UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Análisis de la estructura de la familia de proteínas Cold Shock y reconstrucción de estados ancestrales

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestra en Ciencias

PRESENTA:
Mariana Muñoz Argott

TUTOR PRINCIPAL
Dr. León Patricio Martínez Castilla
Facultad de Química, UNAM

MIEMBROS DEL COMITÉ TUTOR
Dr. Arturo Carlos II Becerra Bracho
Facultad de Ciencias, UNAM

Dra. Alejandra Hernández Santoyo
Instituto de Química, UNAM

Ciudad de México. Abril, 2019

CIENCIAS Bioquímicas
Programa de Maestría y Doctorado

PMDCB/1025/2018

Muñoz Argott Mariana
Estudiante de Maestría en Ciencias Bioquímicas
P r e s e n t e

Los miembros del Subcomité Académico en reunión ordinaria del día 21 de mayo del presente año, conocieron su solicitud de asignación de **JURADO DE EXAMEN** para optar por el grado de **MAESTRÍA EN CIENCIAS**, con la réplica de la tesis **"Análisis de la estructura de la familia de proteínas Cold Shock y reconstrucción de estados ancestrales"**, dirigida por el/la Dr(a). **Martínez Castilla León Patricio**.

De su análisis se acordó nombrar el siguiente jurado integrado por los doctores:

PRESIDENTE	Rodríguez Sotres Rogelio
VOCAL	Rojo Domínguez Arturo
VOCAL	Alcaraz Peraza Luis David
VOCAL	Arciniega Castro Marcelino
SECRETARIO	Sosa Peinado Alejandro

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

A t e n t a m e n t e
"POR MI RAZA, HABLARÁ EL ESPÍRITU"
Cd. Universitaria, Cd. Mx., a 21 de mayo de 2018.
COORDINADORA

Dra. ANA BRÍGIDA CLORINDA ARIAS ÁLVAREZ

contacto: mdcbq@posgrado.unam.mx Tel. 5623 7006

Agradecimientos Institucionales

A la Universidad Nacional Autónoma de México por otorgarme una educación pública de alta calidad.

Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico que me brindó mediante el programa de becas nacionales (CVU: 740751).

Agradezco a la Técnica Académica Titular B, M. en C. Fabiola Ramírez Corona por permitirme hacer uso de las instalaciones y el equipo de cómputo del Taller de Sistemática y Biogeografía de la Facultad de Ciencias de la UNAM para la realización de este proyecto.

Agradecimientos

Agradezco a mi tutor el Dr. León Patricio Martínez Castilla por el apoyo, la orientación, confianza y los consejos. Gracias por brindarme una valiosa experiencia de trabajo científico.

A los miembros de mi comité tutor, la Dra. Alejandra Hernández Santoyo y el Dr. Arturo Carlos II Becerra Bracho por las valiosas aportaciones a este proyecto.

A la M. en C. Fabiola Ramírez Corona por todo su apoyo y confianza a lo largo de mi formación académica y científica, sin su ayuda no hubiera sido posible la realización de este proyecto.

Al M. en C. Octavio Abraham Moreno Guillén por las aportaciones y sugerencias para mejorar la calidad del escrito de este trabajo de Tesis.

A los miembros del jurado de titulación, Dr. Rogelio Rodríguez Sotes, Dr. Arturo Rojo Domínguez, Dr. Alejandro Sosa Peinado, Dr. Marcelino Arciniega Castro, y al Dr. Luis David Alcaraz Peraza, por sus correcciones, aportaciones y comentarios.

A mis compañeras y compañeros del Laboratorio de Bioinformática por sus consejos y observaciones.

A toda mi familia por el apoyo que me ha permitido seguir el camino de mi elección.

A mis amigas Fabiola, Nelly y Verónica por todos estos años de amistad, por brindarme varios de los mejores momentos de mi vida y por darme ánimo en los momentos de incertidumbre.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Abreviaturas

AMBER	Assisted Model Building with Energy Refinement
ASR	Ancestral Sequence Reconstruction
DNA	Deoxyribonucleic acid
ssDNA	Single-stranded DNA
CSD	Cold Shock Domain
Csp	Cold Shock Protein
GFP	Green Fluorescent Protein
HSP	Heat Shock Protein
Indels	Insertion-deletion
IRES	Internal ribosome entry site
ITAF	IRES trans-actin factor
MAFFT	Multiple sequence alignment based on fast Fourier transform
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NMR	Nuclear magnetic resonance
OB-fold	Oligonucleotide/oligosaccharide-binding fold
PDB	Protein Data Bank
PROMALS	PROfile Multiple Alignment with predicted Local Structure
RNA	Ribonucleic acid
mRNA	Messenger ribonucleic acid
rRNA	Ribosomal ribonucleic acid
SD	Shine-Dalgarno
SNAP	Screening for Non-Acceptable Polymorphisms
SNP	Single nucleotide polymorphisms
UTR	Untranslated Transcribed Region

Resumen

Los análisis computacionales de genomas nos han facilitado el estudio detallado de la materia prima en ellos contenido, los genes que codifican para proteína. Diversos métodos de tipo predictivo se han desarrollado para responder preguntas concretas sobre tales proteínas, como la predicción de la estructura y función de las proteínas, la predicción de sitios activos, la búsqueda de motivos estructurales conservados, la reconstrucción de secuencias ancestrales (ASR), entre otros.

La reconstrucción computacional de la secuencia y estructura de proteínas ancestrales proporcionan la oportunidad de hacer un estudio aproximado de las características de las proteínas antiguas. Siempre con referencia en la secuencia observada actual de tales proteínas, es posible hacer inferencias no sólo de los cambios ocurridos, sino de las implicaciones o influencia de tales cambios, por lo que esta herramienta ofrece una oportunidad interesante para mirar hacia el pasado y así hipotetizar sobre los cambios ocurridos hasta la proteína actual. Es pues, el análisis evolutivo de una secuencia proteica.

En el presente trabajo se llevó a cabo una reconstrucción de secuencias ancestrales de la familia de proteínas CSP que se encuentra en casi todas las bacterias psicrófilas, mesófilas, termófilas e hipertermófilas estudiadas, así como del dominio CSD que se ha encontrado en varias proteínas de arqueas y eucariontes. Reportes en literatura confirman que varios miembros de esta familia de proteínas son necesarias en condiciones normales de crecimiento y presumiblemente fundamentales para la adaptación a una disminución drástica de la temperatura. De igual forma, en la literatura se ha postulado que el mecanismo de acción para llevar a cabo tales funciones radica en su capacidad de regulación post-transcripcional de genes. La hipótesis de este trabajo se fundamenta en que el mecanismo de acción de las CSP presumiblemente implica interacciones con fines regulatorios hacia DNA y/o RNA, por lo que haciendo uso de técnicas de reconstrucción de secuencia ancestral (ASR) y la subsecuente predicción de la estructura de la misma, se analizó el nivel de cambio y conservación a nivel de secuencia y estructura que ha permeado la evolución de las CSP.

Se concluyó que la estructura y función de las proteínas CSP y el dominio CSD se han conservado prácticamente intactos a lo largo del tiempo, y dado que el plegamiento que adoptan en su forma tridimensional se observa en proteínas muy antiguas que interactúan con DNA y RNA, es posible que las proteínas precursoras de las CSP y el dominio CSD sean tan antiguos como la separación de los dominios Bacteria, Archaea y Eukarya; sin embargo, su distribución filogenética sesgada, en la que se puede apreciar una marcada ausencia en el dominio Archaea, sea evidencia de una historia de pérdidas y ganancias (mediante Transferencia Horizontal de Genes) de la proteína y su dominio.

Introducción

Gracias al creciente implemento de la secuenciación masiva, las bases de datos de secuencias de biomoléculas han tenido un aumento considerable, presentándonos así un nuevo tipo de biodiversidad. Los análisis computacionales de los genomas procariotas y eucarióticos, nos han ayudado a conocer la función de los genes y las proteínas (Chothia & Gough, 2009), así como han confirmado la importancia del dominio proteico como unidad fundamental en la evolución. Entender cómo podría evolucionar la biodiversidad funcional de genes y proteínas es posiblemente la cuestión central en la evolución molecular (C. A. Orengo & Thornton, 2005).

Durante la evolución, las proteínas pueden cambiar sus secuencias y divergir por mutaciones, como sustituciones de un solo nucleótido o también por inserciones y *deleciones* de múltiples residuos (*indels*), dando lugar a grupos de proteínas homólogas. Las proteínas se pueden agrupar en una clasificación jerárquica por la cual los parientes muy cercanos, por lo general con una alta similitud de secuencia (por ejemplo, >40% de identidad de secuencia), se agrupan en familias. Estos parientes cercanos con frecuencia comparten propiedades funcionales comunes. Los homólogos más remotos, que tienen menor similitud de secuencia (<30%), se agrupan en familias o superfamilias, éstas últimas que abarcan relaciones filogenéticas más amplias. Estos términos fueron acuñados por primera vez por Margaret Dayhoff (Dayhoff, 1978) basado en su reconocimiento del grado en que las proteínas pueden divergir dentro de una familia. Es difícil reconocer parientes muy divergentes comparando sus secuencias por sí solas, por lo que los homólogos muy remotos sólo se han podido detectar comparando su estructura. A veces, puede surgir confusión al decidir si un grupo de homólogos representa una familia cercana o una superfamilia más amplia (C. A. Orengo & Thornton, 2005). Por este motivo, en este trabajo se usará el término familia en el sentido más amplio, es decir, que contiene a todos los parientes, ortólogos y parálogos.

Para entender las relaciones evolutivas y poder utilizarlas, se han tenido que emplear modelos evolutivos basados en técnicas bioinformáticas, como la predicción de sitios activos, la predicción de la estructura y función de las proteínas, la búsqueda de motivos estructurales y secuenciales conservados, genómica comparativa, predicción de estructuras de los RNAs, reconstrucción de dendogramas y reconstrucción de secuencia ancestral (ASR, del inglés *ancestral sequence reconstruction*), etc. Sin embargo, al usar modelos evolutivos, debemos tener cuidado al elegir los supuestos del modelo. Hay muchos ejemplos de estudios en los que el uso de modelos poco realistas de la evolución de la secuencia conducen a inferencias erróneas (Pupko, Huchon, Cao, Okada, & Hasegawa, 2002). Actualmente, se está llevando a cabo un gran esfuerzo para desarrollar modelos más realistas de la evolución de proteínas y ácidos nucleicos que tomen en cuenta fenómenos biológicos, además de que emplean técnicas estadísticas para conocer los parámetros de estos modelos y poder hacer predicciones con ellos (Liberles, 2007).

Reconstrucción de secuencias ancestrales

La reconstrucción computacional y la síntesis de secuencias ancestrales proporcionan una oportunidad de estudiar secuencias similares a aquellas que existieron hace millones de años. Éste enfoque nos permite comprender sus interacciones y especificidades en la función, entre otras (Liberles, 2007).

La inferencia de secuencias ancestrales (ASR, del inglés *ancestral sequence reconstruction*) bajo una filogenia dada se ha convirtiendo en un enfoque importante en Biología Molecular (Golding & Dean, 1998). Esto se debe a que uno de los objetivos centrales en la Evolución Molecular es comprender los mecanismos y la dinámica por los cuales los cambios en la secuencia del gen generan cambios en la función de la proteína y, por lo tanto, en el fenotipo (Golding & Dean, 1998; Stern, 2000). Una comprensión completa de este proceso requiere un análisis de cómo los cambios en la estructura de la proteína median los efectos de las mutaciones en la función. Los análisis

comparativos de las proteínas existentes han proporcionado información indirecta sobre la diversificación de la estructura de las proteínas (Kinch & Grishin, 2002), y los estudios de ingeniería de proteínas han dilucidado las relaciones de estructura-función que dan forma al proceso evolutivo (Bershtein, Segal, Bekerman, Tokuriki, & Tawfik, 2006). Sin embargo, para identificar directamente los mecanismos por los cuales las mutaciones acumuladas generaron nuevas funciones, es necesario analizar los cambios ocurridos y acumulados a través de la evolución (Ortlund, Bridgham, Redinbo, & Thornton, 2007). Por ésta razón, la ASR proporciona una alternativa más para comprender dichos mecanismos. Como ejemplo, los considerados como cambios clave durante la evolución de una proteína pueden reintroducirse en su versión ancestral para inferir el efecto (ver más adelante un caso de estudio con las proteínas GFP). Tal cual es el caso de este trabajo, veremos que es posible hacer reconstrucción de secuencias ancestrales en combinación con otras técnicas, como la Biología estructural, para revelar los mecanismos por los cuales los residuos de aminoácidos de las proteínas pueden interactuar con moléculas y ligandos, conduciendo a funciones complejas (Harms & Thornton, 2010).

Los trabajos del laboratorio de Mikhail Matz, ilustran cómo las secuencias ancestrales reconstruidas (ASR) proporcionan un marco teórico robusto para identificar aminoácidos clave y probar el impacto funcional de las mutaciones en proteínas actuales. Mediante el uso de ASR caracterizaron las secuencias antiguas en toda la familia de proteínas fluorescentes similares a la proteína verde fluorescente (GFP, del inglés *green fluorescent protein*) de los corales dentro del suborden *Faviina*. Dichas proteínas son particularmente diversas en el color de su fluorescencia. Mediante los trabajos de Matz y colaboradores, se descubrió que la proteína ancestral fluorescente del suborden *Faviina* tenía una fluorescencia en color verde, seguida de diversificación en una variedad de otros colores (Field & Matz, 2010; Ugalde, Chang, & Matz, 2004). Luego, trataron de identificar las mutaciones responsables del cambio de fluorescencia hacia el color rojo en las proteínas de la estrella coralina *Montastrea cavernosa*, cuyo ancestro presentaba fluorescencia color verde del suborden *Faviina*, y descubrieron que se produjeron 37 cambios de aminoácidos entre la proteína similar a GFP del ancestro del suborden *Faviina* y la proteína de *M. cavernosa* (en comparación con las 108

diferencias entre la proteína de *M. cavernosa* y la proteína verde fluorescente de su pariente más cercano). Posteriormente, generaron una biblioteca de variantes en la que cada proteína contenía aproximadamente la mitad de dichos aminoácidos en el estado ancestral y la mitad en el estado derivado. Luego, evaluaron la fluorescencia de una gran cantidad de clones rojos y verdes de esta biblioteca y analizaron la asociación entre el caracter de estado en cada sitio y la longitud de onda que emitía la proteína. Este enfoque le permitió a Matz y sus colegas identificar el conjunto de mutaciones que probablemente contribuyan al fenotipo derivado. Descubrieron que 12 de las 37 mutaciones se asociaron significativamente con la fluorescencia roja y cuando este conjunto de cambios se introdujo en la proteína verde fluorescente ancestral, se pudo producir una proteína emisora de color rojo, indistinguible de la proteína actual rojo fluorescente (Field & Matz, 2010).

Sin embargo, en la mayoría de los casos, no es tan fácil identificar las mutaciones implicadas en las funciones actuales de las proteínas. La evolución de un cambio ventajoso en la función proteica puede verse facilitada por mutaciones neutras que no otorgan ningún beneficio cuando surgen por primera vez, pero que pueden producir un cambio sutil en la conformación o estabilidad de la proteína y así modificar el efecto funcional de mutaciones posteriores en otros sitios de la misma proteína (Bloom, Labthavikul, Otey, & Arnold, 2006; Gong, Suchard, & Bloom, 2013; Harms & Thornton, 2013; Starr & Thornton, 2016). Por ejemplo, una mutación del aminoácido en el sitio X de una proteína es neutral cuando aparece por primera vez, pero facilita la fijación de una mutación beneficiosa en el sitio Y que altera la función, haciendo entonces, perjudicial revertir el sitio X a su estado ancestral. Sin embargo, la mutación en el sitio Y es beneficiosa en un contexto en el que la mutación en el sitio X ya ha ocurrido, de lo contrario podría ser neutral o nociva. En principio, el cambio compensatorio en el sitio X podría preceder al cambio funcional en el sitio Y (en cuyo caso se llama sustitución permisiva), o podría ocurrir después, en cuyo caso habría una reducción transitoria de la aptitud. Así mismo, las dos mutaciones también podrían fijarse simultáneamente si ocurrieran en el mismo haplotipo de secuencia (Kimura, 1985; Storz, 2018).

Las sustituciones compensatorias son fundamentales para las preguntas sobre el papel de la contingencia histórica en la configuración de las vías y los resultados de la evolución de las proteínas. Si los efectos de la aptitud de las mutaciones de aminoácidos están condicionados a los antecedentes genéticos, entonces las mutaciones pueden tener diferentes efectos, dependiendo del orden secuencial en que se producen (DePristo, Weinreich, & Hartl, 2005). En consecuencia, la historia acumulada de sustituciones en el pasado influirá en el conjunto de mutaciones permisibles en el futuro, y los resultados evolutivos dependerán históricamente de los puntos de partida ancestrales (Starr, Picton, & Thornton, 2017; Starr & Thornton, 2016). Por esta razón, el estudio de la epistasis intramolecular, junto con la reconstrucción de secuencias ancestrales (ASR), puede contribuir significativamente a la disponibilidad de trayectorias evolutivas al restringir el orden y la reversibilidad de las sustituciones de aminoácidos, ya que una mutación dada puede ser neutral, beneficiosa o perjudicial dependiendo del contexto genético en el que se produce (Storz, 2018).

Las reconstrucciones de secuencias ancestrales se llevan a cabo desde los años 70's, Fitch (Fitch, 1971), Sankoff (Sankoff, 1975) y Rousseau (Sankoff & Rousseau, 1975) desarrollaron algoritmos basados en máxima parsimonia con este propósito y hasta los 90's la máxima parsimonia fue el método de elección (Jermann, Opitz, Stackhouse, & Benner, 1995). Sin embargo, la historia de ASR siguió los pasos de la reconstrucción de filogenias moleculares, y con el tiempo se empezaron a reemplazar los modelos de máxima parsimonia por modelos probabilísticos. Ello debido a que numerosos estudios demostraron las deficiencias de los modelos con el enfoque de máxima parsimonia (Holder & Lewis, 2003) y como esta se encuentra intrínsecamente sesgada hacia una sobreestimación del número de cambios comunes a raros (Eyre-Walker, 1998). Además, no proporciona medios estadísticamente robustos para discriminar entre reconstrucciones igualmente parsimoniosas (Yang, Kumar, & Nei, 1995). Debido a lo anterior se introdujo el concepto de reconstrucción de secuencia probabilística (Koshi & Goldstein, 1998; Yang *et al*., 1995) y se desarrollaron algoritmos eficientes para la reconstrucción utilizando el enfoque probabilístico (Koshi & Goldstein, 1998; Pupko, Pe, Shamir, & Graur, 2000).

Al igual que en las reconstrucciones filogenéticas, la consideración de la heterogeneidad de las tasas evolutivas entre sitios (Yang, 1993) le dio mayor robustez a las reconstrucciones de secuencias ancestrales, ya que se demostró que no tener en cuenta la variación de la tasa entre sitios reduce la precisión de la ASR, y da como resultado una menor verosimilitud para las secuencias reconstruidas (Pupko, Pe'er, Hasegawa, Graur, & Friedman, 2002a). Con base en lo anterior, se puede decir que los modelos que se adaptan mejor a los datos para la reconstrucción de filogenias también son mejores para la ASR (Liberles, 2007). Para mejorar los modelos de las ASR se tiene que tomar en cuenta la variación de la tasa de sustitución entre diferentes aminoácidos (la matriz de sustitución), la variación de la matriz de sustitución entre diferentes sitios y entre diferentes ramas del dendograma, por lo que el dendograma y su topología se consideran parte del modelo probabilístico (Liberles, 2007).

ASR basado en el criterio de máxima parsimonia

La estrategia de reconstrucción de secuencias ancestrales basada en máxima parsimonia identifica los estados ancestrales en cada nodo de un dendograma que minimizan el número de cambios de caracteres necesarios para explicar las diferencias observadas entre las secuencias de dicho dendograma (Liberles, 2007). Los algoritmos para ASR basados en este criterio fueron desarrollados por Fitch (Fitch, 1971), Sankoff (Sankoff, 1975) y Sankoff y Rousseau (Sankoff & Rousseau, 1975). Estos algoritmos usan programación dinámica, asegurando una reconstrucción eficiente. El algoritmo de *downpass* (Fitch, 1971), desarrollado con datos de secuencia de nucleótidos, penaliza igualmente cualquier cambio entre los cuatro estados de caracteres (A, C, G y T). Para la reconstrucción de una posición específica, el algoritmo procede asignando a cada nodo del dendogama un conjunto de estados de caracteres que son compatibles con el número mínimo de cambios. El algoritmo procesa el dendograma del exterior al interior; es decir, que primero se asigna un conjunto de caracteres a las secuencias en las puntas del dendograma (secuencias más externas) y cada nodo del dendograma se visita solo después de que se visitan sus descendientes. Si, por ejemplo, una punta está etiquetada por el carácter A, se asigna un conjunto {A} a esa punta (Fig.1). A

continuación, se evalúa un nodo interno para el que ya se visitaron ambos descendientes. El conjunto asignado a este nodo interno es la intersección de los conjuntos en sus dos nodos descendientes si esta intersección no está vacía, o la unión de los dos conjuntos si la intersección está vacía. Si el nuevo conjunto es una unión, se cuenta un cambio para que el número total sea el número de las operaciones de la unión. El siguiente paso, llamado *top-down*, es atravesar el dendograma desde la raíz hasta las puntas, en orden, para determinar los estados ancestrales de los nodos internos. Inicialmente, el estado ancestral en la raíz es igual al estado del carácter en su conjunto. Si este conjunto incluye más de un estado de caracter, entonces existen reconstrucciones diferentes, igualmente parsimoniosas. En la segunda etapa del algoritmo, llamada *bottom-up*, cada descendiente de la raíz se evalúa de la siguiente manera. Si el estado ancestral en la raíz es un miembro del conjunto del nodo descendiente, se asigna el mismo estado ancestral al nodo descendiente; de lo contrario, otro estado del conjunto en el nodo descendente será elegido (Liberles, 2007). Dicho procedimiento se aplica para cada nodo en el dendograma y encontrará algunas de las reconstrucciones más parsimoniosas, pero no todas (Liberles, 2007). Para garantizar que se encuentren todas las reconstrucciones más parsimoniosas, se deben realizar comparaciones que involucren a los grupos externos de un nodo (Harvey & Pagel, 1991).

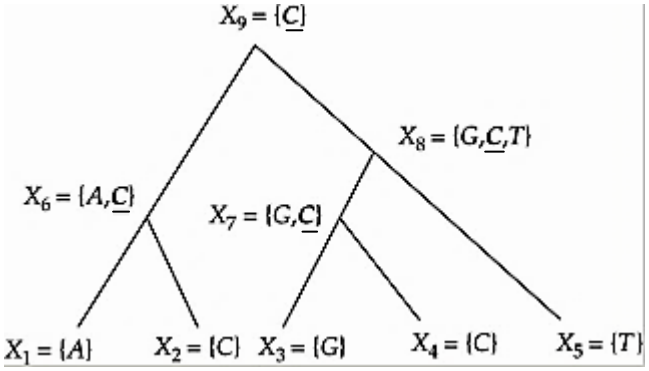


Fig. 1. Ejemplo de reconstrucción de secuencias ancestrales (ASR) con el algoritmo de máxima parsimonia de Fich.
En las puntas del dendograma, los conjuntos de caracteres son: X_1={A}, X_2={C}, X_3={G}, X_4={C}, y X_5={T}. En el nodo X_6, la intersección de los conjuntos de sus dos descendientes, X_1 y X_2 es {A}∩{C}={}, por lo que, el conjunto de unión es: {A}∪{C}={A, C}. Del mismo modo se asigna al nodo X_7 el conjunto de unión de X_3 y X_4; es decir, {G}∪{C}={G, C}. Ahora se puede determinar el conjunto en el nodo X_8, ya que la intersección de los conjuntos X_5 y X_7 está vacía, la unión de estos conjuntos será {G, C}∪{T}={G, C, T}. Finalmente, el conjunto en la raíz (X_9) es la intersección de los conjuntos X_8 y X_6: {A, C}∩{G, C, T}={C}. Se necesitaron tres operaciones de unión; por lo tanto, se necesita un mínimo de tres cambios para esta reconstrucción. Posteriormente, se determinan los estados ancestrales (marcados con una línea) atravesando el dendograma desde la raíz hasta las puntas. Primero, el estado C se determina en la raíz; el estado en X_8 también se establece en C ya que este estado es el estado ancestral en el nodo de origen (X_9) y es un carácter del conjunto de este nodo (X_8). De forma similar, el estado en X_6 es C ya que este estado es el estado ancestral en el nodo de origen, así como un carácter del conjunto en el nodo X_6. Finalmente, se asigna el estado en el nodo X_7, que también es igual a C (Tomado de Liberles, 2007)

El algoritmo de Sankoff (Sankoff, 1975) es una generalización del algoritmo de Fitch. En lugar de suponer que todos los cambios de estado son igualmente probables, permite diferentes costos para diferentes cambios de carácter. De manera similar al algoritmo de Fitch, el dendograma se valora primero tomando en cuenta las secuencias externas y posteriormente toma en cuenta las secuencias desde la raíz hasta las secuencias externas. Ambos algoritmos pueden reconstruir más de un estado ancestral para cada nodo (Liberles, 2007).

ASR usando máxima verosimilitud

La reconstrucción de secuencias ancestrales actual utiliza distribuciones estadísticas para generar probabilidades posteriores de diferentes reconstrucciones a partir de sitios específicos en la alineación de las secuencias. Para cada sitio de la secuencia deducida en un nodo filogenético, los valores posteriores para todos los 20 aminoácidos se calculan y se representa la probabilidad de que un aminoácido particular ocupe un sitio específico en la proteína durante su evolución. Esta distribución de probabilidad posterior se calcula a partir de los patrones de los aminoácidos en las secuencias modernas, descrita a partir de una filogenia, una matriz de probabilidades de sustitución, las frecuencias en equilibrio de los aminoácidos, las longitudes de ramas del dendograma y las tasas de sustitución de sitios específicos. La secuencia ancestral más probable utiliza el aminoácido con la probabilidad posterior mayor en cada sitio dentro de la distribución (Ashkenazy *et al*., 2012).

Para ejemplificar el concepto de ASR usando máxima verosimilitud, consideremos un dendograma simple de cuatro *taxa* (Figura 2) y consideremos sólo dos estados de carácter (0 o 1); por ejemplo, aminoácidos polares y no polares. Las asignaciones de caracteres ancestrales x_5 y x_6 en los nodos internos X_5 y X_6 son desconocidas. Los números sobre las ramas indican longitudes de rama; es decir, el número promedio de sustituciones por sitio de la secuencia. En todos los modelos basados en probabilidad, las probabilidades se expresan en términos de sumatorias y multiplicaciones de factores P*ij*(t). P*ij*(t), expresa la probabilidad de que el carácter *i* sea reemplazado por el carácter *j* a lo largo de una rama de longitud (t). Los factores P*ij*(t) generalmente se expresan en una forma de matriz P(t), de modo que [P(t)]*ij* = P*ij*(t). La matriz P(t) puede calcularse mediante $P(t) = e^{Qt}$, donde Q es la matriz de velocidad instantánea, que determina las probabilidades de sustitución dependiendo si se trata de DNA (codificante o no codificante), RNA, aminoácidos o codones (Felsenstein, 2004). Los modelos de probabilidad también contiene probabilidades iniciales: para cada carácter x, P(x) denota la probabilidad de observar x en la raíz del dendograma.

La verosimilitud de los datos describe la probabilidad de observar los caracteres en las puntas del dendograma dada su topología, las longitudes de las ramas y los factores P*ij*(t). Esta probabilidad es una suma de cuatro términos diferentes, cada uno correspondiente a una asignación de secuencia ancestral específica (x_5=x_6=0, x_5=x_6=1, x_5=0 y x_6=1, y x_5=1 y x_6=0). Para este ejemplo, el nodo interno X_5 se eligió arbitrariamente como la raíz del dendograma (Fig.2). La mayoría de los modelos evolutivos utilizados son reversibles en el tiempo. En términos matemáticos, un modelo es reversible en el tiempo si P(*i*)P*ij*(t)=P*(j*)P*ji*(t) para todos los pares de caracteres, (*i* y *j*). Felsenstein (1981) demostró que para los modelos reversibles en el tiempo, la posición de la raíz del dendograma no afecta el puntaje de verosimilitud. La asignación de caracteres conjunta (x_5, x_6) que más contribuye a la probabilidad anterior se llama reconstrucción de secuencias ancestrales (ASR) conjunta (Liberles, 2007).

