariontes. Para contestar esta interrogante, calculamos los promedios y medianas de longitud para distintos compartimentos celulares en tres organismos modelo: *A. thaliana, H. sapiens* y *S. cerevisiae*.

En el caso de *A. thaliana,* la longitud de proteínas fue significativamente diferente (P-valor = 2.2e-16) entre distintos compartimentos celulares (**Tabla 1.4**). Por ejemplo, la mayor longitud fue encontrada en proteínas de membrana (membrana plasmática, aparato de Golgi y otras membranas), mientras que las proteínas de menor longitud se encontraron en los ribosomas, retículo endoplásmico (RE), proteínas extracelulares y en la categoría de componente desconocido. No hubo diferencias significativas entre las proteínas del citosol, cloroplastos, otros plastidios, núcleo y mitocondria. Cabe señalar que la presencia de péptidos de tránsito en las proteínas de cloroplasto (~10-50 aa) no hace que éstas sean de mayor longitud que las proteínas de citosol. Estos resultados también confirman la hipótesis de que la menor longitud de proteínas en plantas se deba a los plastidios, los cuales no se encuentran presentes en animales y hongos.

Sin embargo, los resultados de la **Tabla 1.4** permiten la interpretación de que en plantas podría haber menor cantidad de proteínas de membrana (de mayor longitud) y más

proteínas de ribosoma, vacuola, extracelular y componente desconocido (todos de menor longitud) en comparación a animales y hongos. Para poner a prueba dicha interpretación, calculamos las longitudes de proteínas en los diferentes compartimentos de *H. sapiens* y *S. cerevisiae*. Las proteínas de membrana en *A. thaliana* representan un 16% del total de su proteoma mientras que en *S. cerevisiae* representan el 21%. Las proteínas de ribosoma representan el 1% del proteoma de *A. thaliana*, mientras que en *S. cerevisiae* representan el 3%. Estos resultados muestran que ambos organismos tienen una distribución similar del número de proteínas que forman sus compartimentos (**Tabla S1.5**). Al comparar la longitud de proteínas entre ambos organismos, encontramos que 9 de 11 compartimentos presentan un sesgo hacia mayor longitud en *S. cerevisiae* (**Figura 1.8**).

**Tabla 1.4. Comparación de longitud de proteínas entre distintos componentes celulares de *A. thaliana*.**

| Categoría | K | Total de proteínas | Mediana aa | Promedio aa |
|---|---|---|---|---|
| Membrana plasmática | a | 2,152 | 446 | 518 |
| Aparato de Golgi | ab | 126 | 431 | 510 |
| Otras membranas | bc | 3,084 | 390 | 453 |
| Citosol | cd | 1,700 | 367 | 444 |
| Pared celular | cd | 549 | 400 | 436 |
| Cloroplasto | cde | 3,446 | 359 | 431 |
| Plastidios | cdef | 1,228 | 361 | 419 |
| Otros componentes intracelulares | cdef | 3,969 | 354 | 425 |
| Otros componentes citoplasmáticos | cdef | 3,063 | 348 | 411 |
| Núcleo | cdef | 1,810 | 325 | 416 |
| Mitocondria | def | 770 | 341 | 389 |
| Otros componentes celulares | ef | 4,612 | 322 | 376 |
| Extracelular | ef | 477 | 316 | 375 |
| Retículo endoplásmico | f | 213 | 321 | 372 |
| Componente celular desconocido | g | 4,468 | 237 | 298 |
| Ribosoma | g | 356 | 206 | 248 |

Las letras diferentes en la columna K indican diferencias significativas entre grupos filogenéticos. Grupos que comparten la misma letra no mostraron diferencias significativas (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05). aa = aminoácidos.

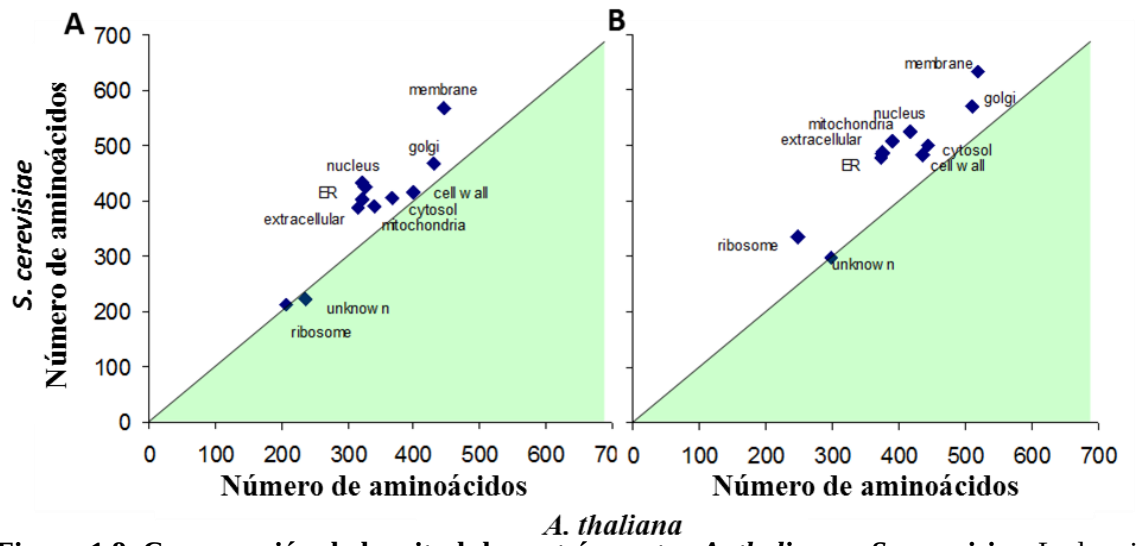



**Figura 1.8. Comparación de longitud de proteína entre *A. thaliana* y *S. cerevisiae*.** La longitud fue calculada para cada compartimento en base a las categorías GO Slim de componente celular. La línea diagonal representa correspondencia exacta entre categorías. Los puntos arriba de la línea indican mayor longitud en *S. cerevisiae*. A) Valores de medianas. B) Valores de promedios.

Por otra parte, en comparación con *A. thaliana*, *H. sapiens* presentó un sesgo hacia proteínas de mayor longitud en varios compartimentos como matriz extracelular (413aa vs 755aa), aparato de Golgi (435aa vs 495aa), citoplasma (339aa vs 491aa), citosol (343aa vs 451aa) y núcleo (328aa vs 472aa). En cambio, las proteínas de mitocondria (329aa vs 312aa) y ribosoma (201aa vs 180aa) fueron similares entre ambas especies (**Tabla S1.6**). En conclusión, al comparar los tamaños de distintos compartimentos celulares de plantas, animales y hongos, encontramos que las plantas presentan un sesgo general hacia proteínas de menor tamaño.

**Las proteínas de plantas tienen menos dominios que las proteínas de animales**

De acuerdo con los resultados mostrados en las secciones previas, las proteínas de las plantas son de menor tamaño que las de los animales y hongos. Para investigar si las proteínas de las plantas tienen menor número de dominios, consultamos la base de datos de UniprotKB (The-UniProt-Consortium, 2017) y obtuvimos los dominios INTERPRO predichos de los proteomas de *H. sapiens* y *O. sativa*. Al comprar el número de dominios encontramos que *H. sapiens* contiene más dominios por proteína (2.8) que *O. sativa* (1.7) (Prueba de Wilcox, P-valor < 0.001). Más del 35% de las proteínas de humanos contiene arriba de cuatro dominios, mientras que el valor en arroz es de solamente 16% (**Figura**

**1.9**). Similarmente, el 5% de proteínas en humanos contiene más de 8 dominios, mientras que en arroz es menos del 2%. Estos resultados concuerdan con reportes previos en los que se encontró que los eucariontes (con proteínas más largas) contienen mayor número de dominios en comparación a los procariontes (Ekman, et al., 2005). De forma similar, *C. elegans* del grupo Metazoa presentó más dominios por proteína en comparación con *S. pombe* del grupo Fungi (Koonin, et al., 2000). Al comparar la longitud de las proteínas y el número de dominios entre plantas y animales es claro que las proteínas presentan mayor longitud debido a que contienen más dominios. Una relación similar entre número de dominios y longitud de proteína fue reportada en un estudio previo que incluyó 49 especies de animales (Middleton, et al., 2010).

Como se mostró en una sección anterior, en animales la mayor longitud de proteína también está asociado a mayor número de exones. Se ha sugerido que el aumento en la longitud de las proteínas se debe tanto a un proceso de duplicación de genes y su fusión posterior (Buljan, et al., 2010) como a un proceso de adquisición de nuevos dominios (exon-shuffling) conocidos como dominios móviles (Nagy and Patthy, 2011; Patthy, 1985; Patthy, 1990). La evidencia que apoya esta última idea se encuentra en que los límites de los dominios en animales correlacionan con los límites de sus exones (Liu and Grigoriev, 2004; Liu, et al., 2005; Patthy, 1999; Rogozin, et al., 2005). Una diferencia notable entre proteínas de animales con respecto a plantas, hongos y protistas es que proteínas multidominio originadas por exon-shuffling son prácticamente inexistentes fuera de genomas de animales (Basu, et al., 2009; França, et al., 2012; Patthy, 1999). Debido a lo anterior, es de suponer que la existencia de dominios móviles codificados en exones individuales ha promovido un incremento en la longitud de las proteínas de animales con respecto a las de plantas, hongos y protistas.

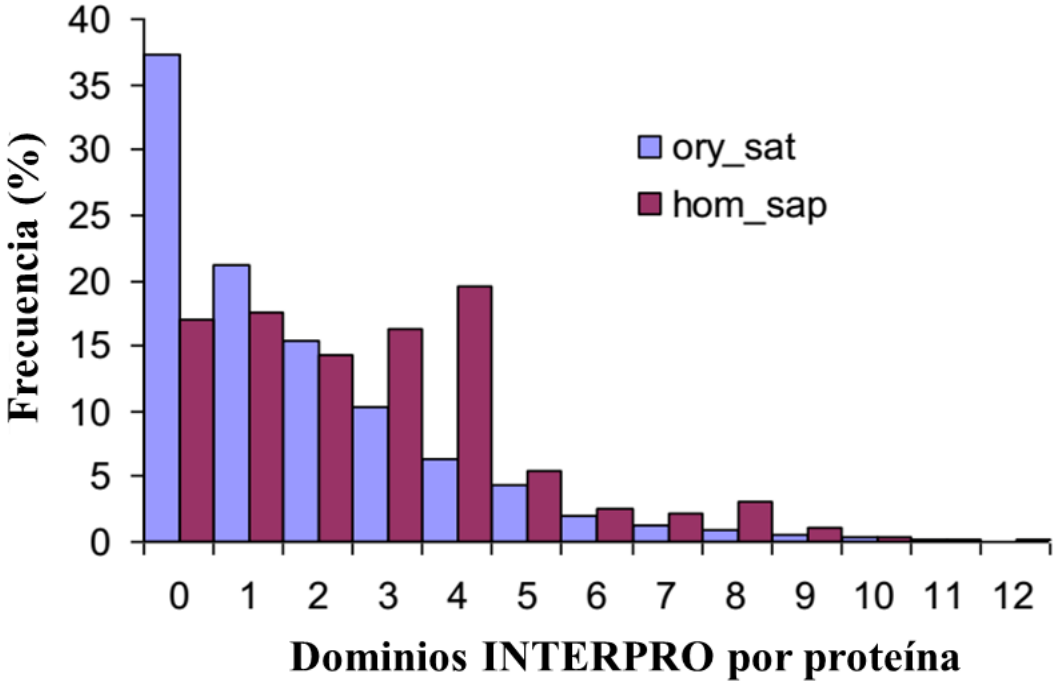



**Figura 1.9. Número de dominios INTERPRO por proteína.** Hom_sa*p* = *H. sapiens*, ory_sat = *O. sativa*.

## Conclusiones

En este trabajo encontramos que existe una gran diversidad de longitudes de proteínas entre distintos linajes de eucariontes. Por ejemplo, las proteínas de plantas son de menor longitud que las de otros eucariontes (animales y hongos). Exploramos distintas hipótesis sobre el posible origen del menor tamaño de proteínas en plantas. De acuerdo a nuestros resultados, se pueden descartar las hipótesis de migración de genes del cloroplasto al núcleo de plantas después de la endosimbiosis debido a que las proteínas que migraron son de mayor tamaño en comparación con el proteoma nuclear de *A. thaliana*. Aunque existen diferencias de longitud entre las proteínas de distintos compartimentos celulares, se descarta la hipótesis de compartimentos específicos con proteínas más grandes debido a que existe un sesgo sistemático hacia proteínas de mayor longitud en animales y hongos independientemente del compartimento en cuestión. En cambio, el número de exones está fuertemente correlacionado con la longitud de proteína en la mayoría de los grupos de eucariontes particularmente en animales. De esta manera, concluimos que las proteínas en plantas son

de menor longitud debido a que están constituidas con un menor número de exones con respecto a animales. Dado que las plantas codifican mayor número de genes en sus genomas con respecto a animales, las proteínas de animales son presumiblemente más complejas ya que éstas contienen mayor número de proteínas multidominio. Está abierto a debate si las proteínas en hongos son más o menos complejas con respecto a las proteínas de animales y plantas. El hecho es que los hongos codifican para un número mucho menor de proteínas en sus genomas con respecto a plantas y animales, pero sus proteínas son de mayor tamaño que las de plantas y contienen aún menos exones. Por otra parte, los protistas al ser un grupo polifilético presentan gran diversidad en el tamaño de sus proteínas y de acuerdo a ello se pueden clasificar en tres categorías: i) los grupos Ichthyosporea, Choanoflagellida, y Nucleariida tienen proteínas de longitud mayor al resto de eucariontes; ii) los grupos Amoebozoa, Alveolata y Stramenopila tienen proteínas de longitud intermedia; iii) mientras que las proteínas de menor longitud en eucariontes se encuentran en los grupos de Cryptophyta y Haptophyta.

# Capítulo 2. Reducción de la longitud de las proteínas y selección por proteínas grandes: dos fuerzas evolutivas opuestas actuando en genomas de endosimbiontes

## Objetivos

### General

Estudiar el efecto de la simbiosis sobre la longitud de las proteínas en el genoma de los endosimbiontes y en las proteínas que migraron al núcleo del hospedero.

### Particulares

1. Determinar la longitud de las proteínas en los genomas de organismos endosimbiontes de distintos grupos de eucariontes.
2. Identificar si existe un cambio en la longitud de las proteínas que migraron del cloroplasto al núcleo de *A. thaliana*.
3. Determinar si existe un proceso de reducción de proteínas en los genomas de los endosimbiontes.

## Métodos

### Conjunto de datos de plástidos

Para construir el conjunto de proteomas de plástidos consultamos la base de datos NCBI Refseq versión 70 (Pruitt, et al., 2007). Los archivos en formato GenBank fueron almacenados en una base de datos relacional (MySQL), previamente procesados mediante scripts de Perl propios (**Programa S2.1** y **Programa S2.2**). Este conjunto resultante contiene 64,714 secuencias pertenecientes a 746 proteomas de plástidos, agrupados en 10 grupos filogenéticos de eucariontes (**Tabla 2.1**). La lista de especies se indica en la **Tabla S2.1.**

**Tabla 2.1. Especies y proteínas del conjunto de datos de plástidos.**

| Grupo | Total de especies | % de especies | Total de proteínas |
|---|---|---|---|
| Alveolata | 13 | 1.74 | 689 |
| Chlorophyta | 57 | 7.64 | 4,700 |
| Cryptophyta | 3 | 0.4 | 375 |
| Excavata | 6 | 0.8 | 373 |
| Glaucophyta | 1 | 0.13 | 149 |
| Haptophyta | 4 | 0.54 | 446 |
| Rhyzaria | 1 | 0.13 | 61 |
| Rhodophyta | 15 | 2.01 | 2,954 |
| Stramenopila | 30 | 4.02 | 4,090 |
| Streptophyta | 616 | 82.57 | 50,877 |
| ***Total*** | **746** | **100** | **64,714** |

**Identificación de ortólogos entre cianobacterias y *Arabidopsis thaliana***

Para identificar los genes de origen endosimbiótico construimos filogenias utilizando las secuencias de *A. thaliana* y las secuencias de especies de archaebacterias, bacterias Gram positivas, cianobacterias, eubacterias, protobacterias y *S. cerevisiae* (Ver métodos **Capítulo 1**). De acuerdo a la metodología propuesta anteriormente (Martin, et al., 2002; Rujan and Martin, 2001), aquellas secuencias de *A. thaliana* que ramificaron junto con las secuencias de cianobacterias fueron seleccionadas como probables genes de origen endosimbiótico (Ver métodos, **Capítulo 1**).

**Identificación de ortólogos entre *E. coli* y endosimbiontes**

A partir del catálogo de simbiontes en la base de datos SymbioGenomesDB (http://symbiogenomesdb.uv.es/, junio de 2016) descargamos los proteomas de 310 bacterias de la base de datos NCBI. La identificación de ortólogos entre *Escherichia coli* y cada uno de los simbiontes la realizamos utilizando el programa BLAST (Altschul, et al., 1997) y el criterio de Reciprocal Best Hit (RBH) (Moreno-Hagelsieb and Latimer, 2008; Ward and Moreno-Hagelsieb, 2014). Se decidió utilizar el criterio de RBH debido a que ha demostrado dar resultados equivalentes a otros algoritmos más sofisticados como OrthoMCL (Altenhoff and Dessimoz, 2009; Kristensen, et al., 2011).

### Análisis estadístico

Las comparaciones de la longitud de las proteínas entre grupos filogenéticos las realizamos utilizando la prueba no paramétrica de Kruskal-Wallis (Kruskal and Wallis, 1952), después de lo cual realizamos las comparaciones por pares (Castellan, 1988). Los P-valores fueron corregidos utilizando el concepto de Tasa de Falsos Descubrimientos (Benjamini and Hochberg, 1995) y los grupos generados en las comparaciones múltiples fueron generados utilizando la librería de R "multcompView" (Spencer, 2012). Todos los análisis estadísticos los realizamos con el paquete estadístico R (R Core Team, 2016).

## Resultados y discusión

### Los genomas de plástidos presentan una reducción considerable en sus proteínas

Los genomas de plástidos se encuentran severamente reducidos y codifican únicamente entre 60 y 200 genes (Martin and Herrmann, 1998), esta cantidad representa solo una fracción de los aproximadamente 3000 genes que tiene una cianobacteria típica como S*ynechocystis* sp. En estos genomas, la mayoría de los genes se han perdido o migrado al núcleo de su hospedero. En *A. thaliana*, se ha reportado que entre 800 y 1700 proteínas migraron del cloroplasto al núcleo (Martin, et al., 2002; Rujan and Martin, 2001). De acuerdo a los resultados presentados en el **Capítulo 1** identificamos 1,339 de posible origen procarionte. Por tanto, es posible que el proceso de reducción genómica también tenga consecuencias reductivas en las proteínas remanentes. Para verificar esta hipótesis, calculamos la longitud de proteínas en plástidos y cianobacterias. Como resultado, el valor de la mediana de longitud en plástidos (156 aa) es significativamente menor que el de cianobacterias (257 aa) (Prueba de Wilcox, P-valor $< 2.2^{-16}$). Esta observación permite concluir que la reducción en la longitud de proteínas, junto con la pérdida masiva de genes, es también una consecuencia de la reducción genómica tras el proceso de simbiogénesis. Dicha reducción podría explicarse debido a que la reducción genómica promueve modificaciones en el genoma tales como inserciones o eliminaciones de nucleótidos que truncan el término amino o el término carboxilo y, a su vez, provocan la reducción en longitud de las proteínas (Lane, et al., 2007).

**Las proteínas de los plástidos primarios, secundarios y terciarios han alcanzado una longitud mínima**

Los cloroplastos se originaron tras el establecimiento de un procarionte fotosintético (cianobacteria) como residente intracelular permanente en un eucarionte heterotrófico, a este proceso se le conoce como endosimbiosis primaria (Margulis, 1981; Mereschkowsky, 1905). En eucariontes, el clado de Archaeplastida (compuesto por plantas, algas verdes, algas rojas y algas verdeazules) se originó por endosimbiosis primaria (Bowman, et al., 2007; Moreira, et al., 2000; Rodríguez-Ezpeleta, et al., 2005). Una endosimbiosis posterior entre un eucarionte fotosintético (p. ej. un alga roja) y otro eucarionte dio lugar a los plástidos secundarios (p. ej. Stramenopiles) (Tirichine and Bowler, 2011). Adicionalmente, la endosimbiosis entre un eucarionte con plástidos secundarios y otro eucarionte dio lugar a los plástidos terciarios (p. ej. Haptophytas) (Archibald, 2009).

En la (**Tabla S2.1**) se muestran los valores de mediana y promedio de la longitud de las proteínas para cada uno de los proteomas de plástidos. Adicionalmente, calculamos las medianas y promedios agrupando los plástidos de acuerdo a los grupos filogenéticos de eucariontes a los que pertenece su hospedero. Al realizar la prueba de comparaciones múltiples encontramos poca variación entre grupos (Prueba de Kruskal-Wallis, P-valor < 0.001) (¡Error! No se encuentra el origen de la referencia. **A**). El grupo de plástidos con mayor longitud de proteínas fue Chlorophyta (167 aa, mediana) mientras que el grupo de Stramenopila tuvo el menor valor (139 aa), la diferencia entre el máximo y el mínimo es de solamente 28 aa. La mediana del 50% de los plástidos se encuentra en un intervalo de entre 152 y 160 aa.

Por otra parte, en un estudio previo se encontró que la longitud (mediana) de los dominios, en la base de datos de dominios Pfam-A, es de 155 aa (Ekman, et al., 2005). En base a este par de observaciones, proponemos que las proteínas de plástidos han alcanzado una longitud mínima debido a limitaciones funcionales, dicho mínimo se encuentra en correspondencia con la longitud de las unidades funcionales que conforman a las proteínas (dominios). En otras palabras, la menor longitud de las proteínas se debe a la menor longitud promedio de dominio, y no al número de dominios por proteínas, como el caso

comparativo entre plantas y animales (**ver Capitulo 1**). Aunque para apoyar esta hipótesis se debe verificar que el número de dominios en proteínas de cloroplastos sea cercano a uno. En consecuencia, especulamos que una vez alcanzado el tamaño de un dominio promedio, éste se mantendrá estable. Sin embargo, si los procesos erosivos son muy fuertes y continúan, los genomas de estos plástidos entrarán en una fase de erosión progresiva de reducción de longitud y pérdida de genes que culminará con la desaparición del protema de dichos plástidos.

**Los plástidos primarios, secundarios y terciarios tienen diferente número de proteínas codificados en su genoma**

Adicionalmente, encontramos que existen diferencias significativas del total de proteínas por genoma entre distintos grupos de plástidos (Prueba de Kruskal-Wallis, P-valor < 0.001). Como se puede observar en la **Figura 2.1B** el grupo Rodophyta presentó mayor número de proteínas (204, mediana), mientras que Excavata, Rhizaria y Alveolata mostraron el menor número de proteínas (62, 61 y 32). Existen diferencias entre cloroplastos de plantas (Streptophyta), algas verdes (Chlorophyta), algas rojas (Rodophyta) y algas verdes-azules (Glaucophyta), todos ellos plástidos primarios. Las diferencias en el contenido de proteínas contrastan con las diferencias de longitudes de las proteínas entre los mismos grupos (**Figura 2.1 A**). Esto significa que el contenido de proteínas en un genoma es más dinámico y varía ampliamente en comparación con la longitud de las proteínas.

