



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
BIOLOGÍA EVOLUTIVA

**ANÁLISIS EVOLUTIVO DE LAS SECUENCIAS SIMPLES Y REGIONES
INTRÍNSECAMENTE DESORDENADAS EN LA EVOLUCIÓN TEMPRANA DE LA VIDA**

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRA EN CIENCIAS BIOLÓGICAS

PRESENTA:

ALMA CAROLINA SÁNCHEZ ROCHA

TUTOR PRINCIPAL DE TESIS: DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM
COMITÉ TUTOR: DRA. NURIA VICTORIA SÁNCHEZ PUIG
INSTITUTO DE QUÍMICA, UNAM
DR. VÍCTOR HUGO ANAYA MUÑOZ
ENES-MORELIA, UNAM

CD. MX.

FEBRERO, 2019



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
BIOLOGÍA EVOLUTIVA

**ANÁLISIS EVOLUTIVO DE LAS SECUENCIAS SIMPLES Y REGIONES
INTRÍNSECAMENTE DESORDENADAS EN LA EVOLUCIÓN TEMPRANA DE LA VIDA**

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRA EN CIENCIAS BIOLÓGICAS

PRESENTA:

ALMA CAROLINA SÁNCHEZ ROCHA

TUTOR PRINCIPAL DE TESIS: DR. ARTURO CARLOS II BECERRA BRACHO
FACULTAD DE CIENCIAS, UNAM
COMITÉ TUTOR: DRA. NURIA VICTORIA SÁNCHEZ PUIG
INSTITUTO DE QUÍMICA, UNAM
DR. VÍCTOR HUGO ANAYA MUÑOZ
ENES-MORELIA, UNAM

MÉXICO, CD. MX. FEBRERO, 2019

OFICIO FCIE/DAIP/0019/2019
ASUNTO: Oficio de Jurado

M. en C. Ivonne Ramírez Wence
Directora General de Administración Escolar, UNAM
Presente

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **26 de noviembre de 2018** se aprobó el siguiente jurado para el examen de grado de **MAESTRA EN CIENCIAS BIOLÓGICAS** en el campo de conocimiento de **Biología Evolutiva** de la alumna **SÁNCHEZ ROCHA ALMA CAROLINA** con número de cuenta **307117157** con la tesis titulada "**Análisis evolutivo de las secuencias simples y regiones intrínsecamente desordenadas en la evolución temprana de la vida**", realizada bajo la dirección del **DR. ARTURO CARLOS II BECERRA BRACHO**:

Presidente: DRA. ALEJANDRA ALICIA COVARRUBIAS ROBLES
Vocal: DR. ALEJANDRO SOSA PEINADO
Secretario: DRA. NURIA VICTORIA SÁNCHEZ PUIG
Suplente: DRA. CLAUDIA ÁLVAREZ CARREÑO
Suplente: M. EN C. RICARDO HERNÁNDEZ MORALES

Sin otro particular, me es grato enviarle un cordial saludo.

ATENTAMENTE
"POR MI RAZA HABLARA EL ESPÍRITU"
Ciudad Universitaria, Cd. Mx., a 11 de enero de 2019


DR. ADOLFO GERARDO NAVARRO SIGÜENZA
COORDINADOR DEL PROGRAMA



AGNS/VMVA/ASR/ipp

Agradecimientos

Al Posgrado en Ciencias Biológicas de la Universidad Nacional Autónoma de México y a la UNAM, la Máxima casa de estudios por brindarme todo el apoyo para mi formación desde el bachillerato.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico otorgado durante la maestría (CVU: 777679). A los apoyos PAEP que me permitieron asistir al 6o Congreso de la Rama de Físicoquímica, Estructura y Diseño de Proteínas en Durango y la *Gordon Research Conference: Intrinsically Disordered Proteins* en Suiza.

A mi tutor principal, el Dr. Arturo Becerra Bracho por todo su apoyo y confianza que me brindó durante la realización de la maestría.

A los miembros de mi Comité Tutor: Dra. Nuria Victoria Sánchez Puig y Dr. Víctor Hugo Anaya Muñoz por el apoyo e interés que me dieron en la maestría y por los comentarios que fueron vitales para la realización de la tesis.

Agradecimientos a título personal

A mi muy querido tutor, el Dr. Arturo Becerra por brindarme su apoyo y confianza en todo momento, gracias por creer en mí y por transmitir la pasión que tienes por la evolución. Pero sobre todo por introducirme al estudio de las secuencias simples, me siento muy afortunada de trabajar contigo en un tema que me parece fascinante y sé que este proyecto de investigación tiene mucho futuro y que trabajaremos juntos. Te admiro profundamente.

Al Dr. Lazcano (The Very Great Dominus et Magister) por todo el apoyo que me has brindado, por todo tu cariño y por abrirme las puertas a su laboratorio. No tengo palabras para expresar el cariño y admiración que te tengo. Gracias por preocuparte genuinamente por los jóvenes de México, por acercar la ciencia a los mexicanos y por representar a nuestro país en todo el mundo. Querido Toño, jamás podré dejar de estar más agradecida contigo, principalmente porque tu libro: “El origen de la vida”, me inspiró a estudiar Biología, además de ayudarme a responder una curiosidad desde niña, explicarme de dónde venimos.

A mi querida y brillante Dra. Claudia Álvarez, gracias ser una fuente de inspiración y apoyo inagotable para mi, esta tesis y la preparación para los congresos, sin duda hubieran sido casi imposibles sin ti.

Al M. en C. Ricardo Hernández. Richard, por todo el apoyo y confianza que me brindó durante la maestría. Te tengo mucho cariño.

A la Dra. Alejandra Covarrubias, por el apoyo para esta tesis y por ser la mejor profesora que he tenido y haberme inspirado para el estudio de las proteínas intrínsecamente desordenadas.

A los miembros de mi comité tutor, la Dra. Nuria Sánchez Puig y al Dr. Víctor Hugo Anaya Muñoz por el apoyo infinito durante la maestría.

Al Dr. Alejandro Sosa Peinado por los comentarios de la tesis y por generar en mí nuevas preguntas en el estudio de las proteínas.

Al M. en C. José Campillo, mi agradecimiento y cariño por ti son infinitos. Muchas gracias por todo tu apoyo y cariño, sabes que estoy muy orgullosa de ti y que te quiero como a un hermano.

A la M. en C. Sara Islas, por toda su confianza, apoyo y por permitirme entender más acerca de LUCA.

A mi querido Dr. Mario Rivas, por todo su apoyo y solucionar todas mis dudas, estoy impresionada de tu habilidad por las máquinas y porque su interés en el estudio de las proteínas intrínsecamente desordenadas.

A los macacos del laboratorio de Origen de la Vida, me siento muy afortunada de trabajar con ustedes. A Wolphie, quien mejoró todos mis días con su amistad, te adoro y admiro por siempre. A Germán, por tu amistad por todos los consejos que me brindaste, sin ti no sé qué hubiera hecho en varias situaciones, te quiero. A Coral, por siempre demostrarme más de una razón para apreciar la vida, te quiero mucho. A Karen, por toda tu amistad y por transmitirme tu pasión por los fósiles, te quiero mucho. A Ervin, gracias por hacer más ameno el trabajo diario y por las charlas con café. A mi querido Isra, por ser tan noble y mostrarme todo su apoyo. A Alex, por ser tan paciente y buen amigo, te quiero mucho. A Rod, por todas las enseñanzas que me ha brindado y por generar momentos de

alegría cuando lleva a sus hijos. A Claudia Sierra, por las buenas conversaciones. A Ale Cisneros por ser muy buena onda y a mi muy querido Amadeo Estrada, con quien viví de los mejores días en la maestría, a Beto por sus buenos comentarios y a Adriana por el buen humor que la caracteriza.

A mi mamá por brindarme todo el amor y apoyo del mundo. Gracias por ser la mejor mamá, te amo con todas mis células.

A mi hermano Armando, por todo el cariño, alegrías y burlas que siempre me hace. Te amo con todo mi ser.

A bebé Julio por ser lo más bello que he visto y por todo el cariño que me da.

A mis amigos y familia, porque con su cariño hicieron posible esta tesis y por ser parte fundamental en mi vida.

A mi madre, Julio y a mi hermano, mis mayores amores

“La biblioteca es ilimitada y periódica. Si un eterno viajero la atravesara en cualquier dirección, comprobaría al cabo de los siglos que los mismos volúmenes se repiten en el mismo desorden (que, repetido, sería un orden: el Orden). Mi soledad se alegra con esa elegante esperanza”.