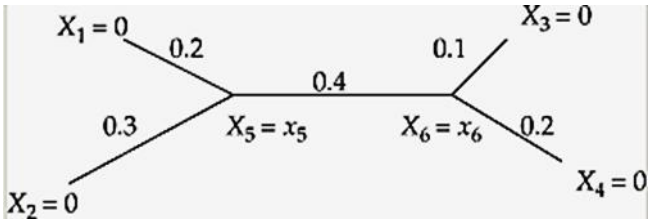


Fig. 2. Ejemplo de reconstrucción de secuencias ancestrales (ASR) usando máxima verosimilitud. El ejemplo es para el caso simple de un alfabeto de dos estados (0 y 1). Los números sobre las ramas reflejan distancias evolutivas. En la reconstrucción conjunta, el par de caracteres que se asigna a los dos nodos internos X_5 y X_6 es el que maximiza la probabilidad de los datos. Esta probabilidad es una suma de cuatro términos diferentes, cada uno correspondiente a una asignación de secuencia ancestral específica (x_5=x_6=0, x_5=x_6=1, x_5=0 y x_6=1 y x_5=1 y x_6=0).

Organización jerárquica del plegamiento espacial de las proteínas

Los proyectos de secuenciación genómica en los diversos organismos que habitan el planeta han logrado que se tenga un amplio repertorio de proteínas cuya secuencia es conocida. El análisis de estos datos no sólo nos ha ayudado en el entendimiento de los procesos por los cuales se forman nuevas proteínas sino también a trazar la relación evolutiva de éstas (Chothia & Gough, 2009). Los análisis computacionales de datos de genomas procariotas y eucarióticos han confirmado la importancia del dominio proteico como unidad fundamental en la evolución, y han revelado la asombrosa diversidad de proteínas que pueden ensamblarse al ocurrir la duplicación de dominios (Apic, Gough, & Teichmann, 2001; Chothia, Gough, Vogel, & Teichmann, 2003). Varias proteínas presentes en los organismos están conformadas por más de un dominio proteico. La importancia de la duplicación de dominios en la evolución ha sido reconocida durante mucho tiempo, pero es gracias al análisis de genomas completos que se ha confirmado hasta qué punto está duplicación está ocurriendo (Apic *et al*., 2001). En células procariotas, al menos el 70% de los dominios se encuentran duplicados, mientras que en eucariotas se llega a alcanzar la cifra de 90% de dominios duplicados (Chothia *et al*., 2003).

Las proteínas pequeñas contienen sólo un dominio con una función particular. Los dominios suelen tener 50-200 residuos, aunque se producen dominios más pequeños y más grandes. Los dominios se pueden agrupar en familias o superfamilias, cuyos miembros descienden de un ancestro común. La capacidad de detectar las relaciones evolutivas de los dominios por similitud de secuencia es limitada porque con frecuencia divergen más allá del punto donde las verdaderas relaciones pueden reconocerse por este medio (Chothia & Gough, 2009). Sin embargo, no mucho después del descubrimiento del DNA y del desarrollo de las tecnologías de secuenciación de proteínas, los métodos para determinar las estructuras tridimensionales se establecieron a fines de los años 70’s. Estos métodos permitieron a los biólogos inspeccionar y probar las interacciones entre los residuos de aminoácidos que determinan el pliegue, y la manera en que las proteínas interactúan con otras proteínas y sustratos en su entorno. A

medida que aumentó el número de estructuras resueltas mediante la técnica de cristalografía de rayos X y la técnica de resonancia magnética nuclear (NMR, del inglés *nuclear magnetic resonance*), quedó claro que la estructura de la proteína está mucho más conservada durante la evolución que la secuencia de la proteína. A diferencia de la secuencia proteica, para la que en algunas familias se han detectado parientes que comparten menos del 5% de residuos idénticos, al menos el 50% de la estructura, principalmente en el núcleo de la proteína, está muy conservada y puede usarse para detectar parientes muy distantes (Orengo & Thornton, 2005).

No obstante, sigue siendo difícil determinar la relación evolutiva entre dos dominios proteicos que no poseen una similitud de secuencia significativa y no hay características indicativas de propiedades funcionales comunes, aunque la similitud estructural sea aparente entre los dos dominios. En estos casos se puede llegar a pensar que estos dos dominios son posiblemente dos parientes muy distantes de la misma familia evolutiva que se distanciaron tanto en la secuencia como en la función, sin embargo también existe la posibilidad de que los dos dominios hayan provenido de diferentes proteínas o dominios ancestrales, pero que con el tiempo convergieron en la misma disposición estructural (Orengo & Thornton, 2005).

Clasificación estructural de las proteínas

Las proteínas son polímeros compuestos por 20 aminoácidos diferentes unidos entre sí mediante enlaces peptídicos. En solución acuosa, bajo condiciones fisiológicas de presión y temperatura estándar, las cadenas polipeptídicas de las proteínas suelen plegarse para formar, en la mayoría de los casos, estructuras globulares. Antes de adquirir su configuración estructural terciaría característica, las proteínas tienen diferentes niveles de organización, incrementado cada vez más la complejidad de las interacciones físicas entre los aminoácidos que las componen. Se dice que las proteínas tienen cuatro niveles de organización: primaria, es la secuencia lineal de aminoácidos ensamblados por uniones químicas covalentes que conforman la proteína; secundaria, se refiere a la disposición particularmente estable de los aminoácidos o trayectoria espacial que adopta la cadena polipeptídica (esqueleto de la proteína), puede describirse mediante los ángulos de torsión de los enlaces químicos sencillos dentro de la cadena y se estabiliza por interacciones locales (entre residuos cercanos en la secuencia) que pueden ser repetitivas (como hélices α y hebras β) o no ser lo (como los giros, asas y estructuras caprichosas o *random coil*); terciaria, engloba todos los aspectos del plegamiento tridimensional de un polipéptido, se estabiliza por interacciones químicas débiles entre residuos de la proteína alejadas en la secuencia y puede incluir enlaces covalente laterales, que suelen ser ocasionales. Cuaternaria, corresponde al ensamblaje de varias cadenas polipeptídicas, ya sean idénticas o distintas que conforman una unidad funcional llamada "la proteína". Se encuentra estabilizada por el mismo tipo de interacciones que la estructura terciaria, pero la diferencia es que este nivel involucra los contactos entre prolipéptidos independientes (Patthy, 2008; Petsko & Ringe, 2004; Voet & Voet, 2011). Para que un polipéptido funcione como una proteína, usualmente debe ser capaz de formar una estructura terciaria estable bajo condiciones fisiológicas. Al mismo tiempo, otra característica importante para la función de las proteínas es que no deben ser demasiado rígidas, se sabe que el universo de estructuras tridimensionales que las proteínas pueden adoptar es grande, aunque finito, ya que se encuentra limitado por restricciones físico-químicas que afectan la estabilidad de las proteínas (Patthy, 2008; Petsko & Ringe, 2004; Voet & Voet, 2011).

Además de reconocer distintos patrones de plegamiento, la mayoría de las clasificaciones de estructura también describen la arquitectura de la estructura de la proteína (C. Orengo *et al*., 1997), esta es la forma general de la proteína, mientras que la topología o grupo de plegamiento describe las orientaciones relativas de las estructuras secundarias en tres dimensiones (3D) y el orden en que están conectadas, la arquitectura es un nivel superior en la clasificación y agrupa estructuras con disposiciones estructurales de nivel secundario similares independientemente de la conectividad. En el nivel más alto en una clasificación estructural, las proteínas se agrupan si pertenecen a la misma clase, es decir, si tienen composiciones de estructura secundaria y empaquetamiento similares. Hay tres clases principales, las que están compuestas principalmente por hélices α, las que están compuestas principalmente por hebras β y las que se componen tanto de hélices α como de hebras β (α–β); (Orengo & Thornton, 2005).

Dominios OB-fold

El dominio de unión a oligonucleótidos y oligosacáridos (OB-fold, del inglés *oligonucleotide/oligosaccharide-binding fold*) presenta un barril cerrado de cinco hebras β y una cara de unión. La cara tiene en su centro las hebras 2 y 3, y está limitada en la parte inferior izquierda por el bucle 1(L1), en la parte superior por el bucle 4 (L4) y en la parte superior derecha por el bucle 2 (L2). A través de las diferentes estructuras de la familia, los bucles 2 y 4 (L2 y L4) son los que presentan mayor variación, tanto en su longitud como en su secuencia. Diferentes proteínas que tienen el dominio OB-fold utilizan esta cara para unirse a RNA (los dominios de unión a anticodones), ssDNA, oligosacáridos (las toxinas bacterianas AB_5) y proteínas (súper antígenos), e incluso forma un sitio catalítico en el caso de pirofosfatasas inorgánicas de levaduras, arqueas y bacterias (Arcus, 2002).

El plegamiento OB-fold presente en la superfamilia de proteínas de unión a ácidos nucleicos tiene representantes dentro de los fagos, las bacterias, las arqueas y los eucariontes. Los miembros de esta superfamilia son proteínas con dominios de unión a anticodones tRNA (Draper, 1999), dominios de unión a ssDNA (Mitton-Fry, Anderson, Hughes, Lundblad, & Wuttke, 2002), dominios de unión al DNA y RNA durante un choque por frío (Csp, del inglés *cold shock protein*), dominio de DNA ligasa (Subramanya, Doherty, Ashford, & Wigley, 1996), subunidad *RuvA* de helicasa de DNA (Nishino, Ariyoshi, Iwasaki, Shinagawa, & Morikawa, 1998) y proteínas de unión a ssDNA de fagos (Shamoo, Friedman, Parsons, Konigsberg, & Steitz, 1995). Esta superfamilia se caracteriza por poseer el pliegue OB-fold y porque en la mayoría de los casos el ligando es un ácido nucleico (Hubbard, Ailey, Brenner, Murzin, & Chothia, 1999; Alexey G. Murzin, Brenner, Hubbard, & Chothia, 1995). Además, hay algunas características de unión al ligando que también son comunes en muchos miembros de esta superfamilia. Por ejemplo, los dominios de unión a los anticodones de Asp-tRNA y Lisil-tRNA sintetasas se unen a su RNA diana utilizando un residuo central de fenilalanina y los bucles L2 y L4 sirven como bucles de unión (Draper, 1999). Este uso de residuos aromáticos centrales y bucles periféricos que llevan residuos cargados positivamente que se unen al esqueleto nucleotídico es también característico de proteínas de unión a ssDNA (Arcus, 2002).

Debido a la conservación de la arquitectura, la topología y una cara de unión del dominio OB-fold, se cree que se trata de un plegamiento antiguo cuya estructura es tolerante a la mutación, capaz de evolucionar para unirse a una amplia gama de secuencias y que posee una función de acuerdo a su ligando. Ésta se conoce como la hipótesis de divergencia, en contraposición a evolución convergente, que propone que proteínas con secuencias muy divergentes pueden adoptar un pliegue similar, sin poseer un ancestro en común resiente, debido a que el número de formas por las cuales se adquiere un pliegue estable son limitadas (Murzin, 1993; Arcus, 2002). Una estimación realizada por Chothia (1992) muestra que el número de pliegues estables es del orden de 10^3, es decir, mucho menor que el número de pliegues estables teóricamente esperados (Chothia, 1992). Si esto es cierto, el pliegue OB y otros pliegues comunes pueden representar los motivos estables de plegamiento que aparecieron en el origen de las

proteínas (Murzin, 1993). La presencia de estos motivos entre las estructuras de proteínas modernas podría haber ocurrido por dos características intrínsecas: (i) un sitio o sitios de unión potenciales, relacionados con el pliegue, que podrían ser editados por la evolución para realizar funciones diferentes y, (ii) una capacidad para acomodar una gran variedad de secuencias bastante diferentes que permitieron su rápida y amplia divergencia (Murzin, 1993).

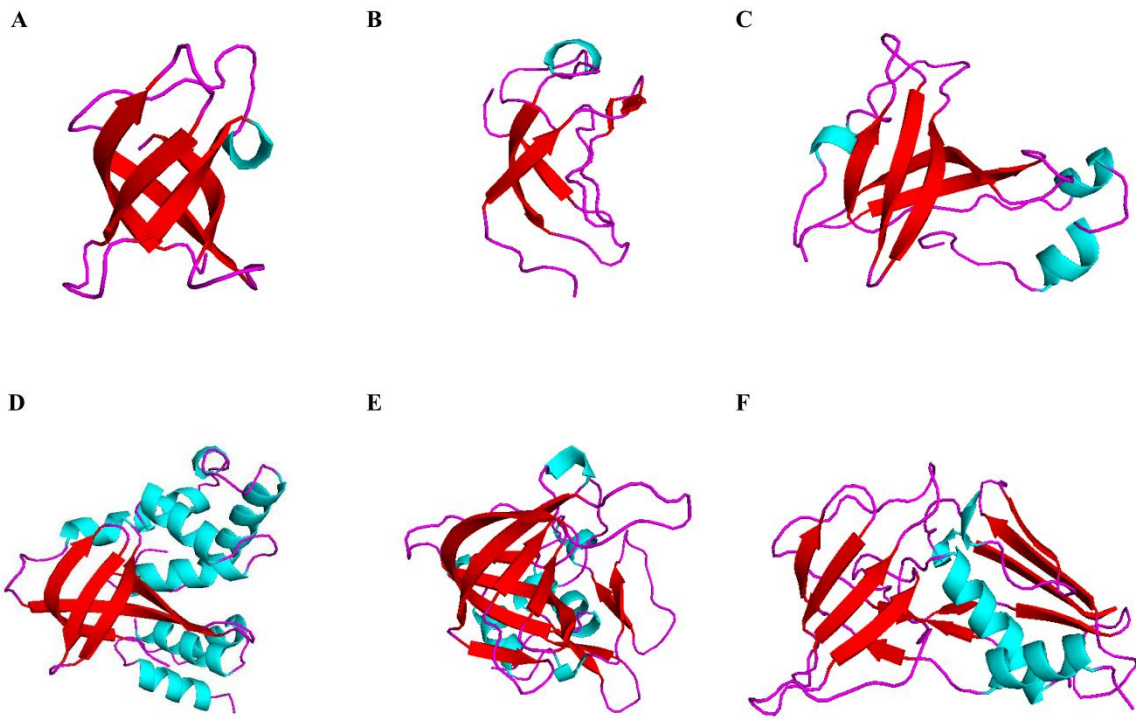


Fig. 3. Ejemplos de proteínas pertenecientes a la súper familia OB-fold.
Las hojas β de las estructuras se encuentran coloreadas de rojo, las hélices α en color cian y los bucles en color magenta. El panel A muestra la estructura de la proteína CspA (1MJC) de *E. coli*, el panel B muestra la estructura de la proteína ribosomal S1 (1SRO) de unión a RNAs de *E.coli*. El panel C muestra la estructura de la proteína Lysyl-tRNA de *E. coli* (1KRS), el panel D muestra la estructura de la proteína DNA helicasa RuvA (1HJP) de *E.coli*. El panel E muestra la estructura de la proteína TIMP-2 de H. sapiens involucrada en la proliferación de células endoteliales y el panel F la estructura de la proteína TSST-1 de *S. aureus* (1AW7), con función de superantigeno.

Familia de proteínas Csp y dominio CSD

Las proteínas de injuria por frío (Csp, del inglés *cold chock protein*) pertenecen a la súper familia de proteínas de unión a oligonucleótidos/oligosacáridos (OB-fold). Las proteínas Csp se encuentran en casi todas las bacterias psicrófilas, mesófilas, termófilas e hipertermófilas estudiadas y varios miembros de ésta familia de proteínas son importantes para la adaptación de las bacterias a una disminución drástica de la temperatura, así como en condiciones normales de crecimiento (Graumann & Marahiel, 1998; Horn, Hofweber, Kremer, & Kalbitzer, 2007). La primera Csp fue identificada en 1987 (Jones, VanBogelen, & Neidhardt, 1987), codificada por uno de los genes de estimulación de injuria por frío (*cold-shock stimulon*) en *Escherichia coli*. Las proteínas Csp se encuentran ampliamente distribuidas en las bacterias Gram-positivas y Gram-negativas y presentan una identidad mayor al 45% (Graumann, Wendrich, Weber, Schröder, & Marahiel, 1997); a pesar de esto, los principales estudios se han realizado en las proteínas Csp de *E. coli*.

Escherichia coli posee 9 genes que codifican para diferentes proteínas Csp, a las cuales se les ha nombrado de forma alfabética desde la CspA a la CspI. Dichas proteínas comparten de 29 a 83% de similitud (Goldstein, Pollitt, & Inouye, 1990; Lee *et al*., 1994; Nakashima, Kanamaru, Mizuno, & Horikoshi, 1996; Sangita *et al*., 2009; Sangita & Severinov, 2009; Yamanaka, Fang, & Inouye, 1998) y se agrupan en dos tipos: las de clase I y las de clase II. Las de la clase I son proteínas que se expresan poco a 37 °C pero su tasa de expresión aumenta drásticamente cuando baja la temperatura (de 15-20°C) (CspA, CspB, CspE, CspG y CspI). Las de la clase II son proteínas que están presentes a 37 °C y que no son funcionales a 15 °C, e incluyen a CspC, CspD, CspF y CspH (Sangita & Severinov, 2009).

Se ha demostrado que a baja temperatura el mRNA de la proteína CspA de *E. coli* experimenta un cambio estructural dependiente de la temperatura, éste cambio estructural provoca que el mRNA se traduzca más eficientemente y sea menos propenso a la degradación, en comparación con la estructura del mRNA de *cspA* a 37 ºC

(Giuliodori *et al*., 2010). Asimismo, se ha visto que el gen *ttcsp*2 inducido por frío de *Thermus thermophilus* actúa como un sensor térmico adoptando una estructura secundaria estable debido a una caída de temperatura (Mega *et al*., 2010). Sólo las CspA, B, E, G e I de *E. coli* son inducidos por choque en frío (Etchegaray, Jones, & Inouye, 1996; Nakashima *et al*., 1996; Uppal, Akkipeddi, & Jawali, 2008; Wang, Yamanaka, & Inouye, 1999) y se descubrió que en *E. coli* cuatro de los nueve genes *csp* (*cspA, cspB, cspE* y *cspG*) debían suprimirse para obtener un fenotipo sensible al frío. Además, la *deleción* de uno o dos genes *csp* aumenta y prolonga la expresión de los genes *csp* restantes inducidos por el frío, por lo que se puede decir que las funciones de los miembros de la familia CspA se superponen y pueden compensarse entre sí (Xia, Ke, & Inouye, 2001). De modo interesante, el fenotipo sensible al frío se revierte por la expresión en *trans* de un homólogo estructural de CspA, el dominio S1 de unión a RNA, el cual forma parte da la súper familia OB-fold y está ampliamente distribuido en distintas proteínas de diversos organismos (Xia *et al*., 2001).

De manera análoga, *Bacillus subtilis* posee 3 genes que codifican proteínas homólogas a las Csp de *E. coli* (CspB, CspC y CspD) (P. Graumann, Schröder, Schmid, & Marahiel, 1996), las cuales son esenciales para el crecimiento a temperatura óptima así como para la adaptación a bajas temperaturas y sobrevivencia durante la fase estacionaria de dicha especie (Graumann *et al*., 1997). Las proteínas CspB y CspD incrementan su expresión ante el descenso de la temperatura mientras que la CspC muestra un patrón de expresión asociado exclusivamente a bajas temperaturas (Schindler *et al*., 1999). Las Csp de *B. subtilis* muestran entre un 72% y 80% de similitud entre sí, son capaces de complementarse unas a otras (Schindler *et al*., 1999) y se ha demostrado que la presencia de al menos una de las proteínas Csp es esencial para la supervivencia de *B. subtilis*, incluso bajo condiciones óptimas de crecimiento (Weber & Marahiel, 2002).

Las funciones de Csp no inducibles por frío todavía no se conocen bien. Estudios con proteínas Csp en *T. thermophilus* han demostrado que juegan un papel importante en el control de la traducción de algunas proteínas y que diferentes factores

de estrés ambiental podrían alterar la afinidad con la que las proteínas Csp se unen a nucleótidos. Así mismo, los estudios en la CspA de *Brucella melitensis* demostraron que ésta proteína juega un papel en la virulencia y la regulación del metabolismo, pero los mecanismos reguladores mediados por CspA no se entienden bien hasta la fecha (Liu *et al*., 2015). Por esta razón se necesitan más investigaciones para entender por qué las proteínas Csp funcionan durante el crecimiento normal y en respuestas de estrés no relacionadas con el frío.

En *Escherichia coli*, la proteína CspD inhibe la replicación del DNA y se induce durante la fase estacionaria al detenerse el crecimiento. Estudios muestran que la sobreproducción de la proteína CspD es tóxica para las células (Uppal, Shetty, & Jawali, 2014; Kunitoshi Yamanaka, Zheng, Crooke, Wang, & Inouye, 2001). CspD también se ha relacionado con la formación de *biofilms* (Y. Kim *et al*., 2010; Y. Kim & Wood, 2010). Recientemente se demostró que las Csp y otras chaperonas de RNA podrían tener un papel en el amortiguamiento de mutaciones que afectan la estructura y el plegamiento del RNA, ya que la sobreexpresión de CspA mejoró la aptitud de las cepas de *E. coli* que habían acumulado mutaciones deletéreas durante experimentos de laboratorio a largo plazo (Rudan, Schneider, Warnecke, & Krisko, 2015). Al unirse con baja especificidad al RNA, CspA puede evitar la formación de estructuras secundarias no deseadas (Jiang, Hou, & Inouye, 1997) ayudando al RNA mal plegado a adoptar una conformación funcional y por lo tanto suprimir el fenotipo de mutaciones perjudiciales (Rudan *et al*., 2015).

Estructura de las Csp

Las Csp son proteínas de bajo peso molecular, aproximadamente de 7.4 kDa (Lim, Thomas, & Cavicchioli, 2000; Perl *et al*., 1998) y constan de un barril constituido por 5 hebras β plegadas anti-paralelas (β1-β5). Dos de dichas hebras (β2 y β3 a partir del extremo N-terminal) incluyen 2 motivos altamente conservados que han sido nombrados como: RNP1 presente en la cadena β2 correspondiente a la secuencia consenso KGYGFIEV y RNP2 presente en la cadena β3 correspondiente a la secuencia

consenso VFVHF (Fig. 4). Estas hebras forman el dominio típico conocido como dominio de injuria por frío (CSD, por sus siglas en inglés *cold shock domain*); (Max, Zeeb, Bienert, Balbach, & Heinemann, 2007; Murata & Los, 1997; Weber & Marahiel, 2002). Los residuos aromáticos presentes en ambos motivos son esenciales para mediar la interacción con moléculas de RNA de cadena simple, que en condiciones de baja temperatura, estabilizan sus estructuras secundarias para que interactúen con el ribosoma de forma habitual, como los mRNA codificantes para las proteínas de inducción en frío (CIP, del inglés *cold inducible protein*), entre ellas las propias Csp (Max *et al*., 2007). Así mismo, se ha demostrado que las Csp también se unen a moléculas de DNA de cadena simple con mayor afinidad que a las moléculas de RNA (Weber, Klein, Müller, Niess, & Marahiel, 2001).

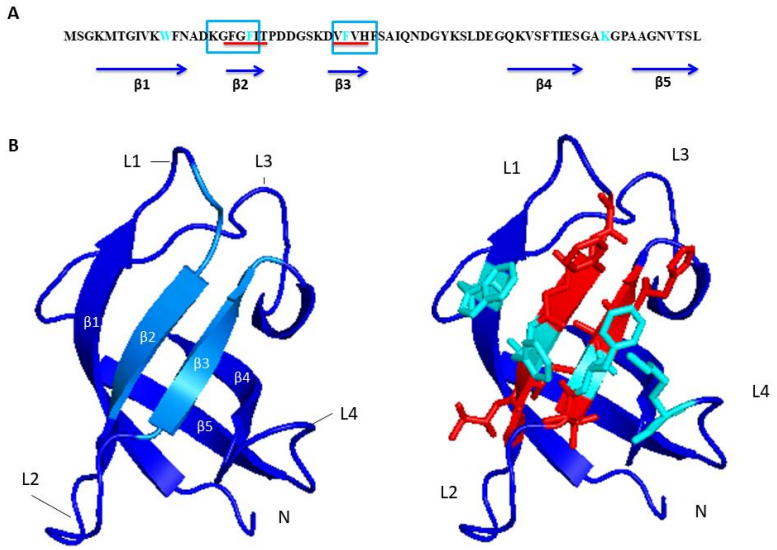


Fig. 4. Secuencia primaria y estructura terciaria de la CspA.
En el panel A podemos observar la secuencia de aminoácidos de la CspA de *E. coli*. Las 5 hebras, β1, β2, β3, β4 y β5 se encuentran indicadas con una flecha azul en la parte inferior de la secuencia, los motivos RNP1 Y RNP2 están dentro de las cajas azul claro, los sitios de unión a RNA se encuentran señalados con una línea roja en la parte inferior de la secuencia y los aminoácidos de unión a DNA son los que tienen un color cian. El panel B representa la estructura terciaria de CspA de *E. coli* (PDB: 1MJC). En la estructura de la izquierda se aprecian las 5 hebras β (β1, β2, β3, β4 y β5) y los 4 bucles (L1, L2, L3 y L4), y los motivos RNP1 Y RNP2 se encuentran coloreados en azul claro. En la estructura de la derecha se indican los sitios de unión a RNA en color rojo y los aminoácidos de unión a DNA en color cian. La N representa el extremo amino terminal de la estructura.

Como en el caso de las proteínas Csp de *Escherichia coli*, la proteína CspB de *B. subtilis* cuenta con los motivos RNP1 (Lys13 a Val20) y RNP2 (Val26 a Phe30) ricos en residuos hidrofóbicos y básicos localizados uno al lado del otro en la superficie de la proteína. Los residuos de dichos motivos, incluyendo a los aminoácidos Lys7, Trp8, Lys13, Phe15, Phe17, Phe27, His29, Phe30, Phe38 y Arg56, forman una plataforma hidrofóbica que empieza en la hebra β1, prosigue en las hebras β2 y β3 y finaliza en el bucle L3 y son importantes para la unión de oligonucleótidos (Kloks *et al*., 2002; Lopez, Yutani, & Makhatadze, 1999, 2001; Perl *et al*., 1998; Schröder, Graumann, Schnuchel, Holak, & Marahiel, 1995; Zeeb & Balbach, 2003).

La propiedad de unirse con alta afinidad a ácidos nucleicos de cadena simple y que funcionen como chaperonas de RNA hace que las proteínas Csp deban estar presentes en prácticamente todos los pasos de la expresión genética incluyendo la transcripción, la traducción y la entrega de RNA (Mihailovich, Militti, Gabaldón, & Gebauer, 2010). En consecuencia, su función es importante para las bacterias y los estudios muestran que, en algunos casos, la presencia de al menos una de las Csp es esencial para la supervivencia de la bacteria, incluso bajo condiciones óptimas de crecimiento (Weber & Marahiel, 2002).

Respuesta de las Csp al choque frío

En *E. coli* son cuatro las principales proteínas Csp que se expresan principalmente en respuesta a un choque frío, CspA, CspB, CspG y CspI (Sangita Phadtare & Severinov, 2005), siendo la más abundante la CspA, la cual alcanza niveles superiores al 13% del total de proteínas sintetizadas. CspA (Fig. 4) es una proteína citoplasmática con un peso de 7.4 kDa (Goldstein *et al*., 1990; P. Jones *et al*., 1987). Esta proteína es más abundante en el choque frío, pero también se encuentra presente durante el crecimiento exponencial a 37 °C, ya que la transcripción de los genes de CspA es abundante, sin embargo, la traducción es ineficiente (Bae, Jones, & Inouye, 1997; Mitta, Fang, & Inouye, 1997) debido a que la vida media de su mRNA es inferior a 12 s, a diferencia de cuando se encuentra en un choque frío que el mRNA es altamente estable y su vida

media es mayor a 20 min (Brandi, Spurio, Gualerzi, & Pon, 1999). Además, el transcrito de CspA consta de una 5´-UTR, que es inusualmente larga (159 bp), tiene una secuencia altamente conservada, llamada *cold-box*, lo que le facilita su transcripción en estas condiciones en comparación a los otros mRNAs celulares (Bae *et al*., 1997; Jiang *et al*., 1997; Mitta *et al*., 1997).