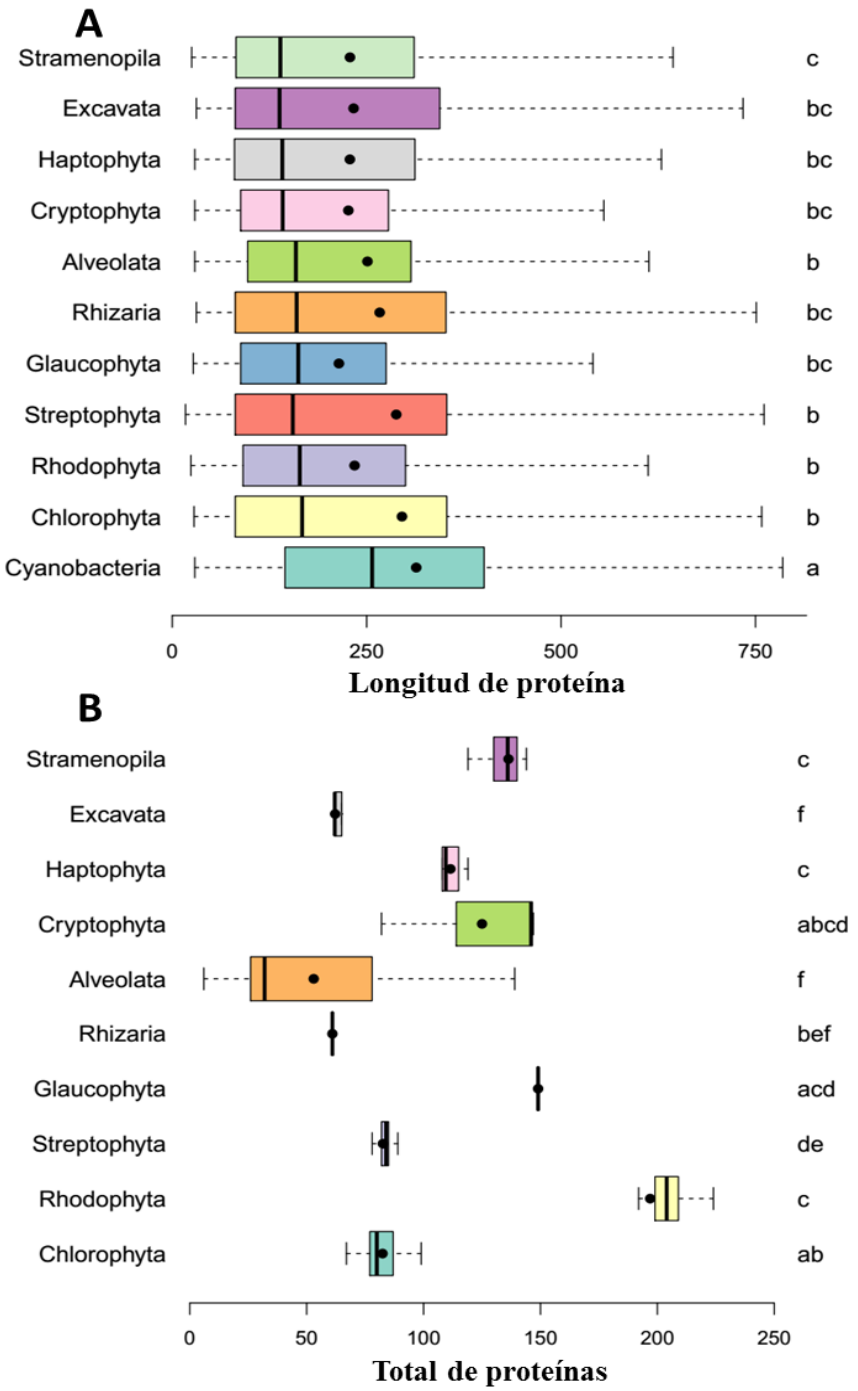



**Figura 2.1. Comparación de longitud/total de proteínas en proteomas de plástidos (organizados de acuerdo a los grupos filogenéticos de sus hospederos). A)** Longitud de proteínas en plástidos. La longitud de las proteínas de cianobacterias fue incluida por comparación. **B)** Total de proteínas en plástidos. Los puntos al interior de los diagramas de cajas y bigotes representan los valores promedios y las barras verticales los valores de medianas. Grupos con la misma letra indican que no son significativamente diferentes (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05).

**Las proteínas que migraron del cloroplasto al núcleo de *A. thaliana* han incrementado su longitud**

Una vez que las proteínas migraron del cloroplasto al núcleo de su hospedero, éstas se encuentran sujetas a las mismas presiones evolutivas características del núcleo. A la fecha, se desconoce si las proteínas de origen endosimbiótico se mantuvieron iguales o si aumentaron su longitud. Para contestar esta interrogante, identificamos aquellos genes que migraron del cloroplasto hacia el núcleo en *A. thaliana.*

Como describimos en el **capítulo 1**, obtuvimos 1,339 ortólogos entre *A. thaliana* y el panproteoma de cianobacterias (conformado por *Nostoc PCC, Prochlorococcus marinus,* y *Synechocystis PCC*). Los valores de las medianas fueron 349 aa para el proteoma nuclear de *A. thaliana*, 160 aa para el proteoma de su cloroplasto, 257 aa para el panproteoma de cianobacterias, 396 aa para los ortólogos identificados en *A. thaliana*, y 334 aa para los ortólogos identificados en el panproteoma de cianobacterias. Como se puede observar en la **Figura 2.2**, las proteínas que migraron del cloroplasto hacia el núcleo de *A. thaliana* incrementaron su longitud (~140 aa) de manera significativa (paired Wilcox test, p-value < 2.2e-16).

Una posible explicación al aumento en longitud puede ser la adición de péptidos de tránsito en el N-terminal de las proteínas que permiten redirigirlas hacia el cloroplasto. Se sabe que dichos péptidos varían considerablemente en longitud entre 20 y 150 aa (Bionda, et al., 2010; Soll and Schleiff, 2004). Sin embargo, a) no todas las proteínas de cloroplasto contienen péptidos de tránsito (Bionda, et al., 2010; Wise and Hoober, 2007), y b) muchas de las proteínas de origen cianobacteriano realizan su función en compartimentos distintos al cloroplasto (Abdallah, et al., 2000; Martin, et al., 2002). Otra posible explicación puede ser la fusión con otros genes o duplicaciones locales de sus dominios. Los mecanismos exactos mediante los cuales las proteínas, cuyos ancestros son las cianobacterias, aumentaron su longitud sigue siendo una pregunta abierta.

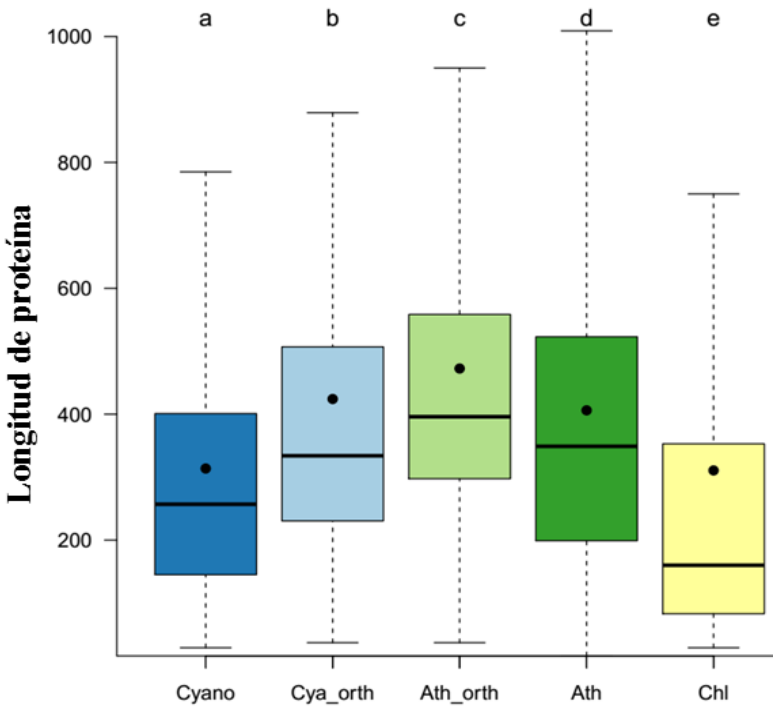



**Figura 2.2. Comparación de la longitud de las proteínas nucleares de *A. thaliana* y cianobacterias, sus ortólgos y las proteínas del cloroplasto de *A. thaliana*.** Cya_orth y Ath_orth son los ortólogos entre cianobacterias y *A. thaliana* (ver texto). Cyano = Cianobacterias (*Nostoc PCC, Prochlorococcus marinus,* and *Synechocystis PCC),* Ath = Proteoma del núcleo de *A. thaliana*. Chl = Proteoma del cloroplasto de *A. thaliana*. Los puntos internos a los diagramas de cajas y bigotes representan promedios de cada conjunto. Letras distintas indican diferencias significativas (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05).

Las fuerzas que moldean la longitud de las proteínas son diferentes entre eucariontes y procariontes (Tiessen, et al., 2012); **Capítulo 1**). Mientras que los eucariontes tienen proteínas más grandes que los procariontes, los plástidos tienen proteínas aún de menor tamaño que ambos. Contrariamente, las proteínas del cloroplasto que migraron hacia el núcleo después de la endosimbiosis primaria son de mayor tamaño que las de cianobacterias (**Figura 2.3**). No es claro si este aumento en tamaño se debe a la adquisición de dominios eucariontes, a la fusión con otras proteínas eucariontes, a la duplicación de sus propios dominios procariontes o a adquisición de más aminoácidos.

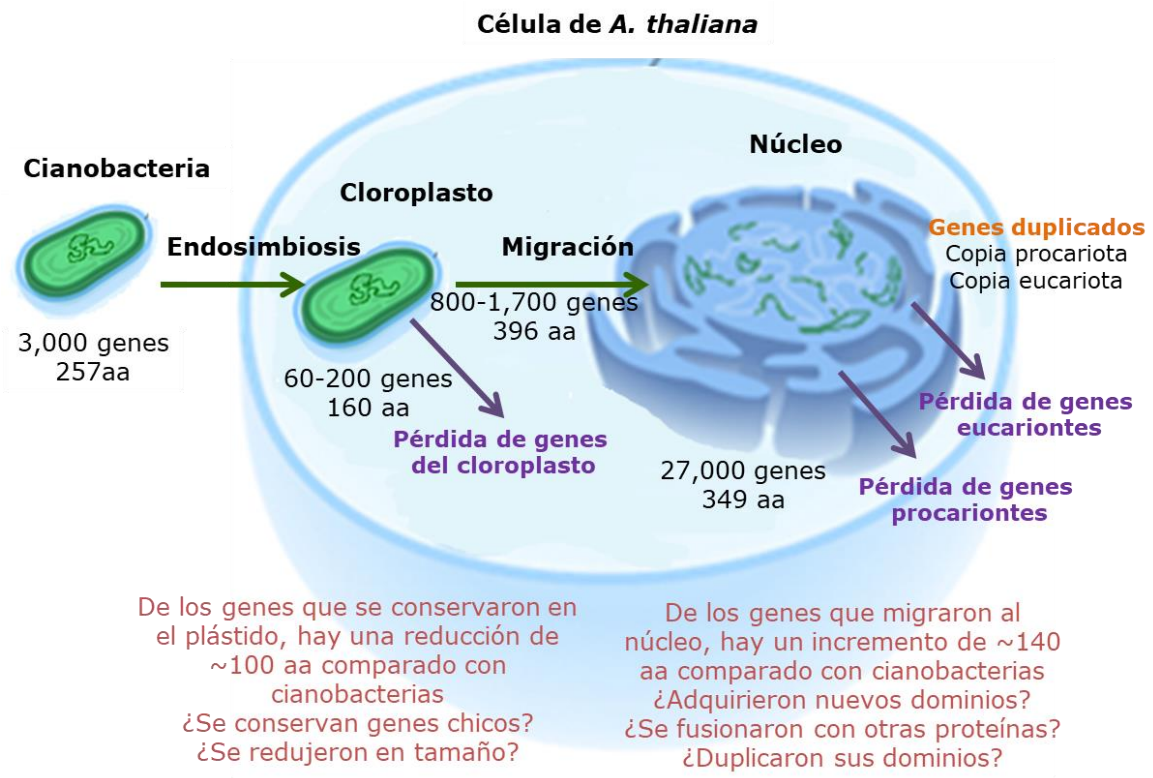



**Figura 2.3. Proceso de evolución en la longitud de proteínas de cianobacterias, después de la endosimbiosis y después de migrar del cloroplasto hacia el núcleo.** Las proteínas que se conservan en el genoma del cloroplasto son de menor tamaño con respecto a las de sus ancestros en cianobacterias. De forma contraria, las proteínas que migraron del cloroplasto al núcleo eucarionte son de mayor longitud con respecto a sus ancestros en cianobacterias. Figura modificada a partir de Kelvinsong (2013) que se encuentra bajo la licencia de Creative Commons (https://commons.wikimedia.org/wiki/File:Chloroplast_secondary_endosymbiosis.svg).

### Las proteínas que permanecen en los endosimbiontes son de mayor tamaño, pero se encuentran en un proceso de reducción

Como es bien sabido, el número de genes se reduce significativamente en los genomas de endosimbiontes debido a un proceso de erosión en el genoma. En estudios previos se ha encontrado que la longitud de proteína se ha reducido en genomas extremadamente reducidos (<300 genes) como es el caso de unos pocos endosimbiontes, mitocondrias, plástidos y nucleomorfos (Lane, et al., 2007; McCutcheon and Moran, 2012) (**Tabla 2.2**). Sin embargo, esta tendencia no ha sido reportada para endosimbiontes cuyos genomas no han experimentado reducción extrema.

Para investigar si existe una relación entre el número de genes en los genomas de simbiontes y la longitud de sus proteínas, analizamos 310 genomas de bacterias endosimbiontes. Aun cuando se ha reportado una relación casi lineal entre la reducción del genoma y la pérdida de genes (~1000 genes por cada Mbp) (McCutcheon and Moran, 2012), en este estudio no encontramos correspondencia entre el número de genes y su longitud **Figura 2.4 A** (Prueba de Spearman, valor-p = 0.9687, Información mutua = 0.17).

Con el resultado anterior, se podría concluir que no existe relación entre en número de genes y la longitud de las proteínas en genomas de endosimbiontes. Sin embargo, al aumentar la resolución de los datos (**Figura 2.4 B**) se puede observar que en genomas reducidos (>300 y <1,000 genes) la longitud de las proteínas se incrementa ligeramente por encima de la mediana general en endosimbiontes (264 aa). En contraste, para genomas extremadamente reducidos (<300 genes) se puede observar que la longitud de las proteínas se encuentra muy por debajo de la mediana general de los endosimbiontes con genomas reducidos y con genomas no reducidos.

Por lo tanto, en endosimbiontes, la relación positiva que ha sido reportada entre tamaño de genoma y número de genes (McCutcheon and Moran, 2012) no se conserva entre el número de genes y el tamaño de las proteínas. Claramente, los genomas reducidos muestran un sesgo hacia proteínas de mayor longitud. ¿Dicho sesgo tienen algún significado biológico? En trabajos previos realizados con bacterias que presentan genomas reducidos se encontró que, para compensar la pérdida de genes, las proteínas restantes adoptan nuevas funcionalidades (moonlighting) e interactúan con más proteínas formando redes más complejas en comparación con sus ortólogos en organismos de vida libre (Catrein and Herrmann, 2011; Kelkar and Ochman, 2013; Wong and Houry, 2004). Por otra parte, como mostramos en el **Capítulo 1** para el caso de organismos eucariontes, las proteínas de mayor longitud poseen mayor número de dominios que les permite participar en mayor número de funciones. En este sentido, suponemos que el sesgo hacia proteínas grandes en simbiontes con genomas reducidos se debe a la necesidad de compensar funciones por aquellos genes que se pierden debido a las fuerzas reductivas características del proceso de endosimbiosis.

Por el contrario, los genomas de endosimbiontes que han sufrido una reducción extrema (<300 genes) poseen proteínas de longitud mucho menor en comparación a los genomas

que han sufrido reducción de genoma en menor grado. La conclusión al respecto es que la reducción del genoma, además de tener como consecuencia la pérdida masiva de genes, también disminuye la longitud de las proteínas. Sin embargo, este fenómeno no es evidente debido a un sesgo de selección hacia proteínas de mayor longitud cuando la reducción no es extrema. A medida que la integración entre endosimbionte y hospedero se vuelve más profunda y el proceso de erosión del genoma continúa, la reducción de la longitud en las proteínas restantes es inevitable, muy posiblemente debido a la acumulación de inserciones o eliminaciones que truncan el marco de lectura en los genes.

**Tabla 2.2. Algunos ejemplos de genomas no reducidos, reducidos y extremadamente reducidos. Modificado de McCutcheon and Moran, 2012.**

| Organismo | Taxonomía | Tamaño genoma (Mbp) | Contenido de GC (%) | Número de proteínas | Longitud (aa)* |
|---|---|---|---|---|---|
| **Simbiontes con genomas no reducidos** | | | | | |
| *Escherichia coli* | Gammaproteobacteria | 4.6 | 51 | 4 144 | 282 |
| *Bacillus subtilis* | Mollicutes | 4.2 | 44 | 4 297 | 244 |
| **Simbiontes con genomas reducidos** | | | | | |
| *Rickettsia prowazekii* | Alphaproteobacteria | 1.1 | 29 | 834 | 283 |
| *Candidatus B. floridanus* | Gammaproteobacteria | 0.7 | 27 | 582 | 294 |
| *Wigglesworthia glossinidia* | Gammaproteobacteria | 0.7 | 24 | 631 | 279 |
| *Candidatus Baumannia cicadellinicola str. Hc* | Gammaproteobacteria | 0.7 | 33 | 591 | 285 |
| *Buchnera aphidicola (Aphis glycines)* | Gammaproteobacteria | 0.6 | 26 | 562 | 275 |
| *Mycoplasma genitalium* | Mollicutes | 0.6 | 32 | 476 | 295 |
| *Candidatus Moranella endobia PCIT* | Gammaproteobacteria | 0.5 | 44 | 411 | 271 |
| *Buchnera aphidicola Str. BCc* | Gammaproteobacteria | 0.4 | 20 | 366 | 278 |
| **Simbiontes con genomas extremadamente reducidos** | | | | | |
| *Candidatus Sulcia muelleri* | Flavobacteria | 0.3 | 23 | 224 | 329 |
| *Candidatus Zinderia insecticola CARI* | Betaproteobacteria | 0.2 | 14 | 206 | 250 |
| *Candidatus Carsonella ruddii HT* | Gammaproteobacteria | 0.16 | 15 | 180 | 218 |
| *Candidatus Hodgkinia cicadicola Dsem* | Alphaproteobacteria | 0.14 | 58 | 169 | 193 |
| *Candidatus Tremblaya princeps* | Betaproteobacteria | 0.14 | 58 | 116 | 160 |
| **Organelos con genomas grandes** | | | | | |
| *Cucurbita pepo* | Mitocondria | 0.98 | 43 | 38 | 219 |
| *Floydiella terrestris* | Cloroplasto | 0.52 | 35 | 74 | 183 |
| *Porphyra purpurea* | Cloroplasto | 0.19 | 33 | 172 | 209 |
| *Reclinomonas Americana* | Mitocondria | 0.69 | 26 | 67 | 197 |
| **Nucleomorfos y plástidos** | | | | | |
| *Guillardia theta* | Nucleomorfo | 0.55 | 45 | 485 | 232 |
| *H. andersenii* | Nucleomorfo | 0.57 | 40 | 451 | 257 |
| *Bigelowiella natans* | Nucleomorfo | 0.37 | 50 | 283 | 211 |
| *Bigelowiella natans* | Plástido | 0.07 | 30 | 61 | 160 |
| *Lotharella oceanica* | Nucleomorfo | 0.69 | 33 | 608 | 160 |
| *Lotharella oceanica* | Plástido | 0.70 | 31 | 60 | 160 |

Mbp = Pares de Mega bases, aa = aminoácidos, * valores de medianas.

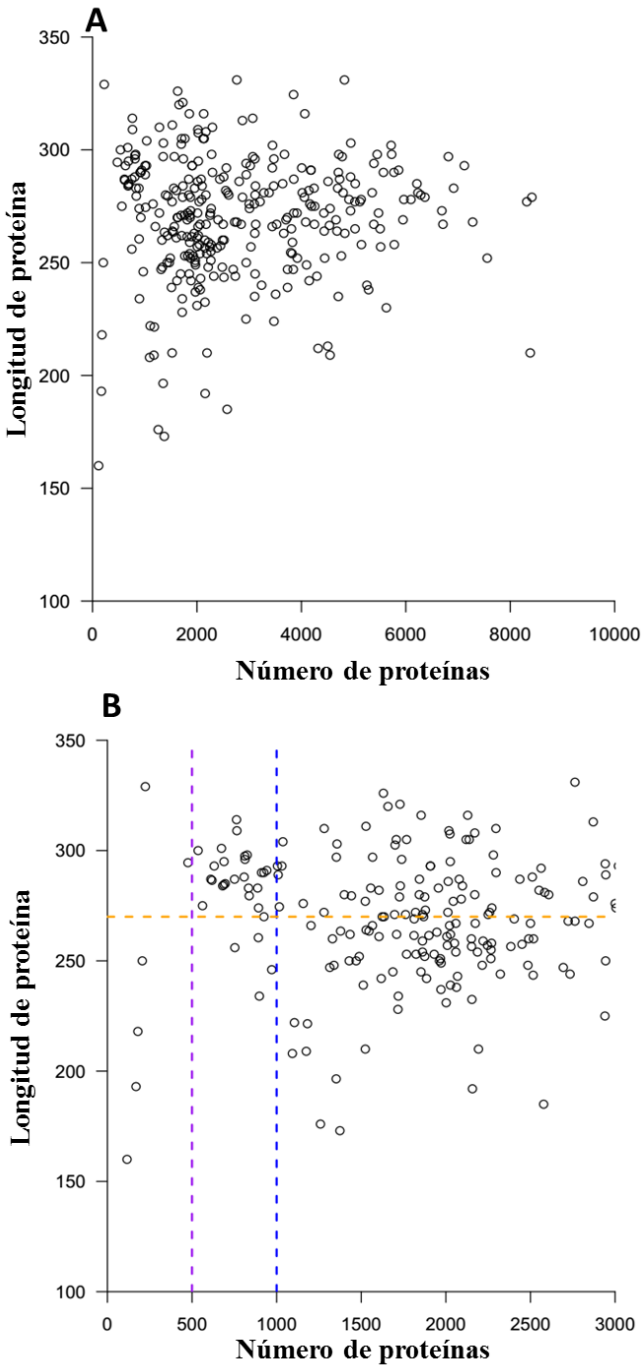



**Figura 2.4. Relación entre la longitud y el número de proteínas por proteoma en endosimbiontes.** En ambas figuras cada punto representa una especie diferente. A) Se muestran los valores longitud y número de proteínas por proteoma para 310 especies de endosimbiontes. B) Se muestran los mismos datos que en A pero con un intervalo de número de proteínas entre 0 y 3000 aa. El valor de la línea horizontal (264 aa) representa la mediana de longitud de las proteínas en endosimbiontes. Las líneas verticales solo se graficaron como apoyo visual.

**El 72% de los genes en endosimbiontes se encuentran reducidos con respecto a sus ortólogos en el genoma no reducido de *E. coli*.**

Una manera alternativa de comprobar si las proteínas que conservan los simbiontes están reduciendo su tamaño es obtener los ortólogos con respecto a otra bacteria que no presente reducción de su genoma. Para ello, obtuvimos los ortólogos entre *Escherichia coli* y cada una de las 315 especies del conjunto de endosimbiontes. Los valores de tamaño de proteoma, número de ortólogos, medianas de longitud y P-valores de cada especie están disponibles para el lector interesado (**Tabla S2.2**). Es importante señalar que, aunque *E. coli* es una bacteria comensal, ésta no ha sufrido reducción en su genoma y su longitud mediana de proteínas es de 282 aa, que es comparable con la mediana global en bacterias (267 aa). Al comparar la longitud entre ortólogos, encontramos que 62% de los proteomas de simbiontes están reducidos con respecto a *E. coli* (Prueba de Wilcox para datos pareados, P-valor < 0.05). Más aun, muchos de los genomas de los endosimbiontes analizados no presentan reducción genómica. De manera que, al considerar solamente a simbiontes con menos de 4,144 genes (tomando como referencia a *E. coli*) encontramos reducción de longitud en 72% de los endosimbiontes (**Figura 2.5**). Con estos resultados concluimos que sí existe reducción en las proteínas de simbiontes pero que no es evidente debido a que proteínas de longitud mayor son seleccionadas para compensar la pérdida de genes.