Jorge Luis Borges, 1941

‘Borges’s infinite, and unsearchable, library is reflected in the Maynard-Smith collection of all possible protein sequences’

Frances H. Arnold, 2011

Índice

Resumen.....	1
Abstract.....	2
Introducción.....	3
Evolución temprana de la vida.....	3
Las primeras proteínas.....	4
El ultimo ancestro común.....	4
Secuencias simples.....	5
Regiones intrínsecamente desordenadas.....	8
Objetivos.....	11
Antecedentes.....	11
Metodología.....	12
Base de datos de proteomas completos.....	12
Búsqueda de secuencias simples (LCRs).....	14
Búsqueda de regiones intrínsecamente desordenadas (IDRs).....	14
Clasificación funcional de proteínas que contienen LCRs e IDRs.....	15
Contenido de GC en proteomas de Bacteria y Archaea.....	15
Composición de aminoácidos de LCRs e IDRs.....	15
Asignación de dominios proteínicos de proteínas con LCRs e IDRs.....	15
Búsqueda de homólogos.....	15
Análisis filogenético de proteínas con LCRs e IDRs.....	16
Alineamientos múltiples de secuencias con LCRs e IDRs.....	17
Clasificación de IDRs con amplia distribución filogenética.....	17
Resultados.....	18
Análisis de LCRs.....	18
Análisis de IDRs.....	27
Secuencias simples con amplia distribución filogenética.....	37
Conservación de las LCRs con altos niveles de distribución.....	41
Regiones intrínsecamente desordenadas con amplia distribución filogenética.....	47
Conservación de IDRs con altos niveles de distribución.....	48
Discusión.....	55
Secuencias simples en Bacteria y Archaea.....	55
Regiones intrínsecamente desordenadas en Bacteria y Archaea.....	56
LCRs con amplia distribución filogenética.....	58
IDRs con amplia distribución filogenética.....	60
La contribución de las secuencias simples en el origen de genes.....	66
El papel de las LCRs e IDRs en la evolución temprana de la vida.....	67
Conclusiones.....	70
Literatura citada.....	71
Anexos.....	80

Índice de abreviaturas

LCA: Último Ancestro Común de todos los seres vivos (Last Common Ancestor o cenancestro).

LCRs: Regiones de baja complejidad (RBC o LCRs del inglés 'Low complexity regions'). También se conocen como secuencias simples.

IDRs: Regiones intrínsecamente desordenadas (RIDs o IDRs del inglés 'Intrinsically disordered regions').

Resumen

Las secuencias simples son regiones de proteínas y ácidos nucleicos que se caracterizan por un sesgo composicional y frecuentemente se localizan dentro de regiones intrínsecamente desordenadas. Su conservación en proteínas antiguas sugiere su presencia en el último ancestro común de todos los seres vivos (LCA por sus siglas en inglés). Es probable que las polimerasas primitivas, al igual que las actuales, inevitablemente se deslizaran sobre el DNA, generando secuencias simples durante la evolución temprana de la vida y posiblemente contribuyendo en el origen de genes, la formación de materia prima, incremento de tamaño del genoma y en la variabilidad genética; no obstante la evidencia de esto es escasa, ya que la preservación del sesgo composicional característico de las secuencias simples se debe principalmente a motivos funcionales o estructurales. Con el objetivo de indagar sobre el posible papel que tuvieron las secuencias simples y regiones intrínsecamente desordenadas en etapas tempranas de la vida, se analizaron proteomas completos de bacterias y arqueas. La composición de aminoácidos de las secuencias simples y de las regiones desordenadas, además de su localización en dominios funcionales de las proteínas o en los extremos amino o carboxilo terminales, sugiere que ya se encontraban presentes en el LCA y que algunas de ellas pudieron tener un papel significativo desde el mundo de RNA/proteínas.

Palabras clave: secuencias simples (LCRs), regiones intrínsecamente desordenadas (IDRs), deslizamiento de la polimerasa, evolución temprana de la vida, último ancestro común (LCA), origen de genes.

Abstract

Simple sequences are segments of proteins and nucleic acids which are biased in residue composition and are frequently present in intrinsically disordered regions. Their conservation in ancient proteins, suggest their presence in the last common ancestor (LCA). It is likely that primitive polymerases, as the current ones, had slipped-strand mispairing (slippage) as an unavoidable characteristic, generating simple sequences in early evolution of life and possibly contributing to the origin of genes, promoting the formation of raw material, the increase of genome size and genetic variability. Nevertheless, the evidence is limited because the preservation of the characteristic compositional bias of simple sequences is due to functional or structural reasons. With the aim of inquire about the possible roles of simple sequences and intrinsically disordered regions in early evolution of life, complete proteomes of Bacteria and Archaea domains were analyzed. Amino acid composition of simple sequences and disordered regions, their localization in protein functional domains or in carboxy- and amino- terminal regions of proteins, suggests that they were already present in the LCA and some of them may have played a significant role since the RNA/protein world stage.

Keywords: simple sequences (LCRs), intrinsically disordered regions (IDRs), slipped strand mispairing, early evolution of life, last common ancestor (LCA), origin of genes.

Introducción

Evolución temprana de la vida

El genoma es similar a una biblioteca, almacena información (o libros) de distintas etapas evolutivas, algunas tan recientes provenientes de la última replicación o más antiguas, que anteceden la divergencia de los linajes celulares. ¿Existe en la biblioteca algún rastro de la evolución temprana de la vida?.

Emile Zuckerkandl y Linus Pauling propusieron en 1964 que “De todos los sistemas naturales, la materia viva es la que, ante grandes transformaciones, preserva inscrita en su organización la mayor cantidad de su propia historia pasada”; es por ello que la comparación de las secuencias de DNA, RNA y proteínas, permite obtener información evolutiva (Zuckerkandl & Pauling, 1965). La genómica comparada ha permitido acercarnos a etapas tempranas en la evolución, teniendo como límite el mundo de RNA/proteínas (Fig.1).

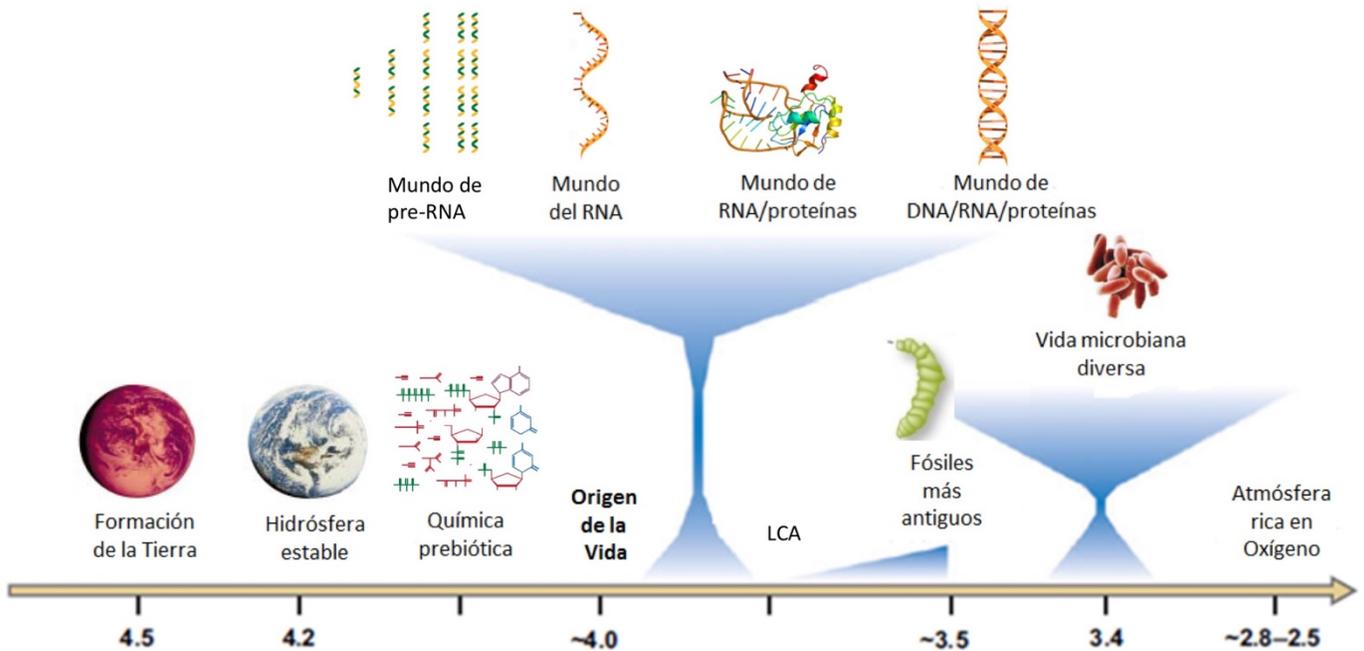


Figura 1: Línea de tiempo que representa a los eventos principales en el origen y evolución temprana de la vida. En la imagen, cada billón de años equivale a 1000 millones de años. Modificado de Becerra *et al.* (2007) y Joyce (2002).

El hecho de que generalmente el DNA es el portador de la información hereditaria en los seres vivos podría dar la primera impresión de que el DNA siempre ha sido el protagonista; sin embargo, este es descendiente de un mundo en donde el RNA prevalecía. Una de las pruebas más convincentes de ello es que para llevar a cabo la síntesis de desoxirribonucleótidos, se necesitan ribonucleótidos y las enzimas ribonucleótido reductasas (Lazcano *et al.*, 1988).

Las células portan diversas evidencias de que somos descendientes del último ancestro común (LCA por sus siglas en inglés) de todos los seres vivos (Lazcano, 1992): “El mismo código genético, las mismas características esenciales de la replicación y expresión genética, reacciones anabólicas básicas y la producción de energía por ATPasas en las membranas; así como, las variaciones mínimas se explican fácilmente como el resultado de procesos divergentes de una vida ancestral” (Becerra *et al.*, 2007). Si bien es cierto que el LCA no representa a los primeros seres vivos, proporciona cierta información acerca de cómo era la vida hace más de 3800 millones de años.