La estabilización de estructuras secundarias de RNA inducida por la baja temperatura interfiere tanto con la etapa de elongación de la transcripción como con la etapa de elongación de la traducción. Se ha propuesto que las proteínas Csp actúan como chaperonas de RNA facilitando la transcripción y la traducción a baja temperatura, en virtud de su capacidad para "fundir" las estructuras secundarias en ácidos nucleicos (Bae, Xia, Inouye, & Severinov, 2000; Jiang *et al*., 1997; Sangita Phadtare, Inouye, & Severinov, 2002). El control de la formación de estructuras secundarias en el mRNA permite el inicio de la traducción y la terminación de la transcripción en respuesta a señales externas. La modulación de la terminación de la transcripción por proteínas de unión a RNA implica la formación de estructuras alternativas (Stülke, 2002). El ejemplo mejor caracterizado es el del mRNA de *cspA* de *E. coli*. Cuando se transcribe a 10 ºC, el mRNA adopta una estructura de pseudonudo, que expone la secuencia SD y el codón de inicio y mejora la traducción de la proteína CspA en el frío. Por el contrario, la transcripción a 37 °C favorece una conformación en la que la secuencia SD está enmascarada dentro de una región bicatenaria, dificultando la traducción del mRNA (Duval, Simonetti, Caldelari, & Marzi, 2015). Por esta razón, se postuló que el mRNA de *cspA* experimenta una redisposición estructural dependiente de la temperatura, que resulta en la estabilización en el frío de su mRNA (Duval *et al*., 2015).

Los genes que muestran una dependencia estricta a las proteínas Csp para su expresión a 15 ° C, incluyen a *male* y *malk* (funciones relacionadas con la membrana), *mopa* y *mopb* (chaperonas), *dps*, *katg*, *rpos*, *uspa* (respuesta al estrés), así como varios otros. Varios de estos genes poseen secuencias promotoras que contienen estructuras de tallos-asa, que se estabilizan a baja temperatura y bloquean el proceso de elongación de

la transcripción. Estas secuencias, denominadas: "terminadores de la transcripción sensibles a proteínas Csp", se proponen como dianas fisiológicas de las proteínas Csp (Sangita & Severinov, 2009).

Regulación de la expresión del gen csp

El gen *cspA* tiene un promotor que se activa tanto a 37 °C como a 15 °C (Fang, Jiang, Bae, & Inouye, 1997; Goldenberg *et al*., 1997; Mitta *et al*., 1997), y está equipado con dos secuencias características del gen. La primera es una secuencia rica en AT ubicada inmediatamente río arriba de la región -35, que se llama elemento "*UP*", la cual es reconocida directamente por la subunidad α de la RNA polimerasa (Goldenberg *et al*., 1997; Mitta *et al*., 1997), lo que confiere una fuerte actividad de promotor de la transcripción (Ross *et al*., 1993). La segunda, es una secuencia TGn situada inmediatamente río arriba de la región -10, esta secuencia, junto con la región -10, constituye lo que se llama la región -10 extendida (Kumar *et al*., 1993) (Fig. 5). Estas dos secuencias son las responsables de que el gen *cspA* cuente con una actividad tanto a 37 °C como a 15 °C (Yamanaka, 1999). Estas secuencias se encuentran presentes en los genes *cspA*, *cspG* y *cspI*, los cuales también se activan en un choque frío (Wang *et al*., 1999).

Sin embargo, es importante tener en cuenta que el mRNA de los genes *cspA*, *cspB*, *cspG* y *cspI* tiene una región 5'-UTR (*Untranslated Transcribed Region*) inusualmente larga, de 159, 161, 156 y 145 bases respectivamente (Tanabe, Goldstein, Yang, & Inouye, 1992; Wang *et al*., 1999); (Fig. 5) y que es extremadamente inestable a 37 °C, teniendo una vida media de aproximadamente 12 segundos (Brandi, Pietroni, Gualerzi, & Pon, 1996; Fang *et al*., 1997; Goldenberg, Azar, & Oppenheim, 1996). Una vez que se presenta el choque frío, el mRNA se torna estable, teniendo ahora una vida media de aproximadamente 20 minutos. La inestabilidad del mRNA de las proteínas Csp se ha visto que es provocada por la presencia de un sitio de reconocimiento a RNasas E, situado río arriba junto a la secuencia SD (Fang *et al*., 1997). Sin embargo, el mRNA de *cspA* forma estructuras secundarias a 15 °C, que impide la unión de estas

enzimas al mRNA y, en consecuencia, se bloquea la degradación del mRNA de la proteína CspA (Hankins, Zappavigna, Prud'homme-Généreux, & Mackie, 2007).

La 5' UTR de varios mRNA codificantes para proteínas Csp (*cspA*, *cspB*, *cspG* y *cspI*) cuenta también con una secuencia única altamente conservada llamada *cold box* (Fig. 5), de 11 pares de bases (Bae *et al.*,, 1997), la cual forma una estructura estable de tallo-asa (Jiang, Fang, & Inouye, 1996) y está involucrada en la regulación de la expresión del gen *cspA*, ya que funciona como el sitio de unión de un represor (Jiang *et al*., 1996). La unión del represor a la secuencia *cold box* en los mRNA bloquea la transcripción del gen *cspA*, por lo que al eliminar dicha región en *cspA*, causa la activación de la expresión del gen *cspA*, dando como resultado una alta expresión constitutiva de *cspA* en células que crecen exponencialmente a 15ºC (Fang, Hou, & Inouye, 1998).

La expresión del gen *cspA* también es regulada a nivel de su traducción (S Phadtare, Alsina, & Inouye, 1999). El mRNA contiene un sitio de unión al ribosoma llamado *Downstream Box* (DB) (Lee *et al*., 1994), de 14 pares de bases localizada río abajo del codón de inicio (Etchegaray & Inouye, 1999; Mitta *et al*., 1997). La secuencia DB es complementaria a la región llamada secuencia anti-DB de la subunidad 16S del rRNA, la formación del complejo es responsable de la iniciación de la traducción de proteínas CspA (Sprengart, Fuchs, & Porter, 1996). Se ha comprobado que la DB es esencial para la traducción de proteínas del choque frío durante la fase de aclimatación, mientras que la traducción de las otras proteínas están bloqueadas (J.-P. Etchegaray & Inouye, 1999; Mitta *et al*., 1997); además de que dicha secuencia también se encuentra presente en otros genes inducidos en choque frío de clase I (*cspB*, *cspG* y *cspI*) (Fig. 5); (Mitta *et al*., 1997).

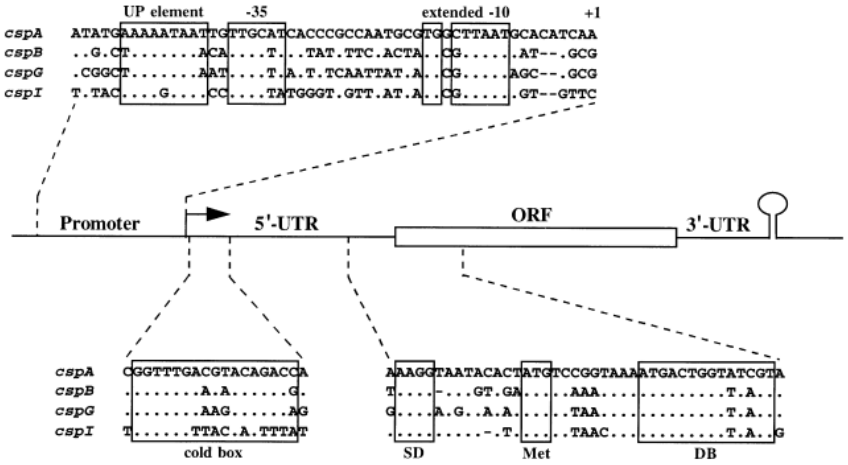


Fig. 5. Comparación de las secuencias de los genes de cspA, cspB, cspG y cspI.
Las estructuras características (*UP element*, región -35, región -10 y *cold box*) del gen *cspA* y para la transcripción (*Shine-Dalgarno*, codón de inicio y *Downstream Box*), se encuentran dentro de las cajas (tomado de Yamanaka, 1999).

Dominio CSD

Las proteínas Csp bacterianas están conformadas por un solo dominio, al cual se le ha denominado como dominio de injuria por frío (CSD del inglés *cold shock domain*) y se ha encontrado en varias proteínas de eucariontes y arqueas, por lo que se cree que, probablemente, estuvo presente en el último ancestro en común, hace aprox. 3,500 millones de años (Graumann & Marahiel, 1998).

En Arqueas se han identificado, hasta el año 2010, 16 especies que poseen proteínas con secuencias homólogas a las proteínas Csp de las bacterias y son, principalmente, del clado *Halobacteriaceae* (Mihailovich *et al*., 2010). La similitud entre las secuencias de las proteínas Csp de Arqueas y la proteína CspA de *E. coli* varía entre el 36 y el 59%, siendo la proteína de *Methanogenium frigidum* con la que tiene mayor similitud (59%), sin embargo las Csp de Arqueas son más similares entre sí (del 68 al 91%); (Giaquinto *et al*., 2007).

Estudios muestran que las Csp de Arqueas tienen una función en células que crecen a 4 °C (Goodchild, Raftery, Saunders, Guilhaus, & Cavicchioli, 2004); esto, junto con el hecho de que las especies de Arqueas que tienen esta proteína son en su mayoría mesófilas y psicrófilas, refuerza la idea de que son importantes para el crecimiento a bajas temperaturas (Giaquinto *et al*., 2007).

Dominio CSD en Eucariontes

La lista de proteínas que contienen el dominio CSD (secuencias ortólogas) en eucariontes ha aumentado con el tiempo; dicha lista contiene proteínas que están implicadas en varios procesos relacionados con el metabolismo del RNA, esta diversidad funcional está proporcionada en parte por la combinación del dominio CSD con otros dominios funcionales, como los dedos de zinc, así como por la alta gama de secuencias que son reconocidas por este dominio. Esta capacidad de combinación junto con la flexibilidad del dominio CSD para el reconocimiento de ácidos nucleicos de cadena simple, da como resultado un repertorio creciente de proteínas implicadas en múltiples aspectos de la fisiología celular (Mihailovich *et al*., 2010).

Familia Y-box

La familia Y-box se encuentra en vertebrados e invertebrados, y está compuesta por las proteínas denominadas DBPA (DNA *binding protein* A, también conocida como CSDA), DbpB (comúnmente conocida como factor de unión Y-box-1, YB-1) y DBPC (*frog germ-cell-specific* o Y-box *protein* 2, FRGY2). Se ha demostrado que DBPA reprime la transcripción de varios genes, en ocasiones en conjunto con DbpB o YB-1 (Coles, Diamond, Occhiodoro, Vadas, & Shannon, 1996). YB-1 es un componente principal del mensajero citoplasmático acentuado en ribonucleoproteínas (mRNPs) y tiene funciones en el empalme (*splicing*) de su pre-mRNA, en la estabilidad del mRNA y en la traducción de estas proteínas (Matsumoto & Wolffe, 1998). YB-1 contiene un único dominio CSD con preferencia por la unión a secuencias de hebra sencilla ricas en

pirimidinas (Kloks *et al*., 2002) (Fig. 6). Se ha visto que el dominio CSD de YB-1 cuenta con un sitio de unión a ssDNA que comprende los residuos Lys8, Lys14, Trp15, Arg19, Tyr22, Phe24, Phe35, His37 y Lys68, ubicados entre los bucles β1-β2 y β4-β5 (motivos RNP1 y RNP2); (Kloks *et al*., 2002). Este sitio de unión a ssDNA implica residuos aromáticos y cargados positivamente característicos de los motivos RNP1 y RNP2 encontrados también, en CspA de *Escherichia coli* y CspB de *Bacillus subtilis* (Newkirk *et al*., 1994; H Schindelin, Jiang, Inouye, & Heinemann, 1994; Hermann Schindelin, Marahiel, & Heinemann, 1993; Schnuchel *et al*., 1993). Asimismo, la afinidad del dominio CSD de YB-1 no es muy fuerte (K_D de 10^{-5} a 10^{-6} M) y la unión del dominio CSD y el ssDNA es poco especifica (Kloks *et al*., 2002) de igual forma como ocurre en las proteínas Csp bacterianas (Max, Zeeb, Bienert, Balbach, & Heinemann, 2006; Max *et al*., 2007; M. Zeeb & Balbach, 2003; Markus Zeeb *et al*., 2006).

El dominio CSD es necesario para la importación nuclear de la proteína, así como su reconocimiento, pero no parece tener otras funciones para la proteína (Jurchott *et al.*, 2003). La proteína YB-1 puede estimular o reprimir la traducción, dependiendo de su cantidad con respecto al RNA diana. Cuando existe una relación baja entre YB-1 y el mRNA, YB-1 promueve la iniciación de la traducción, mientras que un aumento de esta relación inhibe fuertemente la traducción (Evdokimova *et al*., 1998; Pisarev *et al*., 2002), relación parecida a la de las Csp en bacterias.

Proteína Lin28

La proteína Lin28 (*abnormal cell lineage* 28) es una proteína pequeña (aproximadamente de 28 kDa) expresada a partir de dos *loci* en vertebrados: Lin28 (Lin28a) y Lin28b. Lin28 se expresa durante el desarrollo embrionario y se encuentra principalmente en el citoplasma. Ambas formas de Lin28 contienen tres dominios de unión a RNA: un dominio en el extremo N-terminal de injuria por frío (CSD), seguido de dos dominios de dedos de zinc (ZnF) de tipo nudillo (tipo CCHC). Tanto el dominio

CSD como los dedos de zinc tienen un papel en la unión al RNA y en el procesamiento de micro-RNA de la familia Let-7 (Moss & Tang, 2003) (Fig. 6).

Los micro-RNA (miRNA) son RNA pequeños de aproximadamente 22 nucleótidos, que regulan la expresión genética mediante la hibridación a sitios que por lo general están ubicados en el extremo 3' del UTR de los RNA diana. Let-7 es una familia grande y altamente conservada de miRNA que controla la diferenciación y proliferación celular mediante la regulación de los genes involucrados en la oncogénesis, como KRAS, HMGA2, c-MYC, Cdc25a y Cdk6 (Büssing, Slack, & Großhans, 2008; Johnson *et al*., 2005; C. Mayr, Hemann, & Bartel, 2007). La expresión de Let-7 se pierde con frecuencia en ciertas células cancerosas como resultado de eliminaciones y silenciamiento epigenético, lo que sugiere que let-7 es un supresor de tumores (Büssing *et al*., 2008; Johnson *et al*., 2005; C. Mayr *et al*., 2007).

El dominio CSD de Lin28b revelan el pliegue de unión a oligosacárido/oligonucleótido (OB-fold) típico del dominio CSD (Arcus, 2002; Chaikam & Karlson, 2010; Horn *et al*., 2007; Mihailovich *et al*., 2010; H Schindelin *et al*., 1994), que consiste en un barril β compuesto por cinco cadenas β antiparalelas. En general, la arquitectura general es bastante similar a la de los dominios CSD de otras proteínas; sin embargo, los dominios CSD de proteínas Lin28b muestran una mayor similitud estructural con las Csp bacterianas que con las proteínas que contienen el dominio CSD de organismos eucarióticos (F. Mayr, Schütz, Döge, & Heinemann, 2012). De acuerdo con esto, el dominio CSD de Lin28b se asemejan a sus homólogos bacterianos en tener un fuerte carácter anfipático. Mientras que un lado de la proteína muestra un potencial de superficie más bien negativo, la superficie opuesta forma una plataforma hidrófoba intercalada con residuos aromáticos expuestos altamente conservados (Trp39, Phe48, His68, Phe66, Phe77) que están rodeados por residuos básicos y polares (Mayr *et al*., 2012). Se sabe que la mayoría de estos residuos interactúan con ácidos nucleicos monocatenarios a través de contactos hidrófobos y puentes de hidrogeno y, por lo tanto, están altamente conservados tanto en secuencia como en estructura (Horn *et al*., 2007; Max *et al*., 2007).

Proteína UNR

El gen *unr* (*upstream of N-ras*) es esencial, estudios demuestran que los ratones que carecen de éste mueren a los 10 días del desarrollo embrionario (Doniger, Landsman, Gonda, & Wistow, 1992) y se encuentra ubicado aproximadamente 150 bp río arriba del locus N-ras en múltiples especies (Jeffers, Paciucci, & Pellicer, 1990). La proteína UNR se expresa de forma ubicua y se encuentra principalmente en el citoplasma, ya sea en forma soluble o asociada a las membranas, en particular al retículo endoplásmico (Jacquemin-Sablon *et al*., 1994). También, se ha encontrado en el núcleo aunque en pequeñas proporciones, sin embargo, el papel de la proteína UNR en el núcleo sigue siendo poco claro (Ferrer, Garcia-Espana, Jeffers, & Pellicer, 1999; Patalano, Mihailovich, Belacortu, Paricio, & Gebauer, 2009).

La proteína UNR es única entre las proteínas eucariontes que contienen el dominio CSD ya que posee cinco copias del dominio (Doniger *et al*., 1992); (Fig. 6). La proteína UNR tiene un alto grado de conservación entre los eucariontes, sólo está ausente en *Caenorhabditis elegans*. Lo anterior muestra que la tasa de conservación del dominio CSD, especialmente de CSD1, es más alta con respecto a la proteína Csp bacteriana (Mihailovich *et al*., 2010).

Los dominios CSD de la proteína UNR presentan una función tanto en la proteína como en el ácido nucleico al que se une la proteína. Además, UNR muestra preferencia por DNA de una sola cadena (ssDNA) y RNA (Ferrer *et al*., 1999; Jacquemin-Sablon *et al*., 1994). Esta proteína ha sido encontrada en los complejos que intervienen en la regulación de la traducción y la estabilidad del mRNA. En la traducción, la UNR se comporta como un factor IRES *trans-actin* (ITAF), que regula la actividad de varios IRES (*Internal ribosome entry site*) virales y celulares (Brown & Jackson, 2004). También, se ha propuesto que funciona como una chaperona de RNA, pero no está claro si la UNR permanece unida al RNA o puede disociarse de él como lo hacen normalmente las otras proteínas chaperonas (Evans *et al*., 2003; Hunt, Hsuan, Totty, & Jackson, 1999; Mitchell, Spriggs, Coldwell, Jackson, & Willis, 2003).

Proteína rica en glicina

La proteína rica en glicina (*Glycine-rich*) está conformada por el dominio de injuria por frío (CSD) asociado a una región rica en glicina intercalada con dedos de zinc (CCHC) en el extremo c-terminal (Karlson, Nakaminami, Toyomasu, & Imai, 2002); (Fig.6). La proteína rica en glicina se encuentra ampliamente distribuida entre las plantas vasculares (Karlson & Imai, 2003) y cuenta con una estructura muy parecida a la de la proteína Lin-28 (Sasaki & Imai, 2011). La proteína rica en glicina puede tener de 2 a 7 dedos de zinc (CCHC) dependiendo de la especie, las monocotiledóneas contienen por lo general 2 ó 4, mientras que las dicotiledóneas contienen 1, 2, 4,5 o 7 de estos dedos de zinc (CCHC). Las regiones ricas en glicina y los dedos de zinc (CCHC) probablemente están implicadas en la unión de ácidos nucleicos y a la unión de otras proteínas (Karlson *et al*., 2002; Nakaminami, Karlson, & Imai, 2006; Sasaki, Kim, & Imai, 2007) ya que las diversas combinaciones son necesarias para dichos procesos (Sasaki & Imai, 2011).

Mientras que se han realizado considerables avances en la comprensión de la función del dominio CSD en proteínas de bacterias y animales, se sabe poco de sus funciones en plantas. La primera proteína caracterizada en plantas que poseía el dominio fue la proteína WCsp1 del trigo (Karlson *et al*., 2002). La proteína WCsp1 está compuesta por el dominio CSD y por una región rica en glicina intercalada con tres dedos de zinc (CCHC). Su mRNA se encuentra regulado por una respuesta al frío y la proteína se acumula considerablemente en el tejido de la corona durante la aclimatación al frío prolongado. Además, de que sus niveles de transcripción no son modulados por otras injurias ambientales como la salinidad excesiva, la sequía y el calor, o por el tratamiento con ácido abscísico (Karlson *et al*., 2002; Nakaminami *et al*., 2005, 2006). Se cree que la función de la proteína WCsp1 es específica a la adaptación al frío en plantas (Sasaki & Imai, 2011). También, se ha demostrado que mutantes de Csp de *E. coli* con WCsp1 complementan el fenotipo sensible al frío de esta bacteria (Nakaminami *et al*., 2006); esto último y el hecho de que la proteína WCsp1 se una al DNA y al RNA (Karlson *et al*., 2002) sugiere que la WCsp1 comparte una función conservada con la

CspA de *E. coli* y está implicada en la regulación de la aclimatación al frío (Sasaki & Imai, 2011).

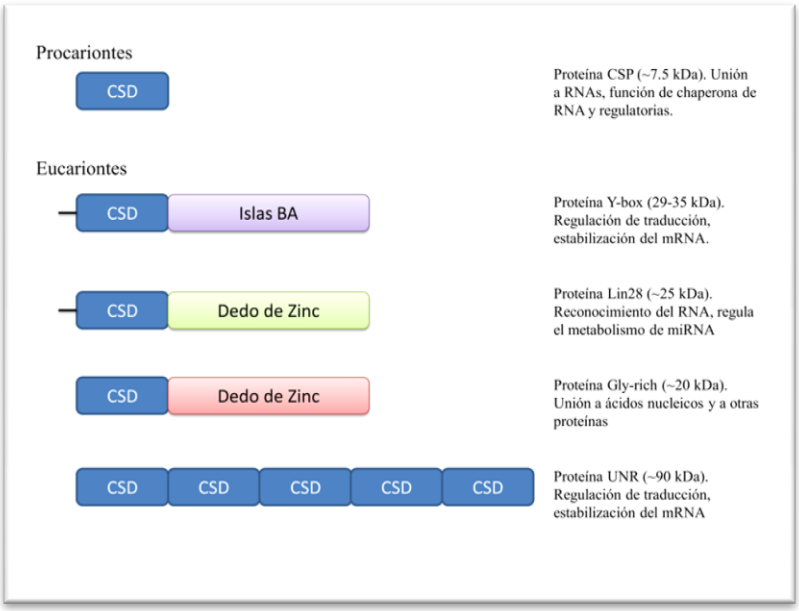


Fig. 6. Proteínas que contienen el dominio CSD.
Las proteínas Csp consisten en un solo dominio CSD, el cual se une a RNA y ssDNA. La proteína Y-box está constituida por un dominio CSD unido a una isla BA (*basic/aromatic*), las proteínas Lin28 y Gly-rich por un dominio CSD y diversos números de dedos de zinc, mientras que la proteína UNR consta de 5 dominios CSD.

Evolución de la proteína Csp y el dominio CSD

El dominio CSD y el dominio S1 son estructuras que contienen al pliegue de unión a oligosacárido/oligonucleótido (OB-fold). Aunque no existe similitud detectable en la secuencia primaria, los dominios CSD son estructuralmente y funcionalmente similares a los dominios S1 que se encuentran en varias copias en la proteína ribosomal S1 de las bacterias. S1 consta al menos de dos dominios distintos; un dominio de unión al ribosoma en la región N-terminal y un dominio de unión a ácido nucleicos en la región C-terminal. Este dominio se une a RNA y es esencial para la iniciación de la traducción

de varios mRNAs, principalmente de los que contienen 5´-UTRs largas (Tedin, Resch, & Bläsi, 1997). Así mismo, se ha demostrado que el dominio S1 tiene un pliegue similar a la proteínas Csp (Bycroft, Hubbard, Proctor, Freund, & Murzin, 1997) y otras proteínas de unión a oligosacáridos (Murzin, 1993).

Aunque el parentesco de los dominios CSD y S1 no se puede demostrar mediante la secuencia, su similitud estructural sugiere que los dos dominios se originaron a partir de una proteína ancestral posiblemente presente en el último ancestro común (Bycroft *et al*., 1997), a partir de la cual, los dominios CSD y S1 actuales se separaron (Mihailovich *et al*., 2010). Esto es apoyado por el hecho de que CSD y S1 están ampliamente distribuidos en los dominios Bacteria, Archaea y Eukarya (Graumann & Marahiel, 1998; Mihailovich *et al*., 2010). Sin embargo, el aspecto más importante de la evolución de las proteínas que contienen el dominio CSD y el dominio S1, es la evidencia que nos pueden dar sobre algunos procesos, como la mezcla de dominios y la duplicación (Mihailovich *et al*., 2010).

La presencia del dominio CSD en proteínas de bacterias, arqueas y eucariontes, sugiere que apareció antes de la divergencia de éstos grupos; sin embargo, los genomas completamente secuenciados de algunas especies de arqueas, como *Methanococcus jannaschii* y *Methanococcus thermoautotrophicum* no contienen ningún gen detectable homologo al de las proteínas Csp. Los genes para Csp que se han encontrado en las secuencias de arqueas son escasos y pertenecen principalmente a especies del clado *Halobacteriaceae*, por lo que se ha llegado a pensar que las arqueas adquirieron los genes de las proteínas Csp por transferencia horizontal (Graumann & Marahiel, 1998; Mihailovich *et al*., 2010). De forma interesante, los datos metagenómicos obtenidos por Spang y colaboradores (2015), sugieren la presencia de genes para las proteínas Csp en el *Phylum Lokiarchaeota* (actualmente considerados como ancestros de los eucariontes) (Spang *et al*., 2015), implicando que quizás, los eucariontes, no sólo heredaron los genes CSP de sus ancestros (posiblemente en el *Phylum Lokiarchaeota*), sino que los mantuvieron y diversificaron acorde a su estilo de vida y presiones de selección. Así mismo, la ausencia de genes para las proteínas Csp en la mayoría de los linajes de

Archaea podrían ser el resultado de pérdidas específicas en los *Phyla* de Archaea, como sucede en los casos de varias proteínas que tienen homólogos en los genomas de todas las ramas principales de los eucariontes (animales, plantas, hongos y protozoos) y no tienen algún homólogo en los genomas bacterianos y en la mayoría de los genomas de arqueas, pero si en el *Phylum Lokiarchaeota* (Spang *et al*., 2015).

También, los análisis transcriptómicos de las arqueas *Methanolobus psychrophilus* R15 y *Methanococcoides burtonii* han encontrado que a bajas temperaturas se inducen unos genes que codifican proteínas que poseen un dominio TRAM (ya que dicho dominio se ha encontrado en el extremo N-terminal de las proteínas TRM2 metilasas y en extremo C-terminal de la familia de proteínas MiaB, TRM2 *and* MiaB); (Campanaro *et al*., 2011; Chen *et al*., 2012). Aunque no hay una homología en la secuencia entre las proteínas TRAM y la proteínas Csps, ambas son proteínas de aproximadamente 7 kDa, que poseen aminoácidos cargados positivamente alrededor de una plataforma hidrófoba y presentan una conformación similar de varias cadenas β antiparalelas que contienen dos motivos de unión a RNA. Así mismo, se ha determinado que la proteína TRAM de *Methanococcoides burtonii* se une preferentemente a los tRNAs y al rRNA 5S (Taha *et al*., 2016), implicando que posee una actividad de chaperona de RNA, semejante a las proteínas Csp. Por lo anterior, se ha sugerido que las proteínas TRAM podrían participar en la adaptación al frío en organismos del dominio Archaea, aunque las funciones de las proteínas TRAM aún no son claras (Zhang *et al*., 2017).

Hipótesis

Comparar la secuencia y estructura inferidas de las proteínas CSP ancestrales contra su versión actual permitirá corroborar que el bajo nivel de sustitución observado en este dominio podría ser evidencia de la fuerte presión de selección a la que está sometido el plegamiento OB-fold que permite interactuar con ácidos nucleicos.

Objetivo general

Inferir la estructura y secuencia de las proteínas Csp ancestrales mediante la reconstrucción de secuencias y la predicción de la estructura terciaria a partir de templados modernos para cuantificar el nivel de cambio observado que tales proteínas han sufrido en su historia.

Objetivos particulares

- Inferir la filogenia de las proteínas Csp y el dominio CSD.
- Hacer la reconstrucción de estados ancestrales para las Bacterias, las Arqueas y los Eucariontes.
- Hacer la predicción de la estructura terciaria de los estados ancestrales de las proteínas del dominio CSD y compararlas con las proteínas actuales.
- Determinar mediante acoplamientos moleculares *in silico* si las proteínas CSD ancestrales pueden o no interactuar con los mismos ligandos que sus versiones modernas.

Métodos

Reconstrucción filogenética del dominio CSD.