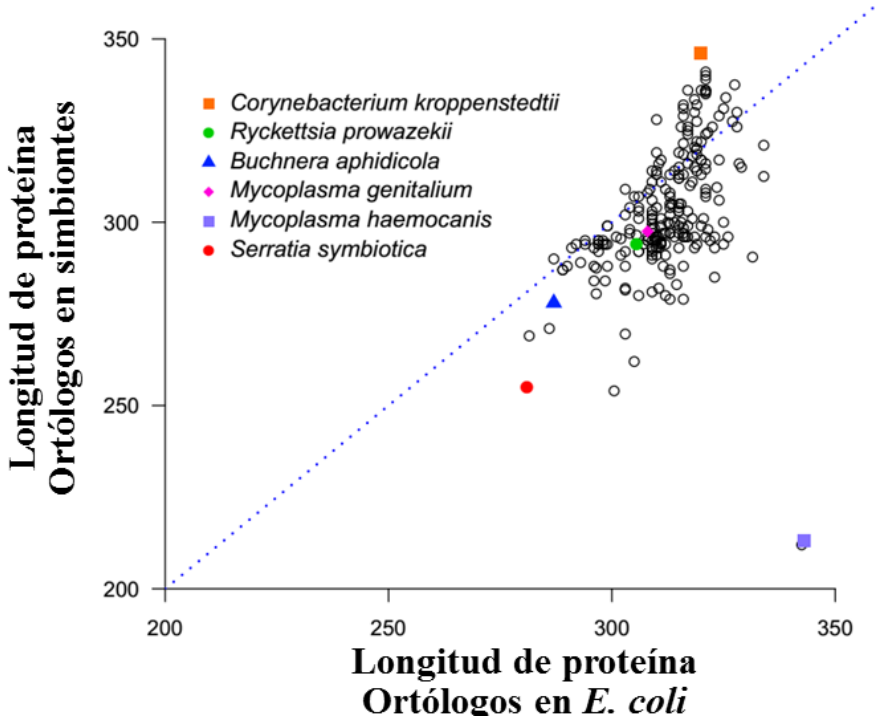



**Figura 2.5. Comparación de la longitud de las proteínas de *E. coli* y sus ortólogos en endosimbiontes con genoma reducido.** Los puntos debajo de la diagonal representan especies de endosimbiontes con proteínas ortólogas a *E. coli* que han sufrido reducción en su longitud de proteínas. Los puntos arriba de la diagonal representan especies de simbiontes con proteínas ortólogas a *E.coli* que han incrementado su longitud. Se puede ver que hay más puntos debajo que arriba de la diagonal.

**La falta de splicing alternativo en proteínas de organismos procariontes podría ser la causa de su limitado tamaño con respecto a las proteínas de eucariontes**

El hecho de no haber encontrado una correlación entre el número de genes y su longitud utilizando las 310 especies de endosimbiontes confirma la observación previa que en procariontes la relación entre número y tamaño de proteínas es ligeramente positiva y diferente que en eucariontes donde la relación es negativa (**Figura 1.4**) (Tiessen et al., 2012). Esto pudiera significar que en eucariontes la fusión y fisión de proteínas es un mecanismo evolutivo más frecuente que en procariontes. Las bacterias absorben ADN ajeno y agregan proteínas a su genoma, pero también pierden proteínas sin afectar el tamaño de las restantes, es decir, no es común la fusión de proteínas. La **Figura 2.3** muestra que en endosimbiontes el número de proteínas se reduce sin afectar mucho su tamaño promedio. Mientras que las proteínas bacterianas cuando migran al núcleo, si cambian su tamaño, posiblemente por fusión con ciertos ORFs que le agregan dominios y

funcionalidades a proteínas. Es tentador especular que la razón por la cual las proteínas eucariotas si evolucionan por fusión-fisión se debe precisamente a los procesos de splincing de mARN mientras que las bacterias que no tienen splicing, no pueden evolucionar ni aumentar su tamaño por medio de splicing. Esto a su vez, implica que existe un mecanismo evolutivo del ADN que no solo se genera a través de replicación de ADN -> ADN sino a través de la transcripción ADN-ARN y posterior síntesis de cADN e incorporación del híbrido duplex ARN/cADN al cromosoma nuclear compuesto de ADN solamente. Posiblemente existe un mecanismo ADN -> preARN que por medio de transplicing y transcripción reversa da origen a un nuevo gen y a una nueva proteína fusionada. Existen varios casos documentados sobre el origen de nuevos genes por un mecanismo de transcripción reversa (Ejima and Yang, 2003; Long, et al., 2003; Wang, et al., 2006), de ellos el más estudiado ha sido el gen jingwei en Drosophila (Long and Langley, 1993; Long, et al., 1999). Es decir, en ciertos casos el ADN de una generación hija no se generó del ADN de la célula madre sino del pre-ARN-mARN de la generación hija (**Figura 2.6**).

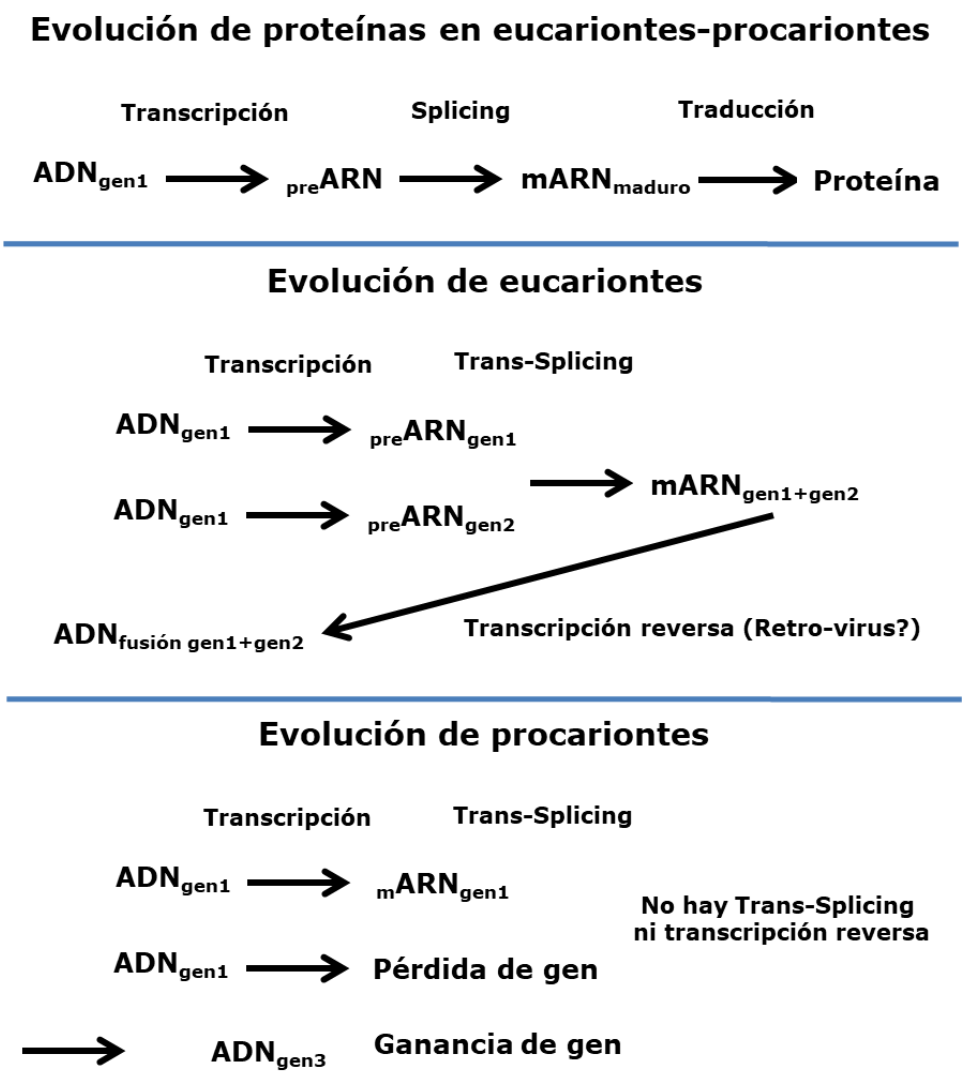



**Figura 2.6. Proceso propuesto de evolución del tamaño en proteínas de eucariontes y procariontes.** La maquinaria de splicing en conjunto con la de transcripción reversa les permite a eucariontes incrementar el tamaño de sus proteínas (ver texto).

## Conclusiones

Durante la endosimbiosis, los genes originales de la cianobacteria (~3000) se quedan en el plastido (~200) o se pierden de su genoma (~2800). Su función puede ser reemplazada por proteínas eucariontes, o bien, los genes migran al núcleo y reemplazan su copia original del plástido. Una vez en el núcleo, las proteínas siguen evolucionando tal vez por la exposición a los mecanismos nucleares de trans-splicing y transcripción reversa. Los datos que obtuvimos indican que hay una reducción en el tamaño de proteínas en el genoma residual del plástido, mientras que los genes que migraron al núcleo tienen una tendencia a

incrementar su tamaño. La mayoría de simbiontes presentan un proceso de compactación de sus genomas lo que también implica la pérdida de genes. En este trabajo encontramos que dicha pérdida es selectiva hacia genes de mayor longitud. Sin embargo, los genes que permanecen en el simbionte comienzan a acumular mutaciones que en muchos de los casos ocasionan la reducción del marco de lectura. Es interesante ver como dos fuerzas evolutivas actúan en direcciones opuestas en los genomas de endosimbiontes. Mientras que genes de mayor longitud tienden a ser seleccionados durante etapas tempranas de endosimbiosis, su progresiva reducción es inevitable a medida que acumulan mutaciones que truncan su marco de lectura. Por otra parte, aquellos genes que migran hacia el núcleo del hospedero aumentan de longitud aunque las fuerzas evolutivas que promovieron dicho aumento no están claras.

## Perspectivas

1) Para investigar los posibles mecanismos mediante los cuales las proteínas, que migraron del cloroplasto al núcleo eucarionte de las plantas, aumentaron su longitud después de migrar hacia el núcleo del organismo eucarionte es necesario realizar alineamientos entre las proteínas provenientes de las cianobacterias y sus 1339 ortólogos que actualmente residen en el núcleo. Con estos alineamiento sería posible determinar en qué partes están insertados los aminoácidos adicionales. En un alineamiento estas inserciones saldrían como gaps entre las dos secuencias proteicas. Con estos alineamientos se pueden contestar las siguientes preguntas: ¿En dónde hay más gaps? ¿En el término-amino, a la mitad de la proteína o en el término-carboxilo? Si las inserciones son primordialmente en el término-amino tal vez sea por el péptido señal para dirigir la proteína al plástico. Si las inserciones son a la mitad o en el término-carboxilo, tal vez sea por la adición de dominios regulatorios a la proteína. Una posibilidad más es que las inserciones sean aleatorias a lo largo de la proteína como consecuencia de un mecanismo de selcción neutral.
2) Para investigar la vía por la cual las proteínas de origen endosimbionte que permanecen en el plástido se reducen, es necesario también realizar alineamientos entre las proteínas originales de las cianobacterias y sus ortólogos más pequeños

que actualmente todavía residen en el genoma plastidico. Con estos alineamiento sería posible determinar en qué partes están deletados los aminoácidos que en un alineamiento aparecen como gaps entre las dos secuencias proteicas. Con estos alineamientos se pueden contestar la pregunta: ¿Hay más indels en el término-amino, a la mitad o en el término-carboxilo o en las regiones flexibles entre dominios?

3) Dos hipótesis de cómo los genes se pueden duplicar en un genoma son: a ) Por medio de ADN que se copia por replicación (DNA polimerasa) en otra parte del genoma, y b) ésta hipótesis implica la transcripción y un intermediario duplex preARN/cDNA que después se inserta en el genoma y se convierte en una hebra doble de ADN/ADN. Para distinguir entre ambos procesos, sería necesario hacer alineamientos entre exones pero también entre intrones de genes duplicados. ¿Qué tanto se conservan los intrones de genes duplicados? ¿Existe un mecanismo de inserción de intrones en genes nuevos que se generaron por transcripción reversa de un preRNA?
4) Sería interesante contar el número de exones y de intrones que tienen los genes de origen cianobacteriano comparados con los demás genes eucariontes. Bajo la hipótesis del mecanismo de trans-splicing y transcripción reversa, dichos genes "nuevos" deberían presentar menor número de exones en comparación con el promedio del número de exones de los genes antiguos del núcleo.
5) Sería relevante contar el número de exones e intrones en genes relativamente nuevos (de reciente duplicación) comparados con genes más antiguos. También sería interesante contar el número de exones e intrones en genes de plantas comparados con los mismos genes ortólogos de animales. Resulta interesante que los animales tienen menos copias de un mismo gen, mientras que las plantas tienen familias de parálogos más numerosos. Las plantas tienen más genes en total pero proteínas más chicas (ver resultados Capitulo 1). Esto coincide con el hecho de que las plantas tienen menos exones mientras que los animales tienen más exones por cada gen. ¿Significa que los genes duplicados de plantas "son más nuevos" y por eso tienen menos exones que los genes de animales que son "más antiguos"?

¿Significa que primero se inicia con un gen duplicado sin intrones, y durante la evolución se le van agregando intrones a la secuencia? ¿Permite la adición de un intrón hacer que la proteína pueda ser más grande?

6) Otro experimento interesante sería estudiar una endosimbiosis artificial (Agapakis, et al., 2011). Por ejemplo, quitarle su cloroplasto y reemplazarlo por una cianobacteria nueva usando microcapilares de inyección celular. Cultivar esta alga hibrida por cientos de generaciones y observar cómo se va perdiendo el ADN de la cianobacteria y cuales proteínas van cambiando.
7) Otra pregunta interesante por contestar es: de las proteínas que migraron del plástido al núcleo eucarionte, ¿cuántas se mantienen duplicadas entre el genoma eucarionte y el genoma del plástido? Para contestarla se tiene que un análisis de ortología entre genes de *A. thaliana*, genes de plástidos y genes de otros eucariontes.

# Capítulo 3. Predicción funcional de proteínas huérfanas de *Arabidopsis thaliana* utilizando ensambles de SVMs

## Objetivos

### Objetivo general

Desarrollar un método de predicción de función independiente de homología para la anotación de proteínas.

### Objetivos particulares

1. Desarrollar un ensamble de SVMs en base a los resultados de los algoritmos de selección de características.
2. Seleccionar los alfabetos reducidos más prometedores para la construcción de ensambles.
3. Desarrollar un conjunto de clasificadores a partir de ensambles SVMs que permitan discriminar entre las siguientes categorías funcionales:
    - Proteínas de unión a ADN.
    - Actividad de Transporte.
    - Actividad Enzimática.
    - Diferentes localizaciones celulares.
    - Diferentes familias de transportadores.
    - Diferentes familias de enzimas.
4. Anotar funcionalmente el conjunto de proteínas denominadas “huérfanas” en el proteoma de *A. thaliana*.
5. Desarrollar un sitio web donde se implemente el método desarrollado.

## Planteamiento del problema

Las proteínas son fundamentales para la vida, están formadas por cadenas lineales de aminoácidos y son las biomoléculas más diversas y versátiles. Éstas, están presentes en prácticamente todos los procesos biológicos de un organismo, incluyendo sus funciones vitales como crecer, reproducirse y reaccionar ante estímulos externos.

Entender el papel biológico que desempeña cada uno de los miles de genes que constituyen a cada genoma es una tarea que desafía a la ciencia actual. Conocer la función de un gen implica determinar distintos aspectos del mismo como por ejemplo: en qué parte de la célula desempeña su actividad, cuál es su plegamiento tridimensional, cuál es su actividad

bioquímica, en qué procesos biológicos participa, cuál etapa de desarrollo o qué estímulos del ambiente afectan su expresión (Petsko, 2004).

Dados los retos para estudiar la función de proteínas en laboratorio y la facilidad con que se secuencian más genomas, desde hace años ha ido en aumento una brecha entre los genes que han sido secuenciados y aquellos de los cuales que se conoce su función (Liolios, et al., 2010; Radivojac, et al., 2013). Sin embargo, en comparación con la experimentación, los métodos basados en modelos computacionales son mucho más baratos y rápidos de implementar.

La técnica computacional más usada para asignar función es el alineamiento de secuencias. A través del alineamiento, dos o más secuencias son comparadas entre sí para determinar si existe homología entre ellas. Por definición, dos o más secuencias (de nucleótidos o aminoácidos) son homólogas cuando se derivaron del mismo ancestro (Fitch, 1970). Por lo general, secuencias homólogas muestran la misma o similar función biológica. En términos sencillos, cuando dos secuencias comparten un alto grado de similitud (típicamente >80%) tienen alta posibilidad de ser homólogas. Se han propuesto varios algoritmos de alineamiento entre los que destaca el algoritmo heurístico de BLAST (Altschul, et al., 1990) debido a su velocidad en tiempos de computo.

En base a los principios de homología y alineamiento de secuencias, se han desarrollado numerosas herramientas y bases de datos para identificar motivos (PRINTS) (Attwood, et al., 2003), firmas de aminoácidos (PROSITE) (Sigrist, et al., 2010), dominios (Pfam, PROSITE) (Sigrist, et al., 2010; Sonnhammer, et al., 1997), entre otras. Es común transferir la función de las proteínas caracterizadas experimentalmente a sus homólogos que han sido identificados mediante estas herramientas (Benso, et al., 2013; Dorden and Mahadevan, 2015).

Sin embargo, una década atrás se mostró que para más de la mitad de las secuencias reportadas no se han detectado homólogos caracterizados experimentalmente, y por tanto permanecen anotadas con las etiquetas de “función desconocida”, “hipotéticas” y “putativas” (Hawkins and Kihara, 2007). En la actualidad dicho problema permanece vigente, por ejemplo, en uno de los genomas mejor estudiados como *A. thaliana*, alrededor

de 8,000 secuencias (30%) permanecen anotadas con la etiqueta de función desconocida (análisis propio no publicado).

Debido a que gran cantidad de proteínas no presentan plegamientos, motivos ni dominios conocidos, se desconoce la función de la gran mayoría de estos (Arendsee, et al., 2014). La falta de homología con otros genes implica que el uso de cualquier herramienta bioinformática basada en homología resulte inviable para su caracterización, por lo que la caracterización funcional de dichas proteínas representa un reto importante tanto para la bilogía experimental como para la bioinformática.

Debido a la gran cantidad de secuencias para las cuales no es posible identificar homólogos, se han propuesto diversos métodos que no dependen del alineamiento de secuencias para la asignación de función. Dichos métodos provienen del campo conocido como aprendizaje de máquina o también llamado aprendizaje estadístico.

Los métodos de anotación no basados en homología tienen en común la generación de covariables predictoras (también llamadas características) a partir de la estructura primaria de las proteínas. Dichas características pueden ser simples: tales como usar la frecuencia de monómeros y pares de aminoácidos o sustituir el alfabeto de 20 aa por alfabetos reducidos (Peterson, et al., 2009). O más elaboradas: construidas a partir de las propiedades físico-químicas de los aminoácidos (Li, et al., 2006). En general, a pesar de ser fáciles de obtener, el uso de este tipo de características han demostrado su utilidad para predecir la función de gran cantidad de categorías funcionales (Chou, 2005; Ebrahimie, et al., 2011; Liu, et al., 2013; Qiu, et al., 2014; Yuan, et al., 2010).

Desde el punto de vista del aprendizaje de máquina, la tarea de asignar función (anotar) a las proteínas corresponde a un problema clasificación cuyo objetivo es identificar a qué categoría (de un conjunto de categorías o subpoblaciones) pertenece una nueva observación. Al algoritmo encargado de realizar dicha tarea se le conoce como clasificador. Para construir un clasificador se utiliza un conjunto de datos de entrenamiento formado por instancias (observaciones) cuya categoría es previamente conocida. Los problemas de clasificación representan un campo del aprendizaje de máquina conocida como aprendizaje supervisado debido a que se dispone de un conjunto de instancias correctamente identificadas (Abu-Mostafa, et al., 2012; Hastie, et al., 2009).

Las instancias que forman parte del conjunto de datos de entrenamiento están compuestas por una seria de propiedades cuantificables llamadas variables explicativas o características.

Muchos de los algoritmos de clasificación pueden ser descritos en términos de una función lineal que asigna una puntuación a cada categoría al combinar de forma lineal las características con un vector de coeficientes. La categoría predicha será aquella que obtenga mayor puntuación. Este tipo de función es conocida como función lineal de predicción y tiene la siguiente forma general:

$$score(X_i, k) = \beta_k X_i ,$$

Donde $\boldsymbol{x_i}$ es el vector de características de la instancia *i*, $\boldsymbol{\beta_k}$ es el vector de coeficientes correspondientes a la categoría *k*, y $score(X_i, k)$ es la puntuación asociada con la asignación de la instancia *i* a la categoría *k*. A los algoritmos que usan una función lineal se les conoce como clasificadores lineales. Ejemplos de tales algoritmos son: regresión logística, el algoritmo del perceptrón, análisis discriminante lineal y máquinas de soporte vectorial (SVM).

Debido a que el aprendizaje de máquina es un campo muy dinámico que implementa mejoras constantemente, el presente trabajo propone un nuevo algoritmo independiente de homología basado en aprendizaje de máquina para la predicción de diversas categorías funcionales. El algoritmo propuesto combina un sistema de codificación (frecuencias de monómeros, dímeros-*n* y ventanas-*n*), el uso de alfabetos reducidos, selección de características y ensambles de SVMs.

## Marco teórico

### Máquinas de Soporte Vectorial (SVM)

Las SVMs son un tipo específico de algoritmo de aprendizaje de máquina que debido a su robustez han sido utilizadas en diversos problemas de clasificación como detección de correos no deseados, clasificación de texto, reconocimiento de escritura, reconocimiento de rostros y objetos, entre otros.

Las Máquinas de Soporte Vectorial se basan en el principio de minimización estructural de riesgo de acuerdo con la teoría de aprendizaje estadístico propuestas por (Vapnik, 1995).

Fueron diseñadas como una herramienta para resolver problemas de clasificación binaria. Dado un conjunto de datos pertenecientes a dos clases mutuamente excluyentes (etiquetadas +1 y -1), el objetivo del algoritmo SVM es representar los ejemplos (datos de entrenamiento) como puntos en un espacio *D*-dimensional y encontrar un hiperplano tal que de un lado contenga todos los datos de la categoría *y = +1*, mientras que el lado opuesto contenga los datos de la categoría *y = -1* (**Figura 3.1**¡Error! No se encuentra el origen de la referencia.).

Aunque pueden existir infinito número de hiperplanos separadores, el algoritmo de SVM identifica aquel que maximiza la distancia entre categorías. Asumiendo que los ejemplos son separables linealmente, según el lado en el que se encuentren los puntos del hiperplano, se cumplirá:

$$\pi_1: \boldsymbol{w^T x_i} + b \geq +1$$
$$\pi_2: \boldsymbol{w^T x_i} + b \leq -1,$$

donde $\pi_1$ y $\pi_2$ son los hiperplanos que delimitan a la región + y la región – respectivamente, ambos son paralelos y la distancia entre ellos $(2/\ ||\boldsymbol{w}||\ ||\boldsymbol{w}||)$ es la norma euclideana de *w*. De todos los posibles hiperplanos ($\pi_1$, $\pi_2$) se busca maximizar la distancia $2/\ ||w||$ entre ellos, o equivalentemente minimizar $||\mathbf{w}||^2/2$, esto es:

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} ||\boldsymbol{w}||^2$$
$$s.a.\, y_i(\boldsymbol{w^T x_i} + b) - 1 \geq 0, i = 1, \ldots, n$$

Los vectores de soporte son aquellos ejemplos que definen los hiperplanos de separación $\pi_1$ y $\pi_2$. Una vez identificado el hiperplano óptimo, el algoritmo puede hacer predicciones verificando de qué lado se ubican los nuevos datos en base a la función discriminante:

$$h(x) = sign(\boldsymbol{w^T} X + b),$$

donde *w* es un vector de coeficientes, y *b* la ordenada al origen.