Las primeras proteínas

En un mundo de RNA debieron haber evolucionado las primeras proteínas (Lazcano *et al.*, 1992), las cuales pudieron surgir completamente *de novo*. Es posible que estuvieran involucradas en el mantenimiento estructural y en la replicación del RNA, antes de presentar actividad catalítica; es posible que las proteínas ancestrales tuvieran baja especificidad y actividad tipo chaperona (Poole *et al.*, 1998). Otros papeles posiblemente importantes en la evolución temprana son la participación en el incremento de la actividad catalítica de las ribozimas, estabilizando sus estructuras ribonucleotídicas, actividad de chaperona que puede ser el equivalente primitivo de la actividad actual de la porción proteica de la RNasa P (Delaye & Lazcano, 2000) y otras ribozimas (Ivanyi-Nagy *et al.*, 2008), contribuyendo con la disminución de la dependencia de iones metálicos como el Mg²⁺, necesarios para la estabilización de la estructura terciaria de las ribozimas (Poole *et al.*, 1998).

A la actividad catalítica de las proteínas, posiblemente le antecedió la formación de una estructura, realizando la función de un andamio. En general, cualquier proteína, ribozima o compuesto inorgánico que realice catálisis, casi siempre requiere de una estructura previa (Di Mauro *et al.*, 2012).

El último ancestro común

La búsqueda de genes con distribución filogenética universal ha reemplazado la ausencia del registro paleontológico del LCA (Delaye *et al.*, 2005). El LCA es el resultado de etapas evolutivas anteriores, es la parte superior de un tronco cuya longitud no se conoce; es decir, incluye secuencias

que se originaron en distintas épocas pre-cen ancestrales, que antecedieron al LCA (Delaye & Becerra, 2012).

Se han desarrollado una gran diversidad de metodologías que realizan búsquedas de los genes altamente conservados en los dominios celulares Archaea, Bacteria y Eucarya; los cuales muy probablemente fueron heredados del ancestro común de todos los seres vivos, formando parte de su complemento génico, es decir, el conjunto de genes que poseía el LCA (Delaye & Becerra, 2012). La diversidad de metodologías que abarca la reconstrucción del LCA se ha realizado considerando dominios proteínicos (Delaye *et al.*, 2005), grupos de genes ortólogos (Harris *et al.*, 2003; Mirkin *et al.*, 2003), superfamilias de plegamientos proteínicos (Yang *et al.*, 2005), superfamilias de dominios proteínicos (Ranea *et al.*, 2006) y en la actualidad con estructuras terciarias de proteínas (Islas *et al.*, in prep).

La forma (método) para determinar el complemento génico del LCA se ha renovado, y en la actualidad, los genes conservados en Bacteria y Archaea constituyen una mejor aproximación (Fig. 2). Al ser los eucariontes un grupo más reciente que los procariontes y también un grupo hermano del grupo TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) de arqueas, se ha propuesto una clasificación de dos dominios celulares (Williams *et al.*, 2013; Raymann *et al.*, 2015).

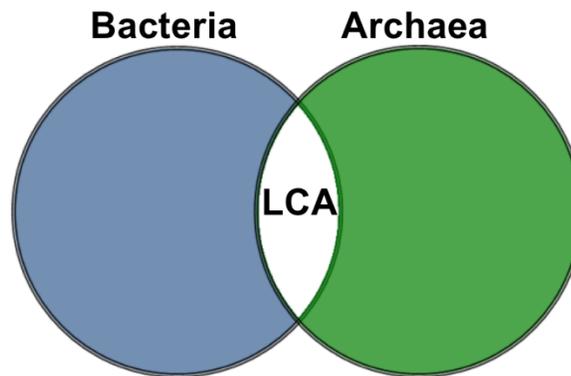


Figura 2: El contenido génico del LCA, de acuerdo con la topología de dos dominios celulares. La intersección de los círculos, corresponde a los genes compartidos en Bacteria y Archaea.

Secuencias simples

Las secuencias simples, son regiones de nucleótidos o aminoácidos con un sesgo en su composición (Fig. 3) (Ellegren, 2004; Velasco *et al.*, 2013). Las secuencias simples en las proteínas

también se conocen como regiones de baja complejidad (LCRs por sus siglas en inglés) (Radó-Trilla, 2013). Las LCRs pueden presentar diversas configuraciones, desde repeticiones de un solo aminoácido (conocidas como homopolímeros) hasta motivos repetidos de más de un aminoácido (Toll-Riera *et al.*, 2012; Chaudhry *et al.*, 2018).

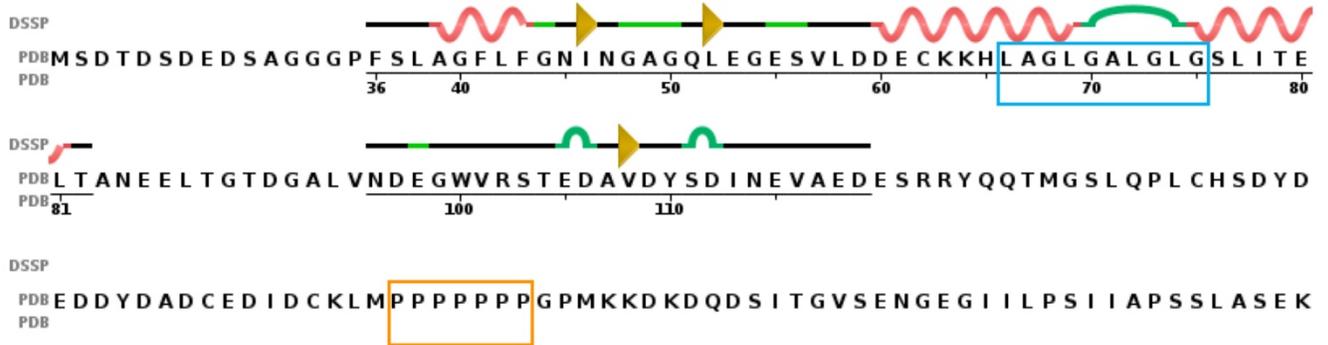


Figura 3: Ejemplos de secuencias simples presentes en el factor de iniciación de la transcripción TFIID (TAF1) de *Homo sapiens*. Las regiones de baja complejidad (LCRs; recuadro de color azul) y el Homopolímero (recuadro de color naranja) fueron detectados con el algoritmo SEG (Wootton & Federhen, 1993). La imagen de la secuencia de la proteína fue obtenida de PDB con el código 6MZD (Patel *et al.*, 2018).

Los motivos de repeticiones de aminoácidos en las secuencias simples son tan variables que abarcan desde repeticiones cortas de aminoácidos, como las regiones de poliglutaminas, causantes de la enfermedad de Huntington, hasta largas repeticiones con múltiples dominios, como la titina, la proteína más larga conocida (Andrade *et al.*, 2001).

El mecanismo principal de origen de las secuencias simples es el deslizamiento de la polimerasa (slippage en inglés) (Fig. 4). Este fenómeno genera mutaciones que ocurren durante la replicación del DNA (Bebenek & Kunkel, 1990; Ellegren, 2004; Radó-Trilla, 2013). Las secuencias de DNA se caracterizan por formar estructuras secundarias, como los loops que bloquean la DNA polimerasa y provocan la disociación del complejo de replicación. Si la estructura del loop se forma en la hebra naciente, va a originar una expansión, si es en la hebra parental será una delección (Levinson & Gutman, 1987; Viguera *et al.*, 2001).

El deslizamiento de la polimerasa puede ocurrir en todos los genomas celulares, virales y posiblemente, de mitocondrias y cloroplastos. Las secuencias simples también pueden originarse por otros procesos como la recombinación y el entrecruzamiento desigual, además de incrementarse por

eventos de duplicación genética o transferencia lateral de genes (Lozada-Chávez, 2004). Aunque las secuencias simples estén presentes en todos los genomas, los eucariontes poseen una mayor proporción de estas secuencias en comparación con los procariontes (Sim & Creamer, 2002).

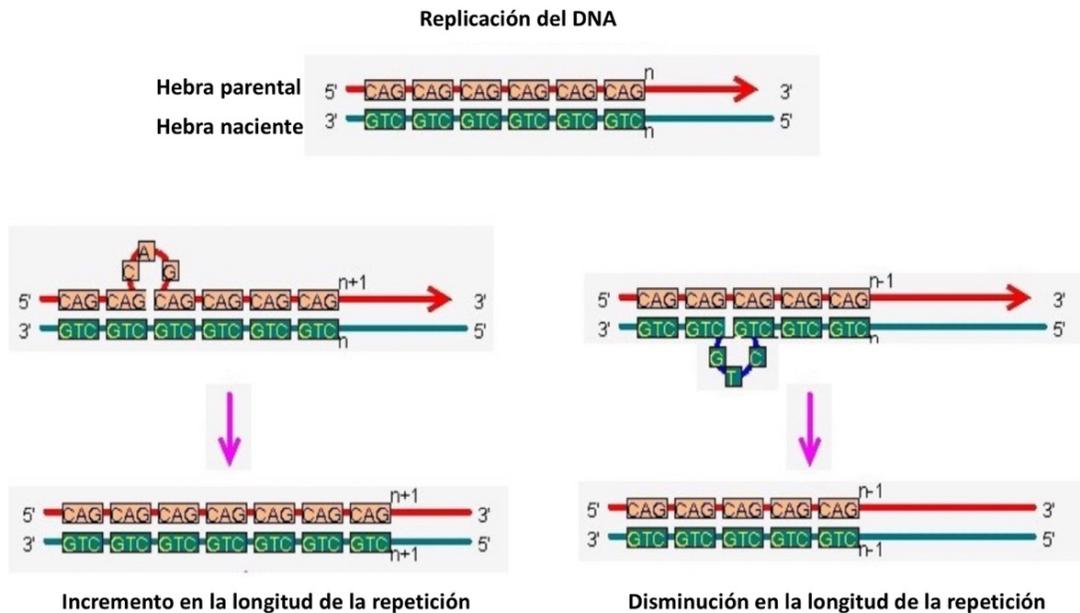


Figura 4: Mecanismo del deslizamiento de la polimerasa (slippage).