Se realizó una búsqueda de secuencias homólogas a la secuencia de la proteína CspA de *Escherichia coli* (NCBI: AAC76580) contra bases de datos de secuencias de Bacterias: *Firmicutes*, *Fusobacteria*, *Planctomycetes*, *Spirochaetes*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria*, *Chloroflexi*, *Deinococcus*, *Cyanobacteria*, *Aquificae* y *Thermotogae*; Arqueas y Eucariontes, del portal del NCBI (*National Center for Biotechnology Information*). La búsqueda se llevó a cabo con la ayuda de la herramienta PSI-BLAST con una matriz de sustitución BLOSUM62 y un valor de inclusión PSI-BLAST *e-value* de 0.001, contra la base de datos de proteínas de referencia (*refseq-protein*). Se llevaron a cabo iteraciones de la búsqueda hasta que no se encontraron secuencias nuevas. Para las búsquedas de secuencias de *Proteobacteria*, *Firmicutes*, *Bacteroidetes* y *Actinobacteria* se realizaron 11 iteraciones; 3 iteraciones para las secuencias de *Fusobacteria* y *Planctomycetes*; 2 iteraciones para las secuencias de *Spirochaetes*, *Chloroflexi*, *Deinococcus*, *Cyanobacteria*, *Aquificae* y *Thermotogae*; 3 iteraciones para la búsqueda de las secuencias de Arqueas y 7 iteraciones para la búsqueda de secuencias de Eucationtes. Para la inclusión de las secuencias en el proyecto, se revisó que éstas tuvieran un *e-value* menor a $1X10^{-20}$ y un porcentaje de covertura mayor al 30%, a continuación se excluyeron las secuencias redundantes para cada organismo, dejando una secuencia representativa de cada grupo con similitud superior al 80%. Además, se excluyeron las secuencias que eran hipotéticas, predichas o putativas. En el caso particular de *E. coli* se quitaron todas las secuencias arrojadas por el PSI-BLAST y se trabajó con las secuencias de las nueve proteínas Csp (CspA-CspI) utilizadas en trabajos anteriores (Etchegaray *et al*., 1996; Goldstein *et al*., 1990; Lee *et al*., 1994; Nakashima *et al*., 1996; Wang *et al*., 1999; Yamanaka & Inouye, 1997), con la finalidad de eliminar redundancias y quedarse con las secuencias que contaban con un mayor número de estudios.

Una vez obtenidas todas las secuencias se hicieron los alineamientos de estas, haciendo un alineamiento de las secuencias de Bacterias, Arqueas y Eucariontes por separado y un alineamiento que contenía todas las secuencias de Bacterias, Arqueas y Eucariontes, tomando como referencia la secuencia de la proteína CspA para cada alineamiento. Para esto se utilizó la herramienta bioinformática MAFFT (*Multiple alignment program for amino acid or nucleotide sequences*) versión 7 (Katoh & Standley, 2013), ya que este programa parte de la idea de que la frecuencia de sustitución entre 2 aminoácidos depende en gran medida de la diferencia de las propiedades fisicoquímicas, particularmente del volumen y polaridad de dichos aminoácidos (Miyata, Miyazawa, & Yasunaga, 1979), por lo que la sustitución entre aminoácidos con propiedades físico-químicas similares tienden a preservar la estructura de la proteína, asumiéndolas así, como sustituciones neutrales que se han ido acumulando en las moléculas a lo largo de la evolución (Kimura, 1983).

Así mismo, se realizó un alineamiento con el programa PROMALS (Pei & Grishin, 2007) con los parámetros de alineación y las opciones para ejecutar una búsqueda con PSI-BLAST descritas en la tabla 1. El programa utiliza un método progresivo para realizar un alineamiento múltiple de secuencias de proteínas haciendo su propia búsqueda en las bases de datos y una predicción de estructura secundaria, con el fin de conocer cuáles son las secuencias representativas de las secuencias del conjunto de datos que se le dio anteriormente al programa.

Tabla 1. Parámetros de alineamiento con el programa PROMALS.

Parámetros de alineación			Parámetros de búsqueda con PSI-BLAST			
Peso para las puntuaciones de aá.	Peso para estructura secundaria predichas	Identidad máxima alineación rápida	# Iteraciones	*e-value* de inclusión	Identidad mínima	# máximo de homólogos por alineación
0.8	0.2	0.6	3	0.001	0.2	1000

Después de hacer los alineamientos se probó, mediante el programa ProtTest (Abascal, Zardoya, & Posada, 2005), qué modelo de matriz de sustitución era el más adecuado para las secuencias obtenidas. Posteriormente, se hicieron 4 árboles de máxima verosimilitud, uno sólo con secuencias de Bacterias, uno sólo con secuencias de Arqueas, uno con secuencias de Eucariontes y otro con todas las secuencias de Bacterias, Arqueas y Eucariontes. Para lo cual se usó el programa PhyML 3.0 (Guindon *et al*., 2010), con un árbol inicial construido con BIONJ (Gascuel, 1997) y un modelo de sustitución LG (el cual incorpora la variabilidad de las tasas de evolución a través de los sitios para la estimación de matriz); (Le & Gascuel, 2008), según los resultados obtenidos con el programa ProtTest (Abascal *et al*., 2005) y un soporte de ramas obtenido mediante la prueba *aLRT SH-like*.

Predicción de la secuencias ancestral del dominio CSD

A continuación, se llevó a cabo la reconstrucción de secuencias ancestrales utilizando el programa FastML, el cual usa un algoritmo que reconstruye las secuencias ancestrales poniendo énfasis en una reconstrucción exacta de sitios de inserción o *deleción* (Ashkenazy *et al*., 2012) y usa el método de máxima verosimilitud para reconstruir las secuencias de los nodos de un dendograma ya construido. La reconstrucción de secuencias ancestrales se hizo con un modelo de sustitución LG y un árbol inicial obtenido anteriormente con el programa PhyML 3.0. El programa aplica dos métodos para hacer la reconstrucción de secuencias ancestrales: la reconstrucción denominada *joint*, la cual toma en cuenta el conjunto de todas las secuencias de nodos internos, y la reconstrucción denominada *marginal*, la cual infiere la secuencia más probable en un nodo interno específico. Ambos métodos se basan en algoritmos de máxima verosimilitud (ML) y tienen un enfoque empírico bayesiano que toma en cuenta la tasa de variación entre los sitios del alineamiento múltiple de secuencias (MSA) (Pupko, Pe'er, Hasegawa, Graur, & Friedman, 2002b; Pupko *et al*., 2000). Sin embargo, los resultados no son necesariamente los mismos. En los casos en los que el programa FastML obtuvo más de 1 secuencia ancestral reconstruida por nodo, igualmente

verosímil, se hizo un consenso de estas secuencias y la secuencia consenso fue la que se ocupó como la secuencia ancestral.

Se tomó la reconstrucción de las secuencias ancestrales de los 3 nodos (N1, N2 y N3) más cercanos a la base del dendograma construido con las 655 secuencias de Bacterias, Arqueas y Eucariontes (Fig. 7) y se denominaros a las secuencias ancestrales reconstruidas como Ance_3D8_1, Ance_3D8_2 y Ance_3D8_3, correspondiendo a los nodos N1, N2 y N3, respectivamente. Para el dendograma de las secuencias representativas obtenido con el programa PROMALS (Fig. 8) se tomó la reconstrucción de las secuencias ancestrales del nodo más cercano a la base (nodo N1), que incluye a todas las 56 secuencias representativas de Bacterias, Arqueas y Eucariontes y los dos nodos que separan al dendograma en dos clados, uno que incluye a la mayoría de las secuencias, 51 secuencias, denominado como N2 y otro con cinco secuencias, principalmente secuencias del *Phylum Planctomycetes*, denominado como N3. Las secuencias ancestrales reconstruidas con el dendograma construido con el programa PROMALS se nombraron como Ance_R5_1, Ance_R5_2 y Ance_R5_3 que corresponden a los nodos N1, N2 y N3, respectivamente (Fig. 8).

Predicción de las estructuras de las secuencias ancestrales del dominio CSD.

Se realizó la predicción de la estructura secundaria con la ayuda del programa PSIPRED (Jones, 1999) de las secuencias ancestrales obtenidas con la reconstrucción denominada *joint* y con la reconstrucción denominada *marginal* del programa FastML con la finalidad de evaluar cuál de estos dos modelos sería más útil para el estudio de las características distintivas de las proteínas Csp y del dominio CSD.

Una vez elegido el método de reconstrucción de secuencias ancestrales, se llevó a cabo la predicción de estructura secundaria con las secuencias ancestrales obtenidas con la reconstrucción denominada *joint*, tanto de la construcción de la filogenia obtenida con el alineamiento con el programa MAFFT (Ance_3D8_1, Ance_3D8_2 y Ance_3D8_3), como en la construcción de la filogenia obtenida con el alineamiento con el programa PROMALS (Ance_R5_1, Ance_R5_2 y Ance_R5_3).

De forma similar a la predicción de estructura secundaria, se llevó a cabo la predicción de la estructura terciaria mediante los programas SWISS MODEL, ROBETTA (D. E. Kim, Chivian, & Baker, 2004) y QUARK (D. Xu & Zhang, 2012) con la finalidad de evaluar los métodos de predicción de estructura terciaria de los programas SWISS MODEL, ROBETTA y QUARK. Las estructuras resueltas con cada método se evaluaron mediante el programa MolProbity (Chen *et al*., 2010) y se decidió utilizar las resueltas con el programa ROBETTA (D. E. Kim, Chivian, & Baker, 2004).

El programa ROBETTA (D. E. Kim *et al*., 2004), hace la predicción de estructura terciaria mediante modelos *ab initio* (Simons *et al*.,,1997; Bradley *et al*.,,2005) y modelos comparativos de dominios de proteínas, que se construyen a partir de estructuras homólogas encontradas en el PDB (*Protein Data Bank*); (Berman *et al*., 2000) y se generan utilizando el protocolo RosettaCM (Song *et al*., 2013), y QUARK (D. Xu & Zhang, 2012), que ocupa un algoritmo *ab initio* de predicción de plegamiento y estructura.

En el caso del programa ROBETTA, el algoritmo eligió como mejor molde a la estructura del PDB 3I2Z, cadena A, que pertenece a la estructura cristalográfica resuelta de la proteína CspE de *Salmonella typhimurium*, la cual es un homodímero.

La validación de las estructuras terciarias se llevó a cabo mediante el programa MolProbity (Chen *et al*., 2010) que proporciona una evaluación de amplio espectro de la

calidad del modelo, tanto a nivel global como local, para proteínas o ácidos nucleicos. Se basa en gran medida en la optimización de la colocación de hidrógeno y el análisis de los contactos de todos los átomos, complementada por geometría covalente y criterios de ángulo de torsión.

Se llevó a cabo un alineamiento estructural de las proteínas ancestrales predichas Ance_3D8_1, Ance_3D8_2, Ance_3D8_3, Ance_R5_1¸ Ance_R5_2 y Ance_R5_3 contra las estructuras tomadas del PDB (Berman *et al*., 2000) de las siguientes proteínas: CspA de *Escherichia coli* (2L15), CspB de *Bacillus subtilis* (2ES2), Csp de *Thermotoga maritima* (1G6P), Csp de *Thermus thermophilus* (2A0J), CspE de *Salmonella typhimurium* (3I2Z), Lin28a de *Mus musculus* (3TRZ), Lin-28b de *Xenopus tropicalis* (4ALP) y Y-box1 de *Homo sapiens* (1H95). Con este propósito se utilizó el algoritmo TM-align (Zhang & Skolnick, 2005) que sirve para identificar el mejor alineamiento estructural entre dos proteínas. El algoritmo TM-align combina una matriz de superposición óptima TM-score (Zhang & Skolnick, 2004), que evalúa la similitud topológica entre las estructuras de dos proteínas, tomando en cuenta la longitud de la proteína diana, el número de residuos equivalentes entre las dos proteínas y la distancia entre los pares de residuos equivalentes entre las dos estructuras proteicas, e iteraciones de programación dinámica. TM-score es más sensible a la topología global que a las variaciones estructurales locales, se normaliza según la proteína diana y tiene valores entre 0 y 1, siendo 1 el valor más alto de similitud estructural entre dos estructuras proteicas (Zhang & Skolnick, 2004). El algoritmo TM-align sólo emplea las coordinadas de los carbonos alfa (Cα) del esqueleto de las estructuras proteicas dadas; sin embargo, la metodología se generaliza fácilmente a cualquier tipo de átomo (Xu & Zhang, 2010).

Inferencia de la función de las proteínas ancestrales del dominio CSD

Predicción de los sitios de unión a ligandos

Se llevó a cabo una predicción de los sitios de unión a ligandos de las secuencias de las proteínas ancestrales (Ance_3D8_1, Ance_3D8_2, Ance_3D8_3, Ance_R5_1, Ance_R5_2 y Ance_R5_3) empleando el programa RaptorX (Peng & Xu, 2011). RaptorX toma en consideración la correlación entre las características de las proteínas y hace una predicción de la estructura secuenciaría y terciaria, el desorden y la accesibilidad al solvente y con base en dichas características asume que la similitud en el pliegue de dos proteínas puede indicar la existencia de una relación evolutiva, que a su vez puede implicar un papel funcional compartido. La base de datos de clasificación estructural de proteínas (SCOP) proporciona una descripción de la relación estructural y evolutiva de la mayoría de las proteínas en el PDB 45,46. Cada vez que se construye un modelo de estructura, RaptorX proporciona una estadística de distribución de clase, pliegue, superfamilia, familia y tipo de proteína de algunas o todas las proteínas usadas como plantillas para la predicción de la estructura de la secuencia dina, utilizando sólo como plantillas proteínas con una calidad de alineación predicha de al menos el 85% contra estructuras de la base de datos SCOP 1.75 (Hubbard *et al*., 1999; Alexey G. Murzin *et al*., 1995). El programa RaptorX proporciona una idea inicial de la naturaleza de la proteína que se está modelando y, por lo tanto, proporcionará un punto de partida para una exploración adicional de la estructura en cuestión (Källberg *et al*., 2012; Peng & Xu, 2011).

Acoplamiento molecular

Se llevó a cabo un acoplamiento molecular (*Molecular Docking*) con las mejores estructuras resueltas de las proteínas ancestrales predichas (Ance_3D8_1 y Ance_R5_1) y con las estructuras tomadas del PDB de las proteínas CspB de *Bacillus caldolyticus* (2HAX) y *Bacillus subtilis* (2ES2 y 3PF5), las estructuras de la proteínas Lin28 de *Homo sapiens* (2LI8), *Xenopus tropicalis* (4ALP) y *Mus musculus* (3TRZ) y la proteína de unión al extremo terminal de telómeros de *Oxytricha nova* (1OTC). El acoplamiento molecular (*Molecular Docking*) es un procedimiento computacional que intenta predecir la unión no covalente de macromoléculas o, más frecuentemente, de una macromolécula (receptor) y una molécula pequeña (ligando) de manera eficiente. El objetivo es predecir las conformaciones del complejo receptor-ligando y la afinidad de unión, mediante la suma de las interacciones intermoleculares e interacciones intramoleculares, la sumatoria se da entre todos los pares de átomos que pueden interaccionar entre sí, excluyendo normalmente, átomos separados por tres enlaces covalentes consecutivos (Trott & Olson, 2009). Primero se muestrean las conformaciones del ligando en el sitio activo de la proteína; luego se clasifican estas conformaciones a través de una función de puntuación. Idealmente, los algoritmos de muestreo deberían poder reproducir el modo de enlace experimental y la función de puntuación también debería clasificarlo como el más alto entre todas las conformaciones generadas (Meng, Zhang, Mezei, & Cui, 2011).

El programa AutoDock vina (Trott & Olson, 2009) incorpora métodos de optimización como los algoritmos genéticos y el algoritmo de recocido anillado (*simulated annealing*), entre otros, para modelar la flexibilidad del ligando y mantener el receptor rígido. Los algoritmos evolutivos son métodos de optimizaciones globales simples, estocásticos y de propósito general, que simulan la evolución natural de los sistemas biológicos. Inicialmente este algoritmo genera una población inicial al azar, luego mediante procesos de selección/reproducción, se genera la población de la siguiente generación y se evalúa a cada miembro de la población calculando la aptitud para ese individuo. El valor de la aptitud se calcula según cuán bien se ajusta el

complejo a los requisitos deseados. Una vez obtenidos los valores de aptitud, se selecciona los individuos más aptos, queremos mejorar constantemente la aptitud general de la población, por lo que la selección ayuda a hacer esto, descartando los malos diseños y manteniendo solo a los mejores individuos de la población. Posteriormente, a los individuos seleccionados se les aplican operadores de búsqueda: la recombinación (*crossover*), esto para crear nuevos individuos combinando aspectos de los individuos seleccionados y la mutación, con el fin de agregar un poco de aleatoriedad a la genética de la población, de lo contrario, cada combinación de soluciones que se puedan crear estaría en la población inicial. Con esto se obtiene la próxima generación y se comienza de nuevo hasta que se cumpla el criterio de terminación (Du & Swamy, 2016). Por otro lado, el algoritmo de recocido anillado (*simulated annealing*) se encuentra inspirado en el proceso de recocido de la metalurgia, el cual consiste en calentar un material hasta una determinada temperatura para después dejar que se enfríe lentamente con el fin de alterar sus propiedades físicas debido a los cambios en su estructura interna. Dicho algoritmo mantiene una variable de temperatura para simular este proceso de calentamiento y de forma análoga al proceso de recocido, la variable de temperatura inicial es alta y luego desciende lentamente a medida que se ejecuta el algoritmo, junto con asensos aleatorios que le permiten salir de mínimos locales. Inicialmente el algoritmo puede aceptar soluciones que son peores que la solución actual, pero a medida que la temperatura decrece, el algoritmo se enfoca gradualmente en un área del espacio de búsqueda en la que puede encontrar una solución de baja energía, ya que la probabilidad de pasar a una mejor solución se mantiene en 1 o tiende hacia valores positivo y la probabilidad de pasar a una peor solución tiende a cero progresivamente (Du & Swamy, 2016).

Los ligandos se extrajeron de las estructuras resueltas por cristalografía o NMR. La hexatimidina (dT6) se tomó de la estructura 2HAX de la proteína CspB de *Bacillus caldolyticus*. La hexauridina (rU6) se tomó de la proteína Lin-28b de *Xenopus tropicalis* (4ALP). El let-7d miRNA *pre-elemen* se tomó de la proteína Lin28a de *Mus musculus* (3TRZ). El ligando AGGAGAU de la familia de miRNA Let-7 se tomó de la proteína Lin-28 de *Homo sapiens* (2LI8). Y el ssDNA se tomó de la proteína de unión al extremo terminal de telómeros (OnTEBP) de *Oxytricha nova* (1OTC), ver Tabla 2.

Tabla 2. Ligandos

Nombre	Proteína	PDB ID	Organismo	Formula
Hexatimidina (dT6)	CspB	2HAX	*Bacillus caldolyticus*	5'-D(*TP*TP*TP*TP*TP*T)-3'
Hexauridina (rU6)	Lin-28b	4ALP	*Xenopus tropicalis*	5'-R(*UP*UP*UP*UP*TP*U)-3'
Let-7d miRNA pre-elemen	Lin-28a	3TRZ	*Mus musculus*	5'-R(*GP*GP*GP*CP*AP*GP*GP*GP*AP*UP*UP*UP*UP*GP*CP*CP*CP*GP*GP*AP*G)-3'
AGGAGAU	Lin-28	2LI8	*Homo sapiens*	5'-R(*AP*GP*GP*AP*GP*AP*U)-3'
ssDNA	OnTEBP	1OTC	*Oxytricha nova*	5'-D(*GP*GP*GP*GP*TP*TP*TP*TP*GP*GP*GP*G)-3'

Los ligandos y las estructuras terciarias de las proteínas ancestrales predichas (Ance_R5_1 y Ance_3D8_1), las estructuras de las proteínas CspB de *Bacillus caldolyticus* (2HAX) y *Bacillus subtilis* (2ES2 y 3PF5), las estructuras de la proteínas Lin28 de *Homo sapiens* (2LI8), *Xenopus tropicalis* (4ALP) y *Mus musculus* (3TRZ) y la proteína de unión al extremo terminal de telómeros de *Oxytricha nova* (1OTC) se prepararon con el programa *AutoDockTools*, agregándole átomos de hidrógeno, solo los polares y las cargas con el método de *Gasteiger* PEOE (Gasteiger & Marsili, 1980). Todas las coordenadas se verificaron manualmente para cada proteína y los experimentos se repitieron 3 veces para cada proteína y ligando.

Se utilizó el programa PLIP (*protein–ligand interaction profiler*); (Salentin, Schreiber, Haupt, Adasme, & Schroeder, 2015) para analizar los patrones de interacción entre los resultados del acoplamiento molecular las proteínas ancestrales predichas y los diferentes ligandos obtenidos de las proteínas modernas (Tabla 2). PLIP utiliza cuatro pasos para detectar e informar las interacciones relevantes: preparación de la estructura, caracterización funcional, comparación basada en reglas y filtrado de las interacciones. En la etapa de preparación, a la estructura de entrada se le añaden hidrógenos y los ligandos se extraen junto con sus sitios de unión. Con este fin, PLIP utiliza el programa OpenBabel (O'Boyle, Morley, & Hutchison, 2008) para la representación interna de moléculas y la mayoría de los cálculos quimioinformáticos. En el siguiente paso, se detectan los posibles grupos que puedan llevar a cabo una interacción entre la proteína y el ligando y se caracterizan funcionalmente, es decir, se detectan los átomos hidrófobos, los átomos aceptadores o donadores para puentes de hidrógeno y halógeno, de igual

forma, se buscan anillos aromáticos y centros de carga en proteínas y ligandos con el fin de detectar la formación de apilamiento-π, interacciones de catión-π o puentes salinos. A continuación, los grupos interactivos putativos se combinan aplicando criterios mayormente geométricos. Dependiendo del tipo de interacción, esto puede incluir restricciones de distancia o ángulo entre la disposición de los átomos (Salentin *et al*., 2015).

Por último, los pasos de filtrado se utilizan para eliminar las interacciones redundantes o superpuestas. Esto es especialmente importante para los contactos hidrófobos, que pueden formarse entre cualquier parte apolar cercana del ligando y la proteína. PLIP busca automáticamente los contactos más relevantes (la distancia interatómica más corta dentro del entorno) para informar. Algunos tipos de interacción (por ejemplo, puentes salinos y puentes de hidrógeno) son muy similares en sus características. En el caso de la detección de ambos tipos de interacción para el mismo par de átomos, solo se informa uno de ellos (por ejemplo, un puente salino); (Salentin *et al*., 2015).

Predicción de los efectos de las mutaciones en las proteínas ancestrales.

Los efectos funcionales de las mutaciones se predicen con SNAP2 (*Screening for non-acceptable polymorphisms*); (Hecht, Bromberg, & Rost, 2015). SNAP2 es un clasificador entrenado que se basa en un dispositivo de aprendizaje automático llamado "red neuronal". Distingue entre el efecto de las variantes genéticas neutrales y los polimorfismos de un solo nucleótido (SNPs) no sinónimos. SNAP2 toma en cuenta una variedad de características: características globales (composición de aminoácidos, estructura secundaria y longitud de la proteína), perfiles PSI-BLAST (Altschul *et al*., 1997) llevando a cabo 4 o más iteraciones con un *e-value* igual a 0.001 contra las bases de datos de UniProt y PDB, perfiles PSIC (*position-specific independent counts*); (Sunyaev *et al*., 1999) que es una forma particular de compilar patrones de similitud

entre las secuencias de proteínas de un alineamiento. El alineamiento se hace con las secuencias obtenidas con PSI-BLAST que posean una identidad mayor al 30% y un *e-value* menor o igual al 0.001, con el programa CLUSTAL W (Hecht, 2015).

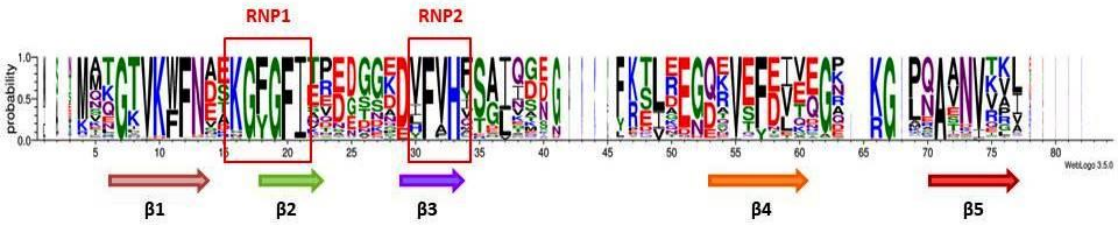
Así mismo, toma en cuenta la flexibilidad de residuos predicha con el programa PROFbval, desorden, la estructura secundaria y accesibilidad relativa al solvente obtenidas con el programa PROFacc, propiedades fisicoquímicas (carga, hidrofobicidad y volumen), perfiles de potencial de contacto, posiciones correlacionadas y regiones de baja complejidad, todo para la secuencia dada y sus variantes obtenidas con el PSI-BLAST. La información sobre la familia de la proteína es otra característica que se considera en la evaluación de la importancia de cambios en posiciones particulares, para lo cual, se toma un conjuntos de información relacionada con el dominio o familia de Pfam al que pertenece la proteína dada, como los sitios conservados y si existe la presencia o ausencia de otros dominios en las áreas circundantes de la secuencia de la proteína (Hecht, 2015).

Los efectos se predijeron para las proteínas ancestrales Ance_R5_1, 2 y 3, Ance_3D8_1, 2 y 3 y, para las proteínas modernas CspA, CspB, CspD y CspC de *Escherichia coli*, Lin28 de *Xenopus tropicalis*, Y-box de *Homo sapiens* y Gly-rich de *Cicer arietinum* con la ayuda del servidor PredictProtein (Rost & Liu, 2003).

Resultados

El alineamiento de las 655 secuencias de proteínas de Bacteria, Arquea y Eucarionte muestra que los motivos RNP1 y RNP2 se encuentran bien conservados, así como los sitios de unión a RNA y ssDNA y la estructura secundaria conformada por 5 hebras beta (Fig.7). En la construcción del dendograma con 655 secuencias derivado del alineamiento con la herramienta MAFFT se puede observar que el cado más basal del dendograma está ocupado por miembros del grupo de las *Fusobacteria* seguido por un grupo de bacterias *Gama-Proteobacteria* y *Delta-Proteobacteria*, los integrantes de estos grupos son principalmente anaerobios, quimioorganotrofos y habitan en su mayoría en ambientes marinos. Así mismo se puede observar que el grupo de las *Proteobacteria* se encuentra disperso en la topología del dendograma, aunque agrupándose según el tipo de *Proteobacteria* (Alfa, Beta, Gamma, Delta y Zeta). Las secuencias de Arqueas se encuentra repartidas en 2 clados, el de las *Euriarqueota* y el de las *Thaumarqueota* y que están representadas principalmente por organismos halófilos (Fig.8).

A



B

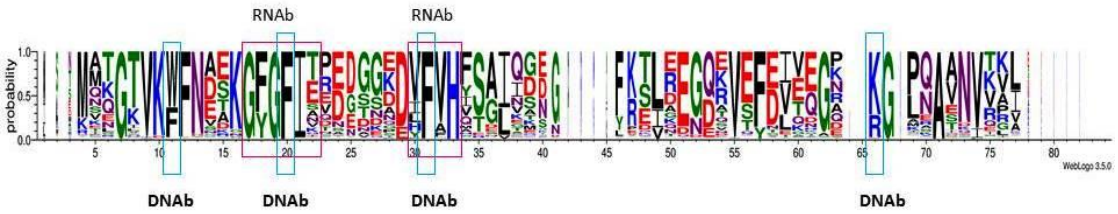


Fig. 7. Logo del alineamiento de las 655 secuencias de las proteínas Csp y el dominio CSD presente en los tres dominios de la vida, Bacteria, Arquea y Eucarionte.
En la figura A se encuentran indicados los motivos RNP1 y RNP2 con un recuadro color rojo, y la predicción de las hebras *β*, se encuentran indicadas con flechas. En la figura B los recuadros de color rojo señalan los motivos de unión a RNA y los aminoácidos de unión a DNA se encuentran remarcados en azul.

Las secuencias de los Eucariontes se encuentran agrupadas principalmente en un clado que a su vez se encuentran divididas según el tipo de proteína del que deriva el dominio CSD, Gly-rich, Y-box y Lin-28 y se puede observar que el clado de las secuencias de eucariontes y el clado de las Arqueas son grupos hermanos. Las secuencias correspondientes al *Phylum* de los *fungi* se encuentran en un clado aparte del de las demás secuencias de Eucariontes más cercano a las *Beta-Proteobacteria* (Fig.8). En los clados más externos del dendograma podemos encontrar al *Phylum* de los *Firmicutes* donde también se encuentran las secuencias de *Aquifex aeolicus* y *Thermotoga marítima* (Fig.8).

A partir del alineamiento con el programa PROMALS se obtuvieron 55 secuencias representativas (respecto a la estructura secundaria de las proteínas) con representantes de los *Phyla Fusobacteria*, *Planctomycetes*, *Spirochaetes*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria* (Alfa, Beta, Gamma, Delta y Zeta), *Chloroflexi*, *Deinococcus*, *Aquificae*, Arqueas (*Euriarqueota* y *Thaumarqueota*), Eucariontes (*Metazoo*, *Rhizaria*, *Viridiplantae*, *Fungi* y *Stramenopliles*). En la reconstrucción filogenética de estas secuencias representativas se obtuvo que el clado de los *Planctomycetes*, junto con representantes de las *Fusobacteria* y las *Spirochaetes* se encuentran en las ramas más basales del dendograma, estas bacterias son principalmente alcalinófilas y halófilas. También, se puede observar que las estructuras de las secuencias de los *Bacteroidetes* deriva de las de las Arqueas y que las estructuras de las secuencias de los Eucariontes derivan principalmente de las de las *Proteobacterias* y que se encuentran divididas principalmente en dos clados separados, (Fig. 9).