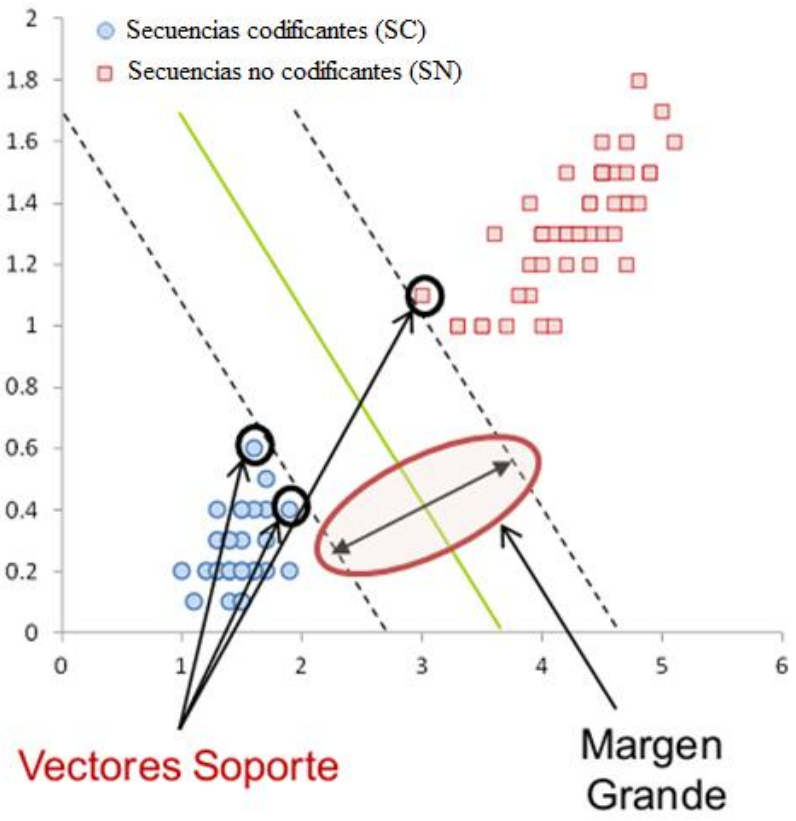



**Figura 3.1. Máquinas de soporte vectorial (SVM).** Las SVMs buscan aquel hiperplano (verde) que maximiza la distancia entre dos categorías (rojo) a partir de los vectores de soporte (círculos negros). En este caso, los ejemplos positivos [+1] están representados como SC y los ejemplos negativos [-] como SN. Figura modificada a partir de Estrada-Perea y Mera Banguero (2015) (https://es.scribd.com/document/91845845/SVM-maquinas-de-vectores-de-soporte)

En la mayoría de los problemas de clasificación los datos no son linealmente separables. Para resolver este problema, las SVMs implementan el llamado "truco del kernel", el cual involucra reemplazar el vector de productos de la formulación original con una nueva función no lineal (kernel). Como resultado, los puntos son comparados en un espacio vectorial de alta dimensión (denominado espacio de características) donde sí es posible la separación lineal. Los dos kernels más usados son el kernel polinomial $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x} \cdot \boldsymbol{x}')^d$ y el kernel radial o Gaussiano $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-c\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$.

En el campo de la bioinformática, el uso de SVMs se encuentra ampliamente extendido. Algunos ejemplos de categorías funcionales en las que las SVMs han sido utilizadas con éxito incluyen: interacciones proteína-proteína (Zhang, et al., 2014), localización celular (Chou and Shen, 2010; Hasan, et al., 2017; Sperschneider, et al., 2016; Yu, et al., 2006)

categorías funcionales (Li, et al., 2016; Saraç, et al., 2010), proteínas de unión al ADN (Kumar, et al., 2009; Kumar, et al., 2007; Lin, et al., 2011; Liu, et al., 2015), predicción de canales de iones (Lin and Ding, 2011) y proteínas de unión a lípidos (Bakhtiarizadeh, et al., 2014), entre muchas otras.

***SVM* multicategoría**

Los clasificadores SVM son de naturaleza binaria, es decir, solo pueden separar entre dos categorías (Vapnik, 1995). Sin embargo, muchos problemas de clasificación incluyen más de dos categorías. Para resolver este problema se construyen varios clasificadores binarios y posteriormente se integran para obtener una decisión final (Mohamed, 2005). Existen dos mecanismos de integración que a continuación se describen:

a) *Uno-contra-uno*: consiste en entrenar $k(k\text{-}1)/2$ clasificadores binarios, donde $k$ es el número de categorías y cada clasificador es entrenado con un par de categorías a la vez. Supongamos que se desea construir un clasificador para 6 categorías, entonces se tienen que entrenar 15 clasificadores binarios y la decisión final se obtiene por un esquema votación mayoritaria.

b) *Uno-contra-resto*: consiste en entrenar un clasificador por categoría utilizando los ejemplos de dicha clase como positivos y los ejemplos de las demás clases son colapsados en una sola categoría y se toman como ejemplos negativos. Una vez entrenados los $k$ clasificadores, para clasificar una nueva muestra la decisión final se toma considerando la categoría que obtuvo la mayor probabilidad. La principal ventaja es una reducción en la complejidad pues solo se tienen que generar tantos clasificadores como categorías en el problema a resolver. La desventaja es la creación de categorías desbalanceadas y, en consecuencia, se generan clasificadores sesgados hacia las categorías con mayor número de ejemplos. A pesar de existir una gran variedad de soluciones al desbalance como remuestreo (sobremuestreo y submuestreo) muchas veces se utilizan junto con ensambles (bagging y boosting), costo-sensibles, modificación del kernel y modificación en el umbral de decisión (He and Garcia, 2009; He and Ma, 2013); en la práctica se prefiere utilizar el método de *uno contra uno* (Meyer, et al., 2014).

**Ensambles de SVMs**

Los ensambles se basan en construir un conjunto de clasificadores para posteriormente combinar sus decisiones mediante algún método. Se han desarrollado bases teóricas sólidas que demuestran que el desempeño de los ensambles es mejor que el de los clasificadores individuales usados para su construcción (Zhou, 2012). Sin embargo, existen dos restricciones para los clasificadores individuales: i) desempeño superior al 50% y ii) diversidad (independencia) entre los errores que cometen.

En los ensambles, los errores cometidos por los clasificadores individuales son eliminados por algún mecanismo de ensamblaje; la mayoría son variantes de voto mayoritario (*bagging*), voto ponderado (*boosting*) e implementación de un nuevo clasificador para procesar las salidas de los clasificadores individuales (*stacking*) (Zhou, 2012) (**Figura 3.2**).

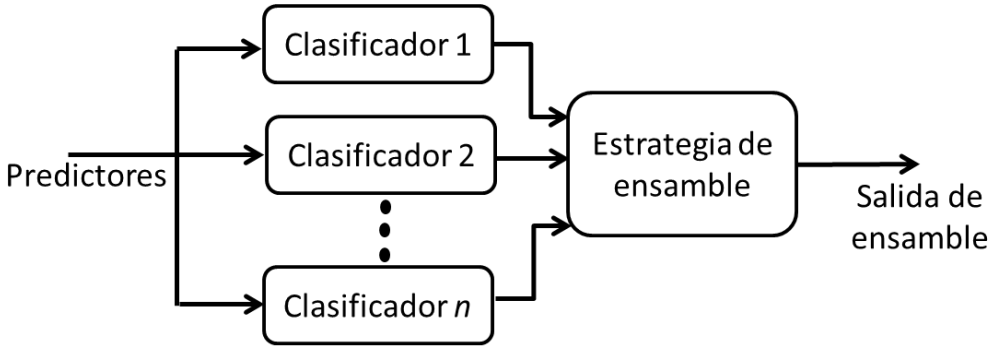



**Figura 3.2. Ensambles de clasificadores.** Los ensambles se basan en consensar las decisiones de los varios clasificadores individuales a través de un método como *bagging*, *boosting* o *estaking*.

**Codificación de secuencias**

Para generar un modelo de clasificación usando métodos de aprendizaje estadístico, se necesita de un número fijo de atributos o características tomados de las propiedades de la estructura primaria de las proteínas. Sin embargo, debido a que la cantidad de aminoácidos que componen una secuencia varía enormemente entre 20 y 25 mil (Petsko, 2004), es necesario utilizar un sistema de codificación que comprima y represente la información que contienen las secuencias a través de un número fijo de características.

Las secuencias de proteínas pueden ser representadas a través de un vector de longitud 20 al calcular la frecuencia de cada aminoácido de la secuencia. Además, para capturar información referente al ordenamiento original de aminoácidos, la frecuencia de pares o

dímeros de aminoácidos también es incorporada generando un vector de $20^2 = 400$ pares distintos. De manera similar, la frecuencia de tripletes de aminoácidos o trímeros incorporaría más información referente al orden de aminoácidos en la secuencia mediante un vector adicional de $20^3 = 6000$. Sin embargo, se sabe que un elevado número de atributos con respecto al número de ejemplos produce un sobreajuste en los modelos que se refleja en una baja capacidad de predicción. A este problema se le conoce en la literatura como "la maldición de la dimensión". Por este motivo, no es recomendable el uso de tripletes como predictores durante la construcción de los modelos (Saeys, et al., 2007).

*Dimeros-n*

Como alternativa a los trímeros, información con respecto al orden de aminoácidos puede ser capturada calculando la frecuencia de "dímeros-*n*" (Yu, et al., 2006). Este tipo de dímeros se generan al ignorar cierto número de aminoácidos entre pares. En otras palabras, dimeros-1 son pares que contienen un aminoácido cualquiera intermedio, dimeros-2 tienen dos aminoácidos cualesquiera intermedios, etc. Como se puede observar en la ¡Error! No se encuentra el origen de la referencia., en este sistema los dimeros-0 corresponden a los dímeros simples previamente descritos.

**Tabla 3.1. Representación de dímeros-*n*.**

| Nombre | Representación |
|---|---|
| Monómero | A |
| Dímero-0 | AA |
| Dímero-1 | AxA |
| Dímero-2 | AxxA |

*Ventanas-n*

Además de la codificación por dímeros-n, con la incorporación información presente a nivel local, las secuencias pueden ser divididas *n* subsecuencias de igual tamaño (ventanas-*n*) y posteriormente cada ventana es codificada usando las frecuencias de monómeros.

*Alfabetos reducidos*

En los alfabetos reducidos, aquellos aminoácidos que comparten propiedades similares de acuerdo a determinado criterio (peso, tamaño, hidrofobicidad, polaridad, etc.) pueden ser

vistos como equivalentes entre sí. Por ejemplo, aminoácidos polares, neutrales e hidrofóbicos generan un alfabeto de 3 letras (A = RKEDQN, B = GASTPHY, C= CVLIMFW). El cálculo del número de características para cada alfabeto se puede obtener de la siguiente forma:

$$C = 1 + a + (3a^2) + (2a + 3a + 4a),$$

donde *C* = número de características, 1 corresponde a la longitud de secuencia y *a* = tamaño de alfabeto. Por ejemplo, en el caso mencionado de un alfabeto de 3 letras: *C* = 1 longitud + 3 monómeros + 3 x 9 dímeros-(0, 1 y 2) + (6+9+12) ventanas-(2, 3 y 4) = 58.

La ventaja de usar alfabetos reducidos es la incorporación de trímeros. El número de características para cada alfabeto se calcula de la siguiente forma:

$$C = a + l\,(3a^2) + (2a + 3a + 4a) + a^3\ ,$$

Por ejemplo, para un alfabeto de tamaño 10 tenemos *C* = 10 + 1 + 300 + 90 + 1000 = 1401.

De esta manera, la cantidad de características puede ser reducida significativamente manteniendo al mismo tiempo información referente al orden.

**Selección de características**

En ocasiones, algunas de las características utilizadas para entrenar los modelos pueden ser redundantes o no significativas, por lo que pueden ser eliminadas sin pérdida considerable de información (Hastie, et al., 2009). En consecuencia, se han desarrollado distintos métodos de selección de características que tienen el objetivo de seleccionar un subconjunto de características relevantes o informativos para los modelos (Saeys, et al., 2007). Acoplar estos métodos durante la construcción de los modelos tiene dos ventajas considerables a tomar en cuenta: disminución de los tiempos de entrenamiento y disminución del riesgo de sobreajuste (Hastie, et al., 2009).

Una manera de seleccionar aquellas características relevantes al modelo es utilizando métodos de regresión penalizada (James, et al., 2014). Dichos métodos imponen una penalización sobre el valor de los coeficientes y provocan que algunos de ellos se contraigan hasta tomar valores de cero. De esta manera, aquellos coeficientes diferentes de cero son conservados para formar el subconjunto de características relevantes. En problemas de clasificación (de dos clases) se puede aplicar el método de penalización

LASSO sobre los coeficientes utilizando el modelo de regresión logística, donde el objetivo es maximizar:

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[\boldsymbol{y_i}(\beta_0 + \boldsymbol{\beta^T x_i}) - log\left(1 + e^{\beta_0 + \boldsymbol{\beta^T x_i}}\right)\right] - \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Donde, $\boldsymbol{y_i}$ toma valores de 0 para una clase y 1 para la otra, $\boldsymbol{x_i}$ representa las características o predictores, $\beta_0$ y $\boldsymbol{\beta}$ son los coeficientes del modelo y $\lambda$ es el parámetro que controla la cantidad de contracción en los coeficientes.

Desde el punto de vista bayesiano se asume que el vector de coeficientes $\boldsymbol{\beta}$ posee alguna distribución *a priori*, digamos $p(\boldsymbol{\beta}^*)$, donde $\boldsymbol{\beta}^* = (\beta_0, \beta_1, \dots, \beta_p)^T$. La verosimilitud de los datos puede ser descrita como una función $f(\boldsymbol{y}|X, \boldsymbol{\beta})$, donde $X = (x_1, \dots, x_p)$. La distribución a *posteriori* se puede obtener multiplicando la distribución a priori por la verosimilitud, que de acuerdo con el teorema de Bayes toma la siguiente forma:

$$p(\boldsymbol{\beta}|X, \boldsymbol{y}) = f(\boldsymbol{y}|X, \boldsymbol{\beta})p(\boldsymbol{\beta})$$

Asumiendo que $p(\boldsymbol{\beta}) = \prod_{j=1}^{p} g(\beta_j)$, para alguna función de densidad *g*, la penalización tipo LASSO es un caso especial en la cual *g* sigue una distribución doble-exponencial (Laplace). Otras distribuciones para *g* pueden ser consideradas: Gaussiana, t-escañada y BayesC (Pérez-Rodríguez and de los Campos, 2014). Entre dichas distribuciones, la doble-exponencial y BayesC pueden ser utilizadas para realizar selección de características.

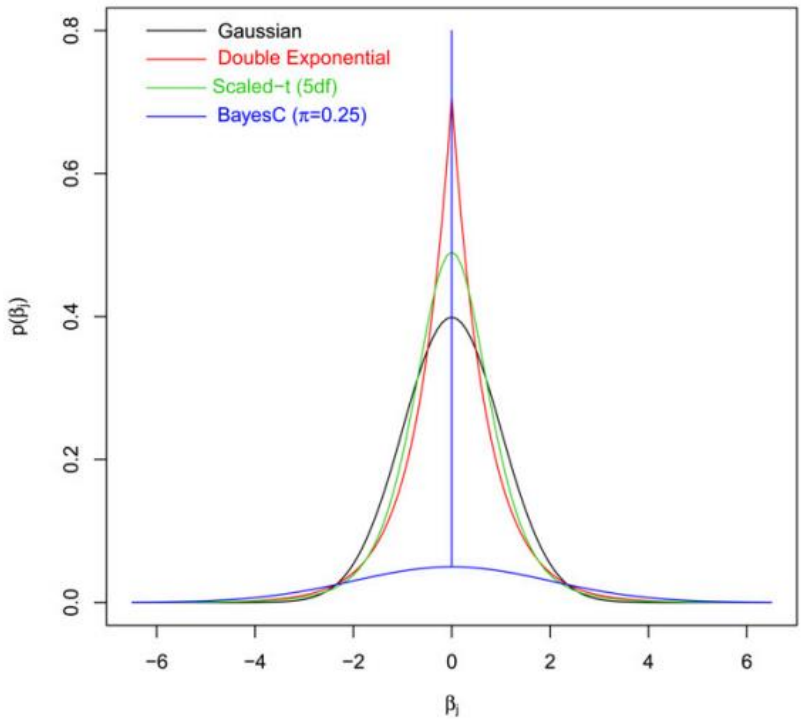



**Figura 3.3. Distintas densidades a *priori* de los coeficientes de regresión utilizadas en el Modelo Bayesiano Lineal Generalizado (Pérez-Rodríguez and de los Campos, 2014).**

### Validación cruzada

Para conocer el desempeño de un clasificador es necesario distinguir entre el error de entrenamiento y el error de prueba. El primero se obtiene al medir el error de las predicciones del modelo con respecto a los mismos datos que se usaron durante el entrenamiento. Mientras que el segundo tipo de error se obtiene usando datos no vistos previamente por el clasificador (datos independientes) durante el entrenamiento. Es fundamental conocer el error de prueba ya que éste provee una estimado de la capacidad de generalización que tendrá el clasificador ante datos que no ha visto antes (Abu-Mostafa, et al., 2012). Es decir, el error de prueba nos dice que tan bueno o malo será nuestro clasificador cuando se le presenten nuevos datos.

Para estimar el error de prueba la técnica más usada es la de validación cruzada de *k* iteraciones (Hastie, et al., 2009). Dicha técnica consiste en dividir los datos disponibles para el entrenamiento en *k* subconjuntos de igual tamaño. Para cada iteración, uno de los *k* subconjuntos se utiliza como conjunto de prueba y el resto (*k*-1) como conjunto de entrenamiento (**Figura 3.4**). Este procedimiento es repetido *k* veces y el error de prueba se

obtiene calculando la media aritmética de las *k* iteraciones. El valor de *k* más comúnmente usado en la literatura es de 5.

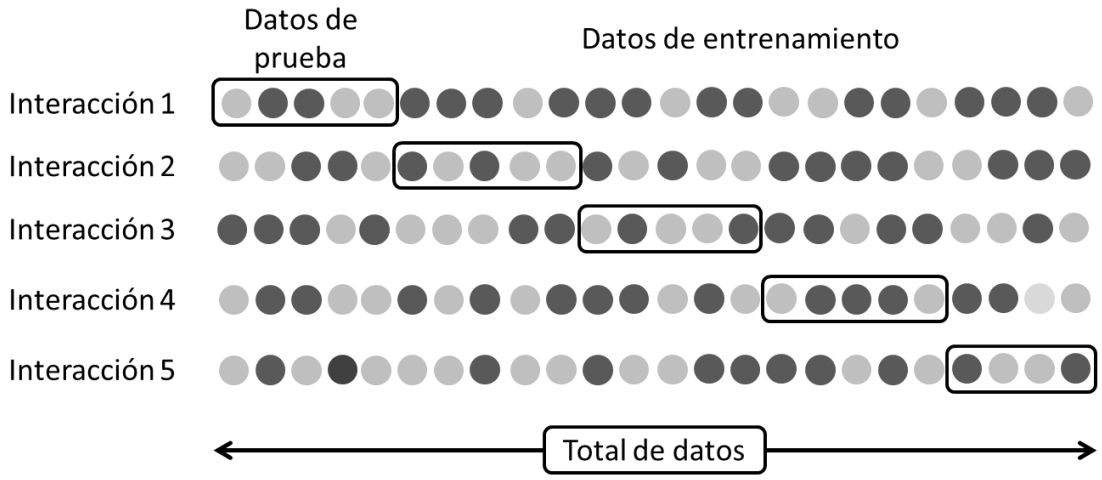



**Figura 3.4. Validación cruzada de k-grupos.** Para estimar el desempeño de los clasificadores sobre datos no usados durante el entrenamiento, el total de datos son divididos en *k* partes aproximadamente iguales. Durante cada iteración, una parte *k* es usada como datos de prueba y el resto es usado para entrenamiento del clasificador.

## Descripción del método propuesto

### Descripción global del algoritmo

En primer lugar, cada secuencia es traducida a un alfabeto diferente. Posteriormente, para obtener vectores de longitud fija, cada secuencia es codificada en un vector de características que incluye monómeros, longitud, ventanas-(2, 3 y 4), dímeros-(0, 1 y 2) y trímeros (para alfabetos de tamaño $\leq$ 12). En el siguiente paso se realiza la selección de características para cada alfabeto (este paso se describe en detalle en la siguiente sección). En base a las características seleccionadas, para cada alfabeto se construye un clasificador utilizando el algoritmo SVM. Finalmente, las predicciones generadas por cada SVM son utilizadas como características durante el entrenamiento de un nuevo clasificador SVM. Este último clasificador SVM actúa como ensamble y tiene la función de generar las predicciones definitivas (**Figura 3.5**).

### Selección de características

Una vez que cada secuencia es codificada en un vector de características que incluye monómeros, longitud, ventanas-(2, 3 y 4), dímeros-(0, 1 y 2) y trímeros (para alfabetos de

tamaño ≤ 12), utilizando el algoritmo BayesC se obtiene la probabilidad de que cada característica sea distinta a cero. Posteriormente, se inicializa un punto de corte (PC) en 0.3 y aquellas características con probabilidad mayor que PC son utilizadas para entrenar un clasificador SVM. El modelo SVM generado en este paso es evaluado mediante validación cruzada. El PC es incrementado gradualmente hasta el valor de 1, disminuyendo al mismo tiempo el número de características a evaluar, hasta encontrar la exactitud (por sus siglas en inglés ACC) máxima (**Figura 3.6**). Las características que generan mejor ACC son utilizadas para entrenar el modelo que será utilizado durante la construcción del ensamble.