La mayor parte de las secuencias simples se encuentran en DNA no codificante, sin embargo, también están en regiones codificantes que, a nivel de secuencia primaria, generan un sesgo composicional de aminoácidos. Es importante mencionar que el sesgo en las proteínas puede deberse o no a la simplicidad en la secuencia de nucleótidos en el DNA, debido a la degeneración del código genético (Radó-Trilla, 2013).

Las secuencias simples han contribuido a la evolución del tamaño de los genomas procariontes y eucariontes (Tautz *et al.*, 1986; Wootton & Federhen, 1993; Hancock 1995; Hancock, 1996; Katti *et al.*, 2001), por lo que además de aportar regiones nuevas a las proteínas, generan materia prima para nuevas funciones. También son consideradas como una fuente de variabilidad en procariontes patógenos (Britten & Kohne, 1968; Tautz & Renz, 1984) y en virus (Lozada-Chávez, 2004; Velasco *et al.*, 2013).

Regiones intrínsecamente desordenadas

Las proteínas intrínsecamente desordenadas (PIDs o IDPs del inglés ‘intrinsically disordered proteins’) han desafiado la visión tradicional de la estructura y función en las proteínas, ya que carecen de una estructura tridimensional única en su estado nativo (Uversky, 2014) y son altamente flexibles (Tompa, 2010). Aquellos segmentos en las proteínas que no adoptan ninguna estructura terciaria se definen como regiones intrínsecamente desordenadas (RIDs o IDRs del inglés ‘intrinsically disordered regions’). En realidad, las proteínas en el genoma pueden considerarse como modulares, porque están constituidas de combinaciones de regiones estructuradas y desordenadas (Fig. 5) (van der Lee *et al.*, 2014). A diferencia de las proteínas y dominios ordenados, las IDPs no tienen una estructura bien definida en equilibrio, y existen como un conjunto heterogéneo, altamente dinámico de conformeros (Uversky, 2014).

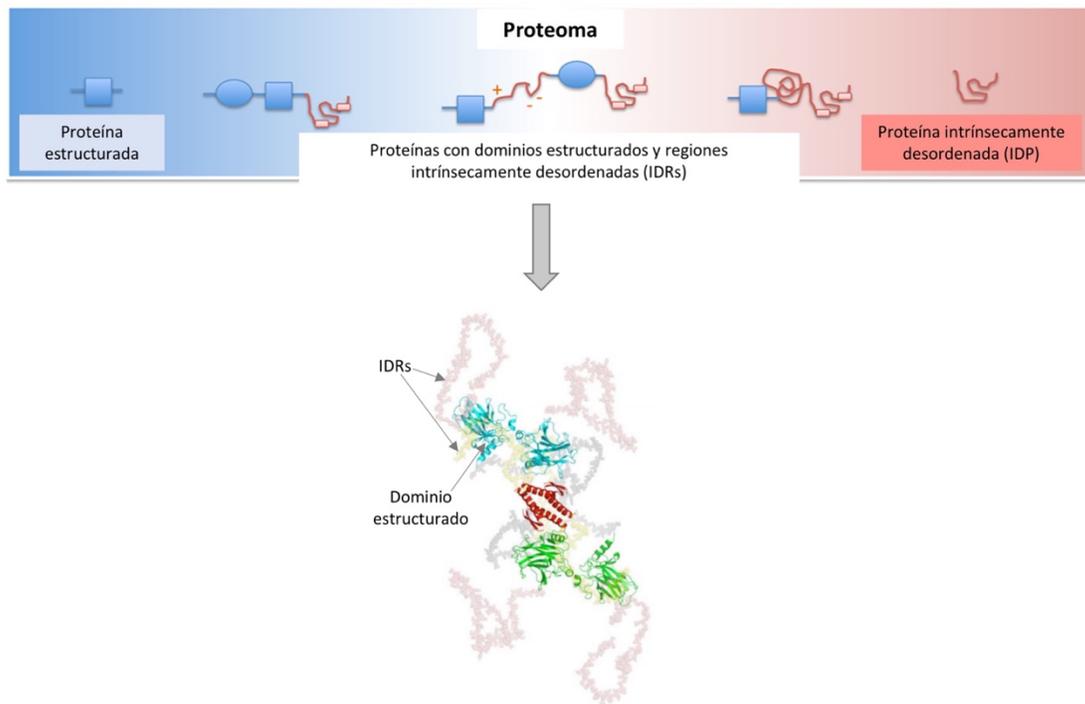


Figura 5: Las proteínas en el genoma, pueden considerarse modulares, porque están constituidas de combinaciones de regiones estructuradas y desordenadas. Un ejemplo es la proteína p53, en la imagen se muestra el modelo SAXS de la proteína (Tidow *et al.*, 2007). Modificado de van der Lee *et al.*, 2014.

Se ha propuesto que las IDPs han cambiado el paradigma de estructura y función de las proteínas. Tradicionalmente, se consideraba necesaria la estructura para que una proteína llevara a

cabo cierta función, a tal grado que se postuló que “para que una proteína sea funcional debe poseer una estructura tridimensional bien definida”. El descubrimiento de nuevas proteínas y la caracterización de algunas ya descritas, ha revelado la existencia de las IDPs, que presentan una estructura flexible o un plegamiento azaroso y son capaces de llevar a cabo una o más funciones. Esto ha sido un reto para ampliar el concepto clásico de estructura-función (Vázquez-Cuevas & Covarrubias-Robles, 2011).

Uno de los aspectos más relevantes de las IDRs es que por mucho tiempo fueron poco analizadas, ya que, al ser regiones altamente flexibles, no pueden ser identificadas por cristalografía de rayos X y aparecen publicadas como regiones con ausencia de densidad electrónica (missing electron density regions) es decir, no aparecen en el patrón de difracción de rayos X. La identificación experimental de proteínas desordenadas se lleva a cabo mediante diversas técnicas como la espectrometría de resonancia magnética nuclear (NMR), la espectroscopia infrarroja (IR), espectroscopia UV, espectroscopia de fluorescencia (Ball *et al.*, 2014) y la difracción por rayos X de ángulos pequeños o SAXS (del inglés Small Angle X-ray Scattering), las cuales permiten analizar la flexibilidad conformacional de las mismas (Siliqi *et al.*, 2018). Adicionalmente, la aplicación de la bioinformática ha sido fundamental para el estudio de las proteínas intrínsecamente desordenadas pues ha permitido demostrar la presencia de IDPs e IDRs en todos los dominios celulares; así como, establecer algunas de sus características comunes (Uversky, 2014; Dunker *et al.*, 2010; Peng *et al.*, 2015; Necci *et al.*, 2016; Basile & Elofsson, 2018).

La flexibilidad de las IDPs e IDRs les confiere un amplio dinámismo conformacional, pues pueden pasar de tener un estado coil-like, completamente desordenado, a ser pre-molten globule, con estructuras secundarias no fijas, hasta tener estados globulares compactos (van der Lee *et al.*, 2014). La flexibilidad estructural de las IDPs también les proporciona la capacidad de adoptar diferentes conformaciones dependiendo de las condiciones microambientales en la que se encuentren y de la presencia de ligandos a partir de los cuales se reflejaría su función. La conformación puede depender del estado metabólico o de desarrollo de las células, del tipo celular o de la presencia de una condición estresante (Vázquez-Cuevas & Covarrubias-Robles, 2011). Algunas IDRs pueden adquirir una conformación estructural específica cuando se unen a su sustrato (Dunker *et al.*, 2002), al cambiar el pH, o en presencia de sales u otra proteína.

Las IDRs e IDPs se componen de aminoácidos promotores del desorden: Lys, Glu, Pro, Ser y Gln, los cuales poseen baja hidrofobicidad y son polares. Se caracterizan por ser pobres en

aminoácidos promotores del orden: Cys, His, Trp, Ile, Tyr, Phe, Leu, Met y Asn. Los aminoácidos Arg, Asp, Ala, Gly y Thr, son neutrales, es decir, pueden estar en regiones ordenadas o desordenadas (Olfield & Dunker, 2014). Debido a que las IDRs carecen de aminoácidos hidrofóbicos son incapaces de formar un núcleo hidrofóbico, que se conformaría como un dominio estructurado, y, por lo consiguiente, su funcionalidad difiere de la visión clásica de la estructura y función de las proteínas estructurales globulares (van der Lee *et al.*, 2014).

Frecuentemente, la secuencia de las proteínas y regiones desordenadas presentan LCRs (Romero *et al.*, 2001; Tompa & Fersht, 2009; Kumari *et al.*, 2014), por lo tanto, no poseen una estructura terciaria estable; sin embargo, algunas LCRs pueden formar estructuras secundarias y se ha notado que la mayoría de las LCRs pueden conformarse en alfa hélices (Saqi, 1995; Kumari *et al.*, 2014).

Objetivos

Objetivo general

Identificar la presencia y posible contribución de secuencias simples y regiones intrínsecamente desordenadas en la evolución temprana de la vida.