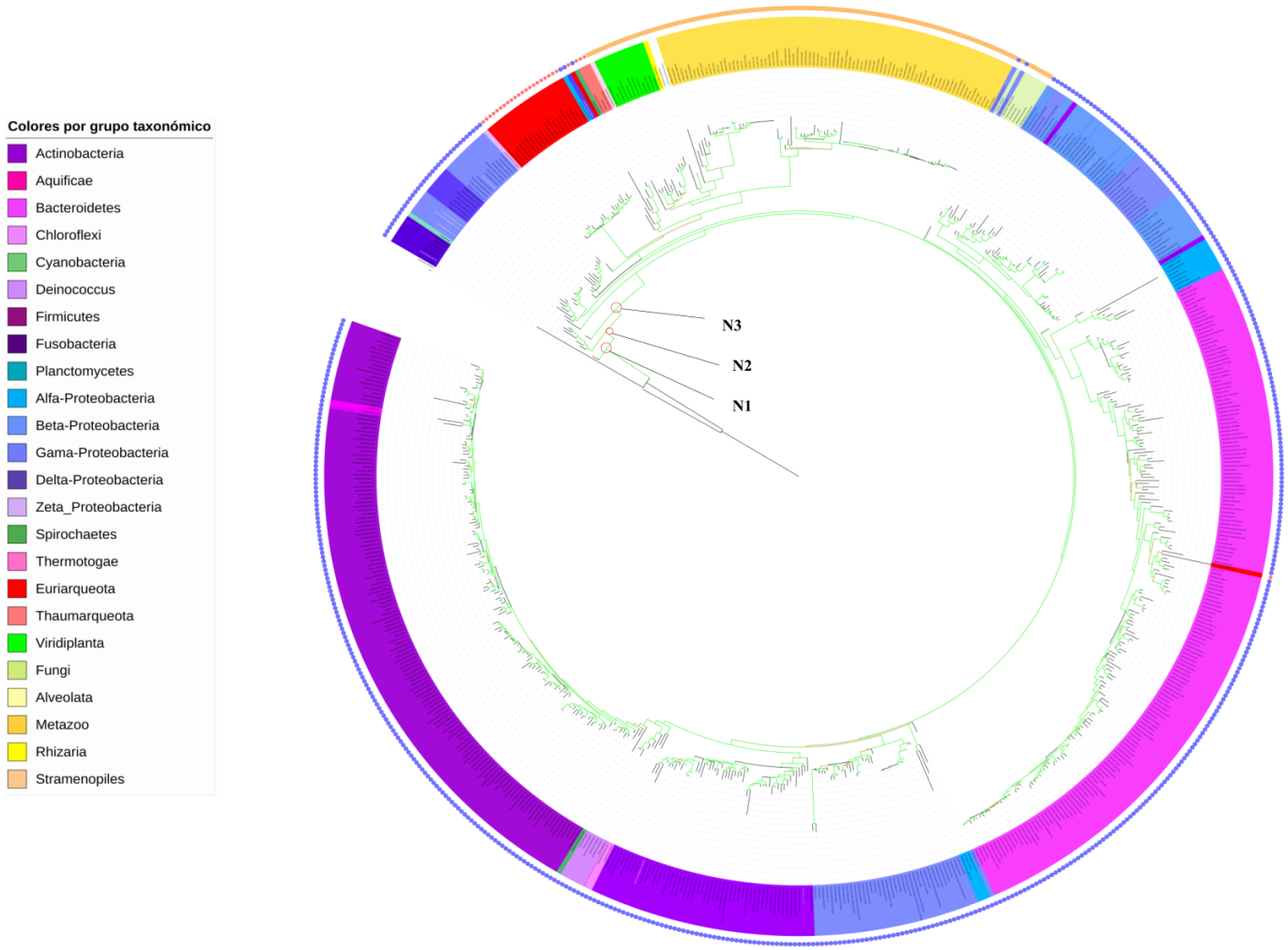


Fig. 8. Reconstrucción filogenética del dominio CSD y las proteínas Csp.
Los colores de las ramas representan los valores de soporte de las ramas, los colores verdes representan los valores cercanos a 1. Los colores que enmarcan el nombre de la especie corresponden al color del *Phylum* y, en el caso de los *fungi*, al *Phylum* correspondiente a los organismos de los que provienen las secuencias, como se puede apreciar en la leyenda del lado izquierdo de la figura. Los círculos azules indican las secuencias pertenecientes al dominio Bacteria, las estrellas rojas indican las secuencias del dominio Arquea y los cuadros amarillos indican las secuencias del dominio Eucarionte. Los nodos que se ocuparon para hacer la reconstrucción de secuencias ancestrales se encuentran indicados con un círculo rojo (N1, N2 y N3).

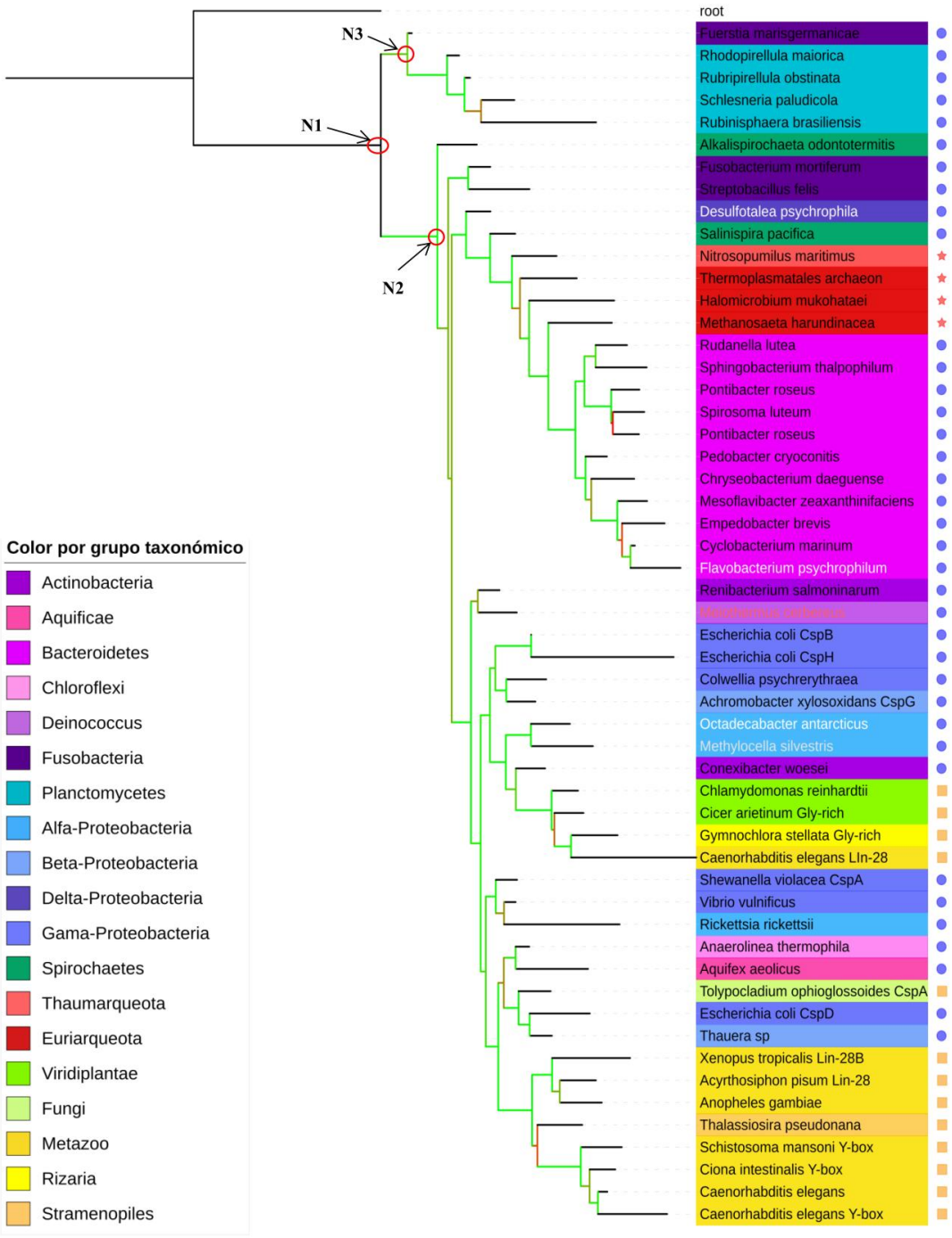


Fig. 9. Dendograma construido a partir del alineamiento con PROMALS de las 55 secuencias representativas.
Los colores de las ramas representan los valores de soporte de las ramas, los colores verdes representan los valores cercanos a 1. Los colores que enmarcan el nombre de la especie corresponden al color del *Phylum* y, en el caso de los *fungi*, al *Phylum* correspondiente a los organismos de los que provienen las secuencias, como se puede apreciar en la leyenda del lado izquierdo de la figura. Los círculos azules indican las secuencias pertenecientes al dominio Bacteria, las estrellas rojas indican las secuencias del dominio Arquea y los cuadros amarillos indican las secuencias del dominio Eucarionte. Los nodos que se ocuparon para hacer la reconstrucción de secuencias ancestrales se encuentran indicados con un círculo rojo (N1, N2 y N3).

Predicción de la secuencia ancestral del dominio CSD

Se obtuvieron 6 secuencias ancestrales del dominios CSD, las tres que se obtuvieron a partir de la reconstrucción filogenética con 655 secuencias se nombraron Ance_3D8_1,Ance_3D8_2 y Ance_3D8_3; las tres con 58 residuos de aminoácidos (Tabla 3) y las tres que se obtuvieron de la reconstrucción filogenética de las 55 secuencias representativas obtenidas a partir del alineamiento con el programa PROMALS se denominaron Ance_R5_1, Ance_R5_2 y Ance_R5_3, con 65 residuos para Ance_R5_1 y 66 residuos para Ance_R5_2 y Ance_R5_3 (Tabla 3). Las secuencias ancestrales predichas conservaron los motivos RNP1 y RNP2, así como los sitios de unión putativos a RNA y DNA (Fig.10) y la formación de 5 hebras beta de la estructura secundaria de los representantes del dominio CSD actuales.

Tabla 3. Secuencias ancestrales predichas.

Nombre	Nodo	Longitud	Secuencia
Ance_3D8_1	N1	58	MGTVKWFNKDKGFGFISGEDGDYFVHYSNIKGFR SLEEGQVSFEVTEGKGPVASNVVA
Ance_3D8_2	N2	58	MGTVKWFNEDKGFGFISGEDGDYFVHFSQIEGFK TLEEGQVTFEITQGKGPQASNVAV
Ance_3D8_3	N3	58	MGTVKWFNQEKGFGFITSEDGDVFVHFSQIPGFKT LEEGQVTFEITQGKGPQASNVAV
Ance_R5_1	N1	65	MAQGTVKRITDRGFGFIATDEGQDMFVHCSNIDG ENFESLQEGQRVSYSVGSGPKGPRAENVRSL
Ance_R5_2	N2	66	MAQGTVKWFNDERGFGFITKEDGNDIFVHYSAIN GEGFRSLDEGQRVSFEIEEGPKGPQAANVTPL
Ance_R5_3	N3	66	MAQGTVKWFNDEKGFGFITREDGNDIFVHYSAIN AEGFRTLDEGQRVSFEIEEGAKGPQAANVTAL

La secuencia ancestral Ance_R5_1 (Tabla3) es la más diferente en comparación con las secuencias de la proteínas Csp de *Escherichia coli* (CspA, *Ec*CspA*;* CspB, *Ec*CspB*;* CspC, *Ec*CspC*;* CspE, *Ec*CspE*;* CspD, *Ec*CspD), *Bacillus subtilis* (*Bs*CspB), *Bacillus caldolyticus* (*Bc*CspB), *Thermotoga maritima* (*Tm*Csp), *Thermus thermophilus* (*Tt*Csp), *Neisseria meningitidis* (*Nm*CspA), *Salmonella typhimurium* (St*Cspe*), siendo la secuencia de *N. meningitidis* (*Nm*CspA) la más parecida a esta con 0.507 de identidad (Tabla 4).

Tabla 4. Identidad entre las secuencias por pares.

\	Ance_R5_1	Ance_R5_2	Ance_R5_3	Ance_3D8_1	Ance_3D8_2	Ance_3D8_3
*Ec*CspA	0.485	0.6	0.614	0.528	0.542	0.557
*Ec*CspB	0.436	0.521	0.563	0.507	0.549	0.535
*Ec*CspC	0.42	0.579	0.608	0.536	0.594	0.608
*Ec*CspD	0.351	0.472	0.472	0.472	0.445	0.445
*Ec*CspE	0.434	0.608	0.666	0.565	0.623	0.637
*Nm*CspA	0.507	0.626	0.656	0.53	0.56	0.56
*Tm*Csp	0.426	0.573	0.588	0.5	0.575	0.59
*Bc*CspB	0.447	0.626	0.641	0.537	0.611	0.626
*Bs*CspB	0.455	0.661	0.676	0.558	0.617	0.632
*Tt*Csp	0.364	0.54	0.594	0.486	0.486	0.513
*St*CspE	0.422	0.605	0.661	0.549	0.605	0.619
*Cr*CSD	0.369	0.476	0.464	0.433	0.397	0.397
*Ca*GlyRich	0.352	0.492	0.492	0.45	0.45	0.464
*Xt*Lin28B	0.31	0.391	0.378	0.378	0.364	0.351
*Ce*YB	0.293	0.413	0.426	0.346	0.333	0.32
*Ce*Lin28	0.352	0.436	0.436	0.414	0.357	0.357
*Hs*YB1	0.324	0.432	0.432	0.405	0.351	0.364

Mediante el alineamiento (Fig. 10) de las secuencias ancestrales reconstruidas (Ance_3D8_1, Ance_3D8_2, Ance_3D8_3, Ance_R5_1, Ance_R5_2 y Ance_R5_3) con las secuencias de proteínas Csp de *E. coli* (CspA), de *B. subtilis* (BsCspB), de *B. caldolyticus* (BcCspB), de la proteína rica en glicina de *C. arietinum* (CaGlyRich), de la proteína Lin28B de *X. tropicalis* (XtLin28B) y de la proteína Y-box de *H. sapiens* (HsYB1) podemos observar que hay residuos altamente conservados como en el caso de la lisina en la décima posición en el alineamiento, correspondiente a K7 en las secuencias ancestrales Ance_R5_1, Ance_R5_2 y Ance_R5_3 y en la secuencia de las proteínas CspB de *B. subtilis* y *B. caldolyticus* y la proteína Lin28B de *X. tropicalis*, a K5 en las secuencias ancestrales Ance_3D8_1, Ance_3D8_2 y Ance_3D8_3 y a K14 en

la secuencia de las proteínas YB1 de *H. sapiens*; así como las glicinas en la séptima, decimoséptima, decimonovena y septuagésima segunda posición; las felnilalaninas en la decimotercera, decimoctava, vigésima, trigésima octava y sexagésima cuarta posición; entre otras. En el caso particular de la secuencia ancestral Ance_R5_1 podemos observar que tiene una arginina (R8) en la onceava posición, una isoleucina en la doceava y una treonina en la treceava posición del alineamiento en vez de del triptófano, fenilalanina y asparagina, respectivamente, presentes en las demás secuencias del alineamiento (Fig. 10).

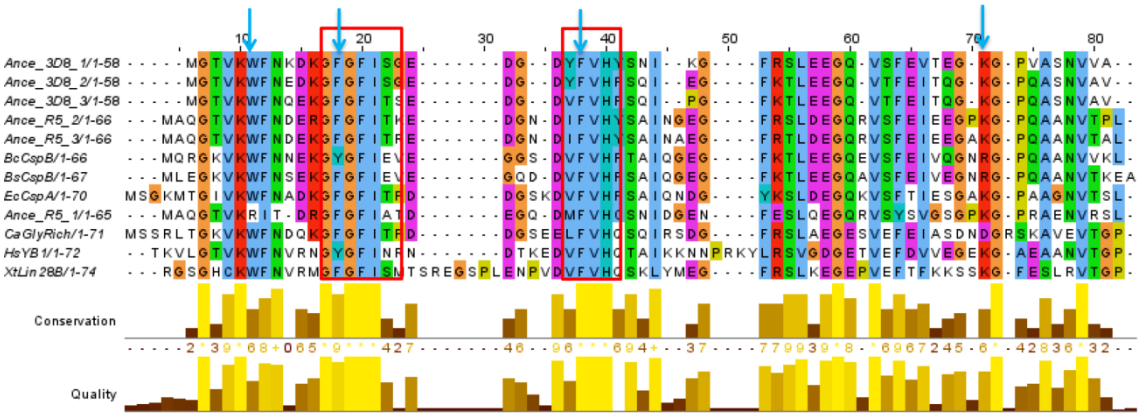


Fig. 10. Alineamiento de las secuencias ancestrales reconstruidas y secuencias actuales. Secuencias ancestrales reconstruidas (Ance_R5_1, Ance_R5_2, Ance_R5_3, Ance_3D8_1, Ance_3D8_2 y Ance_3D8_3) y las secuencias de proteínas Csp de *E. coli* (CspA), de *B. subtilis* (*Bs*CspB), de *B. caldolyticus* (*Bc*Csp), de la proteína Gly-rich de *C. arietinum* (*Ca*GlyRich), de la proteína Lin28B de *X. tropicalis* (*Xt*Lin28B) empezando por el residuo 33 y de la proteína Y-box de *H. sapiens* (*Hs*Ybox) empezando por el residuo 7. Dentro de los recuadros color rojo se encuentran señalados los motivos de unión a RNA y los aminoácidos de unión a DNA se encuentran indicados con una flecha de color azul.

La predicción de los sitios de unión de las secuencias ancestrales (Ance_3D8_1, Ance_3D8_2, Ance_3D8_3, Ance_R5_1, Ance_R5_2 y Ance_R5_3) concuerdan con los sitios de unión observados en las proteínas modernas CspB de *B. subtilis*, YB1 de *H. sapiens* y Lin28B de *X. tropicalis* y *H. sapiens* (Tabla 8). Los residuos K5, W6, F13, F15 y F24 de las proteínas ancestrales Ance_3D8_1, 2 y 3, los residuos F14, F16 y F26 de la proteína Ance_R5_1 y los residuos K7, W8, F14, F16 y F26 de las proteínas ancestrales Ance_R5_2 y 3 son homólogos a los residuos K7, W8, F15, F17, F27 y R56 que interactúan con oligonucleótidos de la proteína CspB, los residuos K14, W15, F24

Y F35 de YB1 y los residuos W39, F48, F66 y F77 de Lin28B (Tabla 8). En el caso particular de la secuencia de la proteína ancestral Ance_R5_1 los residuos R8, T10 y R12 son diferentes a los residuos W8, N10 y K13 de la proteína CspB (Fig. 10), a pesar de estos cambios, se predice que dicho sitio interactúa con oligonucleótidos, en especial con uridina.

Tabla 5. Predicción del sitio de unión de las proteínas ancestrales

Ance_3D8_1				Ance_R5_1			
Hueco	Multiplicidad*	Ligando	Residuos de unión	Hueco	Multiplicidad*	Ligando	Residuos de unión
1	52	DT	K5 W6 F13 F15 D22 F24 P51 V52	**1**	58	DT	F14 F16 D24 F26 G53 K55 G56 R58
2	39	U	W6 F7 N8	**2**	44	U	R8 T10 R12
3	38	A	K9 G12 Y27 F33 R34	**3**	34	A	D11 G13 C29 F37 E38
4	33	G	N8 K9 D10	**4**	27	G	F14 F26 H28 S30
5	28	G	F13 F24 H26 Y27	**5**	18	G	R12 F14
6	28	G	D10 K11 F13	**6**	15	U	I9 T10 D11
7	20	G	N8 F13 F24				
8	14	C	I30 K31 R34				

*La multiplicidad representa la frecuencia con la que se encontró el sitio de unión seleccionado en un conjunto de estructuras de proteínas.

Predicción de la estructura de la proteína ancestral del dominio CSD

La predicción de estructura secundaria realizada con el programa PSIPRED arrojo, para las seis secuencias ancestrales predichas, una estructura compuesta por 5 hebras, en el caso de la proteína ancestral Ance_R5_1 las hebras 1, 2, 3 y 4 poseen tamaños similares mientras que la hebra 5 es considerablemente más pequeña y en comparación con la estructura de la proteína Ance_3D8_1 esta hebra se encuentra más cercana al extremo carboxilo terminal. Estas dos estructuras también se diferencian porque en el caso de las proteína Ance_3D8_1 la hebra más corta es la número 4 (Fig.11 A y B).

Al comparar las estructuras terciarias de las proteínas ancestrales Ance_3D8_1 y Ance_3R5_1 se obtiene un TM-score igual a 0.79 (Fig. 11 C), así como se puede observar que el bucle 3 (L3), la hebra 3 (β3) y la hebra 4 (β4) de la proteína Ance_R5_1 son más largos que los de la Ance_3D8_1, mientras que la hebra 5 es más corta en el caso de la proteína ancestral Ance_R5_1 (Fig. 11 C).

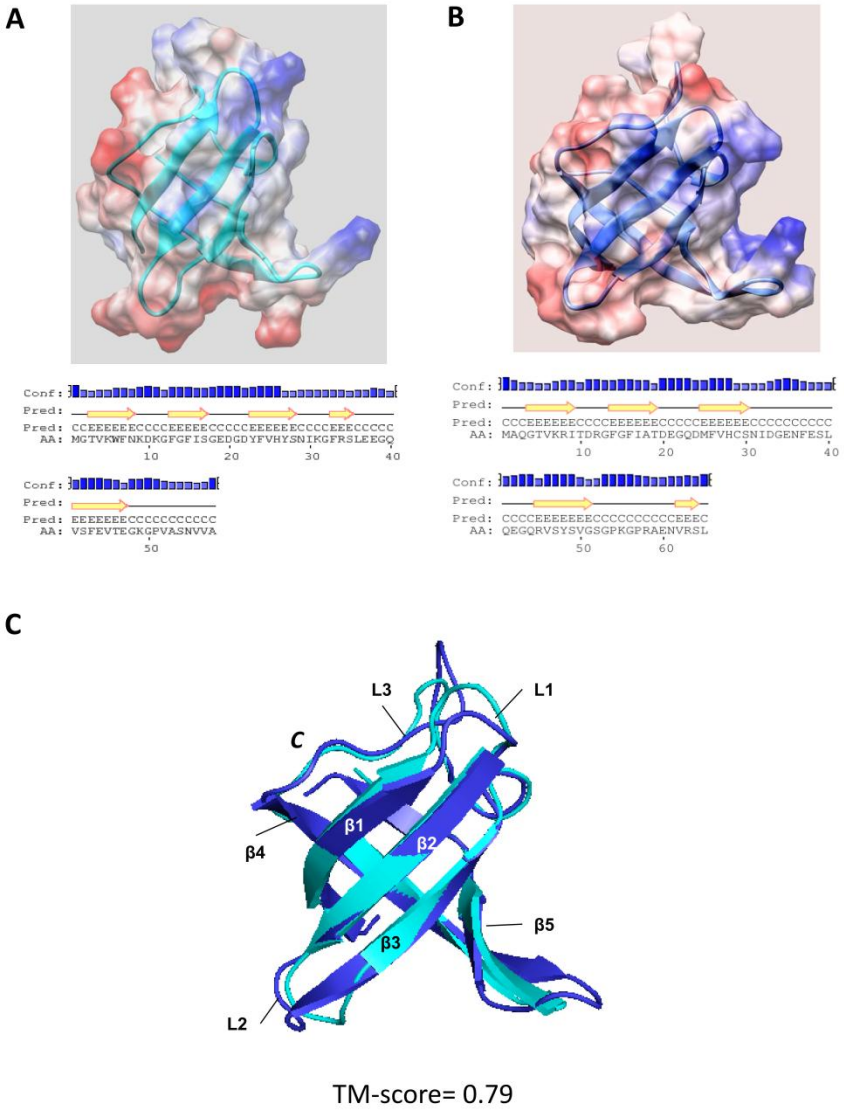


Fig. 11. Modelos esquemáticos de las estructuras terciarias probables para las proteínas ancestrales.
En el panel A podemos observar la estructura terciaria predicha con el programa ROBETTA de la proteína Ance_3D8_1 y en la parte inferior se encuentra la secuencia y su estructura secundaria predicha con PSIPRED. En el panel B podemos observar la estructura terciaria predicha con el programa ROBETTA de la proteína Ance_R5_1 y en la parte inferior se encuentra la secuencia y su estructura secundaria. El panel C muestra el alineamiento de la estructura de la secuencia ancestral Ance_R5_1 (color azul) contra la estructura de la secuencia ancestral Ance_3D8_1 (color cian), con un TM-score=0.79, en el cual se puede apreciar el alineamiento estructural de las 5 hebras β (β1, β2, β3, β4 y β5) y los 3 bucles (L1, L2 y L3) de cada estructura. La C señala el extremo carboxilo terminal de las proteínas.

En cuanto a la estructura terciaria todas las estructura predichas de las secuencias ancestrales son un monómero que conserva la estructura de barril β plegado compuesto por 5 hebras β antiparalelas. De los 5 modelos de estructura terciaria de cada una de las secuencias ancestrales (Ance_3D8_1, Ance_3D8_2, Ance_3D8_3, Ance_R5_1, Ance_R5_2 y Ance_R5_3) arrojados por el programa ROBETTA, el modelo 5 de la secuencia Ance_3D8_1 fue el que obtuvo los mejores valores de validación de MolProbity como se puede observar en la Tabla 5, en el caso de la secuencia Ance_R5_1 el modelo 1 obtuvo los mejores valores (Tabla 6), así mismo el modelo 1 fue el mejor de las secuencias Ance_3D8_2 y Ance_3D3, para el caso de la secuencia Ance_R5_2, los modelos1, 3 y 5 son igual de buenos y para la secuencia Ance_R5_3 los mejores modelos son el 2 y el 5.

Tabla 6. Valores de validación de MolProbity de la estructura 3ª de la secuencia ancestral Ance_3D8_1

Ance_3D8_1												
		Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5		
Contactos entre todos los átomos	Superposiciones estéricas de todos los átomos (> 0.4 Å) por 1000 átomos	0		0		0		0		0		100º percentile (N=1784. todas las resoluciones)
Geometría	Rotameros pobres	0	0%	0	0%	0	0%	0	0%	0	0%	Valor ideal: <0.3%
	Rotameros favorables	47	100%	47	100%	47	100%	47	100%	47	100%	Valor ideal: >98%
	Valores atípicos de Ramachandran	0	0%	0	0%	0	0%	0	0%	0	0%	Valor ideal: <0.05%
	Valores favorables de Ramachandran	55	98.21 %	54	96.43 %	54	96.43 %	55	98.21 %	56	100%	Valor ideal: >98%
	Puntaje de MolProbity	0.5		0.74		0.74		0.5		0.5		100º percentile (N=27675. 0Å - 99Å)
	Desviación Cβ >0.25Å	0	0%	0	0%	0	0%	0	0%	0	0%	Valor ideal: 0
	Enlaces desfavorables	0 / 462	0%	0 / 462	0%	0 / 462	0%	1 / 462	0.22 %	0 / 462	0%	Valor ideal: 0%
	Ángulos desfavorables	2 / 620	0.32 %	4 / 620	0.65 %	2 / 620	0.32 %	1 / 620	0.16 %	1 / 620	0.16 %	Valor ideal: <0.1%

Péptidos Omegas	Prolinas en Cis	0 / 1	0%	0 / 1	0%	0 / 1	0%	0 / 1	0%	0 / 1	0%	Experado: ≤1 por cadena. o ≤5%
	Otros aminoácidos en Cis	0 / 56	0%	0 / 56	0%	0 / 56	0%	0 / 56	0%	0 / 56	0%	Valor ideal: <0.05%
	Péptidos retorcidos	0 / 57	0%	1¬57	1.75 %	1¬57	1.75 %	1¬57	1.75 %	1¬57	1.75 %	Valor ideal: 0
Criterios de baja resolución	Valores atípicos CaBLAM	1	1.82 %	1	1.82 %	0	0%	1	1.82 %	0	0%	Valor ideal: <1.0%
	Valores geométricos atípicos CA	2	3.64 %	2	3.64 %	3	5.45 %	3	5.45 %	2	3.64 %	Valor ideal: <0.5%

Tabla 7. Valores de validación de MolProbity de la estructura 3ª de la secuencia ancestral Ance_R5_1.