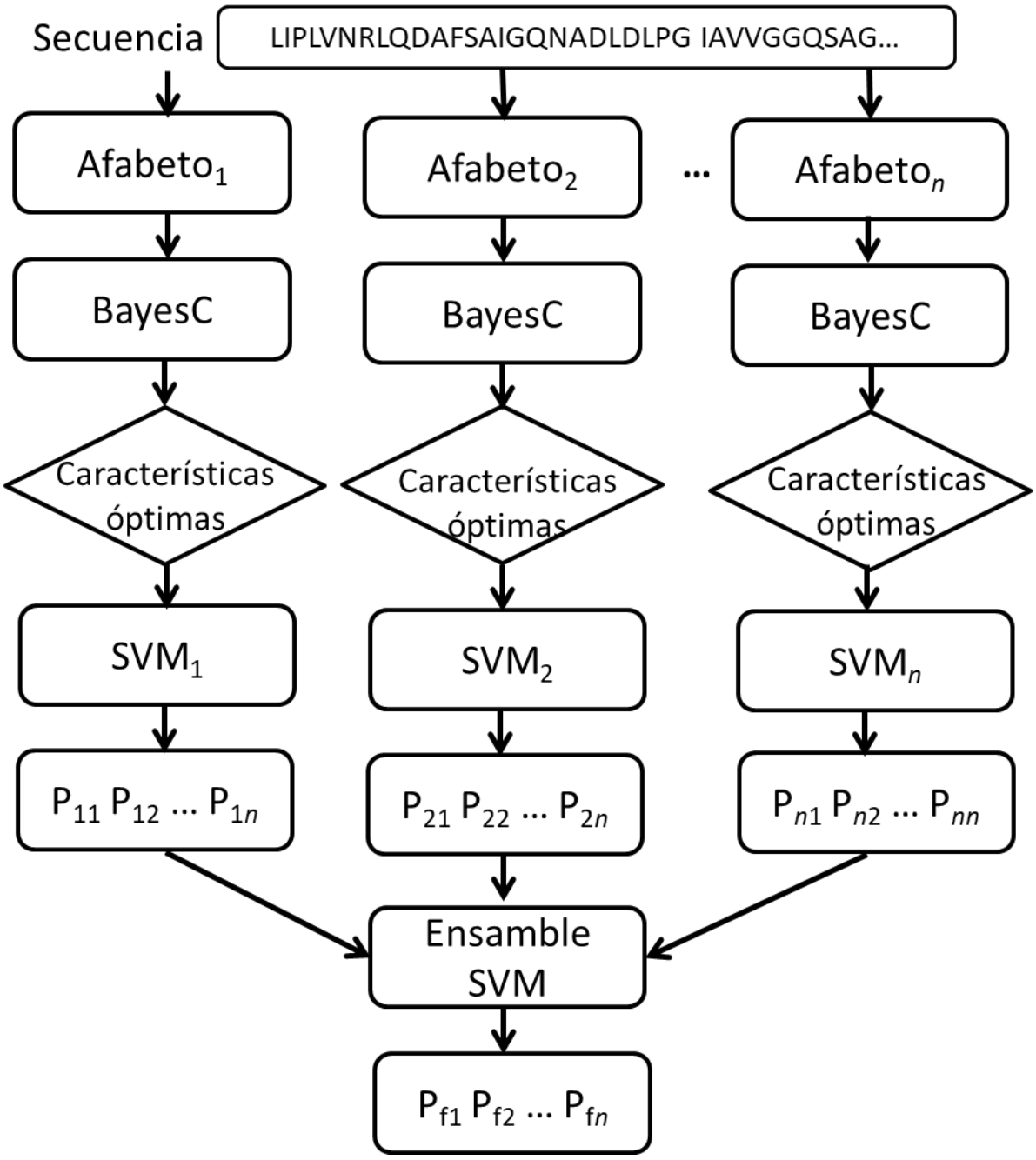



**Figura 3.5. Algoritmo utilizado para la construcción de ensambles.**

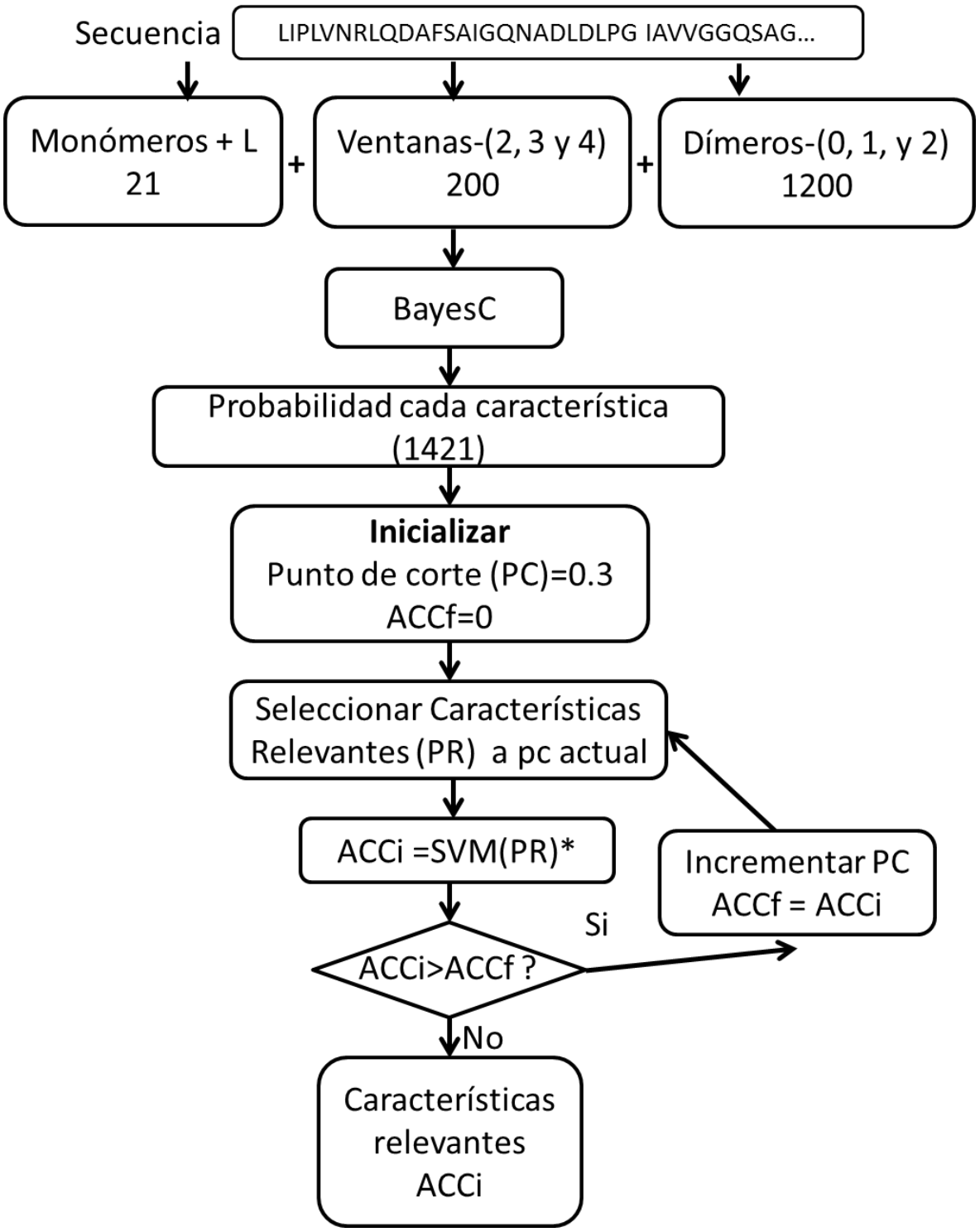



**Figura 3.6. Algoritmo utilizado para la selección de características.**

# Metodología

## Conjuntos de datos de referencia

Un paso clave durante el desarrollo de nuevos clasificadores es construir conjuntos de datos confiables y precisos. Si alguno de los datos usados para entrenar y probar los modelos fuera incorrecto, el clasificador obtenido contendrá errores y la información generada a partir de éste no sería útil.

Para el desarrollo del algoritmo de clasificación utilizamos varios conjuntos de datos de referencia previamente publicados, curados manualmente y construidos bajo estrictos criterios durante la selección de secuencias. En estos conjuntos de datos todas las secuencias anotadas como "fragmentos" o de longitud menor a 50 aa fueron removidas y el nivel de redundancia entre secuencias fue reducido por debajo del 80% utilizando herramientas como CD-HIT (Li and Godzik, 2006).

Para realizar este trabajo utilizamos 4 conjuntos de datos tomados de la literatura: unión a ADN, localización celular, familias de transportadores y familias de enzimas. A continuación se describen brevemente.

a) Proteínas de unión a ADN. Este conjunto de datos está formado por 156 secuencias de proteínas de unión a ADN y 250 secuencias de no-unión a ADN y fue publicado en 2007 (Kumar, et al., 2007). Dicho conjunto fue construido a partir de la base de datos PDB (versión junio-2007) y la identidad entre secuencias es menor al 25%.
b) Actividad catalítica y familias enzimáticas. El conjunto de datos usado en este trabajo, publicado por (Shen and Chou, 2007), contiene 9,832 secuencias de enzimas obtenidas de la base de datos ENZYME (versión mayo-2007), y 9850 secuencias de no-enzimas tomadas aleatoriamente de la base de datos de SWISSPROT (versión 52, marzo-2007). Las secuencias de enzimas en este conjunto se encuentran distribuidas en 6 familias de la siguiente manera: hidrolasas (2791), isomerasas (518), ligasas (776), lyasas (379), oxidoreductasas (1618) y transferasas (3450). En este conjunto, el nivel de redundancia es menor al 40% entre secuencias de cada familia.
c) Actividad de transporte y familias de transportadores. Éste conjunto, publicado en 2014 (Mishra, et al., 2014) con un nivel de identidad menor al 70%, está compuesto por 900 secuencias de transportadores y 600 secuencias de no-transportadores tomados de la

base de datos SWISSPROT (versión marzo-2013). Las secuencias de las 7 familias de transportadores en este conjunto están divididas de la siguiente manera de acuerdo a su sustrato: oligopéptidos (85), aniones (72), cationes (269), electrones (70), ARNm (85), azúcares (72) y otros (220).

d) Localización celular. Este conjunto fue publicado por Park y Kanehisa en el año 2003 y ha sido usado en posteriores publicaciones como punto de comparación. Dicho conjunto de datos contiene 7,589 secuencias pertenecientes a 12 compartimentos celulares: cloroplasto (671), citoplasma (1245), citoesqueleto (41), retículo endoplásmico (RE) (114), extracelular (862), aparato de Golgi (48), lisosoma (93), mitocondria (727), núcleo (1,932), peroxisoma (125), membrana plasmática (1,677) y vacuola (54). El nivel de redundancia en este conjunto es menor al 80%.

**Conjunto de análisis (genes huérfanos de *Arabidopsis thaliana*)**

El conjunto de genes huérfanos que utilizamos para este trabajo comprende 1,786 secuencias de *A. thaliana,* las cuales fueron identificadas en un trabajo previo (Donoghue, et al., 2011).

**Codificación de secuencias**

Para este trabajo codificamos las secuencias de aminoácidos usando monómeros y dímeros-0,1 y 2. Además, dividimos las secuencias en ventanas-2,3 y 4. Adicionalmente, utilizamos 172 alfabetos reducidos, de los cuales, 162 pertenecen a una compilación previa (Peterson, et al., 2009) y cuatro más pertenecen a otras fuentes (Dubchak, et al., 1999; Yu, et al., 2006).

**Selección de características**

Para selección de características utilizamos el modelo de regresión bayesiana recientemente implementado en el paquete BGLR (Bayesian Genaralized Linear Regresion) por (Pérez-Rodríguez and de los Campos, 2014). La distribución a *priori* para los coeficientes fue BayesC. Utilizamos la función liga probit debido a que es un problema de clasificación. El número de iteraciones iniciales descartadas y el número total de iteraciones fueron: Burn-in = 2000 y nIter = 5000, respectivamente.

**Máquinas de Soporte Vectorial (SVM)**

Durante la etapa de construcción y evaluación de clasificadores utilizamos el paquete e1071 (Meyer, et al., 2014) de R (R Core Team, 2016). Para determinar los parámetros C y gamma utilizamos la implementación heurística interna del paquete.

**Medidas de desempeño**

Para evaluar el desempeño de un clasificador la medida comúnmente usada es la exactitud o ACC (por sus siglas en inglés). Dicha medida es el resultado de dividir el número de ejemplos clasificados correctamente entre el total de ejemplos. Sin embargo, es común en bioinformática que las categorías se encuentren bastante desbalanceadas, por lo que usar el valor de exactitud para evaluar el desempeño presenta desventajas. Por tanto, otras medidas como el coeficiente de correlación de Matthews (MCC) y el área bajo la curva (AUC) han sido propuestas (Fawcett, 2006; Matthews, 1975; Powers, 2011). A continuación se describen algunas de ellas:

$$Sensibilidad(SN, TPR) = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$Especificidad(SP, TNR) = \frac{TN}{N} = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{P + N}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Donde P = ejemplos positivos, N = ejemplos negativos en los datos, TP = verdaderos positivos (ejemplos positivos clasificados correctamente), TN = verdaderos negativos (ejemplos negativos clasificados correctamente), FP = falsos positivos (ejemplos negativos clasificados como positivos), FN = falsos negativos (ejemplos positivos clasificados como negativos).

Por otra parte, el área bajo la curva ROC (AUCROC pero comúnmente solo abreviada AUC) se calcula a partir de graficar TPR contra FPR y variar el umbral de decisión. Valores cercanos a 0.5 se obtienen cuando el clasificador en cuestión da resultados aleatorios, mientras que un valor de 1 se obtiene cuando dicho clasificador no comete ningún error.

### Implementación del sitio web

Durante la etapa de implementación en el servidor web se utilizó el módulo scikit-learn (Pedregosa, et al., 2011) basado en Python. Tanto e1071 como scikit-learn usan la implementación de LIBSVM (Chang and Lin, 2011).

El sitio web lo implementamos en un servidor local a partir de una plantilla base y utilizando para ello los lenguajes HTML, JavaScript, PHP y Python.

## Resultados y discusión

### Incorporación de dímeros-*n* y ventanas-*n* en el clasificador SVM

En la **Figura 3.7** se muestran los valores de ACC obtenidos por el clasificador SVM en los seis conjuntos de datos descritos previamente. En la primera barra de cada gráfico se pueden observar los valores de ACC obtenidos al utilizar 20 características, que corresponden a los 20 monómeros del alfabeto completo (aa20). La segunda barra corresponde a los valores de ACC obtenidos al utilizar dímeros-0, que corresponden a los 400 dímeros del alfabeto aa20. Los valores de ACC que se muestran en la tercera barra fueron obtenidos al utilizar 1,401 características: 1 longitud de secuencia, 20 monómeros, 400 dímeros-0, 400 dímeros-1, 400 dímeros-2, 40 monómeros en ventanas-2, 60 monómeros en ventanas-3 y 80 monómeros en ventanas-4.

De manera intuitiva, se espera que al adicionar información de orden (dímeros-*n*) así como información local (frecuencia de monómeros al dividir la secuencia en subsecuencias de igual tamaño (ventanas-*n*)) "ayude" al clasificador a distinguir mejor entre categorías. De acuerdo con los resultados que obtuvimos, al adicionar esta información el valor de ACC se incrementó para los conjuntos de actividad enzimática, localización celular, familias de transportadores y familias de enzimas pero disminuyó ligeramente para los conjuntos de unión a ADN y actividad de transporte.

En trabajos pioneros de predicción de función en base a algoritmos de clasificación se utilizaron como características las frecuencias de monómeros y dímeros. Posteriormente, el concepto de dímeros-*n* se aplicó para una variedad de categorías. No obstante, el concepto de ventanas-*n* ha sido poco explorado. En este trabajó se implementaron todos estos

conceptos para construir una matriz de 1,401 características que en 4 de 6 conjuntos de datos mejoraron la capacidad de predicción de los clasificadores.

**Selección de características**

Aunque el adicionar la codificación de dímeros-*n* y ventanas-*n* puede aportar información útil al clasificador, es muy probable que tal adición incorpore ruido o información redundante y tengan un impacto negativo. Para mitigar este problema, en este trabajo utilizamos un método de selección de características basado en un modelo de regresión bayesiana para datos binarios.

Dicho método no realiza la selección de características de forma implícita. En otras palabras, no devuelve una lista con aquellas características que deben ser incluidas en el modelo. En su lugar, por cada característica devuelve una probabilidad (*P*) de estar incluida en el modelo. La selección del número óptimo de características se obtiene al aplicar el algoritmo descrito previamente (ver métodos). Con dicho algoritmo, obtuvimos distintos subconjuntos de características de acuerdo a su valor *P*, calculado previamente por BayesC. Los subconjuntos de características son utilizados para entrenar clasificadores SVM que son evaluados por validación cruzada. El subconjunto de características óptimas es aquel que genere el mayor valor de ACC en el clasificador SVM.

Al incorporar información de dímeros-*n* y ventanas-*n*, obtuvimos 1,401 características con el alfabeto aa20. Al realizar la selección de características obtuvimos un incremento promedio de 16% en el valor de ACC en los seis conjuntos, con respecto al utilizar únicamente monómeros (**Figura 3.7**). De forma individual el incremento fue de 10%, 7 %, 7%, 14%, 30% y 25% para los conjuntos de unión a ADN, actividad de transporte, actividad enzimática, localización celular, familias de transportadores y familias de enzimas, respectivamente (**Tabla S3.3-S3.8**). Claramente, al aplicar la selección de variables sobre dímeros-*n* y ventanas-*n* se obtienen clasificadores con mejor capacidad predictiva.

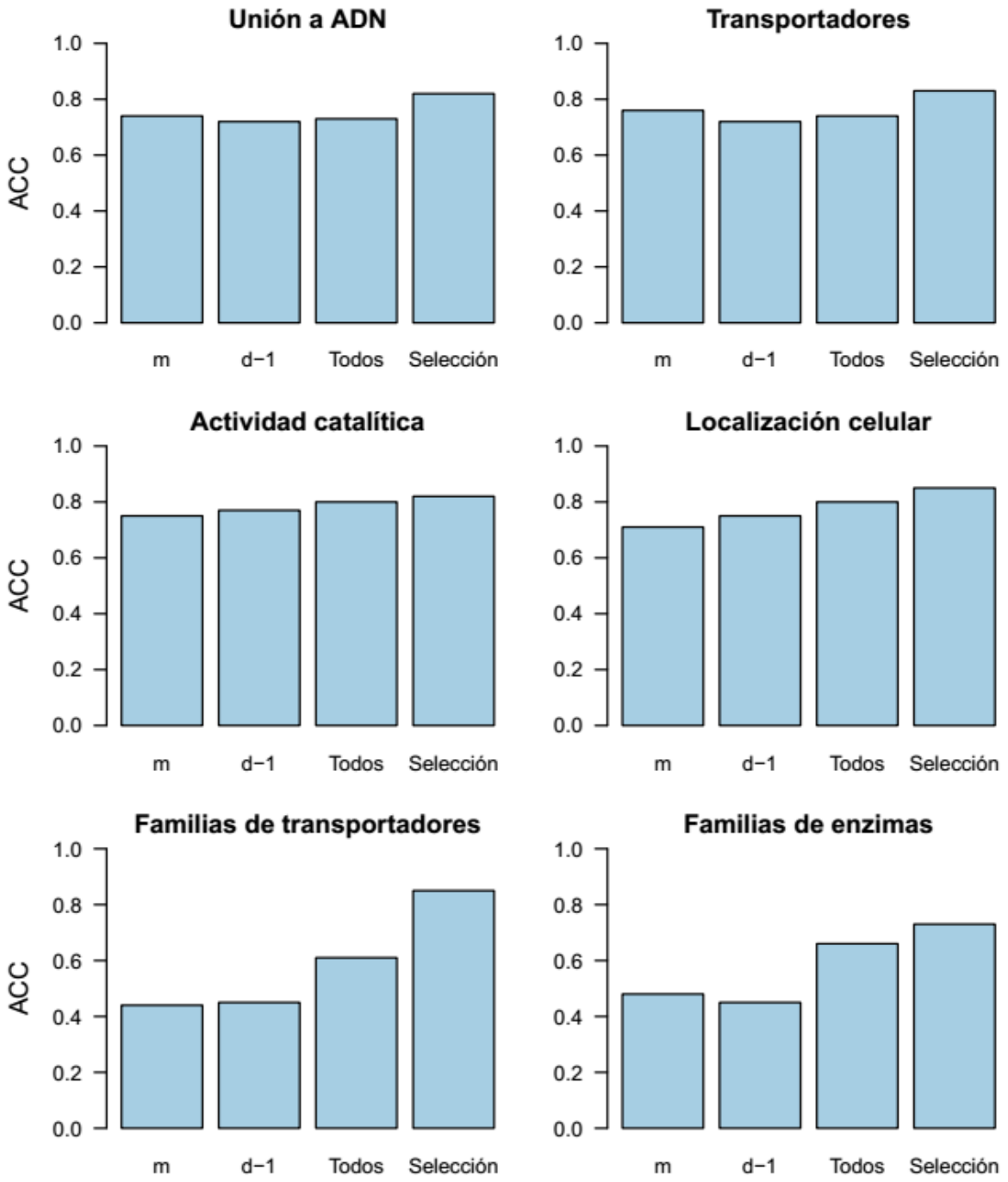



**Figura 3.7. Exactitud de SVM utilizando el alfabeto aa.20 en 6 conjuntos de datos.** m = Monómeros + longitud. d-1 = dímeros-1. Todos= m + ventanas-(2,3,4) + dímeros (1,2,3). Selección = Predictores seleccionados a partir de BayesC. ACC =exactitud.

### Comparación del método de selección de características

El método propuesto en este trabajo se utilizó en un conjunto de datos construido previamente por (Nakariyakul, et al., 2012), el cual consiste en 915 proteínas termófilas y

793 proteínas mesófilas. Las secuencias de este conjunto de datos las obtuvimos a partir de la base de datos UniProt (versión 2012) con un nivel de identidad menor al 40%. Al utilizar monómeros y dímeros para la selección de características obtuvimos un valor de ACC de 94%. En comparación, con un método conocido como IFFS (Improved forward floating selection, por sus siglas en inglés), (Nakariyakul, et al., 2012) obtuvo un ACC de 94%. Los valores de ACC obtenidos con ambos métodos son idénticos. Sin embargo, ambos algoritmos difieren notablemente en el número de características seleccionadas: con BayesC se obtuvieron 93 características mientras que con el algoritmo IFFS se obtuvieron 28. Una inspección detallada mostró que 17 variables coinciden en ambos métodos (**Tabla S3.1**). Por lo tanto, el algoritmo BayesC puede ser considerado como una alternativa a IFSS para el problema de selección de características.

**Evaluación de clasificadores construidos con alfabetos reducidos**

*Dímeros*

Los alfabetos reducidos son resultado de agrupar aminoácidos equivalentes con respecto a cierto criterio (Por ejemplo, polaridad, tamaño, etc.). Al ser subconjuntos del alfabeto original de 20 aa, los alfabetos reducidos siempre tendrán menos de 20 aa y un mínimo de 2.

En este trabajo, recopilamos 173 alfabetos reducidos a partir de diversas fuentes. Sin embargo, una inspección detallada reveló que varios de estos alfabetos se encontraban repetidos, los cuales fueron removidos. En total se evaluaron 146 alfabetos reducidos.

Como primer paso, evaluamos el uso de dímeros de los 146 alfabetos en los seis conjuntos de datos (**Figura 3.8**). Posteriormente, utilizando los alfabetos con mejores valores de ACC procedimos a realizar la selección de características. Como resultado, en cuatro conjuntos de datos obtuvimos un valor de ACC superior con respecto al alfabeto aa20 (**Figura 3.9**).

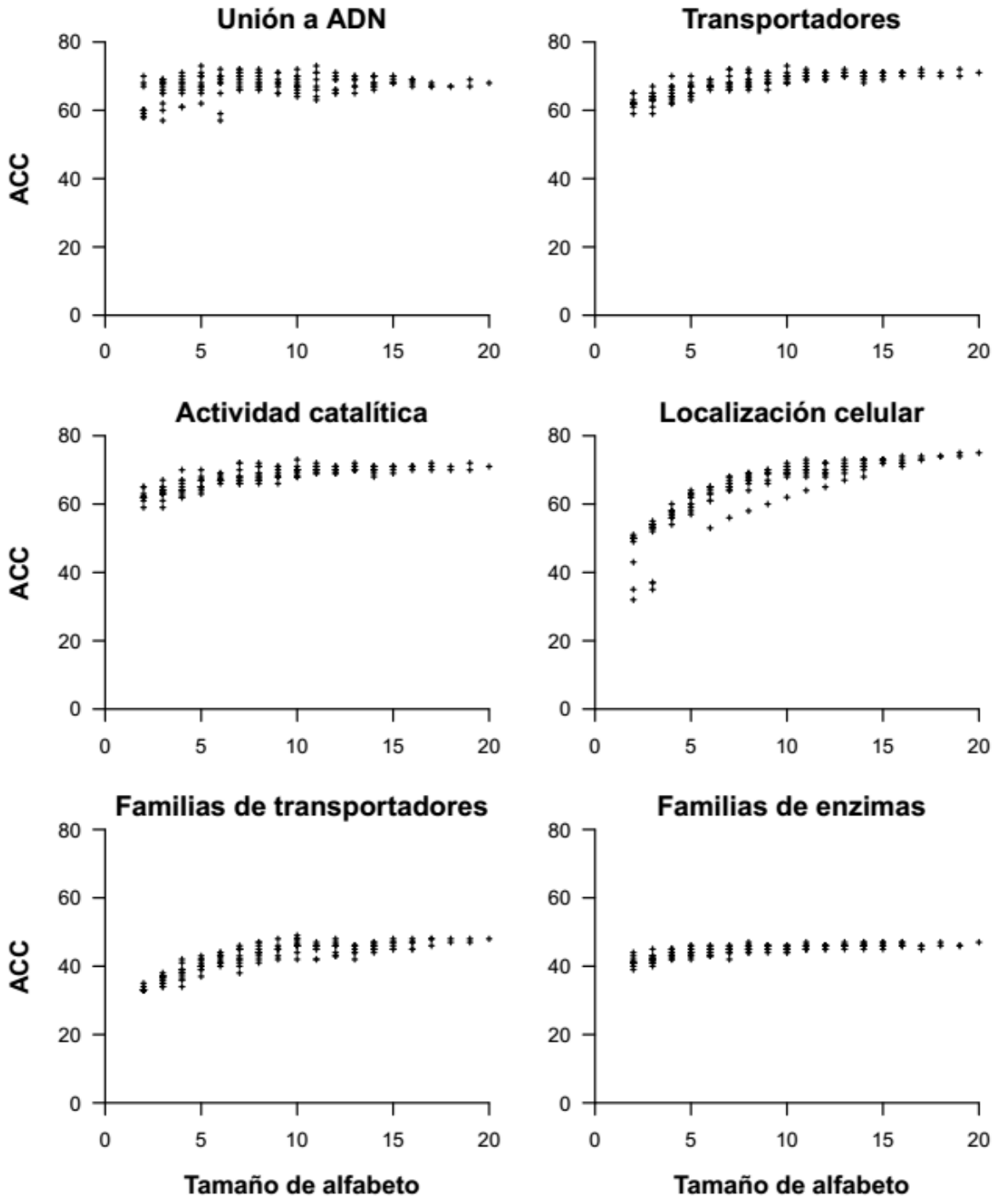



**Figura 3.8. ACC de dímeros-1 utilizando 146 alfabetos en 6 conjuntos de datos.**

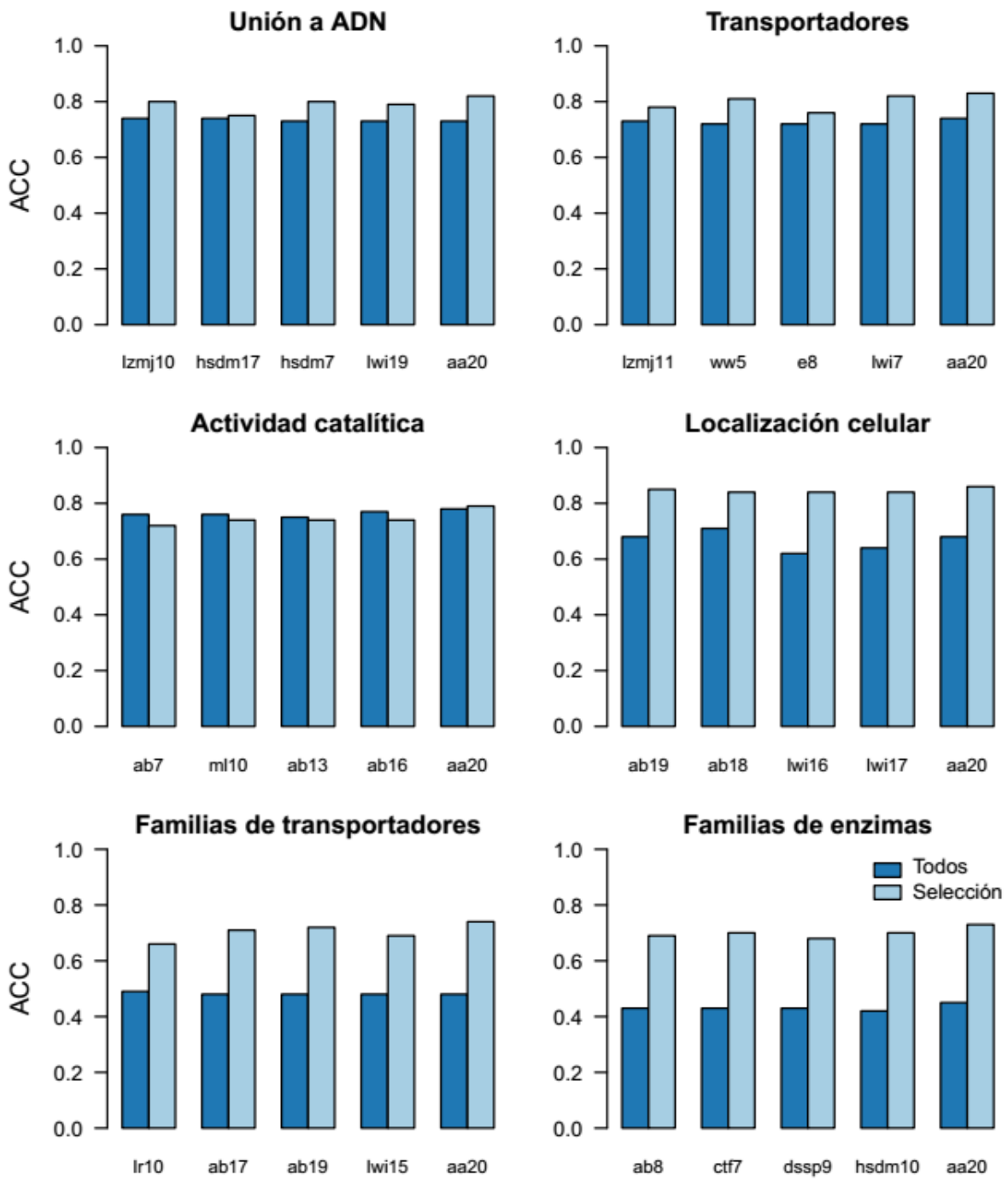



**Figura 3.9. ACC de dímeros de diferentes alfabetos antes y después de la selección de características.**

*Trímeros*

El objetivo de los alfabetos reducidos es encapsular información redundante y capturar mayor información de orden en la secuencia al ser posible incorporar trímeros en los

predictores. Por ejemplo, un alfabeto con 7 letras genera un vector de trímeros de longitud 7 x 7 x 7 = 343, que es de mucho menor tamaño que utilizar 20 x 20 x 20 = 8000. En este trabajo, también evaluamos el uso de trímeros en alfabetos reducidos con tamaño ≤ 12 (**Figura 3.10**).