Objetivos particulares

Buscar secuencias simples y regiones intrínsecamente desordenadas en proteomas de una muestra representativa perteneciente a bacterias y arqueas.

Identificar las proteínas con secuencias simples y regiones intrínsecamente desordenadas que presenten una amplia distribución filogenética en Bacteria y Archaea e inferir su posible presencia en el LCA.

Analizar la composición de aminoácidos de las secuencias simples y regiones intrínsecamente desordenadas antes mencionadas.

Antecedentes

Las secuencias simples o regiones de baja complejidad (LCRs por sus siglas en inglés) y las regiones intrínsecamente desordenadas (IDRs) son abundantes en los proteomas celulares, sin embargo, se desconoce qué tan antiguas son y qué papel pudieron haber tenido en la evolución temprana de la vida y, a la fecha, este fenómeno ha sido poco estudiado. Asimismo, se pretende obtener indagar sobre la relación evolutiva entre LCRs e IDRs, ya que anteriormente solo se habían estudiado las secuencias simples en el Laboratorio de Origen de la Vida. El análisis de proteomas de Bacteria y Archaea permitirá inferir la antigüedad relativa de las LCRs e IDRs y aportará información acerca de las relaciones entre la estructura y la función de estas proteínas peculiares desde una perspectiva evolutiva.

Las IDRs y LCRs se caracterizan por presentar una alta variabilidad en la secuencia y se ha reportado que, en las LCRs la huella de la simplicidad en la secuencia desaparece con el tiempo, ya que el sesgo composicional en la secuencia primaria solo se conserva si tiene importancia estructural o funcional en las proteínas, por lo que el análisis de la conservación a nivel de secuencia de LCRs e IDRs resulta fundamental.

Metodología

Base de datos de proteomas completos

La selección de los proteomas se realizó con base en la clasificación actual del NCBI de categorías de genomas de referencia y representativos. Los genomas de referencia son los de mejor calidad en la anotación, respaldada por la curación del staff del NCBI, además son de importancia médica, de calidad en el ensamblaje del genoma y la clasificación también se basa en la disponibilidad de evidencia experimental. Los genomas representativos también son de alta calidad y se identifican por métodos de agrupación de genomas y selección por consideración de clasificación a nivel de especie, calidad de ensamblaje y diversidad taxonómica (Tatusova *et al.*, 2014).

Tabla I, Phyla de Bacteria analizados.

Bacteria

Acidobacteria	Epsilonproteobacteria
Actinobacteria	Fibrobacteres
Alphaproteobacteria	Firmicutes
Aquificae	Fusobacteria
Armatimonadetes	Gammaproteobacteria
Bacteroidetes	Gemmatimonadetes
Betaproteobacteria	Nitrospirae
Caldiseptica	Other proteobacteria
Chlorobi	Planctomycetes
Chloroflexi	Spirochaetes
Chrysiogenetes	Synergistetes
Cyanobacteria	Tenericutes
Deferribacteres	Thermodesulfobacteria
Deinococcus-Thermus	Thermotogae
Deltaproteobacteria	Unclassified Bacteria
Dictyoglomi	Unclassified Terrabacteria group
Elusimicrobia	Verrucomicrobia

Tabla II, Phyla de Archaea analizados.

Archaea

Bathyarchaeota	Lokiarchaeota
Crenarchaeota	Nanohaloarchaeota
Euryarchaeota	Thaumarchaeota
Korarchaeota	Unclassified Archaea

Los proteomas se descargaron de KEGG (Kanehisa *et al.*, 2000). Con el fin de evitar redundancia biológica se eligió a una especie representativa o de referencia por género de cada phylum y para tener representantes de todos los phyla de la base de KEGG (Tabla I y Tabla II). La muestra biológica consiste en 643 proteomas, 561 de Bacteria y 82 de Archaea. Asimismo, los organismos se clasificaron de acuerdo con su estilo de vida en: i) Vida libre, ii) Patógenos, iii) Intracelulares y iv) Extremófilos (Tabla III). La clasificación de estilos de vida se basó en la anotación de los genomas del NCBI, en la descripción de BacMap y la literatura de la anotación de los genomas completos.

Estilo de vida	Bacterias	Arqueas
Vida libre	326	24
Patógenos	99	0
Extremófilos	136	58
Total	561	82

Tabla III: Proteomas analizados de Bacteria y Archaea

Para la identificación de las secuencias simples y regiones intrínsecamente desordenadas en el complemento génico del LCA, no se consideraron especies pertenecientes a la categoría de *Intracelulares* que incluye a endosimbiontes o parásitos (Delaye *et al.*, 2005) ya que pueden presentar grandes pérdidas polifiléticas de genes (Becerra *et al.*, 1997).

Búsqueda de secuencias simples (LCRs)

Las LCRs fueron identificadas mediante el programa SEG (Wootton & Federhen, 1993), que encuentra secuencias de baja complejidad con una alta concentración de segmentos cortos repetidos. Está fundamentado en la entropía de Shannon adaptada a secuencias proteínicas. La entropía es aplicada a dos niveles, el primero como el valor promedio de la entropía sobre la probabilidad de la distribución de los estados de complejidad de una longitud de ventana; el segundo es la entropía promedio de la secuencia y las probabilidades de sus letras. El programa primero calcula la complejidad para localizar segmentos de las secuencias de aminoácidos de un tamaño dado por la entropía de Shannon, los extiende y reduce a una LCR (Tompa & Fersht, 2009). Los valores utilizados en SEG (Anexo I) fueron: longitud de la ventana del disparador (W): 12; complejidad del disparador (K1): 1.9 y complejidad de la extensión (K2): 0.2.

Búsqueda de regiones intrínsecamente desordenadas (IDRs)

En la búsqueda de IDRs en las proteínas es recomendable emplear múltiples predictores de desorden y confiar en una predicción basada en el método de consenso. Es más probable que las IDRs predichas por varios métodos sean verosímiles que aquellas que difieren entre distintos métodos (Atkins *et al.*, 2015; Meng *et al.*, 2017). En este trabajo se analizaron proteomas completos y se utilizó el método de consenso con los siguientes predictores:

- **ESpritz** (Walsh *et al.*, 2012): se basa en redes neurales recursivas y está entrenado en tres tipos de metodologías para desorden: (i) Rayos X, en donde las regiones desordenadas no muestran densidad electrónica en las estructuras del Protein Data Bank (PDB); (ii) Disprot, una base de datos de proteínas desordenadas, la cual se curó manualmente; y (iii) Nuclear Magnetic Resonance (NMR), que proporciona información más confiable sobre la presencia de desorden estructural en proteínas.
- **IUpred2a** (Mészáros *et al.*, 2018): estima la capacidad de los polipéptidos de formar contactos estabilizadores bajo la premisa de que las proteínas globulares presentan muchas interacciones entre los residuos que las componen, proveyendo energía estabilizante, en contraste con las regiones desordenadas que no tienen la capacidad de formar esas interacciones. Ofrece tres tipos de predicción de desorden: desorden largo, corto y dominios estructurados; también predice las regiones de unión a proteínas.
- **DISOPRED3** (Jones & Cozzetto, 2015): este predictor emplea un método de *machine learning*, y está clasificado como uno de los más precisos y se especializa en predecir IDRs largas de manera óptima (Meng *et al.*, 2017). Adicionalmente, este predictor permite asignar

las regiones de unión a otras proteínas dentro de la región intrínsecamente desordenada.

Se ha descrito que las regiones desordenadas largas, de una longitud mayor o igual a 30 aminoácidos, pueden presentar alguna funcionalidad biológica, o formar parte de dominios proteínicos (Ward *et al.*, 2004), por lo que en este estudio se realizó un análisis de aquellas IDRs de longitud mayor o igual a 30 residuos.

Clasificación funcional de proteínas que contienen LCRs e IDRs

La clasificación funcional de las proteínas que presentan LCRs e IDRs se realizó mediante los grupos KO (KEGG Orthology) y los niveles de clasificación correspondientes de KEGG Brite (Kanehisa *et al.*, 2000). Adicionalmente, se utilizaron los números enzimáticos (EC), mediante el esquema de clasificación numérica para las enzimas. Los códigos de grupos KO y los números EC, se obtuvieron a partir de los códigos UniProt de cada proteína (The UniProt Consortium, 2017).

Contenido de GC en proteomas de Bacteria y Archaea

Los valores del contenido de GC fueron obtenidos de la base de datos *GENOME_REPORTS* del NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt).

Composición de aminoácidos de LCRs e IDRs

El análisis de la composición de aminoácidos en LCRs e IDRs se realizó con el programa ProtParam EXPASY tool (Gasteiger *et al.*, 2005).

Asignación de dominios proteínicos de proteínas con LCRs e IDRs

La anotación de los dominios proteínicos fue realizada a partir de los resultados parámetros obtenidos con hmmscan para cada secuencia y usando como referencia la base de datos de perfiles de Pfam. La anotación se automatizó mediante un script (en *perl*) elaborado en el Laboratorio de Origen de la Vida, por la Dra. Claudia Álvarez.

Búsqueda de homólogos

Para la búsqueda de homólogos de las proteínas con LCRs e IDRs se empleó el programa Proteinortho (Lechner *et al.*, 2011). Los valores de BLAST (Altschul *et al.*, 1990) que utiliza el programa son *e-value*= 0.00001, porcentaje de identidad $\geq 25\%$ y *query coverage* $\geq 50\%$.