Ance_R5_1												
		Modelo 1		Modelo 2		Modelo 3		Modelo 4		Modelo 5		
Contactos entre todos los átomos	Superposiciones estéricas de todos los átomos (> 0.4 Å) por 1000 átomos	0		0		1.03		0		0		100º percentile (N=1784. todas las resoluciones)
Geometría	Rotameros pobres	0	0%	0	0%	0	0%	0	0%	0	0%	Valor ideal: <0.3%
	Rotameros favorables	53	100%	53	100%	53	100%	53	100%	53	100%	Valor ideal: >98%
	Valores atípicos de Ramachandran	0	0%	1	1.59 %	0	0%	1	1.59 %	1	1.59 %	Valor ideal: <0.05%
	Valores favorables de Ramachandran	61	96.83 %	60	95.24 %	61	96.83 %	61	96.83 %	59	93.65 %	Valor ideal: >98%
	Puntaje de MolProbity	0.69		0.83		1		0.69		0.92		100º percentile (N=27675. 0Å - 99Å)
	Desviación Cβ >0.25Å	0	0%	0	0%	0	0%	0	0%	0	0%	Valor ideal: 0
	Enlaces desfavorables	0/503	0%	0/503	0%	0/503	0%	0/503	0%	0/503	0%	Valor ideal: 0%
	Ángulos desfavorables	0/674	0%	2/674	0.3%	0/674	0%	3/674	0.45 %	2/674	0.3%	Valor ideal: <0.1%

Péptidos Omegas	Prolinas en Cis	0/2	0%	0/2	0%	0/2	0%	0/2	0%	0/2	0%	Experado: ≤1 por cadena. o ≤5%
	Otros aminoácidos en Cis	0/62	0%	0/62	0%	0/62	0%	0/62	0%	0/62	0%	Valor ideal: <0.05%
	Péptidos retorcidos	0/64	0%	1/64	1.56 %	0/64	0%	1/64	1.56 %	1/64	1.56 %	Valor ideal: 0
Criterios de baja resolución	Valores atípicos CaBLAM	2	3.23 %	3	4.84 %	2	3.23 %	3	4.84 %	3	4.84 %	Valor ideal: <1.0%
	Valores geométricos atípicos CA	2	3.23 %	3	4.84 %	2	3.23 %	3	4.84 %	3	4.84 %	Valor ideal: <0.5%

Mediante el alineamiento con el algoritmo TM-align de las estructuras terciarias de las secuencias ancestrales predichas Ance_3D8_1, Ance_3D8_2, Ance_3D8_3, Ance_R5_1, Ance_R5_2 y Ance_R5_3 contra las estructuras de las proteínas CspA de *Escherichia coli* (PDB:2L15), CspB de *Bacillus subtilis* (PDB:2ES2), Csp de *Thermotoga maritima* (PDB:1G6P), Csp de *Thermus thermophilus* (PDB:3A0J), CspE de *Salmonella typhimurium* (PDB:3I2Z), Lin28A de *Mus musculus* (PDB:3TRZ), Lin-28B de *Xenopus tropicalis* (PDB:4ALP) y Y-box1 de *Homo sapiens* (PDB:1H95), se puede observar que las estructuras de las secuencias ancestrales Ance_R5_1, 2 y 3 presenta valores de TM-score mayores que los de las estructuras de las secuencias ancestrales predicha Ance_3D8_1, 2 y 3 (Tabla 7), por ejemplo, en el caso de la comparación de la estructura de la proteína CspB de *B. subtilis* contra la estructura de la secuencia ancestral Ance_3D8_1 se obtuvo un TM-score igual a 0.76 mientras que contra la secuencia ancestral Ance_R5_1 el TM-score es igual a 0.89 (Fig. 12). Las puntuaciones de TM-score por debajo de 0.17 corresponden a proteínas no relacionadas elegidas al azar, mientras que las estructuras con una puntuación superior a 0.5 suponen generalmente el mismo pliegue en SCOP/CATH (Zhang & Skolnick, 2004).

Tabla 8.Alineamiento estructural con el algoritmo TM-align de las proteínas ancestrales predichas.

Proteína		CspA	CspB	Csp	Csp	CspE	Lin28A	Lin28B	Y-box
	PDB ID	2L15	2ES2	1G6P	3A0J	3I2Z	3TRZ	4ALP	1H95
Ance_3D8_1	TM-score	0.74	0.76	0.68	0.71	0.77	0.59	0.59	0.62
	RMSD	1.12	1.25	1.95	1.24	1.22	1.2	1.28	1.9
Ance_3D8_2	TM-score	0.75	0.75	0.68	0.70	0.77	0.6	0.6	0.62
	RMSD	1.03	1.41	1.98	1.42	1.33	1.29	1.26	1.91
Ance_3D8_3	TM-score	0.74	0.75	0.68	0.70	0.75	0.59	0.59	0.62
	RMSD	1.15	1.35	1.98	1.43	1.31	1.35	1.33	1.86
Ance_R5_1	TM-score	0.85	0.89	0.71	0.79	0.89	0.64	0.68	0.65
	RMSD	1.11	1.06	2.2	1.44	1.03	1.33	1.18	1.85
Ance_R5_2	TM-score	0.85	0.91	0.72	0.81	0.91	0.69	0.68	0.66
	RMSD	1.15	0.92	2.35	1.42	0.99	1.12	1.27	2.06
Ance_R5_3	TM-score	0.84	0.92	0.71	0.82	0.91	0.69	0.68	0.67
	RMSD	1.14	0.90	2.43	1.32	0.97	1.19	1.35	2.12

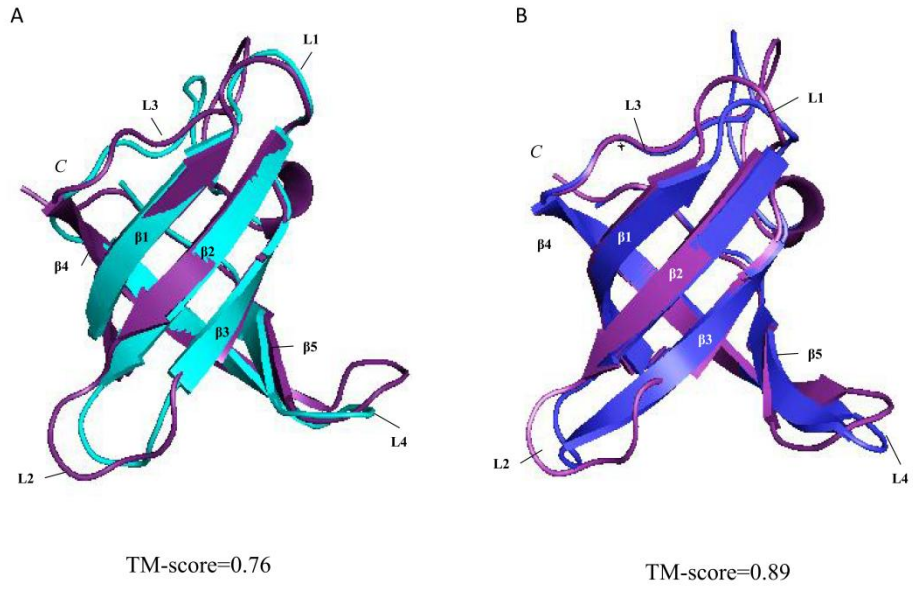


Fig. 12. Alineamiento de la estructura de las secuencias ancestrales y la proteína CspB de *B. subtilis*. En el panel A se puede observar el alineamiento estructural de la proteína ancestral Ance_3D8_1 (color cian) y la proteína CspB de *B. subtilis* (2ES2), en color violeta, obtenido con el algoritmo TM-align, el cual presenta un valor de TM-score de 0.76. En el panel B se puede observar el alineamiento estructural de la proteína ancestral Ance_R5_1 (color azul) y la proteína CspB de *B. subtilis* (2ES2), en color violeta, obtenido con el algoritmo TM-align, el cual presenta un valor de TM-score de 0.89. La C señala el extremo carboxilo terminal de las proteínas y las β1, β2, β3, β4 y β5 señalan a las 5 hebras β.

Acoplamiento Molecular

De acuerdo a los resultados obtenidos con el programa Autodock vina (Trott & Olson, 2009) las proteínas ancestrales predichas presentan mayor afinidad por secuencias de RNA que por secuencias de DNA de cadena sencilla (ssDNA). La proteína Ance_R5_1 obtuvo una energía libre de -12.1 (kcal/mol) de afinidad (Tabla 6) con el ligando AGGAGAU (Tabla 2) y la proteína Ance_3D8_1 obtuvo una energía libre de -12.2 para el ligando AGGAGAU. De acuerdo con las energía calculadas *in silico*, las proteínas Ance_R5_1 y Ance_3D8_1 son menos afines a la hexatimidina (dT6); (Tabla 2) que las proteínas CspB modernas de *Bacillus caldolyticus*, CspB de *B. subtilis* y a la proteína Lin28a de *M. musculus* (Tabla 6). Así mismo, la proteína Ance_R5_1 tiene una mayor afinidad a la secuencia de ssDNA y a la Let-7d miRNA pre-element (Tabla 2); (-9.4 y -10.1 kcal/mol, respectivamente) que la proteína Ance_3D8_1 (-7.8 y -9.1 kcal/mol, respectivamente) pero una menor afinidad a los ligandos restante (rU6 y AGGAGA); (Tabla 2) que la proteína Ance_3D8_1 (Tabla 6). La secuencia ancestral Ance_R5_1 como las secuencias de las proteínas CspB de *B. caldolyticus* y Lin28a de *M. musculus* presentaron menor afinidad por la hexauridina (rU6); (Tabla 2) a diferencia de la proteína ancestral Ance_3D8_1 y las proteínas CspB de *B. subtilis* y Lin28b de *X. tropicalis* (Tabla 6).

Tabla 6. Afinidad de proteínas y ligandos de acuerdo con la energía de enlace (kcal/mol).

	dT6	ssDNA	rU6	AGGAGAU	Let-7D miRNA pre-element
Ance_R5_1	-6.9	-9.4	-8.2	-12.1	-10.1
Ance_3D8_1	-5.8	-7.8	-11.4	-12.2	-9.1
2ES2	-7.4	-	-6.8	-	-
2HAX	-8.5	-8.6	-8.9	-	-
3PF5	0	-6.5	-10.4	-	-0.9
2LI8	-5.1	-5.1	-	-10.0	-
3TRZ	-8.5	-	-8.2	-11.7	-10.3
4ALP	-4.5	-7.6	-14.0	-	-10.6
1OTC	-	-9.3	-	-	-10.3

La proteína ancestral predicha Ance_3D8_1 interacciona con la hexatimidina sólo con sus residuos Y27, S28 y S54 ubicados al final de la hebra β3 principalmente (Fig. 13); presenta interacciones hidrofóbicas, puentes de Hidrógeno, interacciones π-catión y un puente salino con la hexauridina mediante residuos ubicados en las hebras 1, 2 y 3 (β1, β2 y β3) y en el bucle 3 (L3); (Tabla 9 y Fig. 13). La interacción con el ssDNA se lleva a cabo principalmente mediante puentes de hidrógeno con residuos ubicados en el extremo carboxilo terminal y en la hebra 3 (Tabla 9). Finalmente la interacción con los ligandos de la familia de miRNA Let-7 se producen, para el caso de AGGAGAU, mediante puentes de hidrogeno, apilamientos-π e interacciones catión-π con residuos de las hebras 1, 2 y 3 (β1, β2 y β3) y del bucle 1 (L1), mientras que para el caso de Let-7d miRNA *pre-element* se producen interacciones hidrofóbicas y puentes de hidrógeno, principalmente, con residuos también ubicados en las hebras 1, 2 y 3 (β1, β2 y β3) y del bucle 1 (L1); (Tabla 9 y Fig. 13).

En cuanto a la proteína ancestral predicha Ance_R5_1 las interacciones con la hexatimidina (dT6) se dan por puentes de hidrógeno con los residuos R12 y E38 u un puente salino con el residuo H28, ubicados en el bucle 1 (L1), en el bucle 3 (L3) y en la hebra 3 (β3), respectivamente (Fig. 14). Con la hexauridina (rU6) presenta interacciones hidrofóbicas, puentes de hidrógeno y un puente salino mediante aminoácidos ubicados en las hebras 1 y 2 (β1 y β2) y en los bucles 1, 2 y 4 (L1, L2 y L4); (Tabla 9 y Fig. 14). La interacción con el ssDNA se lleva a cabo mediante interacciones hidrofóbicas, puentes de hidrógeno, principalmente, apilamiento-π y puentes salinos residuos ubicados en los bucles 3 y 4 (L3 y L4) y la hebra 5 (β5);(Tabla 9). Finalmente la interacción con los ligandos de la familia de miRNA Let-7 se producen, para el caso de AGGAGAU, mediante interacciones hidrofóbicas, puentes de hidrogeno, apilamientos-π y puentes salinos con residuos de las hebras 1 y 2 (β1 y β2) y del bucle 4 (L4), mientras que para el caso de Let-7d miRNA *pre-element* se producen interacciones hidrofóbicas, puentes de hidrógeno, interacción catión-π y puente salino con residuos ubicados en la hebra 1 (β1) y en los bucles 1, 3 y 4 (L1, L3 y L4); (Tabla 9 y Fig. 14)

Tabla 9. Residuos de interacción de las proteínas ancestrales Ance_3D8_1 y Ance_R5_1.

Ance_3D8_1					
	Ligando				
Enlace no polar	**dT6**	**rU6**	**ssDNA**	**AGGAGAU**	**Let-7d miRNA *pre-element***
Interacciones hidrofóbicas		W6,F13,F24	F24		W6,N8,K11,F13,F24
Puente de hidrógeno	Y27,S28, S54	K11,H26,S28*, N29	V25,S28,T46, K49,G50,S54	K5*,F7,N8,D10, K11,D22,H26, Y27,S28*	W6,F7,N8*,K11+,G12, S28*,G50
Apilamiento-π				F24,H26*,H26	H26
Interacción catión-π		H26	F24	K11*,F15*,H26	
Puente salino		H26	H26*		H26

Ance_R5_1					
	Ligando				
Enlace no polar	**dT6**	**rU6**	**ssDNA**	**AGGAGAU**	**Let-7d miRNA *pre-element***
Interacciones hidrofóbicas		R12,F14,F16, F26	I32*,T48,P57	F14,F16,F26	R8,K55
Puente de hidrógeno	R12,E38	R12,D24,P54*, G56*	T19,H28,S30*, N31^,I32,D33*, G34,E38, S49*,G51,S52, R58,N61,V62	G15,E21,M25, N31,S52,G53, R58	R8^,I9,T10,D11,R12, G13,S30,E38,S39, K55,R58
Apilamiento-π				F26	
Interacción catión-π			H28		R12
Puente salino	H28	K55*	E21,R58,R63	R58,E60	K55

* Dos enlaces.
+ Tres enlaces.
^ Cuatro enlaces.

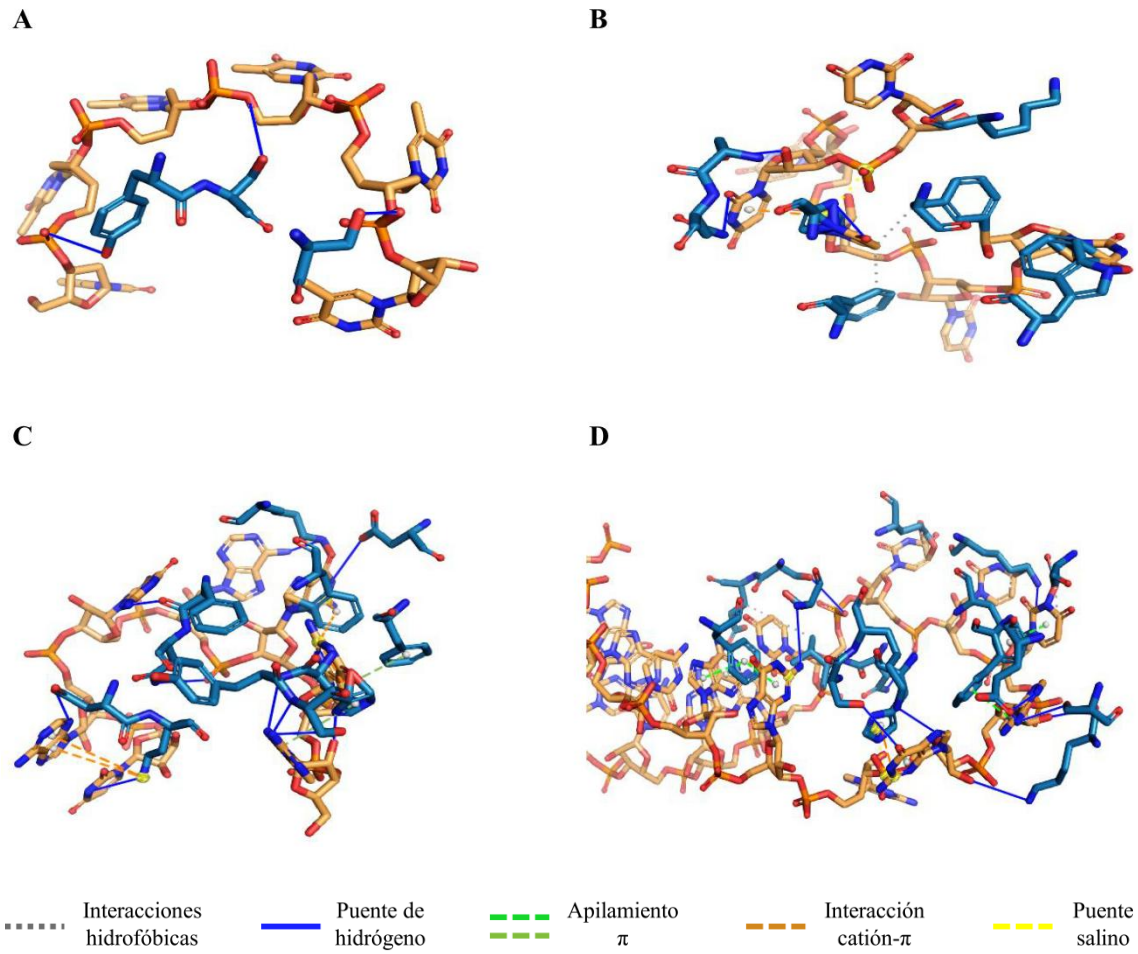


Fig. 13. Interacciones de la proteína ancestral Ance_3D8_1.
Los residuos de la proteína ancestral predicha Ance_3D8_1 que interaccionan con los ligandos se encuentran coloreados en azul y los diferentes ligandos se pueden observar en color naranja. En el panel A se pueden observar los residuos Y27, S28 y S54 que interaccionan con el ligando dT6. En el panel B se pueden observar los residuos F13, F24, H26 y S28 que interaccionan con el ligando rU6. En el panel C se pueden observar los residuos D10, K11, F15, D22, F24, H26, entre otros, que interaccionan con el ligando AGGAGAU. En el panel C se pueden observar los residuos K11, G12, H26, entre otros, que interaccionan con el ligando Let-7d miRNA *pre-element*.

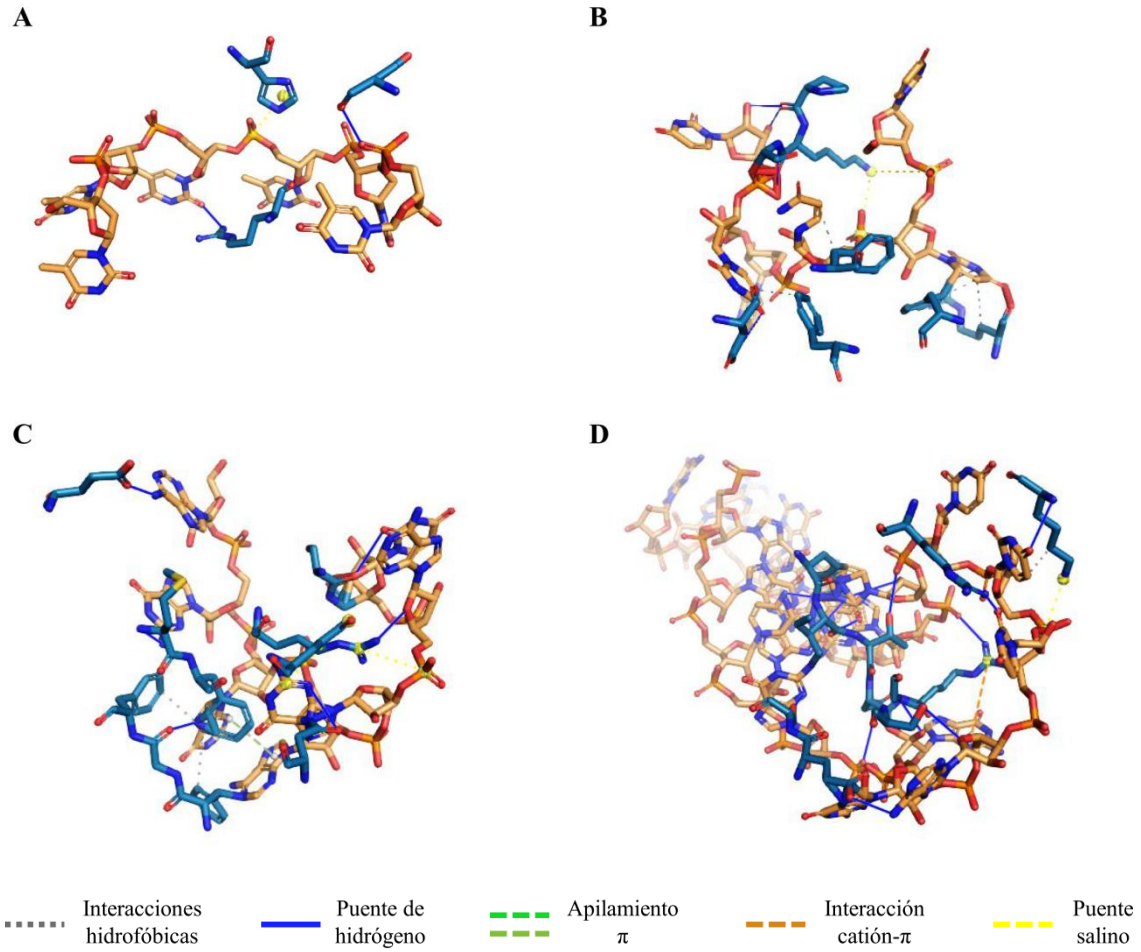


Fig. 14. Interacciones de la proteína ancestral Ance_R5_1.
Los residuos de la proteína ancestral predicha Ance_R5_1 que interaccionan con los ligandos se encuentran coloreados en azul y los diferentes ligandos se pueden observar en color naranja. En el panel A se pueden observar los residuos R12, H28 y E38 que interaccionan con el ligando dT6. En el panel B se pueden observar los residuos R12, F14, F16, F26, P54, K55, ente otros, que interaccionan con el ligando rU6. En el panel C se pueden observar los residuos F14, F16, E21, F26, R58, entre otros, que interaccionan con el ligando AGGAGAU. En el panel C se pueden observar los residuos R8, R12, K55, entre otros, que interaccionan con el ligando Let-7d miRNA *pre-element*.

Predicción del efecto de las mutaciones de las proteínas ancestrales

De acuerdo con las inferencias del programa SNAP2, las mutaciones puntuales tendrían un mayor efecto en las zonas que corresponden a los motivos de unión a RNA, tanto de las secuencias ancestrales predichas como en las secuencias de las proteínas actuales (Fig.15). De acuerdo con las estimaciones teóricas del programa empleado y mostradas en la figura 15 A podemos ver que la proteína ancestral Ance_R5_1 posee una mayor cantidad de sitios que soportan mejor un mayor número de mutaciones sin mucho efecto, a diferencia de la proteína ancestral Ance_3D8_1 (fig.15 B) y de la proteína CspA (Fig.15 C), CspB, CspC y CspD de *Escherichia coli*. Ello, a pesar de que cuentan con mayor número de mutaciones puntuales con poco efecto. Esto mismo pasa al comparar la proteína ancestral Ance_R5_1 contra el dominio CSD de las proteínas de eucariontes Gly-rich, Lin28 y Y-box, siendo sólo la proteína Lin28 la que presenta regiones más marcadas donde las mutaciones tienen poco efecto. Sin embargo, en promedio las proteínas Ance_R5_1, CspA, CspB, CspC y CspD de *E. coli*, el dominio CSD de *Glycin-rich* de *Cicer arietinum*, de Lin28 de *Xenopus tropicalis* y de Y-box de *Homo sapiens* obtienen puntajes asignados por SNAP2 en el rango de -50 a 50, Ance_R5_1=41, CspA=45, CspB=45, CspC=41, CspD=44, Gly-rich=43, Lin28=43 y Y-box=45, estos puntajes están indicando que en promedio las mutaciones tendrían poco efecto sobre estas proteínas y el dominio CSD de las proteínas de eucariontes, a diferencia del puntaje promedio obtenido para la proteína Ance_3D8_1=51, que estaría en el rango donde se estima que las mutaciones causarían un efecto mayor sobre la proteína.

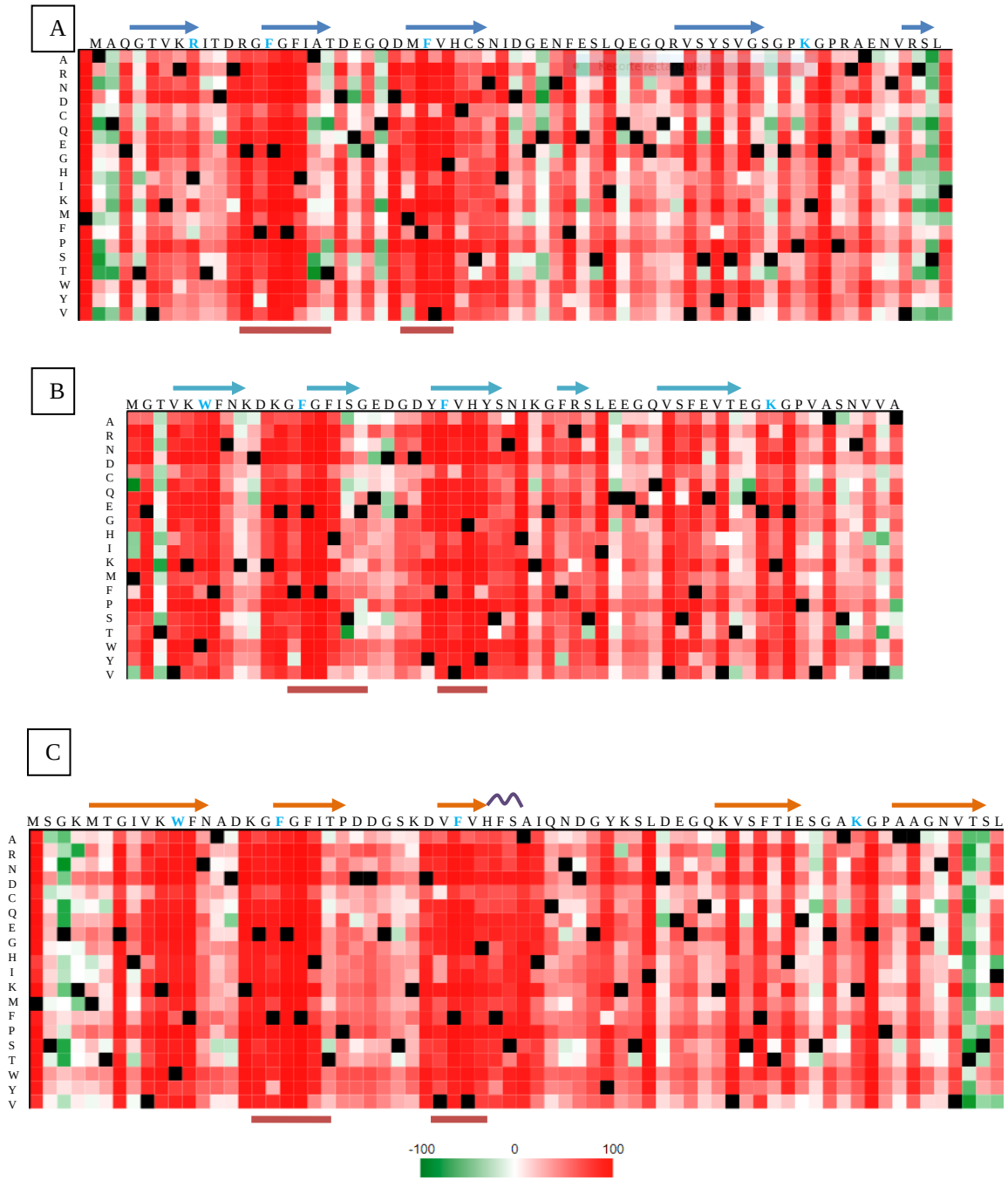


Fig. 15. Mapa de calor para los efectos teóricos de las mutaciones puntuales sobre la funcionalidad de las proteínas Csp.