De forma similar al apartado anterior, seleccionamos aquellos trímeros de alfabetos que mostraron mejor valor de ACC. Posteriormente, con cada uno de ellos realizamos la selección de características. Durante la selección de características incluimos tanto dímeros como trímeros, con excepción del conjunto de actividad de transporte, los alfabetos reducidos mostraron mejores valores en ACC con respecto al alfabeto aa20 (**Figura 3.11**).

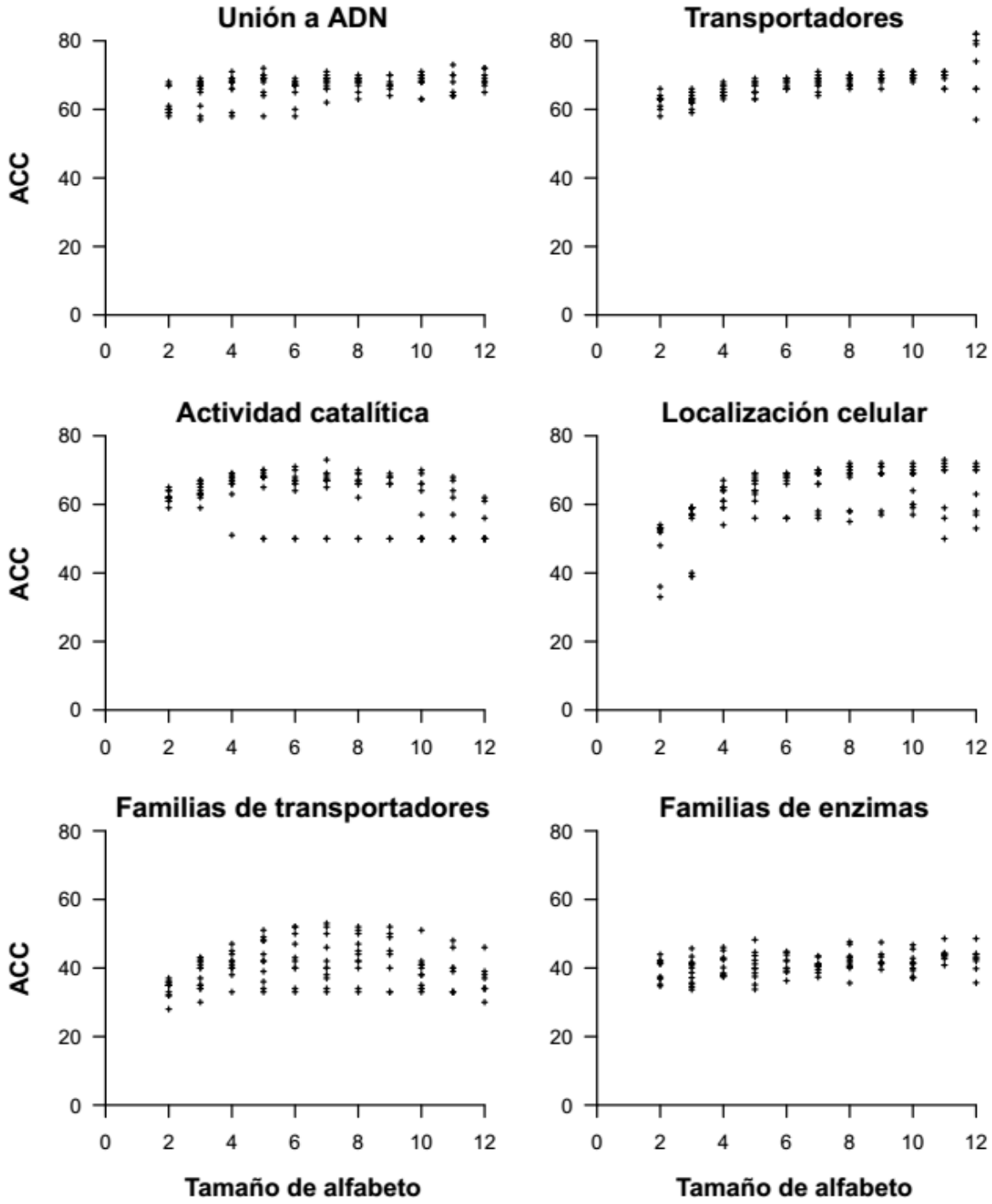



**Figura 3.10. ACC de trímeros utilizando 113 alfabetos reducidos (de tamaño ≤ 12) en 6 conjuntos de datos.**

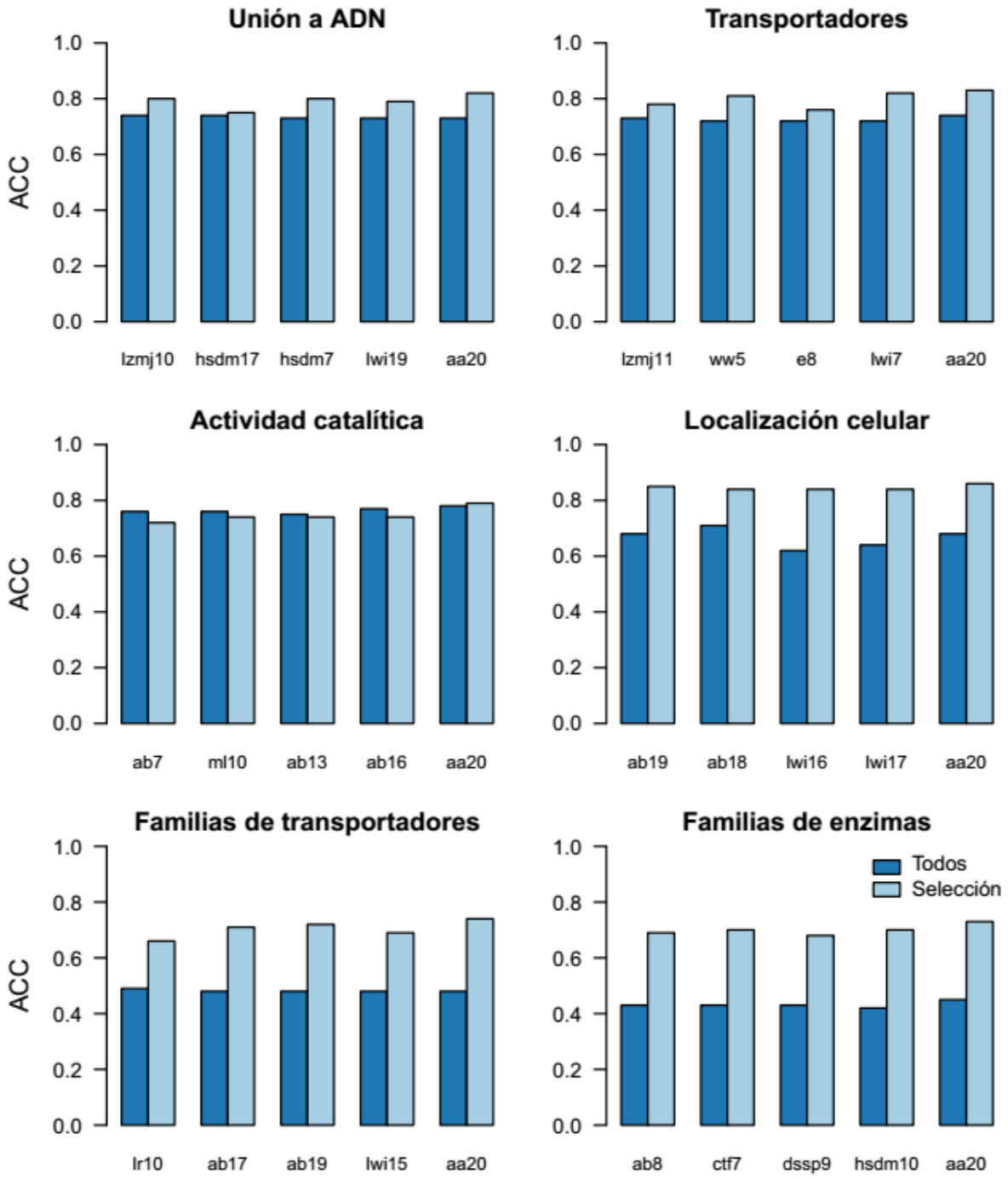



**Figura 3.11. ACC de trímeros de diferentes alfabetos antes y después de la selección de características.**

**Construcción del ensamble SVM y comparación con otros métodos**

Para mejorar la capacidad predictiva de los clasificadores individuales obtenidos en la sección anterior, realizamos un ensamble utilizando el criterio de stacking. De esta manera, el primer nivel del ensamble está compuesto por los mejores clasificadores SVM, cada uno entrenado con las características significativas de un alfabeto en particular. El segundo nivel del ensamble utiliza las predicciones (en forma de probabilidad) de cada clasificador y con ellas se entrena un nuevo clasificador que actúa como jurado SVM, el cual genera la predicción final.

La capacidad predictiva de los ensambles incrementó en promedio un 3% el valor de ACC, con respecto a los mejores clasificadores individuales. En general, utilizar a los mejores 5 clasificadores con mejor valor de ACC fue suficiente para generar cada ensamble; agregar más clasificadores no mostró mejoras en el valor de ACC. A continuación se presentan los resultados obtenidos en cada conjunto de datos:

*Conjunto de unión a ADN*. Los mejores tres alfabetos fueron lzbl11(t), lwi12(t) y aa20(d) con ACC de 82%, 84% y 82% respectivamente. Mientras que el ACC del ensamble (86%) mejoró con respecto a los clasificadores individuales. El ensamble que obtuvimos en este trabajo sobrepasa considerablemente a los métodos más recientes propuestos en la literatura tales como: DNAbinder(Kumar, et al., 2007) DNA-Prot (Kumar, et al., 2009) iDNA-Prot (Lin, et al., 2011) e iDNAPro-PseAAC (Liu, et al., 2015) (¡Error! No se encuentra el origen de la referencia.). A excepción de iDNA-Prot, el resto se basa en el uso de perfiles PSSM. El método de (Liu, et al., 2015) fusiona los perfiles PSSM con el concepto de composición de pseudoaminoácidos (PseAAC) (Chou, 2001), ampliamente usado en el campo de predicción de función (Kuo-Chen, 2009). Al igual que el método que propongo, el método de iDNA-Prot es el único que puede ser considerado independiente de homología. Dicho método utiliza la frecuencia de los 20 aminoácidos y tres valores adicionales derivados de la aplicación de la teoría gris de sistemas.

**Tabla 3.2. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Unión a ADN.**

| Composición | N | Acc | Mcc | AUC | Sn(%) | Sp(%) |
|---|---|---|---|---|---|---|
| lzbl11 (t) | 237 | 0.82 | 0.63 | 0.90 | 83 | 81 |
| lwi12 (t) | 299 | 0.84 | 0.67 | 0.91 | 85 | 83 |
| aa20 (d) | 250 | 0.82 | 0.63 | 0.90 | 82 | 81 |
| Ensamble | **n.a** | **0.86** | **0.70** | **0.93** | **87** | **86** |
| iDNAPro-PseAAC | | 0.77 | 0.53 | 0.84 | 76 | 77 |
| iDNA-Prot | 23 | 0.75 | 0.05 | 0.76 | 84 | 65 |
| DNAbinder | 400 | 0.74 | 0.47 | 0.82 | 66 | 80 |
| DNA-Prot | 400 | 0.73 | 0.44 | 0.79 | 83 | 60 |

N = Número de carácterísticas, Sn = Sensibilidad, Sp = Especificidad, Acc = Exactitud, Mcc = Coeficiente de correlacion de Matthews, AUC = Área bajo la curva ROC, (d) = monómeros + ventanas(2 ,3 y 4) + dímeros-(0, 1 y 2). (t) = (d) + trímeros, n.a = No aplica.

*Actividad de transporte*. Los mejores tres alfabetos fueron aa20(t), lzmj12(t) y ab12(d) con ACC de 89%, 83% y 83% respectivamente. En este conjunto el ensamble no superó al mejor alfabeto completo aa20(t). Sin embargo, superó en 10 % al método propuesto por (Mishra, et al., 2014) que utiliza 400 variables PSSM y 49 variables en base a propiedades bioquímicas (AAI) (¡Error! No se encuentra el origen de la referencia.).

**Tabla 3.3. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Transportadores.**

| Composición | N | Acc | Mcc | AUC | Sn | Sp |
|---|---|---|---|---|---|---|
| aa.20 (t) | 362 | 0.89 | 0.77 | 0.96 | 89 | 89 |
| lzmj12 (t) | 211 | 0.83 | 0.64 | 0.90 | 79 | 86 |
| ab12 (t) | 249 | 0.83 | 0.64 | 0.91 | 83 | 83 |
| Ensamble | **n.a** | **0.89** | **0.63** | **0.91** | **89** | **90** |
| AAI+PSSM | 49+400 | 0.79 | n.d | 0.87 | 80 | 78 |

N = Número de carácterísticas, Sn = Sensibilidad, Sp = Especificidad, Acc = Exactitud, Mcc = Coeficiente de correlacion de Matthews, AUC = Área bajo la curva ROC, AAI = Aminoacid Index Coding. PSSM = Position specific score matrix, n.a. = No aplica, n.d = no disopnnible, (d) = monómeros + ventanas(2, 3 y 4) + dímeros-(0, 1 y 2). (t) = (d) + trímeros

*Actividad enzimática.* Los mejores tres alfabetos fueron aa20(d), ab16(d) y Hsdm16(d) con valores de ACC de 81%, 80% y 80% respectivamente. Mientras que el ACC del ensamble mejoró con respecto a los clasificadores individuales (85%). En este conjunto, el ACC del ensamble propuesto quedó 6% debajo del propuesto por (Shen and Chou, 2007) (¡Error! No se encuentra el origen de la referencia.). Sin embargo, dicho método se basa en previa identificación de dominios Pfam en conjunto con una versión modificada de PSSM. De forma similar métodos más recientes incorporan, además de dominos, perfiles filogenéticos e información estructural (Amin, et al., 2013; Nagao, et al., 2014). La principal ventaja del ensamble propuesto, es no depender de la identificación de dominios o información estructural para su entrenamiento.

**Tabla 3.4. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Actividad enzimática.**

| **Composición** | **N** | **Acc** | **Mcc** | **AUC** | **Sn** | **Sp** |
|---|---|---|---|---|---|---|
| aa20 (d) | 230 | 0.81 | 0.58 | 0.84 | 80 | 82 |
| ab16 (d) | 239 | 0.80 | 0.55 | 0.82 | 77 | 82 |
| Hsdm16 (d) | 206 | 0.80 | 0.52 | 0.82 | 77 | 81 |
| Ensamble | n.a | 0.85 | 0.61 | 0.92 | 84 | 87 |
| FunD + Pse-PSSM | 8958+40 | 0.91 | n.d | n.d | 92 | 90 |

N = Número de características, Sn = Sensibilidad, Sp = Especificidad, Acc = Exactitud, Mcc = Coeficiente de correlacion de Matthews, AUC = Área bajo la curva ROC, AAI = Aminoacid Index coding. PSSM = Position specific score matrix, n.d = No disponible, n.a. = No aplica. (d) = monómeros + ventanas(2,3 y 4) + dímeros-(0, 1 y 2), (t) = (d) + trímeros, funD = Functional domain, Pse-PSSM = Pseudo position-specific scoring matrix.

*Conjunto de Localización celular*. Los mejores tres alfabetos fueron aa20(d), ab19(d) y Lwi18(d) con ACC total de 86%, 85% y 85% respectivamente. Mientras que el valor de ACC del ensamble mejoró con respecto a los clasificadores individuales (88%). En comparación con el método propuesto por (Yu, et al., 2006) entrenado también con estructura primaria, el ensamble mostró una mejora del 3% (¡Error! No se encuentra el origen de la referencia.).

**Tabla 3.5. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Localización Celular.**

| Categoría | aa20 (d) | ab19 (d) | Lwi18 (d) | Ensamble | Cello II |
|---|---|---|---|---|---|
| Cloroplasto | 85 | 84 | 83 | **86** | 80 |
| Citoplasma | 79 | 79 | 79 | **78** | 77 |
| Citoesqueleto | 60 | 65 | 68 | **68** | 68 |
| Retículo endoplásmico | 76 | 70 | 75 | **73** | 68 |
| Extracelular | 89 | 89 | 89 | **90** | 90 |
| Aparato de Golgi | 58 | 53 | 64 | **63** | 53 |
| Lisosoma | 77 | 79 | 78 | **77** | 69 |
| Membrana | 95 | 95 | 94 | **95** | 96 |
| Mitocodria | 75 | 73 | 73 | **73** | 73 |
| Nucleo | 90 | 90 | 90 | **91** | 91 |
| Peroxisoma | 59 | 56 | 54 | **60** | 47 |
| Vacuola | 52 | 52 | 56 | **56** | 52 |
| TOTAL | 86 | 85 | 85 | **88** | 85 |

Valores de ACC en %. (d) = monómeros + ventanas(2, 3 y 4) + dímeros-(0, 1 y 2).

*Conjunto de Familias de transportadores*. Los mejores tres alfabetos fueron aa20(d), ab19(d) y Lwi18(d) con ACC total de 74%, 72% y 72% respectivamente. Mientras que el ACC del ensamble mejoró con respecto a los clasificadores individuales (76%) (¡Error! No se encuentra el origen de la referencia.). Es necesario aclarar que los resultados publicados por (Mishra, et al., 2014) no corresponden a los valores de un clasificador multi categoría. En su lugar, (Mishra, et al., 2014) presenta los valores de Sn, Sp, ACC, MCC y AUC correspondientes a clasificadores binarios. Para tener un punto de comparación más realista, en la ¡Error! No se encuentra el origen de la referencia. presento los resultados del clasificador multi categoría utilizando monómeros aa20. Claramente el ensamble representa una mejora considerable (32%) en la clasificación de este conjunto de datos. Por su parte, el

método propuesto por (Mishra, et al., 2014) alcanza mejores resultados que los monómeros aa20 apenas por un 3% en los clasificadores binarios que presenta.

**Tabla 3.6. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Familias de Transportadores.**

| Categoría/ alfabeto | aa.20 (d) | ab.19 (d) | lw-i.18 (d) | Ensamble | Monómeros aa20 |
|---|---|---|---|---|---|
| Aminoácido | 86 | 80 | 81 | **85** | 24 |
| Anión | 32 | 28 | 32 | **36** | 2 |
| Catión | 87 | 84 | 82 | **88** | 72 |
| Electrón | 68 | 67 | 78 | **76** | 53 |
| Otro | 74 | 72 | 70 | **76** | 38 |
| Proteína | 59 | 64 | 67 | **67** | 36 |
| Azúcar | 73 | 65 | 67 | **74** | 22 |
| TOTAL | 74 | 72 | 72 | **76** | 44 |

Valores de ACC en %. (d) = monómeros + ventanas(2, 3 y 4) + dímeros-(0, 1 y 2).

*Conjunto de familias de enzimas*. Los mejores tres alfabetos fueron aa20(d), Hsdm17(t) y Ab18(d) con ACC total de 73%, 71% y 73% respectivamente. En este conjunto el ensamble no logró incrementar el ACC (¡Error! No se encuentra el origen de la referencia.). Al igual que en la clasificación de actividad enzimática, el ensamble queda por debajo de EzyPred (Shen and Chou, 2007) y otros más recientes (Amin, et al., 2013; Nagao, et al., 2014). De nueva cuenta, todos estos clasificadores utilizan como características la presencia de dominios funcionales, perfiles filogenéticos e información estructural.

**Tabla 3.7. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Familias de Enzimas.**

| Categoría | Aa20 (d) | Hsdm17(t) | Ab18(d) | Ensamble | EzyPred |
|---|---|---|---|---|---|
| Hidrolasas | 80 | 78 | 81 | 82 | **91** |
| Isomerasas | 64 | 65 | 65 | 60 | **95** |
| Ligasas | 83 | 82 | 81 | 80 | **97** |
| Liasas | 55 | 50 | 55 | 60 | **85** |
| Oxidorectasas | 71 | 73 | 70 | 70 | **84** |
| Transferasas | 72 | 65 | 72 | 73 | **97** |
| TOTAL | 73 | 71 | 73 | 73 | **94** |

Valores de Acc en %. (d) = monómeros + ventanas(2, 3 y 4) + dímeros-(0, 1 y 2), (t) = (d) + trímeros.