Análisis filogenético de proteínas con LCRs e IDRs

Con el objetivo de indagar sobre el origen y evolución de las LCRs e IDRs en las proteínas bajo estudio, se realizó un análisis de la distribución filogenética de las mismas. El criterio que se empleó para analizar la distribución filogenética fue la presencia de las proteínas en clados filogenéticos, una estrategia desarrollada en el Laboratorio de Origen de la vida (Fig. 6). A continuación, se muestra la clasificación de clados de Bacteria y Archaea, y los phyla que los constituyen:

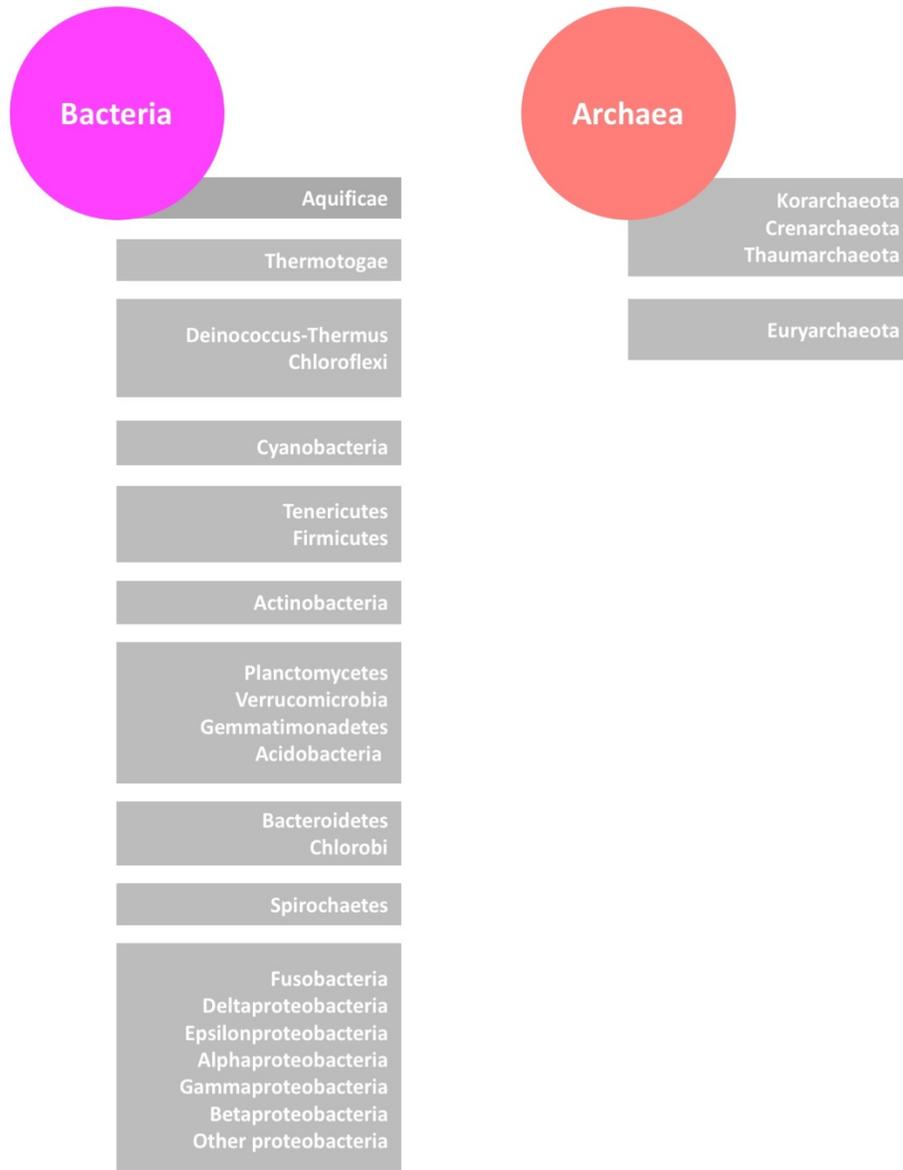


Figura 6: Clasificación de clados de Bacteria y Archaea para el análisis de distribución filogenética. Cada recuadro gris corresponde a un clado.

Cada clado puede agrupar uno o varios phyla. En Bacteria, el superphylum PVC incluye a los phyla Planctomycetes y Verrucomicrobia; el grupo Terrabacteria agrupa a Tenericutes y Firmicutes; el grupo FCB incluye a Bacteroidetes y Chlorobi; mientras que en el caso del Phylum Proteobacteria, esta conformado por las clases Deltaproteobacteria, Epsilonproteobacteria, Alphaproteobacteria, Gammaproteobacteria y Betaproteobacteria. En las arqueas, el grupo TACK se conforma por Korarchaeota, Crenarchaeota y Thaumarchaeota. Aquellas proteínas que contengan LCRs o IDRs y que estén presentes en al menos cinco clados de Bacteria y dos clados de Archaea, se consideraron como de amplia distribución filogenética.

Alineamientos múltiples de secuencias con LCRs e IDRs de amplia distribución filogenética

Con el objetivo de analizar la conservación a nivel de secuencia primaria de las LCRs e IDRs, se llevaron a cabo alineamientos múltiples de las proteínas con secuencias simples y regiones intrínsecamente desordenadas con amplia distribución filogenética. Los alineamientos se realizaron en el programa MAFFT 7 (Katoh & Standley, 2013), utilizando L-INS-i como método iterativo de refinamiento.

Para determinar la posición en la secuencia en la que se localizaban las LCRs e IDRs, se utilizó el método de Coletta *et al.* (2010), el cual las clasifica en ‘N-terminal’ o ‘C-terminal’, si las LCRs o IDRs se encuentran a no más de 25 aminoácidos de cualquiera de estos extremos en la secuencia; o bien las clasifica como ‘Central’, si se encuentran a 50 o más aminoácidos de los extremos de la secuencia.

Clasificación de IDRs con amplia distribución filogenética

Las proteínas con IDRs de amplia distribución filogenética se agruparon de acuerdo con la clasificación de Bellay *et al.* (2011):

- IDR conservada en secuencia (*Constrained*): Regiones en la secuencia de aminoácidos donde el desorden se conserva en el 50% de las especies o más, y la secuencia primaria se conserva en el 50% de los residuos alineados o más. Se asocian a sitios de unión a RNA y actividad como chaperona de proteínas.
- IDR conservada (*Flexible*): Regiones en la secuencia de aminoácidos donde el desorden se conserva en el 50% de las especies o más, y la secuencia primaria no se conserva. Están asociadas a vías de señalización y a la multifuncionalidad.
- IDR no conservada: Regiones en la secuencia de aminoácidos donde ni el desorden ni la secuencia primaria se conservan. No se asocian a ninguna categoría funcional.

Resultados

Análisis de LCRs

A partir de la búsqueda de LCRs en proteomas completos de 561 bacterias y 82 arqueas, se localizaron 435,153 proteínas con LCRs en bacterias y 35,599 proteínas con LCRs en arqueas.

Distribución filogenética de LCRs en Bacteria

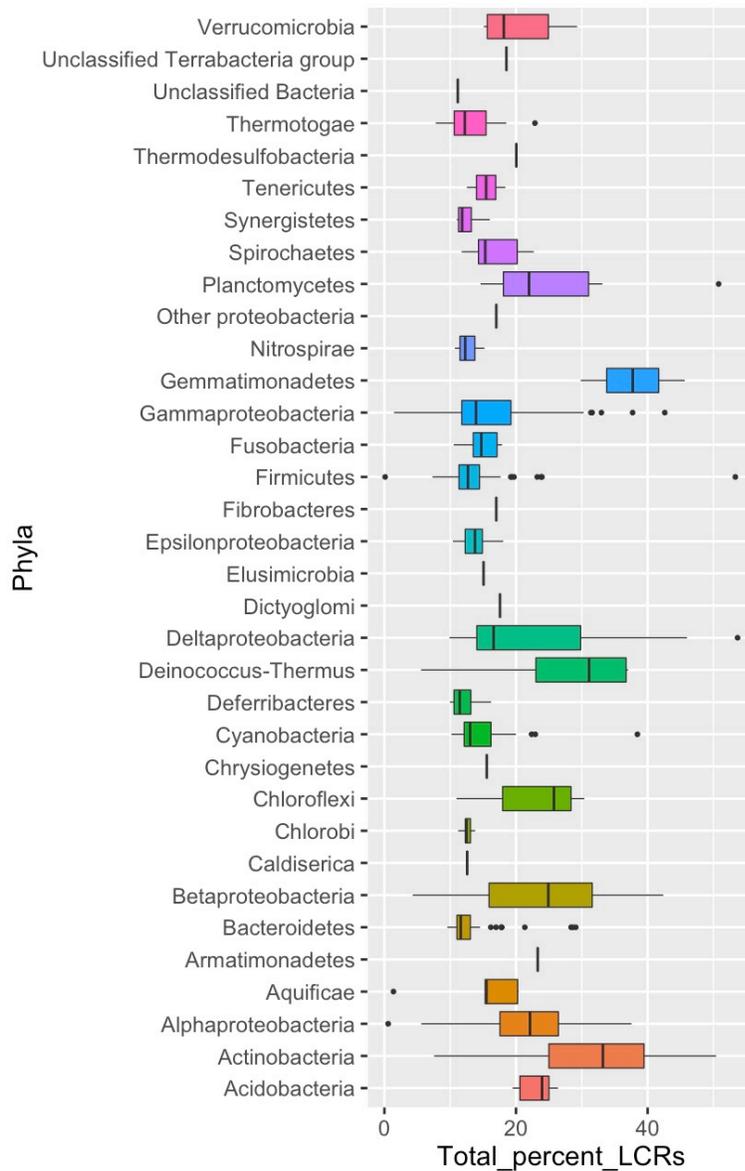


Figura 7: Porcentaje de secuencias simples en bacterias. Los diagramas de caja y bigote muestran el valor mínimo, el primer cuartil, el segundo cuartil (la mediana, señalada con una línea negra dentro de cada caja), el tercer cuartil y el valor máximo.