La predicción se realizó con el software SNAP2 según se describe en métodos. Se muestra la sustitución de cada uno de los 20 aminoácidos de forma independiente para cada posición de la proteína. El color rojo oscuro indica un puntaje alto (puntaje > 50, mutaciones con mayor efecto sobre la función), blanco indica efectos débiles (puntajes de -50 a 50), el color verde un puntaje bajo (puntaje <-50, mutación neutral/sin efecto) y el color negro representa un cambio por el mismo aminoácido. En el panel A de la figura podemos ver la representación del efecto de las mutaciones por cada uno de los 20 aminoácidos en cada posición de la secuencia de aminoácidos de la proteína ancestral predicha Ance_R5_1. En el panel B de la figura podemos ver la representación del efecto de las mutaciones por cada uno de los 20 aminoácidos en cada posición de la secuencia de aminoácidos de la proteína ancestral predicha Ance_3D8_1. En el panel C de la figura podemos ver la representación del efecto de las mutaciones por cada uno de los 20 aminoácidos en cada posición de la secuencia de aminoácidos para la proteína CspA de *E. coli*. Las flechas en la parte superior de la secuencia de aminoácidos representan a las hebras β de las proteínas. Las barras en color rojo a bajo de cada mapa de color señalan los motivos RNP1 y RNP2, mientras que los aminoácidos en color cian en la secuencia indican los aminoácidos de unión a DNA.

Discusión

Los resultados que arroja la reconstrucción filogenética sugieren que el *Phylum Fusobacteria* tiene a las representantes más antiguas de este dominio proteico. Sin embargo, estas no se encuentran bien estudiadas, a diferencia de las proteínas que se encuentran en los organismos como *E. coli*, *B. subtilis* y *Y. pestis*, por lo que sólo se puede suponer que, posiblemente, juegan un papel importante en la adaptación de estos organismos ante un cambio en su ambiente. Lo que si se puede decir es que aunque la secuencia varíe, hasta en un 40% de similitud respecto a las versiones más modernas de las proteínas Csp, los motivos de unión a RNA y de unión a ssDNA se mantienen y conservan las funciones de regulación de la condensación del DNA, la transcripción y de la traducción y, por inferencia, se propone que las proteínas ancestrales, ya poseían dichas funciones. Los dendogramas reconstruidos con máxima verosimilitud, a partir de las 655 secuencias alineadas con el programa MAFFT y el reconstruido a partir del alineamiento con las 55 secuencias representativas proporcionadas por el programa PROMALS están bien respaldados al presentar un soporte de ramas, en su mayoría y sobre todo en los nodos basales, igual o mayor al 0.5, aunque en algunos nodos no fue posible discriminar entre las topologías alternativas.

Actualmente sigue siendo difícil determinar si los caracteres faltantes en las secuencias ancestrales inferidas, en comparación a las secuencias actuales, son realmente estados de carácter ancestrales, ya que esto podría implicar que las secuencias ancestrales solo contaban con ciertos elementos necesarios para llevar a cabo su función y que, posteriormente, sufrieron eventos de inserción o duplicación (Pupko *et al*., 2000; Liberles, 2007; Ashkenazy *et al*., 2012).

Las secuencias ancestrales reconstruidas Ance_3D8_1, 2 y 3 sólo presentan reducciones en la secuencia, en comparación con las secuencias de las proteínas modernas con menor número de residuos, en los extremos amino y carboxilo, así como cerca del motivo RNP2, dos residuos hacia el extremo amino y 2 residuos hacia el extremo carboxilo y un residuo hacia el extremo amino de su K49. Además de la falta

de estos residuos, las proteínas Ance_3D8_1 y 2 presentan una tirosina Y23 en vez de la valina (V23) presente en la secuencias de la proteína ancestral Ance_3D8_3 y en el caso específico de la secuencia de la proteína ancestral Ance_3D8_1 se presenta una tirosina en la posición 27 (Y27) la cual interacciona con los ligandos hexatimidina y AGGAGAU. Las secuencias de las proteínas ancestrales Ance_R5_1, 2 y 3 poseen la misma extensión que la proteína CspB de *Bacillus caldolyticus* y sólo la secuencia de la proteína ancestral Ance_R5_1 presenta, a diferencias de las otras proteínas ancestrales predichas y oras proteínas Csp, una arginina (R8), una isoleucina (I9) y una treonina (T10) en vez de del triptófano, fenilalanina y asparagina presente en la mayoría de las secuencias de proteínas CSP, a pesar de dicho cambio los residuos R8, I9 y T10 pueden interaccionar con el ligando llamado Let-7d miRNA.

En cuanto a la función de las proteínas ancestrales predichas se piensa que estas tenían que ser generalistas al tener la capacidad de reconocer y unirse a una amplia gama de ligandos como en el caso de algunas enzimas actuales que llevan a cabo actividades secundarias, biológicamente relevantes de "pluriempleo" (Copley, 2012, 2015; Jeffery, 1999) y que las proteínas modernas especializadas aparecieron después de la duplicación génica que les dio origen, la distribución y el perfeccionamiento de sus antiguas actividades (James & Tawfik, 2003; O'Brien & Herschlag, 1999; Tawfik, 2010) presentes en las proteínas ancestrales. Dicha hipótesis es ampliamente aceptada ya que se cree que la evolución de una nueva interacción específica o actividad bioquímica puede requerir muchos cambios genéticos y que la pérdida de actividades bioquímicas podría ser genéticamente más simple que la evolución de otras nuevas, ya que es poco probable que se crucen caminos evolutivos tan complicados después de una duplicación genética antes de que las mutaciones perjudiciales eliminen la copia adicional (James & Tawfik, 2003; O'Brien & Herschlag, 1999).

Sin embargo, varios estudios han documentado familias de proteínas en las que las funciones biológicas específicas de las proteínas actuales evolucionaron *de novo* a partir de las proteínas ancestrales que carecían de esas funciones y revelan los mecanismos por los cuales evolucionaron esas nuevas funciones (Anderson *et al*., 2016;

McKeown *et al*., 2014). Mostrando así, que las nuevas actividades específicas a menudo pueden evolucionar con solo una o algunas sustituciones. Esto nos lleva a pensar que es más probable que las proteínas estén continuamente ganando y perdiendo actividades bioquímicas secundarias durante su historia evolutiva debido a los procesos estocásticos de mutación y deriva y que la supuesta dificultad de desarrollar nuevas actividades bioquímicas no es, pues, una explicación plausible de la preponderancia de historias en las que el antepasado ya tenía las funciones de sus descendientes (Siddiq, Hochberg, & Thornton, 2017).

Los resultados del acoplamiento molecular nos indican que la proteína ancestral Ance_3D8_1 posee una mayor afinidad por moléculas de RNA que por secuencias de ssDNA, aunque también pueden unirse a moléculas de ssDNA con una afinidad parecida a la de las proteínas actuales. La proteína ancestral Ance_R5_1 presenta mayor afinidad por RNA de secuencia no canónica y secuencias de ssDNA, pero la diferencia no es mucha. Sin embargo, a partir del acoplamiento molecular no podemos determinar si las proteínas ancestrales inferidas en este trabajo, puedan llevar a cabo alguna función que ayude a los organismos en la adaptación de algún cambio drástico en el ambiente, sólo se puede decir que consiguen unirse a secuencias, tanto de RNA como de ssDNA, por lo que podrían tener funciones durante el desarrollo normal de los organismos y durante un cambio drástico en el ambiente; aunque si se toma en cuenta que tienen mayor afinidad por el RNA se puede suponer que su función primordial es como chaperonas de mRNAs. Para poder conocer mejor su actividad en la célula, se tiene que considerar la temperatura, la concentración osmótica y la accesibilidad al solvente, ya que por ejemplo, las moléculas de agua que se encuentran cercanas a las proteínas Csp ayudan en la formación de puentes de hidrógeno entre la proteína y el ligando (Max *et al*., 2007), así como también lo hacen los iones de Na^+ y Mg.

La secuencia y estructura de las proteínas CSP ancestrales aquí inferidas se encuentra muy conservada en comparación con sus descendientes, es decir, las proteínas actuales. Por lo anterior, se puede asumir que la fuerte presión de selección que en ellas recae es proporcional al nivel de conservación que manifiestan (Kimura, 1983).

Ciertamente, la proteína no manifiesta el mismo nivel de conservación a lo largo de su secuencia, pues según los resultados obtenidos con SNAP2 (Figura 15), algunos sitios son claramente más susceptibles a la mutación, así que podrían corresponder a aquellos que no son funcional o estructuralmente relevantes (Orengo & Thornton, 2005). Por otro lado, se puede postular que los cambios observados son parte del proceso evolutivo de las proteínas CSP, así que las mutaciones ocurren constantemente (y por eso observamos distintos niveles de conservación en las distintas CSP analizadas), independientemente de que sean adaptativas o no (Tóth-Petróczy & Tawfik, 2014). Ya que, la adaptación depende de una o algunas mutaciones que proporcionan una ventaja selectiva a través de una función nueva o mejorada (Maynard Smith, 1970), y la función evolutiva se define como el parámetro que mide sus contribuciones a la aptitud del organismo, las diferencias más marcadas en cuanto a los niveles de conservación de las proteínas CSP se nota sólo cuando se comparan proteínas CSP que han evolucionado en distintos linajes. Esto, explica porque los organismos eucariontes poseen las CSP con mayor divergencia funcional en comparación a las versiones ancestrales de las mismas proteínas, presumiendo que, los cambios observados en las CSP de eucariontes obedecen a las presiones de selección impuestas por el linaje.

Probando y prediciendo la divergencia funcional después de la duplicación del gen, se puede observar la asociación entre la secuencia, la divergencia de la función y la estructura, apoyando la noción de divergencia funcional (Liberles, 2007). A la capacidad de desarrollar *flexibilidad adaptativa* como si fuese un carácter heredable, es decir, que se transmite de generación en generación como si fuese un carácter benéfico independiente, se le ha denominado evolucionabilidad (del inglés: *evolvability*). La evolucionabilidad posee dos componentes cruciales (Kirschner, Gerhart, Otey, & Arnold, 1998; Wagner, 2005): robustez e innovación. La robustez se relaciona con la propiedad para asimilar mutaciones sufriendo un efecto mínimo sobre la función y la estructura (Kirschner *et al*., 1998; Wagner, 2005). Esta característica, junto con la innovación, son críticas para la supervivencia a largo plazo de las proteínas (Wagner, 2005). Sin embargo, mientras más robusta es una proteína (es decir, la mayoría de las mutaciones no tienen, o tienen efectos inconsecuentes) menos innovadora es (Tóth-Petróczy & Tawfik, 2014).

La estructura del dominio CSD presenta entre 40-45% de elementos estructurales secundarios, principalmente hebras *β*, dándoles un grado mayor de orden estructural. Un mayor grado de orden estructural y empaquetamiento confiere tolerancia a las mutaciones, promoviendo la capacidad de evolución de las proteínas, así como indica una mayor estabilidad (DePristo *et al*., 2005; Shakhnovich, Deeds, Delisi, & Shakhnovich, 2005; Tokuriki & Tawfik, 2009), es decir, una proteína altamente ordenada y bien empaquetada proporciona un umbral de estabilidad más alto y permite que se acumulen mutaciones menos conservativas sin causar un gran efecto (Rorick & Wagner, 2011; Shakhnovich *et al*., 2005) pero sólo si la estabilidad es un parámetro global aditivo mediante el cual, las mutaciones estabilizadoras en una región (por ejemplo, en los residuos que mantienen la estructura de la proteína) compensan fácilmente los efectos desestabilizadores de las mutaciones en otros lugares (por ejemplo, en la región del sitio activo), por lo cual, este tipo de robustez es transitorio, una vez que se cruza el límite de mutaciones estabilizadoras (por lo general dentro de pocas mutaciones), las siguientes mutaciones provocan una gran disminución en la funcionalidad de la proteína (Bershtein *et al*., 2006). Por otro lado, el aumento de la estabilidad reduce la plasticidad conformacional y la adquisición de nuevas funciones, a menudo, depende de la plasticidad conformacional, sobre todo de los sitios catalíticos (Tóth-Petróczy & Tawfik, 2014).

La predicción de los efectos de las mutaciones realizados con el programa SNAP2 (Figura 15) sobre las proteínas ancestrales predichas (Ance_3D8_1, 2 y 3 y Ance_R5_1, 2 y 3) muestran que, en general, los motivos de unión a RNA y los residuos de unión a DNA toleran muy poco el cambio y que en las regiones de los bucles y hacia el extremo carboxilo terminal se pueden producir mutaciones neutrales o con poco efecto sobre la proteína. En general las proteínas CSP, el dominio CSD y las proteínas ancestrales predichas podrían sufrir ciertas mutaciones sin que causaran un gran efecto en la función y la estructura, con excepción de las proteínas ancestrales predicas Ance_3D8_1, 2 y 3, sin embargo los resultados no son concluyentes. El grado de conservación del dominio CSD podría explicarse debido a que la correlación entre los residuos, o epistasis intramolecular, ralentiza drásticamente la tasa de evolución ya que, el cambio en un residuo depende por completo de un cambio en otro, u otros

residuos (Tóth-Petróczy & Tawfik, 2014). Sin embargo, es igualmente factible que selección purificadora (selección que mantiene en un valor promedio la divergencia de forma y función, eliminado valores extremos) esté actuando sobre el dominio CSD, así los pocos cambios observados en secuencia y estructura sean debido a la baja tolerancia del dominio CSD a las sustituciones no neutras, lo que apoya y explica la presencia de ramas y distancias evolutivas muy cortas observadas en el dendograma de las secuencias del dominio CSD (Figuras 8 y 9).

Los resultados obtenidos también se pueden tomar como evidencia que soportan las hipótesis previamente planteadas, en las que se ha evidenciado la versatilidad del pliegue OB-fold, pues, aunque se encuentra principalmente en proteínas que interactúan con ssDNA o RNA, también es parte de la arquitectura de proteínas con funciones variadas, como de pirofosfatasas e incluso de moléculas citotóxicas, como la enterotoxina B de *Staphylococcus* (Arcus, 2002). La parte más conservada del OB-fold, que consiste en el arreglo de 5 hojas β, contiene el segmento catalítico (o cara) que puede interactuar con distintos ligandos (primordialmente ácidos nucleicos y ribonucleicos). En contraste, la parte menos conservada del OB-fold, que consiste en los bucles que conectan el arreglo de hojas β, así como una α-hélice en la parte superior del barril de hojas β, es parte primordial para la función de proteínas con funciones no relacionadas con DNA o RNA. Esta dualidad, en la que el OB-fold consiste de segmentos muy conservados y de otros que pueden variar, muy seguramente es la característica que promueve la versatilidad de este plegamiento y con ello su preponderancia, en un conjunto grande de proteínas con diversas funciones.

Las proteínas CSP de bacteria sólo poseen el dominio CSD (son generalmente un monómero y en algunos casos un homodímero), que está sumamente conservado a nivel de estructura; sin embargo, las proteínas CSP de eucariontes muestran una arquitectura multidominio (son poliheterómeros), lo que es perfectamente congruente con el mecanismo molecular por el cual evolucionan las proteínas, en donde el reciclaje y adopción de distintos dominios en una misma proteína (arquitectura) es el mecanismo primordial por el cual surgen nuevas variantes y funciones de proteínas ya existentes.

Siguiendo la misma idea, un caso interesante es la evolución molecular del dominio CSD en el linaje de las Arqueas. En este grupo de organismos, las proteínas CSP y el dominio CSD poseen una distribución en mosaico (refiriéndose a que se encuentran presentes en ciertos grupos y ausentes en otros). Dicha distribución, aparentemente azarosa, podría obedecer a circunstancias distintas. Una propuesta es que la mayoría de las Arqueas perdieron de forma secundaria la o las secuencias de las proteínas CSP. Esto implica que el ancestro de las Arqueas ya poseía la secuencia de la proteína CSP (al menos una copia), así que todos sus descendientes tenían la proteína; sin embargo, con el paso del tiempo, en algunos grupos hubo pérdida de la misma. Aunque las razones pudieron ser varias, aquí se sugirieren dos de las más probables: I) La pérdida de las secuencias de la proteína CSP se debió a que las mismas no representaban beneficio sobre la aptitud de los organismos portadores. Dicha hipótesis está fuertemente apoyada en el hallazgo de que no se ha podido detectar secuencias de proteína CSP en Arqueas termófilas e hipertermófilas (Giaquinto *et al*., 2007). Dado que las CSP responden y se expresan en situaciones de estrés por disminución en la temperatura, la pérdida de estos genes por organismos que viven en ambientes con altas temperaturas podría estar bien justificado; en cuyo caso, la pérdida obedeció a presiones de selección. II) Otra alternativa, es que la pérdida pudo deberse a fenómenos estocásticos, como la deriva genética. Si bien la deriva genética pudo ser un factor, es poco probable, pues para que sus efectos produjeran la distribución filogenética sesgada que se observa de la proteína CSP en el genoma de múltiples Arqueas, la deriva, un fenómeno azaroso, debería haber ocurrido más veces de lo que se esperaría para producir el patrón de distribución observado. Si esto sucede, la deriva genética se transforma en un fenómeno de selección estadísticamente significativo que ya no es azaroso, algo que contradice la misma definición de deriva genética y, por esta razón, no se considerara como la opción más viable.

Una alternativa más e igualmente probable implica que la distribución en mosaico que observamos para las secuencias de la proteína CSP en Arqueas, se debe a la transferencia horizontal de genes entre Arqueas y Bacterias (Fuchsman *et al*., 2017). No obstante, el nivel de conservación de las secuencias de proteína CSP es tan alto que la información disponible para hacer la predicción de la transferencia horizontal es

escaza. Además, se debe considerar que, aún con los métodos disponibles más sofisticados, no es posible distinguir entre el donante y el receptor en un evento de transferencia horizontal, por lo que la transferencia no puede ser sustentada en evidencia cuantificable, al menos hasta que se desarrollen nuevos métodos más sensibles de identificación de dicho fenómeno.

Aunque no ha sido posible delimitar los procesos que han producido la distribución irregular de proteínas CSP en el linaje de las Arqueas, este problema está lejos de ser una limitante para estos organismos, pues algunos estudios han reportado el hallazgo de unas proteínas llamadas TRAM (Campanaro *et al*., 2011; Chen *et al*., 2012) que llevan a cabo las mismas funciones que las proteínas CSP de Bacterias, favoreciendo la adaptación a bajas temperaturas al evitar la sub-regulación de genes de mantenimiento mediante diversos mecanismos que requieren de la interacción con ácidos nucleicos (Taha *et al*., 2016; Zhang *et al*., 2017). Las proteínas TRAM se encuentran ampliamente distribuidas en el linaje de las Arqueas y, aunque a nivel de secuencia, no poseen similitud perceptible con las proteínas CSP de Bacteria, mantienen las características estructurales distintivas de las proteínas CSP y del plegamiento OB-fold, por lo que es factible pensar en un posible origen común.

Debido a lo anterior, se podría decir que la capacidad de sobrevivir al estrés causado por la disminución de la temperatura fue asegurada manteniendo secuencias de proteínas que podían realizar la misma función, así pues, es posible que las proteínas TRAM sean el resultado de un evento antiguo de duplicación y diversificación ocurrido en el linaje de las Arqueas, para, subsecuentemente, perder la secuencia que le dio origen, dando como resultado un remplazo de secuencias ortólogas. Sin embargo, no podemos descartar que alguna arquea haya simplemente perdido la secuencia para subsecuentemente reintegrarla a su genoma como resultado de una transferencia horizontal. Para resolver tal interrogante es necesario analizar a profundidad las características distintivas de la secuencia y estructura de todas las proteínas TRAM y CSP disponibles de Archaea y Bacteria y, aun así, se podrían seguir apoyando diferencialmente todas las hipótesis, por lo que es más probable que la historia de las

CSP en Arqueas sea el resultado conjunto de una compleja historia de pérdidas, duplicaciones (que generaron las proteínas TRAM) y ganancias (mediante transferencia horizontal) cuya influencia no podrá ser individualmente delimitada, pues los tres fenómenos, en conjunto, han moldeado la evolución molecular de las proteínas TRAM y CSP que observamos hoy.

Conclusión

La reconstrucción de secuencias ancestrales, así como la predicción de la estructura de tales secuencias ayudó a reconocer que el pliegue que define al dominio CSD es posiblemente antiguo, ya que presenta una amplia distribución filogenética y se ha conservado con escasa variación en los distintos linajes, incluyendo los presentes en proteínas de organismos eucariontes. En aquellos grupos de organismos en los que la proteína se ha mantenido con muy poca variación, sus propiedades catalíticas tampoco han cambiado, así que en todos los casos depende de la interacción con ácidos nucleicos, ya sea con DNA, DNA de cadena sencilla y/o RNA para realizar siempre la misma función, de mantener activa la regulación de genes de mantenimiento fundamentales. Por lo anterior, se puede decir que las proteínas cuyo plegamiento correspondía con el OB-fold; proteínas que, de hecho, pudieron fungir como ancestros de las CSPs, ya estaban presentes en el último ancestro celular de todos los organismos actuales y, que las proteínas CSP y su componente funcional, el dominio CSD, están bajo una fuerte presión de selección, por lo que, su secuencia, estructura y función se han mantenido intactas en múltiples linajes.

Debido a su nivel de conservación las proteínas CSP y su dominio CSD podrían ser útiles para formar una base de datos de prueba que sirva de control. Cuando algoritmos nuevos aparecen (enfocándonos en los de aplicación bioinformática), necesitan probar mediante análisis de práctica o ensayo sus capacidades y fortalezas. Con frecuencia, los conjuntos de datos para hacer dichas pruebas son complicados y difíciles de analizar (con el fin de enfatizar las virtudes de los nuevos algoritmos), sin embargo, un conjunto de datos real (puesto que muchos son simulados) y sencillo, cuya solución correcta pueda conocerse fácilmente, resultaría igualmente útil. En este sentido, un conjunto de datos que consista de secuencias ortólogas de CSP puede ser empleado como control en estudios de, por ejemplo, variabilidad de la tasa de sustitución en proteínas conservadas o el efecto cuantitativo de la selección natural en proteínas ancestrales.

Referencias.

Abascal, F., Zardoya, R., & Posada, D. (2005). ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*, *21*(9), 2104–2105.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–402.

Anderson, D. P., Whitney, D. S., Hanson-Smith, V., Woznica, A., Campodonico-Burnett, W., Volkman, B. F., King, N., Thornton, J. W., & Prehoda, K. E. (2016). Evolution of an ancient protein function involved in organized multicellularity in animals. *ELife*, *5*, e10147.

Apic, G., Gough, J., & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of Molecular Biology*, *310*(2), 311–325.

Arcus, V. (2002). OB-fold domains: A snapshot of the evolution of sequence, structure and function. *Current Opinion in Structural Biology*. Elsevier Current Trends.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., & Pupko, T. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, *40*(Web Server issue), W580-4.

Bae, W., Jones, P. G., & Inouye, M. (1997). CspA, the major cold shock protein of *Escherichia coli*, negatively regulates its own gene expression. *Journal of Bacteriology*, *179*(22), 7081–8.

Bae, W., Xia, B., Inouye, M., & Severinov, K. (2000). *Escherichia coli* CspA-family RNA chaperones are transcription antiterminators. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(14), 7784–9.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., & Tawfik, D. S. (2006). Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, *444*(7121), 929–932.

Bloom, J. D., Labthavikul, S. T., Otey, C. R., & Arnold, F. H. (2006). Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(15), 5869–74.

Brandi, A., Pietroni, P., Gualerzi, C. O., & Pon, C. L. (1996). Post-transcriptional regulation of CspA expression in *Escherichia coli*. *Molecular Microbiology*, *19*(2), 231–40.

Brandi, A., Spurio, R., Gualerzi, C. O., & Pon, C. L. (1999). Massive presence of the *Escherichia coli* "major cold-shock protein" CspA under non-stress conditions. *The EMBO Journal*, *18*(6), 1653–9.

Brown, E. C., & Jackson, R. J. (2004). All five cold-shock domains of unr (upstream of N-ras) are required for stimulation of human rhinovirus RNA translation. *The Journal of General Virology*, *85*(Pt 8), 2279–87.

Büssing, I., Slack, F. J., & Großhans, H. (2008). Let-7 microRNAs in development, stem cells and cancer. *Trends in Molecular Medicine*, *14*(9), 400–409.

Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M., & Murzin, A. G. (1997). The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell*, *88*(2), 235–42.

Campanaro, S., Williams, T. J., Burg, D. W., De Francisci, D., Treu, L., Lauro, F. M., & Cavicchioli, R. (2011). Temperature-dependent global gene expression in the antarctic archaeon *Methanococcoides burtonii*. *Environmental Microbiology*, 13(8), 2018–2038.

Chaikam, V., & Karlson, D. T. (2010). Comparison of structure, function and regulation of plant cold shock domain proteins to bacterial and animal cold shock domain proteins. *BMB Reports*, *43*(1), 1–8.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*(1), 12–21.

Chen, Z., Yu, H., Li, L., Hu, S., & Dong, X. (2012). The genome and transcriptome of a newly described psychrophilic archaeon, *Methanolobus psychrophilus* R15, reveal its cold adaptive characteristics. *Environmental Microbiology Reports*, 4(6), 633–641.

Chothia, C., & Gough, J. (2009). Genomic and structural aspects of protein evolution. *The Biochemical Journal*, *419*(1), 15–28.

Chothia, C., Gough, J., Vogel, C., & Teichmann, S. A. (2003). Evolution of the protein repertoire. *Science (New York, N.Y.)*, *300*(5626), 1701–3.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, *357*(6379), 543–544.

Coles, L. S., Diamond, P., Occhiodoro, F., Vadas, M. A., & Shannon, M. F. (1996). Cold shock domain proteins repress transcription from the GM-CSF promoter. *Nucleic Acids Research*, *24*(12), 2311–7.

Copley, S. D. (2012). Moonlighting is mainstream: Paradigm adjustment required. *BioEssays*, *34*(7), 578–588.

Copley, S. D. (2015). An evolutionary biochemist's perspective on promiscuity. *Trends in Biochemical Sciences*. Elsevier Ltd.

Dayhoff, M. O. (1978). *Atlas of protein sequence and structure*. Washington, D.C: National Biomedical Research Foundation.

DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews Genetics*, *6*(9), 678–687.

Doniger, J., Landsman, D., Gonda, M. A., & Wistow, G. (1992). The product of unr, the highly conserved gene upstream of N-ras, contains multiple repeats similar to the cold-shock domain (CSD), a putative DNA-binding motif. *The New Biologist*, *4*(4), 389–95.

Draper, D. E. (1999). Themes in RNA-protein recognition. *Journal of Molecular Biology*, *293*(2), 255–270.

Du, K.-L., & Swamy, M. N. S. (2016). *Search and Optimization by Metaheuristics*.Suiza: Springer International Publishing.

Duval, M., Simonetti, A., Caldelari, I., & Marzi, S. (2015). Multiple ways to regulate translation initiation in bacteria: Mechanisms, regulatory circuits, dynamics. *Biochimie*, *114*, 18–29.

Etchegaray, J. P., & Inouye, M. (1999). DB or not DB in translation? *Molecular Microbiology*, *33*(2), 438–439.

Etchegaray, J. P., Jones, P. G., & Inouye, M. (1996). Differential thermoregulation of two highly homologous cold-shock genes, cspA and cspB, of *Escherichia coli*. *Genes to Cells : Devoted to Molecular & Cellular Mechanisms*, *1*(2), 171–8.

Evans, J. R., Mitchell, S. A., Spriggs, K. A., Ostrowski, J., Bomsztyk, K., Ostarek, D., & Willis, A. E. (2003). Members of the poly (rC) binding protein family stimulate the activity of the c-myc internal ribosome entry segment in vitro and in vivo. *Oncogene*, *22*(39), 8012–20.

Evdokimova, V. M., Kovrigina, E. A., Nashchekin, D. V, Davydova, E. K., Hershey, J. W., & Ovchinnikov, L. P. (1998). The major core protein of messenger ribonucleoprotein particles (p50) promotes initiation of protein biosynthesis in vitro. *The Journal of Biological Chemistry*, *273*(6), 3574–81.

Eyre-Walker, A. (1998). Problems with parsimony in sequences of biased base composition. *Journal of Molecular Evolution*, *47*(6), 686–690.

Fang, L., Hou, Y., & Inouye, M. (1998). Role of the cold-box region in the 5' untranslated region of the cspA mRNA in its transient expression at low temperature in *Escherichia coli*. *Journal of Bacteriology*, *180*(1), 90–5.