### Anotación funcional de genes huérfanos de *Arabidopsis thaliana*

Muchos métodos bioinformáticos para asignar función a proteínas se basan en la homología de secuencias. Sin embargo, predecir la función de en genes huérfanos es una tarea difícil debido a la escasa o nula similitud de éstos con proteínas de otras especies. Esta situación se ve reflejada en la anotación GO actual de los genes huérfanos de *A. thaliana*, donde 48%, 88% y 97% están anotados como componente celular desconocido, proceso biológico desconocido y función biológica desconocida (Donoghue, et al., 2011). Para predecir la función de 1,786 genes huérfanos en *A. thaliana*, utilizamos los ensambles SVMs descritos en la sección anterior (**Figura 3.12**). Dichos ensambles están entrenados únicamente a partir de la estructura primaria de proteínas y son, por tanto, totalmente independientes de homología lo que los convierte en una herramienta ideal para la predicción de función de genes/proteínas huérfanos. De los 1,786 genes huérfanos: 1) 771 fueron predichos con actividad de Unión a ADN; 2) 289 con actividad de transporte. De los cuales, de acuerdo a su sustrato, 128 pertenecen a la familia de transportadores de cationes, 127 a transportadores de electrones, 23 a transportadores varios, 10 a transportadores de proteínas y 1 a transportadores de azúcares; 3) 74 con actividad catalítica. De los cuales, 47 fueron predichos como hidrolasas, 4 isomerásas, 1 ligasa, 1 liása, 9 oxidorectasas y 12 transferasas. En base a localización celular, las predicciones que obtuvimos fueron las

siguientes: 51 en cloroplasto, 107 en citoplasma, 645 extracelulares, 107 en membrana celular, 579 en mitocondria y 297 nucleares.

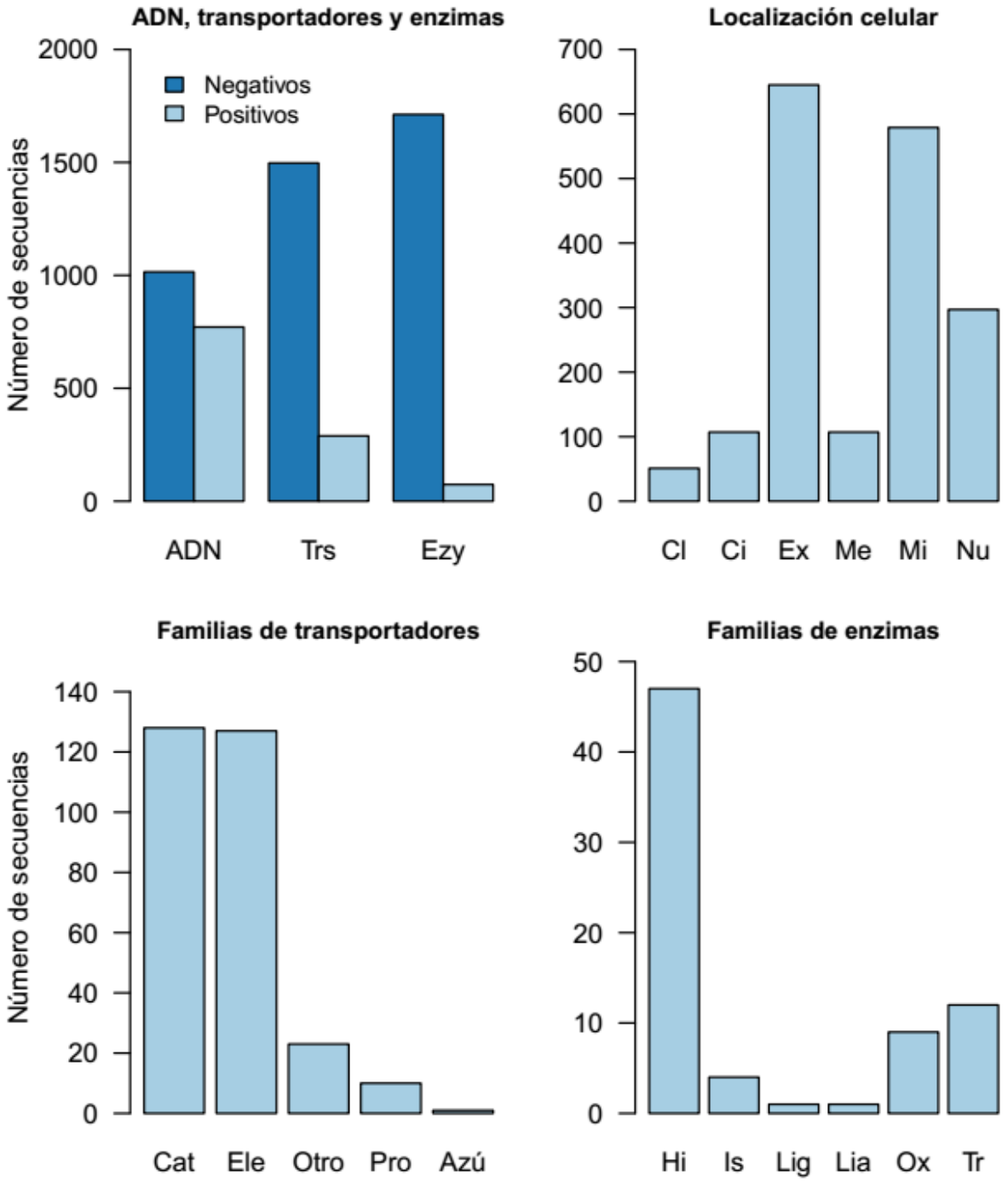



**Figura 3.12. Predicción de función en 1786 genes huérfanos de *A. thaliana*.**

**Guía de la aplicación web para el usuario**

A continuación se hace una breve descripción sobre el uso de la aplicación web.

*Paso 1*

En esta primera interfaz (**Figura 3.13**) el usuario dispone de 4 opciones: a) Dirigirse directamente al predictor, b) revisar el tutorial sobre cómo usar el predictor, c) conocer algunos detalles sobre cómo fue desarrollado el método, d) conocer más sobre la teoría del aprendizaje estadístico.

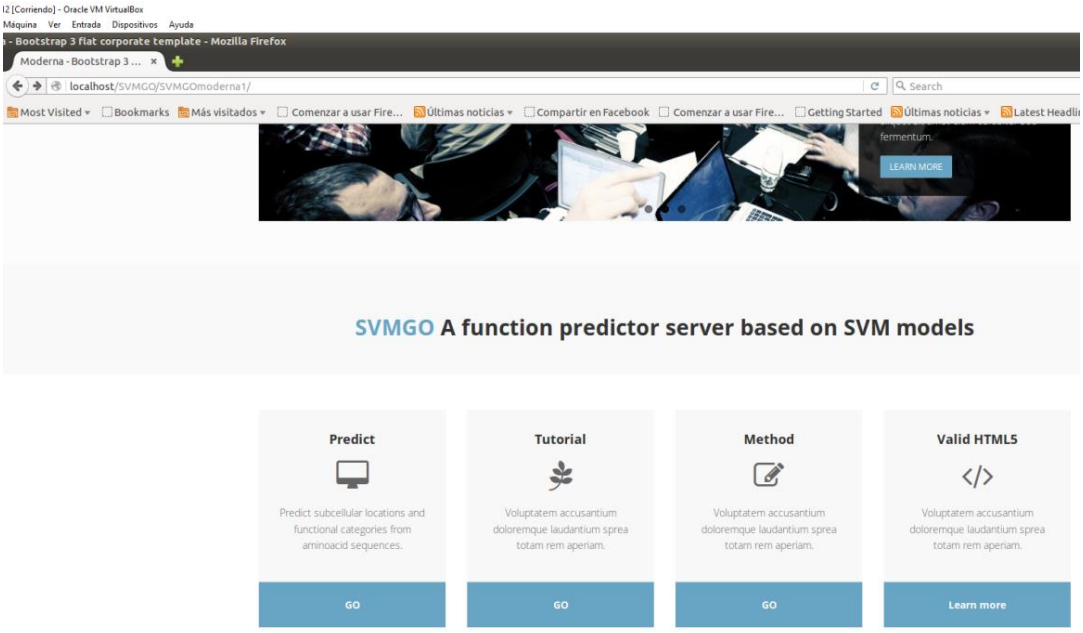



**Figura 3.13. Página inicial del servidor.**

*Paso 2*

En la interfaz perteneciente propiamente al predictor (**Figura 3.14**). El usuario tiene la opción de introducir sus secuencias en formato fasta (A) o seleccionar y subir un archivo en formato fasta desde su equipo (B). Si una dirección de correo electrónico es introducida (C), una liga con los resultados serían enviados a dicho correo. El siguiente paso es seleccionar la categoría que se quiere predecir, en este caso está seleccionada la de unión a ADN (D). Para proceder con la predicción se tiene que seleccionar el botón de ejecutar (E).

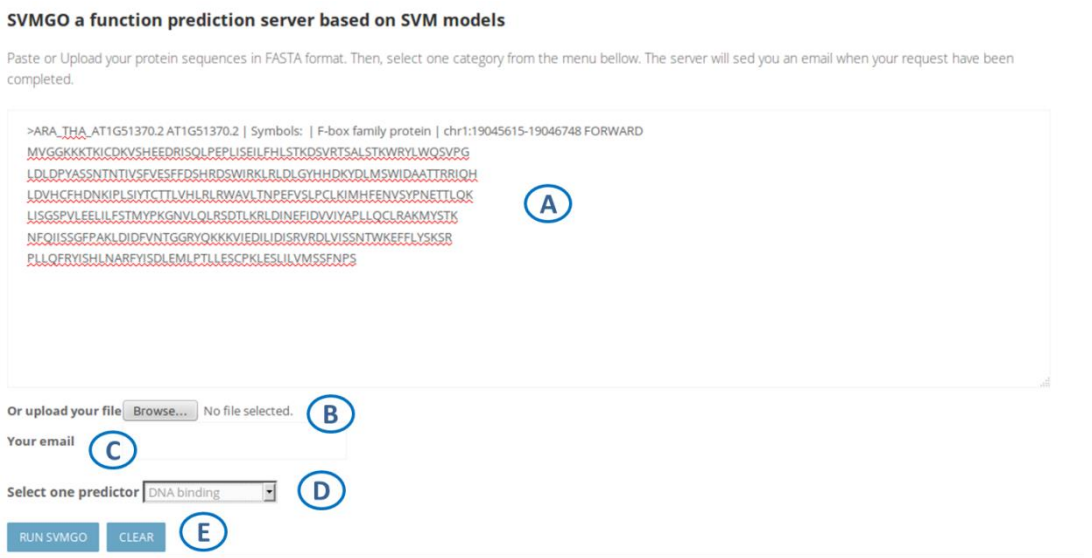



**Figura 3.14. Interfaz para que el usuario introduzca sus secuencias y seleccione las categorías a predecir.**

*Paso 3*

Las predicciones son mostradas en una tabla (**Figura 3.15**) que muestra el id de cada secuencia (A), la categoría predicha y su probabilidad asociada (B). Los resultados son desplegados en bloques de tamaño definido por el usuario (C), y se puede acceder a ellos a través de los botones de previo y siguiente (D). Adicionalmente, se pueden realizar búsquedas por id de secuencia (E).

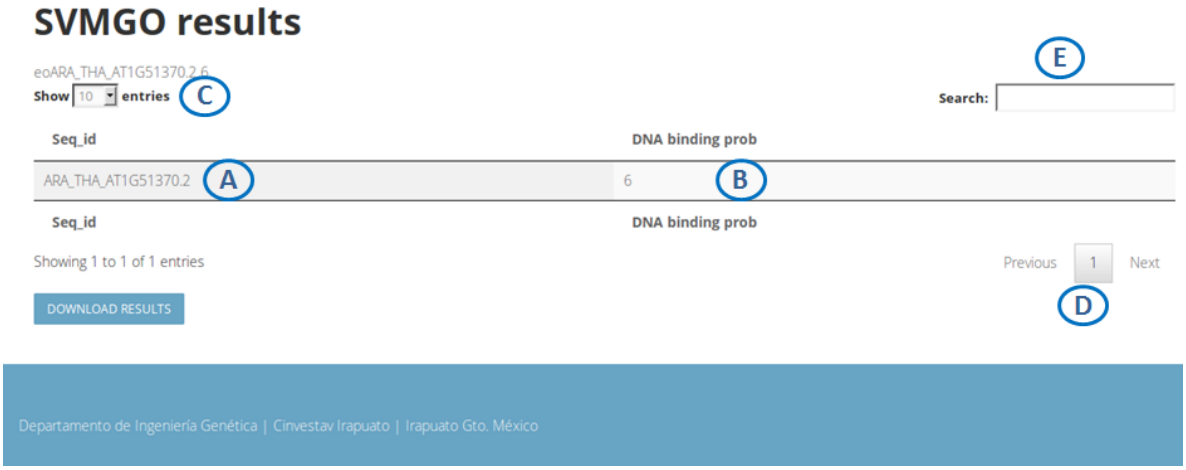



**Figura 3.15. Página de resultados html.**

*Paso 4*

Finalmente, el usuario tiene la opción de descargar a su equipo los resultados de las predicciones (**Figura 3.16**) seleccionando el botón de descarga (A).

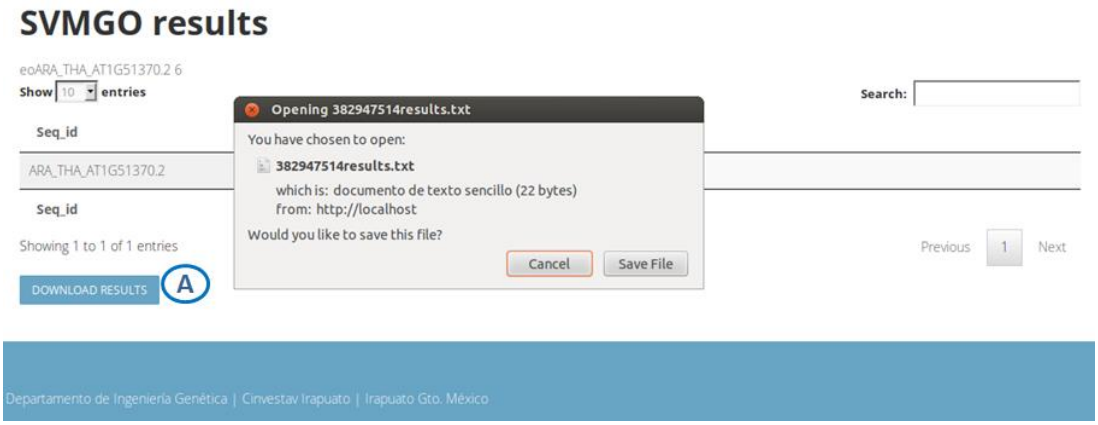



**Figura 3.16. Página de resultados. El usuario puede descargar los resultados a su computadora.**

## Conclusiones

La predicción de función a partir de la secuencia primaria ha sido un reto para la bioinformática desde mediados de la década pasada, sin embargo con el desarrollo de nuevas herramientas teóricas ha sido posible afrontar este problema con mejores resultados.

En este trabajo desarrollamos un clasificador totalmente independiente de homología para la predicción de seis categorías funcionales, a saber: de unión a ADN, actividad de transporte, actividad enzimática, localización celular, familias de transportadores y familias de enzimas. Los conjuntos de datos utilizados en este trabajo fueron elegidos debido a que se encuentran disponibles públicamente y han sido utilizados ampliamente como referencia en diversas publicaciones subsecuentes. Es importante señalar que las funciones de proteínas representadas por estos conjuntos de datos se encuentran muy limitadas. Sin embargo, el método desarrollado en este trabajo puede ser fácilmente aplicado a nuevos conjuntos de datos, particularmente a aquellos que contemplan un mayor número de funciones biológicas y en adición basados en la clasificación de Gene Ontology, por ejemplo el conjunto de datos de GOPred (Saraç, et al., 2010).

Debido a la dificultad para determinar la función en proteínas huérfanas o con homólogos de los que se desconoce su función, los predictores de función que utilizan únicamente información de secuencia representan herramientas valiosas para el estudio de dichas

proteínas. En este trabajo desarrollamos un método de predicción de función independiente de homología construido en base a un ensamble de Máquinas de Soporte Vectorial (SVM), en el que cada clasificador SVM fue construido utilizando un alfabeto reducido. Adicionalmente, la selección de características que se realiza como paso previo al ensamble permite identificar aquellos monómeros y dímeros de aminoácidos que son relevantes para la predicción de función. La identificación de dichos monómeros y dímeros podría beneficiar a la comunidad que investiga las características físico-químicas que determinan la función de las proteínas. En comparación a otros métodos recientes, el método propuesto logró mejores resultados en 4 de los 6 conjuntos de datos en que fue evaluado.

Con esta serie de ensambles realizamos una anotación preliminar de las proteínas huérfanas del proteoma de *Arabidopsis thaliana*. La lista de predicciones se encuentra disponible para la comunidad dedicada al estudio de *A. thaliana*. Al no usar información evolutiva (tales como perfiles PSSM) durante la construcción de los modelos, éstos son más adecuados para predecir la función en proteínas huérfanas. De hecho, en ciertas categorías funcionales, aun cuando no incorporan información evolutiva, dichos modelos superan a los que sí en conjuntos de proteínas para las que no se obtienen hits utilizando PSI-BLAST (Kumar, et al., 2007).

Adicionalmente, el diseño modular de esta herramienta permite que sea sencillo incorporar muchas más categorías en el corto plazo. Esperamos que esta herramienta les permita realizar anotaciones que permitan direccionar los experimentos en laboratorio para determinar la función de proteínas.

## Perspectivas del capítulo

Las siguientes perspectivas podrían abordarse en trabajos futuros:

1) Crear una base de datos y una herramienta de predicción automática que esté abierta al público fuera de Cinvestav Irapuato.
2) Comparar el desempeño de los ensambles de SVMs contra el uso de bosques aleatorios (*random forests*) y determinar cual tiene mayor poder predictivo para cada caso.

3) Determinar las variables, los monómeros y dímeros que tienen mayor capacidad de discriminación entre grupos funcionales.
4) Determinar de forma experimental en proteínas huérfanas de *A. thaliana* las predicciones hechas con el método propuesto.

## Perspectivas Generales

1) Investigar la razón por la que en endosimbiontes hay cambios en el tamaño de las proteínas.
2) Investigar la razón por la que en diferentes organismos unicelulares (fungi, protozoarios y demás) existen diferencias de tamaño de proteínas y de estructura de exones comparados con animales y plantas.
3) Investigar cómo es que las proteínas cambian su tamaño.
4) Encontrar más evidencia de un mecanismo evolutivo de proteínas que implica trans-splicing y transcripción reversa.
5) Como se comportan los intrones durante multiplicación de los miembros de las familias de parálogos en plantas.
6) Aplicar nuestro método de anotación automática a más genomas y de esta forma enriquecer la anotación funcional de las proteínas desconocidas de plantas, hongos y animales.

## Referencias


Abdallah, F., Salamini, F. and Leister, D. (2000) A prediction of the size and evolutionary origin of the proteome of chloroplasts of Arabidopsis, *Trends in Plant Science*, **5**, 141-142.

Abu-Mostafa, Y.S., Magdon-Ismail, M. and Lin, H.-T. (2012) *Learning From Data*. AMLBook.

Adams, K.L. and Wendel, J.F. (2005) Polyploidy and genome evolution in plants, *Current Opinion in Plant Biology*, **8**, 135-141.

Adl, S.M., *et al.* (2012) The Revised Classification of Eukaryotes, *Journal of Eukaryotic Microbiology*, **59**, 429-514.

Agapakis, C.M., *et al.* (2011) Towards a Synthetic Chloroplast, *PLoS ONE*, **6**, e18877.

Altenhoff, A.M. and Dessimoz, C. (2009) Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods, *PLOS Computational Biology*, **5**, e1000262.

Altschul, S.F., *et al.* (1990) Basic local alignment search tool, *Journal of Molecular Biology*, **215**, 403-410.

Altschul, S.F., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Research*, **25**, 3389-3402.

Amin, S.R., *et al.* (2013) Prediction and experimental validation of enzyme substrate specificity in protein structures, *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E4195-E4202.

Archibald, J.M. (2009) The Puzzle of Plastid Evolution, *Current Biology*, **19**, R81-R88.

Arendsee, Z.W., Li, L. and Wurtele, E.S. (2014) Coming of age: orphan genes in plants, *Trends in Plant Science*, **19**, 698-708.

Attwood, T.K., *et al.* (2003) PRINTS and its automatic supplement, prePRINTS, *Nucleic Acids Research*, **31**, 400-402.

Bakhtiarizadeh, M.R., *et al.* (2014) Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology, *Journal of Theoretical Biology*, **356**, 213-222.

Basu, M.K., Poliakov, E. and Rogozin, I.B. (2009) Domain mobility in proteins: functional and evolutionary implications, *Briefings in Bioinformatics*, **10**, 205-216.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289-300.

Benso, A., *et al.* (2013) A combined approach for genome wide protein function annotation/prediction, *Proteome Science*, **11**, S1.

Bionda, T., *et al.* (2010) Chloroplast Import Signals: The Length Requirement for Translocation In Vitro and In Vivo, *Journal of Molecular Biology*, **402**, 510-523.

Bowman, J.L., Floyd, S.K. and Sakakibara, K. (2007) Green Genes. Comparative Genomics of the Green Branch of Life, *Cell*, **129**, 229-234.

Brocchieri, L. and Karlin, S. (2005) Protein length in eukaryotic and prokaryotic proteomes, *Nucleic Acids Research*, **33**, 3390-3400.

Brosch, M., *et al.* (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome, *Genome Research*, **21**, 756-767.

Buljan, M., Frankish, A. and Bateman, A. (2010) Quantifying the mechanisms of domain gain in animal proteins, *Genome Biology*, **11**, R74.
Burki, F. (2014) The Eukaryotic Tree of Life from a Global Phylogenomic Perspective, *Cold Spring Harbor Perspectives in Biology*, **6**.
Cai, J.J., *et al.* (2006) Accelerated Evolutionary Rate May Be Responsible for the Emergence of Lineage-Specific Genes in Ascomycota, *Journal of Molecular Evolution*, **63**, 1-11.
Cancherini, D.V., França, G.S. and de Souza, S.J. (2010) The role of exon shuffling in shaping protein-protein interaction networks, *BMC Genomics*, **11**, S11-S11.
Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics*, **25**, 1972-1973.
Carvunis, A.-R., *et al.* (2012) Proto-genes and de novo gene birth, *Nature*, **487**, 370-374.
Castellan, S.a. (1988) *Non parametric statistics for the behavioural sciences*. MacGraw Hill int., New York.
Catrein, I. and Herrmann, R. (2011) The proteome of Mycoplasma pneumoniae, a supposedly "simple" cell, *PROTEOMICS*, **11**, 3614-3632.
Colbourne, J.K., *et al.* (2011) The Ecoresponsive Genome of Daphnia pulex, *Science (New York, N.Y.)*, **331**, 555-561.
Coulombe-Huntington, J. and Majewski, J. (2007) Intron Loss and Gain in Drosophila, *Molecular Biology and Evolution*, **24**, 2842-2850.
Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: A library for support vector machines, *ACM Trans. Intell. Syst. Technol.*, **2**, 1-27.
Chou, K.-C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition, *Proteins: Structure, Function, and Bioinformatics*, **43**, 246-255.
Chou, K.-C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, *Bioinformatics*, **21**, 10-19.
Chou, K.-C. and Shen, H.-B. (2010) Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization, *PLoS ONE*, **5**, e11335.
Dagan, T., *et al.* (2013) Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids, *Genome Biology and Evolution*, **5**, 31-44.
Darriba, D., *et al.* (2011) ProtTest 3: fast selection of best-fit models of protein evolution, *Bioinformatics*.
Darriba, D., *et al.* (2012) jModelTest 2: more models, new heuristics and parallel computing, *Nat Meth*, **9**, 772-772.
Donoghue, M.T., *et al.* (2011) Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana, *BMC Evolutionary Biology*, **11**, 1-23.
Dorden, S. and Mahadevan, P. (2015) Functional prediction of hypothetical proteins in human adenoviruses, *Bioinformation*, **11**, 466-473.
Dubchak, I., *et al.* (1999) Recognition of a protein fold in the context of the SCOP classification, *Proteins: Structure, Function, and Bioinformatics*, **35**, 401-407.
Ebrahimie, E., *et al.* (2011) Protein attributes contribute to halo-stability, bioinformatics approach, *Saline Systems*, **7**, 1-1.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research*, **32**, 1792-1797.
Ejima, Y. and Yang, L. (2003) Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling, *Human Molecular Genetics*, **12**, 1321-1328.
Ekman, D., *et al.* (2005) Multi-domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions, *Journal of Molecular Biology*, **348**, 231-243.
Ekman, D., *et al.* (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?, *Genome Biology*, **7**, R45-R45.
Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861-874.
Felsenstein, J. (1985) Phylogenies and the Comparative Method, *The American Naturalist*, **125**, 1-15.
Fischer, D. and Eisenberg, D. (1999) Finding families for genomic ORFans, *Bioinformatics*, **15**, 759-762.
Fitch, W.M. (1970) Distinguishing Homologous from Analogous Proteins, *Systematic Biology*, **19**, 99-113.
França, G.S., Cancherini, D.V. and de Souza, S.J. (2012) Evolutionary history of exon shuffling, *Genetica*, **140**, 249-257.
Fu, L., *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, **28**, 3150-3152.
Garland, T., Harvey, P.H. and Ives, A.R. (1992) Procedures for the Analysis of Comparative Data Using Phylogenetically Independent Contrasts, *Systematic Biology*, **41**, 18-32.
Guillou, L., *et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy, *Nucleic Acids Research*, **41**, D597-D604.
Guindon, S., *et al.* (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0, *Systematic Biology*, **59**, 307-321.
Guo, W.-J., *et al.* (2007) Significant Comparative Characteristics between Orphan and Nonorphan Genes in the Rice (Oryza sativa L.) Genome, *Comparative and Functional Genomics*, **2007**, 21676.
Gutiérrez, R.A., Larson, M.D. and Wilkerson, C. (2004) The Plant-Specific Database. Classification of Arabidopsis Proteins Based on Their Phylogenetic Profile, *Plant Physiology*, **135**, 1888-1892.
Hasan, M.A.M., Ahmad, S. and Molla, M.K.I. (2017) Protein subcellular localization prediction using multiple kernel learning based support vector machine, *Molecular BioSystems*.
Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.
Hawkins, T. and Kihara, D. (2007) Function prediction of uncharacterized proteins, *Journal of Bioinformatics and Computational Biology*, **05**, 1-30.
He, H. and Garcia, E.A. (2009) Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering*, **21**, 1263-1284.
He, H. and Ma, Y. (2013) *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press.