Las secuencias simples forman parte de los proteomas de todos los phyla de Bacteria, sin embargo, cada grupo presenta diferentes porcentajes de LCRs (Fig. 7). Los diagramas de caja y bigote permiten explicar el contenido de LCRs en los phyla, al ser una forma óptima de mostrar las semejanzas y diferencias entre los datos. Estos gráficos consisten en una caja central dividida en dos áreas divididas por una línea vertical (mediana) y otras dos áreas horizontales o bigotes; cada una de las cajas engloba el 50 % de los datos y representa la variación en la muestra, por lo que, a mayor tamaño, la variación y la desviación estándar (DE) es mayor. Aquellos puntos que se localizan a la derecha o izquierda de los bigotes son los valores atípicos (Solano & Rojas, 2005).

Los phyla bacterianos que presentan una mayor proporción de secuencias simples o LCRs, son Gemmatimonadetes, Actinobacteria y Deinococcus-Thermus (Fig. 7), como se puede ver con los valores de la mediana, que corresponden a los valores más altos de la mediana. Los phyla con porcentajes más bajos de LCRs son Synergisetales, Bacteroidetes y Deferribacteres.

La Desviación estándar es una medida que muestra la variación en los porcentajes de las LCRs dentro de todos los organismos que conforman cada phylum. El grupo que presenta la mayor variación en el contenido de LCRs es Planctomycetes, como se puede apreciar en el valor de la DE es 18.48 (Fig. 7 y Anexo II). Los phyla que presentan la menor DE son Tenericutes y Spirochaetes, lo que indica que la proporción de secuencias simples en los organismos de estos phyla varía poco.

Los valores de porcentaje mínimo y máximo al interior de cada phylum se aprecian en cada extremo de los bigotes, el de la izquierda es el más bajo y el de la derecha el valor más alto. Asimismo, los puntos negros representan valores atípicos de la muestra.

Un factor importante para considerar es el número de proteomas que se analizó por phylum, el cual difiere en cada caso, lo que podría ser importante al interpretar los diagramas de caja y bigote y, por tanto, la abundancia de estas secuencias en cada phyla.

Distribución de LCRs en Archaea

Los phyla de arqueas que presentan una mayor proporción de secuencias simples o LCRs, son Nanoarchaeota y Euryarchaeota (Fig. 8). En comparación, Bathyarchaeota y Thaumarchaeota presentan el menor porcentaje de LCRs en sus proteomas. Euryarchaeota presenta la mayor variación de porcentaje de LCRs, mientras que Thaumarchaeota la menor.

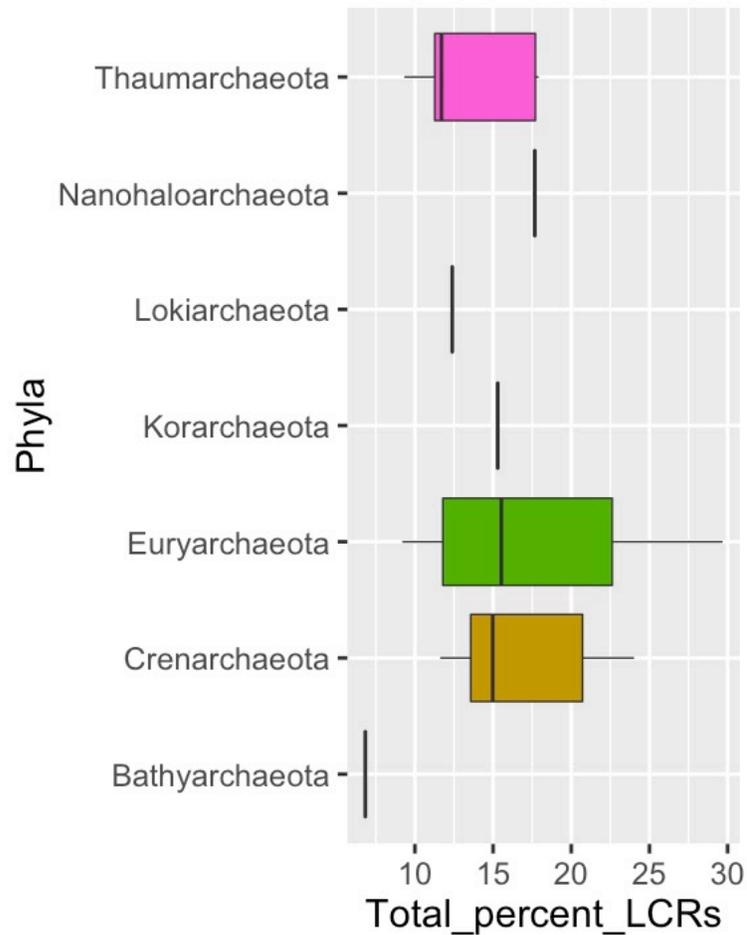


Figura 8: Porcentaje de secuencias simples en arqueas. Los diagramas de caja y bigote muestran el valor mínimo, el primer cuartil, el segundo cuartil (la mediana), el tercer cuartil y el valor máximo.

Tamaño del proteoma vs el porcentaje de LCRs

Con el objetivo de indagar si el tamaño del proteoma es directamente proporcional al porcentaje de secuencias simples en los proteomas analizados, se realizó un análisis estadístico utilizando el coeficiente de correlación de Pearson, que es una medida de la asociación lineal entre dos variables y se denota por r . El coeficiente de correlación de Pearson (r) puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva, es decir, a medida que aumenta el valor de una variable también lo hace el valor de la otra variable. Un valor menor que 0 indica una asociación negativa (Pearson, 1948; Lane *et al.*, 2013).

La asociación entre el tamaño del proteoma y el porcentaje de secuencias simples en Bacteria y Archaea es positiva, como lo indica el coeficiente de correlación de Pearson ($r= 0.33$) (Fig. 9). Sin

embargo, es una relación lineal ascendente débil. La asociación del porcentaje de LCRs con el tamaño del proteoma en Bacteria es $r= 0.30$, mientras que para Archaea corresponde a $r= 0.47$.

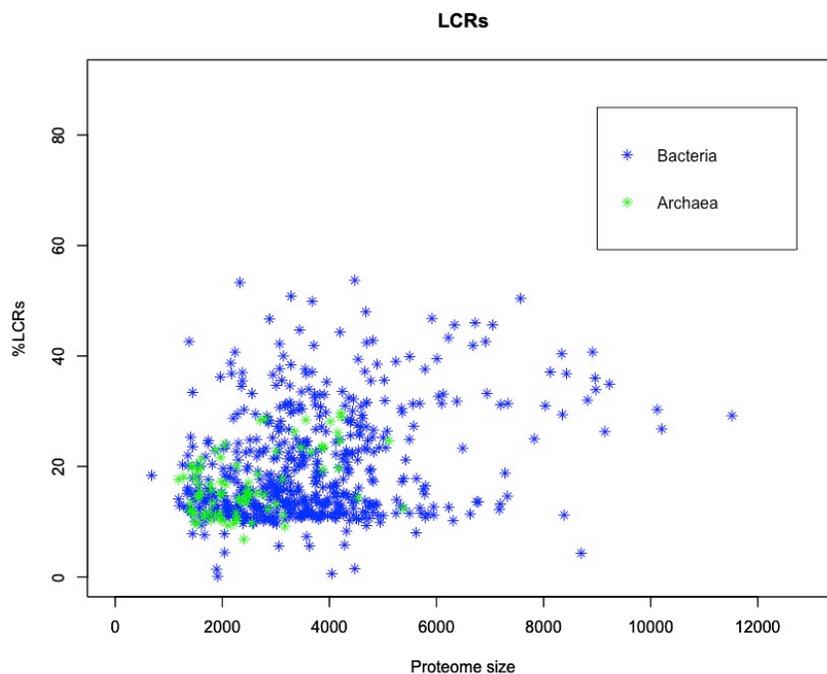


Figura 9: Asociación entre el tamaño del proteoma vs el porcentaje de secuencias simples en Bacteria y Archaea

Clasificación funcional de las secuencias simples

De las 435,153 proteínas con LCRs en bacterias, 192,975 se encuentran anotadas funcionalmente, mientras que para las 35,599 proteínas con LCRs en arqueas, 15,316 tienen anotación. El análisis funcional de las proteínas con secuencias simples o LCRs reveló que tienen un sesgo funcional importante (Anexo III y Fig. 10). La mayor proporción de estas pertenece a la clasificación funcional de *Transporte de membrana*, tanto en Archaea como en Bacteria (Fig. 10). Sin embargo, un gran número de proteínas con LCRs carecen de asignación funcional.

Además de la abundancia de LCRs en proteínas de membrana, también se aprecia que la segunda categoría funcional que prevalece en bacterias y arqueas es la relacionada a *Traducción* (Fig. 10), lo que señala la importancia que pueden tener las LCRs en dichas funciones celulares. Las LCRs también presentan una alta proporción en enzimas del *Metabolismo de carbohidratos* y *Metabolismo de aminoácidos*.