Fang, L., Jiang, W., Bae, W., & Inouye, M. (1997). Promoter-independent cold-shock induction of cspA and its derepression at 37 degrees C by mRNA stabilization. *Molecular Microbiology*, *23*(2), 355–64.

Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland, Massachusetts: Sinauer Associates.
Ferrada, E., & Wagner, A. (2008). Protein robustness promotes evolutionary innovations on large evolutionary time-scales. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1643), 1595–1602.
Ferrer, N., Garcia-Espana, A., Jeffers, M., & Pellicer, A. (1999). The unr gene: evolutionary considerations and nucleic acid-binding properties of its long isoform product. *DNA and Cell Biology*, *18*(3), 209–18.
Field, S. F., & Matz, M. V. (2010). Retracing evolution of red fluorescence in GFP-like proteins from *Faviina corals*. *Molecular Biology and Evolution*, *27*(2), 225–233.
Fitch, W. M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, *20*(4), 406.
Fuchsman, C. A., Collins, R. E., Rocap, G., & Brazelton, W. J. (2017). Effect of the environment on horizontal gene transfer between bacteria and archaea. *PeerJ*, *5*, e3865.
Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–95.
Gasteiger, J., & Marsili, M. (1980). Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, *36*(22), 3219–3228.
Giaquinto, L., Curmi, P. M. G., Siddiqui, K. S., Poljak, A., DeLong, E., DasSarma, S., & Cavicchioli, R. (2007). Structure and function of cold shock proteins in archaea. *Journal of Bacteriology*, *189*(15), 5738–48.
Giuliodori, A. M., Di Pietro, F., Marzi, S., Masquida, B., Wagner, R., Romby, P., Gualerzi, C. O., & Pon, C. L. (2010). The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Molecular Cell*, *37*(1), 21–33.
Goldenberg, D., Azar, I., & Oppenheim, A. B. (1996). Differential mRNA stability of the cspA gene in the cold-shock response of *Escherichia coli*. *Molecular Microbiology*, *19*(2), 241–8.
Goldenberg, D., Azar, I., Oppenheim, A. B., Brandi, A., Pon, C. L., & Gualerzi, C. O. (1997). Role of *Escherichia coli* cspA promoter sequences and adaptation of translational apparatus in the cold shock response. *Molecular & General Genetics : MGG*, *256*(3), 282–90.
Golding, G. B., & Dean, A. M. (1998). The structural basis of molecular adaptation. *Molecular Biology and Evolution*, *15*(4), 355-69
Goldstein, J., Pollitt, N. S., & Inouye, M. (1990). Major cold shock protein of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(1), 283–7.
Gong, L. I., Suchard, M. A., & Bloom, J. D. (2013). Stability-mediated epistasis constrains the evolution of an influenza protein. *ELife*, *2*, e00631.
Goodchild, A., Raftery, M., Saunders, N. F. W., Guilhaus, M., & Cavicchioli, R. (2004). Biology of the cold adapted archaeon, *Methanococcoides burtonii* determined by proteomics using liquid chromatography-tandem mass spectrometry. *Journal of Proteome Research*, *3*(6), 1164–1176.
Graumann, P. L., & Marahiel, M. A. (1998). A superfamily of proteins that contain the cold-shock domain. *Trends in Biochemical Sciences*, *23*(8), 286–90.
Graumann, P., Schröder, K., Schmid, R., & Marahiel, M. A. (1996). Cold shock stress-induced proteins in *Bacillus subtilis*. *Journal of Bacteriology*, *178*(15), 4611–9.
Graumann, P., Wendrich, T. M., Weber, M. H., Schröder, K., & Marahiel, M. A. (1997). A family of cold shock proteins in *Bacillus subtilis* is essential for cellular growth and for efficient protein synthesis at optimal and low temperatures. *Molecular Microbiology*, *25*(4), 741–56.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–21.

Hankins, J. S., Zappavigna, C., Prud'homme-Généreux, A., & Mackie, G. A. (2007). Role of RNA structure and susceptibility to RNase E in regulation of a cold shock mRNA, cspA mRNA. *Journal of Bacteriology*, *189*(12), 4353–8.

Harms, M. J., & Thornton, J. W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Current Opinion in Structural Biology*, *20*(3), 360–6.

Harms, M. J., & Thornton, J. W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nature Reviews Genetics*, *14*(8), 559–571.

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford ; New York: Oxford University Press.

Hecht, M., Bromberg, Y., & Rost, B. (2015). Better prediction of functional effects for sequence variants. *BMC Genomics*, *16*(8), S1.

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, *4*(4), 275–284.

Horn, G., Hofweber, R., Kremer, W., & Kalbitzer, H. R. (2007). Structure and function of bacterial cold shock proteins. *Cellular and Molecular Life Sciences : CMLS*, *64*(12), 1457–70.

Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., & Chothia, C. (1999). SCOP: a structural classification of proteins database. *Nucleic Acids Research*, *27*(1), 254–6.

Hunt, S. L., Hsuan, J. J., Totty, N., & Jackson, R. J. (1999). unr, a cellular cytoplasmic RNA-binding protein with five cold-shock domains, is required for internal initiation of translation of human rhinovirus RNA. *Genes & Development*, *13*(4), 437–48.

Jacquemin-Sablon, H., Triqueneaux, G., Deschamps, S., le Maire, M., Doniger, J., & Dautry, F. (1994). Nucleic acid binding and intracellular localization of unr, a protein with five cold shock domains. *Nucleic Acids Research*, *22*(13), 2643–50.

James, L. C., & Tawfik, D. S. (2003). Conformational diversity and protein evolution - A 60-year-old hypothesis revisited. *Trends in Biochemical Sciences*, *28*(7), 361-8.

Jeffers, M., Paciucci, R., & Pellicer, A. (1990). Characterization of unr; a gene closely linked to N-ras. *Nucleic Acids Research*, *18*(16), 4891–9. t

Jeffery, C. J. (1999). Moonlighting proteins. *Trends in Biochemical Sciences*, *24*(1), 8-11.

Jermann, T. M., Opitz, J. G., Stackhouse, J., & Benner, S. A. (1995). Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, *374*(6517), 57–59.

Jiang, W., Fang, L., & Inouye, M. (1996). The role of the 5'-end untranslated region of the mRNA for CspA, the major cold-shock protein of *Escherichia coli*, in cold-shock adaptation. *Journal of Bacteriology*, *178*(16), 4919–25.

Jiang, W., Hou, Y., & Inouye, M. (1997). CspA, the major cold-shock protein of *Escherichia coli*, is an RNA chaperone. *The Journal of Biological Chemistry*, *272*(1), 196–202.

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D., & Slack, F. J. (2005). RAS Is Regulated by the let-7 MicroRNA Family. *Cell*, *120*(5), 635–647.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, *292*(2), 195–202.

Jones, P. G., VanBogelen, R. A., & Neidhardt, F. C. (1987). Induction of proteins in response to low temperature in *Escherichia coli*. *Journal of Bacteriology*, *169*(5), 2092-2095.

Jones, P. G., & Inouye, M. (1996). RbfA, a 30S ribosomal binding factor, is a cold-shock protein whose absence triggers the cold-shock response. *Molecular Microbiology*, *21*(6), 1207–18.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, *7*(8), 1511–1522.

Karlson, D., & Imai, R. (2003). Conservation of the cold shock domain protein family in plants. *Plant Physiology*, *131*(1), 12–5.

Karlson, D., Nakaminami, K., Toyomasu, T., & Imai, R. (2002). A cold-regulated nucleic acid-binding protein of winter wheat shares a domain with bacterial cold shock proteins. *The Journal of Biological Chemistry*, *277*(38), 35248–56.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–80.

Keedy, D. A., Georgiev, I., Triplett, E. B., Donald, B. R., Richardson, D. C., & Richardson, J. S. (2012). The role of local backrub motions in evolved and designed mutations. *PLoS Computational Biology*, *8*(8), e1002629.

Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, *32*, W526–W531.

Kim, Y., Wang, X., Zhang, X. S., Grigoriu, S., Page, R., Peti, W., & Wood, T. K. (2010). *Escherichia coli* toxin/antitoxin pair MqsR/MqsA regulate toxin CspD. *Environmental Microbiology*, *12*(5), 1105–1121.

Kim, Y., & Wood, T. K. (2010). Toxins Hha and CspD and small RNA regulator Hfq are involved in persister cell formation through MqsR in *Escherichia coli*. *Biochemical and Biophysical Research Communications*, *391*(1), 209–213.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge, UK: Cambridge University Press.

Kimura, M. (1985). The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, *64*(1), 7–19.

Kinch, L. N., & Grishin, N. V. (2002). Evolution of protein structures and functions. *Current Opinion in Structural Biology*, *12*(3), 400–408.

Kirschner, M., Gerhart, J., Otey, C. R., & Arnold, F. H. (1998). Evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(15), 8420–7.

Kloks, C. P., Spronk, C. A., Lasonder, E., Hoffmann, A., Vuister, G. W., Grzesiek, S., & Hilbers, C. W. (2002). The solution structure and DNA-binding properties of the cold-shock domain of the human Y-box protein YB-1. *Journal of Molecular Biology*, *316*(2), 317–26.

Koshi, J. M., & Goldstein, R. A. (1998). Models of natural mutations including site heterogeneity. *Proteins: Structure, Function and Genetics*, *32*(3), 289–295.

Kumar, A., Malloch, R. A., Fujita, N., Smillie, D. A., Ishihama, A., & Hayward, R. S. (1993). The minus 35-recognition region of *Escherichia coli* sigma 70 is inessential for initiation of transcription at an "extended minus 10" promoter. *Journal of Molecular Biology*, *232*(2), 406–18.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, *25*(7), 1307–20.

Lee, S. J., Xie, A., Jiang, W., Etchegaray, J. P., Jones, P. G., & Inouye, M. (1994). Family of the major cold-shock protein, CspA (CS7.4), of *Escherichia coli*, whose members show a high sequence similarity with the eukaryotic Y-box binding proteins. *Molecular Microbiology*, *11*(5), 833–9.

Liberles, D. A. (2007). *Ancestral sequence reconstruction*. Oxford ; New York: Oxford University Press.

Lim, J., Thomas, T., & Cavicchioli, R. (2000). Low temperature regulated DEAD-box RNA helicase from the antarctic archaeon, *Methanococcoides burtonii*. *Journal of Molecular Biology*, *297*(3), 553–67.

Liu, W., Dong, H., Li, J., Ou, Q., Lv, Y., Wang, X., Xiang, Z., He, Y., & Wu, Q. (2015). RNA-seq reveals the critical role of OtpR in regulating *Brucella melitensis* metabolism and virulence under acidic stress. *Scientific Reports*, *5*(4), 417–424.

Lopez, M. M., Yutani, K., & Makhatadze, G. I. (1999). Interactions of the major cold shock protein of *Bacillus subtilis* CspB with single-stranded DNA templates of different base composition. *The Journal of Biological Chemistry*, *274*(47), 33601–8.

Lopez, M. M., Yutani, K., & Makhatadze, G. I. (2001). Interactions of the cold shock protein CspB from *Bacillus subtilis* with single-stranded DNA. Importance of the T base content and position within the template. *The Journal of Biological Chemistry*, *276*(18), 15511–8.

Matsumoto, K., & Wolffe, A. P. (1998). Gene regulation by Y-box proteins: coupling control of transcription and translation. *Trends in Cell Biology*, *8*(8), 318–23.

Max, K. E. A., Zeeb, M., Bienert, R., Balbach, J., & Heinemann, U. (2006). T-rich DNA single strands bind to a preformed site on the bacterial cold shock protein Bs-CspB. *Journal of Molecular Biology*, *360*(3), 702–714.

Max, K. E. A., Zeeb, M., Bienert, R., Balbach, J., & Heinemann, U. (2007). Common mode of DNA binding to cold shock domains: crystal structure of hexathymidine bound to the domain-swapped form of a major cold shock protein from *Bacillus caldolyticus*. *FEBS Journal*, *274*(5), 1265–1279.

Maynard Smith, J. (1970). Natural selection and the concept of a protein space. *Nature*, *225*(5232), 563–564.

Mayr, C., Hemann, M. T., & Bartel, D. P. (2007). Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science (New York, N.Y.)*, *315*(5818), 1576–9.

Mayr, F., Schütz, A., Döge, N., & Heinemann, U. (2012). The Lin28 cold-shock domain remodels pre-let-7 microRNA. *Nucleic Acids Research*, *40*(15), 7492–7506.

McKeown, A. N., Bridgham, J. T., Anderson, D. W., Murphy, M. N., Ortlund, E. A., & Thornton, J. W. (2014). Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell*, *159*(1), 58–68.

Mega, R., Manzoku, M., Shinkai, A., Nakagawa, N., Kuramitsu, S., & Masui, R. (2010). Very rapid induction of a cold shock protein by temperature downshift in *Thermus thermophilus*. *Biochemical and Biophysical Research Communications*, *399*(3), 336–340.

Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer Aided-Drug Design*, *7*(2), 146–157.

Mihailovich, M., Militti, C., Gabaldón, T., & Gebauer, F. (2010). Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(2), 109–18.

Mitchell, S. A., Spriggs, K. A., Coldwell, M. J., Jackson, R. J., & Willis, A. E. (2003). The Apaf-1 internal ribosome entry segment attains the correct structural conformation for function via interactions with PTB and unr. *Molecular Cell*, *11*(3), 757–71.

Mitta, M., Fang, L., & Inouye, M. (1997). Deletion analysis of cspA of *Escherichia coli* : requirement of the AT-rich UP element for cspA transcription and the downstream box in the coding region for its cold shock induction. *Molecular Microbiology*, *26*(02), 321–335.

Mitton-Fry, R. M., Anderson, E. M., Hughes, T. R., Lundblad, V., & Wuttke, D. S. (2002). Conserved structure for single-stranded telomeric DNA recognition. *Science (New York, N.Y.)*, *296*(5565), 145–7.

Miyata, T., Miyazawa, S., & Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution*, *12*(3), 219–36.

Moss, E. G., & Tang, L. (2003). Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. *Developmental Biology*, *258*(2), 432–42.

Murata, N., & Los, D. A. (1997). Membrane fluidity and temperature perception. *Plant Physiology*, *115*(3), 875–879.

Murzin, A. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *The EMBO Journal*, *12*(3), 861–7.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, *247*(4), 536–540.

Nakaminami, K., Karlson, D. T., & Imai, R. (2006). Functional conservation of cold shock domains in bacteria and higher plants. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 10122–7.

Nakaminami, K., Sasaki, K., Kajita, S., Takeda, H., Karlson, D., Ohgi, K., & Imai, R. (2005). Heat stable ssDNA/RNA-binding activity of a wheat cold shock domain protein. *FEBS Letters*, *579*(21), 4887–91.

Nakashima, K., Kanamaru, K., Mizuno, T., & Horikoshi, K. (1996). A novel member of the cspA family of genes that is induced by cold shock in *Escherichia coli*. *Journal of Bacteriology*, *178*(10), 2994–7.

Newkirk, K., Feng, W., Jiang, W., Tejero, R., Emerson, S. D., Inouye, M., & Montelione, G. T. (1994). Solution NMR structure of the major cold shock protein (CspA) from *Escherichia coli*: identification of a binding epitope for DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(11), 5114–8.

Nishino, T., Ariyoshi, M., Iwasaki, H., Shinagawa, H., & Morikawa, K. (1998). Functional analyses of the domain structure in the holliday junction binding protein RuvA. *Structure (London, England)*, *6*(1), 11–21.

O'Boyle, N. M., Morley, C., & Hutchison, G. R. (2008). Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, *2*(1), 5.

O'Brien, P. J., & Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry and Biology*, *6*(4), R91-R105.

Ollikainen, N., Smith, C. A., Fraser, J. S., & Kortemme, T. (2013). Methods in Enzymology: Flexible backbone sampling methods to model and design protein alternative conformations. *Methods Enzymol*, 523, 61-85.

Orengo, C. A., & Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annual Review of Biochemistry*, *74*(1), 867–900.

Orengo, C. A., Michie, A., Jones, S., Jones, D., Swindells, M., & Thornton, J. (1997). CATH – a hierarchic classification of protein domain structures. *Structure*, *5*(8), 1093–1109.

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., & Thornton, J. W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, *317*(5844), 1544–1548.

Patalano, S., Mihailovich, M., Belacortu, Y., Paricio, N., & Gebauer, F. (2009). Dual sex-specific functions of *Drosophila* upstream of N-ras in the control of X chromosome dosage compensation. *Development (Cambridge, England)*, *136*(4), 689–98.

Patthy, L. L. (2008). *Protein Evolution*. Chicester, UK: Wiley-Blackwell.

Pei, J., & Grishin, N. V. (2007). PROMALS: Towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics*, *23*(7), 802–808.

Peng, J., & Xu, J. (2011). Raptorx: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 161–171.

Perl, D., Welker, C., Schindler, T., Schröder, K., Marahiel, M. A., Jaenicke, R., & Schmid, F. X. (1998). Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nature Structural Biology*, *5*(3), 229–35.

Petsko, G. A., & Ringe, D. (2004). *Protein structure and function*. London; UK: New Science Press.

Phadtare, S., Alsina, J., & Inouye, M. (1999). Cold-shock response and cold-shock proteins. *Current Opinion in Microbiology*, *2*(2), 175–80.

Phadtare, S., Inouye, M., & Severinov, K. (2002). The nucleic acid melting activity of *Escherichia coli* CspE is critical for transcription antitermination and cold acclimation of cells. *The Journal of Biological Chemistry*, *277*(9), 7239–45.

Phadtare, S., & Severinov, K. (2005). Nucleic acid melting by *Escherichia coli* CspE. *Nucleic Acids Research*, *33*(17), 5583–90.

Phadtare, S., & Severinov, K. (2009). Comparative analysis of changes in gene expression due to RNA melting activities of translation initiation factor IF1 and a cold shock protein of the CspA family. *Genes to Cells : Devoted to Molecular & Cellular Mechanisms*, *14*(11), 1227–39.

Phadtare, S., Zhu, L., Uemori, T., Mukai, H., Kato, I., & Inouye, M. (2009). Applications of nucleic acid chaperone activity of CspA and its homologues. *Journal of Molecular Microbiology and Biotechnology*, *17*(3), 110–7.

Pisarev, A. V, Skabkin, M. A., Thomas, A. A., Merrick, W. C., Ovchinnikov, L. P., & Shatsky, I. N. (2002). Positive and negative effects of the major mammalian messenger ribonucleoprotein p50 on binding of 40 S ribosomal subunits to the initiation codon of beta-globin mRNA. *The Journal of Biological Chemistry*, *277*(18), 15445–51.

Pupko, T., Huchon, D., Cao, Y., Okada, N., & Hasegawa, M. (2002). Combining multiple data sets in a likelihood analysis: which models are the best? *Molecular Biology and Evolution*, *19*(12), 2294–2307.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D., & Friedman, N. (2002a). A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. *Bioinformatics*, *18*(8), 1116–23.

Pupko, T., Pe'er, I., Hasegawa, M., Graur, D., & Friedman, N. (2002b). A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: application to the evolution of five gene families. *Bioinformatics*, *18*(8), 1116–1123.

Pupko, T., Pe'er, I., Shamir, R., & Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, *17*(6), 890–896.

Rorick, M. M., & Wagner, G. P. (2011). Protein structural modularity and robustness are associated with evolvability. *Genome Biology and Evolution*, *3*(0), 456–475.

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., & Gourse, R. L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* , *262*(5138), 1407–13.

Rost, B., & Liu, J. (2003). The predict protein server. *Nucleic Acids Research*, *31*(13), 3300–3304.

Rudan, M., Schneider, D., Warnecke, T., & Krisko, A. (2015). RNA chaperones buffer deleterious mutations in *E. coli*. *ELife*, *4*, e04745.

Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., & Schroeder, M. (2015). PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, *43*(W1), W443–W447.

Sankoff, D. (1975). Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, *28*(1), 35–42.

Sankoff, D., & Rousseau, P. (1975). Locating the vertices of a steiner tree in an arbitrary metric space. *Mathematical Programming*, *9*(1), 240–246.

Sasaki, K., & Imai, R. (2011). Pleiotropic roles of cold shock domain proteins in plants. *Frontiers in Plant Science*, *2*, 116.

Sasaki, K., Kim, M.-H., & Imai, R. (2007). *Arabidopsis* cold shock domain protein2 is a RNA chaperone that is regulated by cold and developmental signals. *Biochemical and Biophysical Research Communications*, *364*(3), 633–8.

Schindelin, H., Jiang, W., Inouye, M., & Heinemann, U. (1994). Crystal structure of CspA, the major cold shock protein of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(11), 5119–23.

Schindelin, H., Marahiel, M. A., & Heinemann, U. (1993). Universal nucleic acid-binding domain revealed by crystal structure of the *B. subtilis* major cold-shock protein. *Nature*, *364*(6433), 164–168.

Schindler, T., Graumann, P. L., Perl, D., Ma, S., Schmid, F. X., & Marahiel, M. A. (1999). The family of cold shock proteins of *Bacillus subtilis*. Stability and dynamics in vitro and in vivo. *The Journal of Biological Chemistry*, *274*(6), 3407–13.

Schnuchel, A., Wiltscheck, R., Czisch, M., Herrler, M., G. Wllllmsky, G., Graumann, P., Marahiel, M. A., & Holak, T. A. (1993). Structure in solution of the major cold-shock protein from *Bacillus subtilis*. *Nature*, *364*(6433), 169–171.

Schröder, K., Graumann, P., Schnuchel, A., Holak, T. A., & Marahiel, M. A. (1995). Mutational analysis of the putative nucleic acid-binding surface of the cold-shock domain, CspB, revealed an essential role of aromatic and basic residues in binding of single-stranded DNA containing the Y-box motif. *Molecular Microbiology*, *16*(4), 699–708.

Shakhnovich, B. E., Deeds, E., Delisi, C., & Shakhnovich, E. (2005). Protein structure and evolutionary history determine sequence space topology. *Genome Research*, *15*(3), 385–392.

Shamoo, Y., Friedman, A. M., Parsons, M. R., Konigsberg, W. H., & Steitz, T. A. (1995). Crystal structure of a replication fork single-stranded DNA binding protein (T4 gp32) complexed to DNA. *Nature*, *376*(6538), 362–366.

Siddiq, M. A., Hochberg, G. K., & Thornton, J. W. (2017). Evolution of protein specificity: insights from ancestral protein reconstruction. *Current Opinion in Structural Biology*. *47*,113-122.

Song, Y., Dimaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T., Thompson, J., & Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure*, *21*(10), 1735–1742.

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., Van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, 521(7551), 173–179.

Sprengart, M. L., Fuchs, E., & Porter, A. G. (1996). The downstream box: an efficient and independent translation initiation signal in *Escherichia coli*. *The EMBO Journal*, *15*(3), 665–74.

Starr, T. N., Picton, L. K., & Thornton, J. W. (2017). Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, *549*(7672), 409–413.

Starr, T. N., & Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Science*, *25*(7), 1204–1218.

Stern, D. L. (2000). Perspective: evolutionary developmental biology and the problem of variation. *Evolution*, *54*(4), 1079–1091.

Storz, J. F. (2018). Compensatory mutations and epistasis for protein function. *Current Opinion in Structural Biology*, *50*, 18–25.

Stülke, J. (2002). Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures. *Archives of Microbiology*, *177*(6):433-4.

Subramanya, H. S., Doherty, A. J., Ashford, S. R., & Wigley, D. B. (1996). Crystal Structure of an ATP-Dependent DNA Ligase from Bacteriophage T7. *Cell*, *85*(4), 607–615.

Sunyaev, S. R., Eisenhaber, F., Rodchenkov, I. V., Eisenhaber, B., Tumanyan, V. G., & Kuznetsov, E. N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Engineering, Design and Selection*, *12*(5), 387–394.

Taha, Siddiqui, K. S., Campanaro, S., Najnin, T., Deshpande, N., Williams, T. J., Aldrich-Wright, J., Wilkins, M., Curmi, P. M., & Cavicchioli, R. (2016). Single TRAM domain RNA-binding proteins in archaea: functional insight from Ctr3 from the antarctic methanogen *Methanococcoides burtonii*. *Environmental Microbiology*, 18(9), 2810–2824.

Tanabe, H., Goldstein, J., Yang, M., & Inouye, M. (1992). Identification of the promoter region of the *Escherichia coli* major cold shock gene, cspA. *Journal of Bacteriology*, *174*(12), 3867–73.

Tawfik, O. K. and D. S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annual Review of Biochemistry*, *79*(1), 471–505.

Tedin, K., Resch, A., & Bläsi, U. (1997). Requirements for ribosomal protein S1 for translation initiation of mRNAs with and without a 5′ leader sequence. *Molecular Microbiology*, *25*(01), 189–199.

Tokuriki, N., & Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology*, *19*(5), 596–604.

Tóth-Petróczy, Á., & Tawfik, D. S. (2014). The robustness and innovability of protein folds. *Current Opinion in Structural Biology*, *26*, 131–138.

Trott, O., & Olson, A. J. (2010). Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), 455–461.

Ugalde, J. A., Chang, B. S. W., & Matz, M. V. (2004). Evolution of coral pigments recreated. *Science* , *305*(5689), 1433.

Uppal, S., Akkipeddi, V. S., & Jawali, N. (2008). Posttranscriptional regulation of cspE in *Escherichia coli*: involvement of the short 5'-untranslated region. *FEMS Microbiology Letters*, *279*(1), 83–91.

Uppal, S., Shetty, D. M., & Jawali, N. (2014). Cyclic AMP receptor protein regulates cspd, a bacterial toxin gene, in *Escherichia coli*. *Journal of Bacteriology*, *196*(8), 1569–1577.

VanBogelen, R. A., & Neidhardt, F. C. (1990). Ribosomes as sensors of heat and cold shock in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(15), 5589–93.

Voet, D., & Voet, J. G. (2011). *Biochemistry*. Hoboken, NJ: John Wiley & Sons.

Wagner, A. (2005). *Robustness and evolvability in living systems*. Princeton, NJ: Princeton University Press.

Wang, N., Yamanaka, K., & Inouye, M. (1999). CspI, the ninth member of the CspA family of *Escherichia coli*, is induced upon cold shock. *Journal of Bacteriology*, *181*(5), 1603–9.

Weber, M. H., Klein, W., Müller, L., Niess, U. M., & Marahiel, M. A. (2001). Role of the *Bacillus subtilis* fatty acid desaturase in membrane adaptation during cold shock. *Molecular Microbiology*, *39*(5), 1321–9.

Weber, M. H., & Marahiel, M. A. (2002). Coping with the cold: the cold shock response in the Gram-positive soil bacterium *Bacillus subtilis*. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *357*(1423), 895–907.

Xia, B., Ke, H., & Inouye, M. (2001). Acquirement of cold sensitivity by quadruple deletion of the cspA family and its suppression by PNPase S1 domain in *Escherichia coli*. *Molecular Microbiology*, *40*(1), 179–88.

Xu, D., & Zhang, Y. (2012). Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function and Bioinformatics*, *80*(7), 1715–1735.

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, *26*(7), 889–895.

Yamanaka, K. (1999). Cold shock response in *Escherichia coli*. *Journal of Molecular Microbiology and Biotechnology*, *1*(2), 193–202.

Yamanaka, K., Fang, L., & Inouye, M. (1998). The CspA family in *Escherichia coli*: multiple gene duplication for stress adaptation. *Molecular Microbiology*, *27*(2), 247–55.

Yamanaka, K., & Inouye, M. (1997). Growth-phase-dependent expression of cspD, encoding a member of the CspA family in *Escherichia coli*. *Journal of Bacteriology*, *179*(16), 5126–30.

Yamanaka, K., Zheng, W., Crooke, E., Wang, Y. H., & Inouye, M. (2001). CspD, a novel DNA replication inhibitor induced during the stationary phase in *Escherichia coli*. *Molecular Microbiology*, *39*(6), 1572–1584.

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, *10*(6), 1396–1401. Yang, Z., Kumar, S., & Nei, M. (1995). A new method ofinference of ancestral nucleotide and amino acid sequences. *Genetics*, *141*(4), 1641–1650.

Zeeb, M., & Balbach, J. (2003). Single-stranded DNA binding of the cold-shock protein CspB from *Bacillus subtilis*: NMR mapping and mutational characterization. *Protein Science*, *12*(1), 112–123.

Zeeb, M., Max, K. E. A., Weininger, U., Löw, C., Sticht, H., & Balbach, J. (2006). Recognition of T-rich single-stranded DNA by the cold shock protein Bs-CspB in solution. *Nucleic Acids Research*, *34*(16), 4561–4571.

Zhang, B., Yue, L., Zhou, L., Qi, L., Li, J., & Dong, X. (2017). Conserved TRAM domain functions as an Archaeal cold shock protein via RNA chaperone activity. *Frontiers in Microbiology*, *8*, 1597.

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Genetics*, *57*(4), 702–710.

Zhang, Y., & Skolnick, J. (2005). TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.