He, X. and Zhang, J. (2005) Gene Complexity and Gene Duplicability, *Current Biology*, **15**, 1016-1021.
James, G., *et al.* (2014) *An Introduction to Statistical Learning: with Applications in R*. Springer Publishing Company, Incorporated.
Kaplunovsky, A., *et al.* (2009) Statistics of Exon Lengths in Animals, Plants,Fungi, and Protists, *World Academy of Science, Engineering and Technology*, **28**.
Katinka, M.D., *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi, *Nature*, **414**, 450-453.
Kelkar, Y.D. and Ochman, H. (2013) Genome Reduction Promotes Increase in Protein Functional Complexity in Bacteria, *Genetics*, **193**, 303-307.
Koonin, E.V., Aravind, L. and Kondrashov, A.S. (2000) The Impact of Comparative Genomics on Our Understanding of Evolution, *Cell*, **101**, 573-576.
Koralewski, T.E. and Krutovsky, K.V. (2011) Evolution of Exon-Intron Structure and Alternative Splicing, *PLoS ONE*, **6**, e18055.
Kristensen, D.M., *et al.* (2011) Computational methods for Gene Orthology inference, *Briefings in Bioinformatics*, **12**, 379-391.
Kruskal, W.H. and Wallis, W.A. (1952) Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association*, **47**, 583-621.
Kumar, K.K., Pugalenthi, G. and Suganthan, P.N. (2009) DNA-Prot: Identification of DNA Binding Proteins from Protein Sequence Information using Random Forest, *Journal of Biomolecular Structure and Dynamics*, **26**, 679-686.
Kumar, M., Gromiha, M.M. and Raghava, G.P.S. (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles, *BMC Bioinformatics*, **8**, 463-463.
Kuo-Chen, C. (2009) Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology, *Current Proteomics*, **6**, 262-274.
Kuo, C.-H. and Kissinger, J.C. (2008) Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites Plasmodium and Theileria, *BMC Evolutionary Biology*, **8**, 108-108.
Lane, C.E., *et al.* (2007) Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function, *Proceedings of the National Academy of Sciences*, **104**, 19908-19913.
Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, *Bioinformatics*, **22**, 1658-1659.
Li, Y.H., *et al.* (2016) SVM-Prot 2016: A Web-Server for Machine Learning Prediction of Protein Functional Families from Sequence Irrespective of Similarity, *PLoS ONE*, **11**, e0155290.
Li, Z.R., *et al.* (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, *Nucleic Acids Research*, **34**, W32-W37.
Libbrecht, M.W. and Noble, W.S. (2015) Machine learning applications in genetics and genomics, *Nat Rev Genet*, **16**, 321-332.
Lin, H. and Ding, H. (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition, *Journal of Theoretical Biology*, **269**, 64-69.

Lin, M.M. and Zewail, A.H. (2012) Hydrophobic forces and the length limit of foldable protein domains, *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 9851-9856.
Lin, W.-Z., *et al.* (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model, *PLoS ONE*, **6**, e24756.
Liolios, K., *et al.* (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata, *Nucleic Acids Research*, **38**, D346-D354.
Liu, B., Wang, S. and Wang, X. (2015) DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation, *Scientific Reports*, **5**, 15479.
Liu, B., *et al.* (2013) Protein Remote Homology Detection by Combining Chou's Pseudo Amino Acid Composition and Profile-Based Protein Representation, *Molecular Informatics*, **32**, 775-782.
Liu, M. and Grigoriev, A. (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes; evidence of exon shuffling?, *Trends in Genetics*, **20**, 399-403.
Liu, M., *et al.* (2005) Significant expansion of exon-bordering protein domains during animal proteome evolution, *Nucleic Acids Research*, **33**, 95-105.
Long, M., *et al.* (2003) The origin of new genes: glimpses from the young and old, *Nat Rev Genet*, **4**, 865-875.
Long, M., *et al.* (2003) Origin of new genes: evidence from experimental and computational analyses. In Long, M. (ed), *Origin and Evolution of New Gene Functions*. Springer Netherlands, Dordrecht, pp. 171-182.
Long, M. and Langley, C.H. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila, *Science*, **260**, 91.
Long, M., Wang, W. and Zhang, J. (1999) Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in Drosophila, *Gene*, **238**, 135-141.
Margulis, L. (1981) *Symbiosis in Cell Evolution*. New York.
Martin, W. and Herrmann, R.G. (1998) Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why?, *Plant Physiology*, **118**, 9-17.
Martin, W., *et al.* (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus, *Proceedings of the National Academy of Sciences*, **99**, 12246-12251.
Martins, E.P. and Hansen, T.F. (1997) Phylogenies and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data, *The American Naturalist*, **149**, 646-667.
Massange-Sanchez, J.A., *et al.* (2015) The novel and taxonomically restricted Ah24 gene from grain amaranth (Amaranthus hypochondriacus) has a dual role in development and defense, *Frontiers in Plant Science*, **6**, 602.
Matthews, B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et Biophysica Acta (BBA) - Protein Structure*, **405**, 442-451.
McCarrey, J.R. and Thomas, K. (1987) Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene, *Nature*, **326**, 501-505.

McCutcheon, J.P. and Moran, N.A. (2012) Extreme genome reduction in symbiotic bacteria, *Nat Rev Micro*, **10**, 13-26.
Mereschkowsky, C. (1905) *Über Natur und Ursprung der Chromatophoren im Pflanzenreiche*.
Meyer, D., *et al.* (2014) e1071: Misc Functions of the Department of Statistics (e1071), *R package version 1.6-4*.
Middleton, S., Song, T. and Nayak, S. (2010) Length constraints of multi-domain proteins in metazoans, *Bioinformation*, **4**, 441-444.
Mishra, N.K., Chang, J. and Zhao, P.X. (2014) Prediction of Membrane Transport Proteins and Their Substrate Specificities Using Primary Sequence Information, *PLoS ONE*, **9**, e100278.
Mohamed, A. (2005) Survey on multiclass classification methods. Technical Report, Caltech.
Moore, A.D., *et al.* (2008) Arrangements in the modular evolution of proteins, *Trends in Biochemical Sciences*, **33**, 444-451.
Moreira, D., Le Guyader, H. and Philippe, H. (2000) The origin of red algae and the evolution of chloroplasts, *Nature*, **405**, 69-72.
Moreno-Hagelsieb, G. and Latimer, K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits, *Bioinformatics*, **24**, 319-324.
Nagao, C., Nagano, N. and Mizuguchi, K. (2014) Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests, *PLoS ONE*, **9**, e84623.
Nagy, A. and Patthy, L. (2011) Reassessing Domain Architecture Evolution of Metazoan Proteins: The Contribution of Different Evolutionary Mechanisms, *Genes*, **2**, 578.
Nakabachi, A., *et al.* (2006) The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella, *Science*, **314**, 267.
Nakariyakul, S., Liu, Z.-P. and Chen, L. (2012) Detecting thermophilic proteins through selecting amino acid and dipeptide composition features, *Amino Acids*, **42**, 1947-1953.
Ohm, R.A., *et al.* (2012) Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi, *PLoS Pathogens*, **8**, e1003037.
Palmieri, N., Kosiol, C. and Schlötterer, C. (2014) The life cycle of Drosophila orphan genes, *eLife*, **3**, e01311.
Paradis, E., Claude, J. and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language, *Bioinformatics*, **20**, 289-290.
Patthy, L. (1985) Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules, *Cell*, **41**, 657-663.
Patthy, L. (1990) Evolutionary Assembly of Blood Coagulation Proteins, *Semin Thromb Hemost*, **16**, 245-259.
Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling — a review, *Gene*, **238**, 103-114.
Pawlowski, J., *et al.* (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms, *PLoS Biol*, **10**, e1001419.
Pedregosa, F., *et al.* (2011) Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, **12**, 2825--2830.

Pérez-Rodríguez, P. and de los Campos, G. (2014) Genome-Wide Regression and Prediction with the BGLR Statistical Package, *Genetics*, **198**, 483-495.
Peterson, E.L., *et al.* (2009) Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment, *Bioinformatics*, **25**, 1356-1362.
Petsko, G.A., Ringe, D (2004) *Protein Structure and Function*. New Science Press, London.
Powers, D.M.W. (2011) Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation, *Journal of Machine Learning Technologies*, **2**, 37-63.
Pruesse, E., Peplies, J. and Glöckner, F.O. (2012) SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes, *Bioinformatics*, **28**, 1823-1829.
Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Research*, **35**, D61-D65.
Qiu, W.-R., Xiao, X. and Chou, K.-C. (2014) iRSpot-TNCPseAAC: Identify Recombination Spots with Trinucleotide Composition and Pseudo Amino Acid Components, *International Journal of Molecular Sciences*, **15**, 1746-1766.
Quast, C., *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, *Nucleic Acids Research*, **41**, D590-D596.
R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Radivojac, P., *et al.* (2013) A large-scale evaluation of computational protein function prediction, *Nat Meth*, **10**, 221-227.
Ramírez-Sánchez, O., *et al.* (2016) Plant Proteins Are Smaller Because They Are Encoded by Fewer Exons than Animal Proteins, *Genomics, Proteomics & Bioinformatics*, **14**, 357-370.
Rodríguez-Ezpeleta, N., *et al.* (2005) Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes, *Current Biology*, **15**, 1325-1330.
Rogozin, I.B., *et al.* (2005) Analysis of evolution of exon-intron structure of eukaryotic genes, *Briefings in Bioinformatics*, **6**, 118-134.
Rujan, T. and Martin, W. (2001) How many genes in Arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies, *Trends in Genetics*, **17**, 113-120.
Saeys, Y., Inza, I. and Larrañaga, P. (2007) A review of feature selection techniques in bioinformatics, *Bioinformatics*, **23**, 2507-2517.
Saraç, Ö.S., Atalay, V. and Cetin-Atalay, R. (2010) GOPred: GO Molecular Function Prediction by Combined Classifiers, *PLoS ONE*, **5**, e12382.
Schlegela, M. (1994) Molecular phylogeny of eukaryotes, *Trends in Ecology & Evolution*, **9**, 330 - 335.
Schliep, K.P. (2011) phangorn: phylogenetic analysis in R, *Bioinformatics*, **27**, 592-593.
Schmid, K.J. and Aquadro, C.F. (2001) The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes, *Genetics*, **159**, 589-598.
Shen, H.-B. and Chou, K.-C. (2007) EzyPred: A top–down approach for predicting enzyme functional classes and subclasses, *Biochemical and Biophysical Research Communications*, **364**, 53-59.

Sigrist, C.J.A., *et al.* (2010) PROSITE, a protein domain database for functional characterization and annotation, *Nucleic Acids Research*, **38**, D161-D166.
Soll, J. and Schleiff, E. (2004) Protein import into chloroplasts, *Nat Rev Mol Cell Biol*, **5**, 198-208.
Sonnhammer, E.L.L., Eddy, S.R. and Durbin, R. (1997) Pfam: A comprehensive database of protein domain families based on seed alignments, *Proteins: Structure, Function, and Bioinformatics*, **28**, 405-420.
Spencer, G., et al. (2012) multcompView: Visualizations of Paired Comparisons.
Sperschneider, J., *et al.* (2016) LOCALIZER: subcellular localization prediction of plant and effector proteins in the plant cell, *bioRxiv*.
The-UniProt-Consortium (2017) UniProt: the universal protein knowledgebase, *Nucleic Acids Research*, **45**, D158-D169.
Tiessen, A., Perez-Rodriguez, P. and Delaye-Arredondo, L. (2012) Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes, *BMC Research Notes*, **5**, 85.
Tirichine, L. and Bowler, C. (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth, *The Plant Journal*, **66**, 45-57.
Vapnik, V.N. (1995) *The nature of statistical learning theory*. Springer-Verlag New York, Inc.
Wang, D., Hsieh, M. and Li, W.-H. (2005) A General Tendency for Conservation of Protein Length Across Eukaryotic Kingdoms, *Molecular Biology and Evolution*, **22**, 142-147.
Wang, M., *et al.* (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world, *Genome Research*, **17**, 1572-1585.
Wang, W., *et al.* (2006) High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes, *The Plant Cell*, **18**, 1791-1802.
Ward, N. and Moreno-Hagelsieb, G. (2014) Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss?, *PLoS ONE*, **9**, e101850.
Wernegreen, J.J. (2012) Endosymbiosis, *Current Biology*, **22**, R555-R561.
Wilson, G.A., *et al.* (2005) Orphans as taxonomically restricted and ecologically important genes, *Microbiology*, **151**, 2499-2501.
Wilson, G.A., *et al.* (2007) Large-Scale Comparative Genomic Ranking of Taxonomically Restricted Genes (TRGs) in Bacterial and Archaeal Genomes, *PLoS ONE*, **2**, e324.
Wise, R.R. and Hoober, J.K. (2007) *The Structure and Function of Plastids*. Springer Netherlands.
Wissler, L., *et al.* (2013) Mechanisms and Dynamics of Orphan Gene Emergence in Insect Genomes, *Genome Biology and Evolution*, **5**, 439-455.
Wong, P. and Houry, W.A. (2004) Chaperone networks in bacteria: analysis of protein homeostasis in minimal cells, *Journal of Structural Biology*, **146**, 79-89.
Xu, D. and Nussinov, R. (1998) Favorable domain size in proteins, *Folding and Design*, **3**, 11-17.
Xu, L., *et al.* (2006) Average Gene Length Is Highly Conserved in Prokaryotes and Eukaryotes and Diverges Only Between the Two Kingdoms, *Molecular Biology and Evolution*, **23**, 1107-1108.

Yin, Y. and Fischer, D. (2008) Identification and investigation of ORFans in the viral world, *BMC Genomics*, **9**, 24-24.
Yu, C.-S., *et al.* (2006) Prediction of protein subcellular localization, *Proteins: Structure, Function, and Bioinformatics*, **64**, 643-651.
Yuan, Y., *et al.* (2010) Prediction of interactiveness of proteins and nucleic acids based on feature selections, *Molecular Diversity*, **14**, 627-633.
Yue, J., *et al.* (2012) Widespread impact of horizontal gene transfer on plant colonization of land, *Nat Commun*, **3**, 1152.
Zhang, J. (2000) Protein-length distributions for the three domains of life, *Trends in Genetics*, **16**, 107-109.
Zhang, S.-W., Hao, L.-Y. and Zhang, T.-H. (2014) Prediction of Protein–Protein Interaction with Pairwise Kernel Support Vector Machine, *International Journal of Molecular Sciences*, **15**, 3220-3233.
Zhou, Z.-H. (2012) *Ensemble Methods: Foundations and Algorithms*. Chapman \& Hall/CRC.

## Material suplementario

**Programas S2.1 y S2.2. Parser de archivos en formato gen Banck.**

https://www.dropbox.com/s/0ors5utvuqs4efx/count_exons_gbff_2.pl?dl=0

https://www.dropbox.com/s/mdwkxd5e4feq5db/number_and_size_exons.pl?dl=0

**Tabla S1.1. Longitud de proteína y número de exones de 51 especies de eucariontes (conjunto 1).**

https://www.dropbox.com/s/f1lpwnz0ubpbb7v/TableS1.1.xlsx?dl=0

**Tabla S1.2. Longitud de proteínas, número de exones y longitud de exones.** Resumen de estadísticas de 492 especies de eucariontes (conjunto 3).

https://www.dropbox.com/s/14zgiqide6ff5x2/TableS1.2.xlsx?dl=0

**Tabla S1.3. Correlación entre longitud de proteína y número de exones es especies del conjunto 3.** También se proporcionan los coeficientes de regresión $B_0$ y $R^2$. Se ajustó el modelo lineal sin intercepto (a través del origen).

https://www.dropbox.com/s/wku2nq5jd7nj0yx/TableS1.3.xlsx?dl=0

**Tabla S1.4. Lista de especies utilizadas en el análisis de contrastes filogenéticos (PIC).** Del total de 492 especies en el conjunto 3, solo utilicé 233 para el análisis PIC.

https://www.dropbox.com/s/w2ognm0o8qsmv2j/TableS1.4.xlsx?dl=0

**Tabla S1.5. Comparación de longitud de proteína entre distintos componentes celulares de *H. sapiens*.** Letras distintas en la columna K indican diferencias significativas entre componentes. (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05). aa = aminoácidos.

https://www.dropbox.com/s/okv14w61n3ghcc1/TableS1.5.xlsx?dl=0

**Tabla S1.6. Comparación de longitud de proteína entre distintos componentes celulares de *S. cerevisiae*.** Letras distintas en la columna K indican diferencias

significativas entre componentes. (Prueba de Kruskal-Wallis para comparaciones múltiples, P-valor < 0.05). aa = aminoácidos.

https://www.dropbox.com/s/2qw93dn5v199vs0/TableS1.6.xlsx?dl=0

**Tabla S2.1. Longitud de proteína, número de exones y longitud de exones de 746 especies de plástidos.**

https://www.dropbox.com/s/y41kh0qerxxhmpl/TableS2.1.xlsx?dl=0

**Tabla S2.2. Comparación de longitud de proteína entre proteomas de 310 especies de simbiontes contra *E. coli* y comparación de longitud de proteína entre proteínas ortólogas 310 especies de simbiontes y *E. coli*.**

https://www.dropbox.com/s/mcxdvl1exncmaxt/TableS2.2.xlsx?dl=0

**Tabla S3.1. Comparación de variables seleccionadas por los métodos BayesC e IFFS**

https://www.dropbox.com/s/cqjmwvdgmboiase/TableS3.1.xlsx?dl=0

**Tabla S3.2. Predicción de función de 1786 genes huérfanos de *A. thaliana***

https://www.dropbox.com/s/nwuos8jr20qjzyn/TableS3.2.xlsx?dl=0

**Tabla S3.3. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Unión a ADN utilizando el alfabeto aa20.**

| **Composición** | **N** | **Sn** | **Sp** | **Acc** | **Mcc** | **AUC** |
|---|---|---|---|---|---|---|
| Monómeros (m) | 20 | 70 | 84 | 0.80 | 0.55 | 0.85 |
| Dímeros-0 (d-0) | 400 | 60 | 84 | 0.75 | 0.46 | 0.82 |
| M + L + d-(0,1 y 2) + v-(2, 3 y 4) | 1421 | 63 | 85 | 0.77 | 0.48 | 0.83 |
| Selección de características | 141 | 88 | 91 | **0.90** | 0.80 | 0.97 |

N = Número de predictores, Sn = Sensibilidad, Sp = Especificidad, Acc = Exactitud, Mcc = Coeficiente de correlacion de Matthews, AUC = Área bajo la curva ROC, d = Dímeros, v = Ventanas

**Tabla S3.4. Comparación de desempeño de distintos clasificadores en el conjunto de datos de actividad de Transporte utilizando el alfabeto aa20.**

| **Composición** | **N** | **Sn** | **Sp** | **Acc** | **Mcc** | **AUC** |
|---|---|---|---|---|---|---|
| Monómeros (m) | 20 | 79 | 72 | 0.76 | 0.5 | 0.84 |
| Dímeros-0 (d-0) | 400 | 73 | 70 | 0.72 | 0.73 | 0.80 |
| M + L + d-(0,1 y 2) + v-(2, 3 y 4) | 1421 | 73 | 74 | 0.74 | 0.47 | 0.83 |
| Selección de características | **249** | **83** | **83** | **0.83** | **0.65** | 0.91 |

N = Número de características, Sn = Sensibilidad, Sp = Especificidad, Acc = Exactitud, Mcc = Coeficiente de correlacion de Matthews, AUC = Área bajo la curva ROC, d = Dímeros, v = Ventanas

**Tabla S3.5. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Actividad Enzimática utilizando el alfabeto aa20.**

| **Composición** | **N** | **Sn** | **Sp** | **Acc** | **Mcc** | **AUC** |
|---|---|---|---|---|---|---|
| Monómeros (m) | 20 | 71 | 80 | 0.75 | 0.46 | 0.83 |
| Dímeros-0 (d-0) | 400 | 76 | 78 | 0.77 | 0.51 | 0.84 |
| M + L + d-(0,1 y 2) + v-(2, 3 y 4) | 1421 | 71 | 90 | 0.80 | 0.57 | 0.86 |
| Selección de características | **249** | **81** | **83** | **0.82** | **0.60** | 0.91 |

N = Número de características, Sn = Sensibilidad, Sp = Especificidad, Acc = Exactitud, Mcc = Coeficiente de correlacion de Matthews, AUC = Área bajo la curva ROC, d = Dímeros, v = Ventanas

**Tabla S3.6. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Localización Celular utilizando el alfabeto aa20.**

| Categoría | m | d-0 | Todos | Selección de características |
|---|---|---|---|---|
| N | 20 | 400 | 1421 | - |
| Cloroplasto | 55 | 66 | 76 | **84** |
| Citoplasma | 66 | 70 | 73 | **81** |
| Citoesqueleto | 42 | 55 | 25 | **52** |
| Retículo endoplásmico | 39 | 55 | 64 | **66** |
| Extracelular | 71 | 78 | 83 | **88** |
| Aparato de Golgi | 11 | 24 | 67 | **40** |
| Lisosoma | 52 | 69 | 78 | **66** |
| Membrana | 90 | 90 | 95 | **94** |
| Mitocodria | 39 | 45 | 59 | **74** |
| Nucleo | 84 | 84 | 88 | **89** |
| Peroxisoma | 8 | 22 | 64 | **56** |
| Vacuola | 24 | 44 | 50 | **44** |
| TOTAL | 71 | 75 | 80 | **85** |

N = Número de características, m = monómeros, d = Dímeros. Valores de ACC en %.

**Tabla S3.7. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Familias de Transportadores utilizando el alfabeto aa20.**

| Categoría | m | d-0 | Todos | Selección de características |
|---|---|---|---|---|
| N predictores | 20 | 400 | 1421 | - |
| Amino | 24 | 44 | 83 | **86** |
| Anión | 2 | 8 | 5 | **32** |
| Catión | 72 | 65 | 67 | **87** |
| Electrón | 53 | 35 | 65 | **68** |
| Otro | 38 | 43 | 57 | **74** |
| Proteína | 36 | 23 | 59 | **59** |
| Azúcar | 22 | 42 | 75 | **73** |
| TOTAL | 44 | 45 | 61 | **74** |

N = Número de características, m = monómeros, d = Dímeros. Valores de ACC en %.

**Tabla S3.8. Comparación de desempeño de distintos clasificadores en el conjunto de datos de Familias de Enzimas utilizando el alfabeto aa20.**

| Categoría | m | d-0 | Todos | Selección de características |
|---|---|---|---|---|
| N predictores | 20 | 400 | 1421 | - |
| Hydrolasas | 60 | 59 | 74 | **80** |
| Isomerasas | 19 | 17 | 48 | **64** |
| Ligasas | 45 | 40 | 79 | **83** |
| Lyasas | 10 | 10 | 36 | **55** |
| Oxidorectasas | 41 | 34 | 64 | **71** |
| Transferasas | 59 | 57 | 69 | **72** |
| TOTAL | 48 | 45 | 66 | **73** |

N = Número de características, m = monómeros, d = Dímeros. Valores de ACC en %.