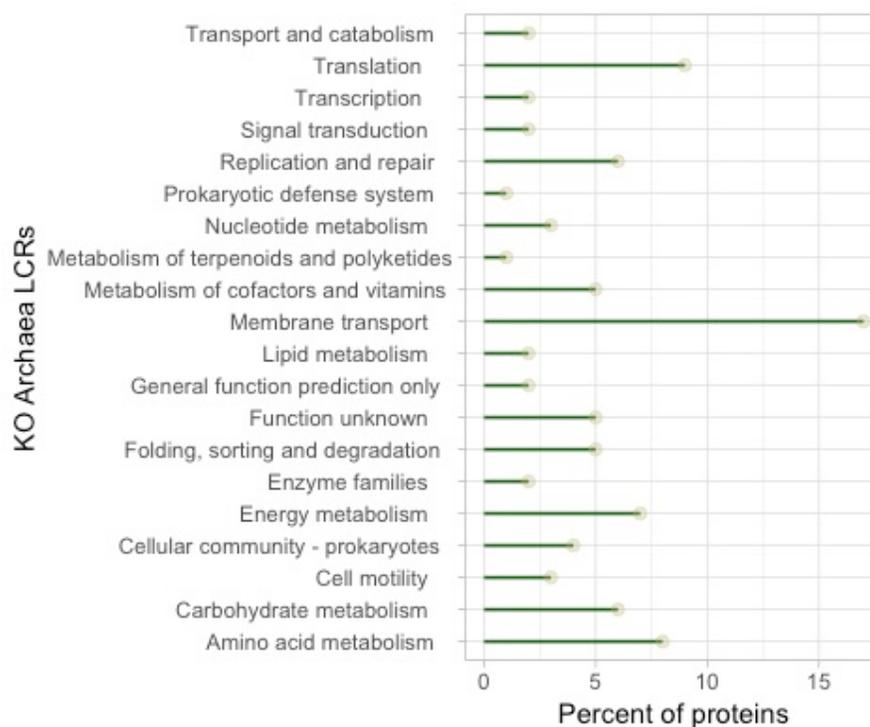
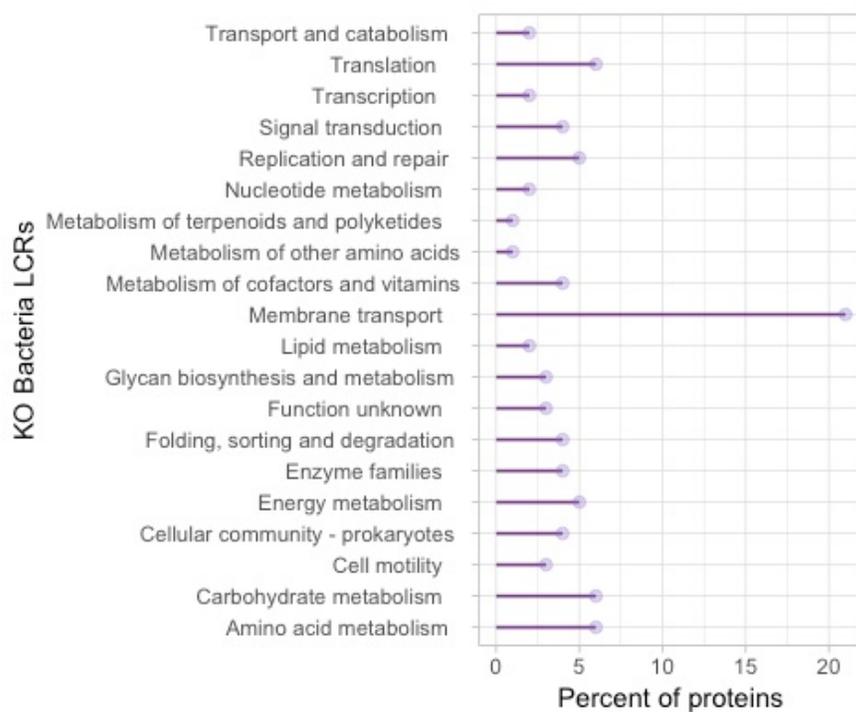


Figura 10: Categorías funcionales del segundo nivel de los grupos *KEGG orthology* (KO) de mayor abundancia en las proteínas con LCRs. En la parte superior se muestran los valores porcentuales pertenecientes a las categorías funcionales más abundantes en Bacteria y en la inferior, de Archaea.

Los phyla que presentan una mayor cantidad de secuencias simples presentan una proporción similar de las categorías funcionales más abundantes (Fig. 11), lo que sugiere que la participación de las secuencias simples en ciertos procesos celulares se mantiene a lo largo de la Evolución.

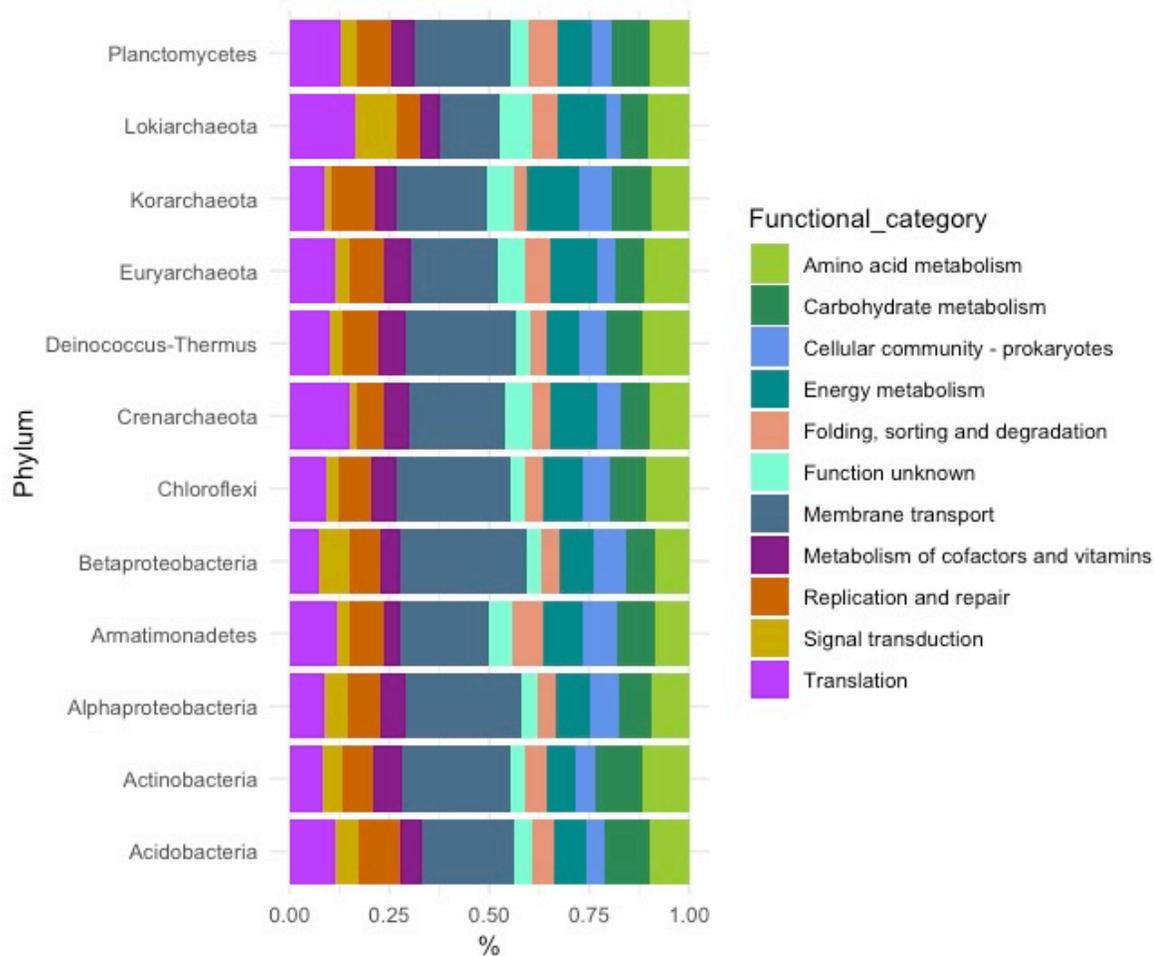


Figura 11: Funciones más abundantes en los phyla con mayor proporción de LCRs en sus proteomas. Se muestran los valores relativos.

Análisis evolutivo de las secuencias simples

Las secuencias simples forman parte de todos los proteomas de bacterias y arqueas analizados, aunque la proporción de LCRs difiere a nivel interespecífico y en el interior de cada phylum. Las barras de color azul claro representan los porcentajes de secuencias simples por proteoma (Fig.12).

Phyla	
■	Acidobacteria
■	Alphaproteobacteria
■	Actinobacteria
■	Armatimonadetes
■	Bacteroidetes
■	Betaproteobacteria
■	Aquificae
■	Chlorobi
■	Chloroflexi
■	Cyanobacteria
■	Deinococcus-Thermus
■	Deltaproteobacteria
■	Epsilonproteobacteria
■	Firmicutes
■	Fusobacteria
■	Gammaaproteobacteria
■	Gemmatimonadetes
■	Nitrospirae
■	Planctomycetes
■	Spirochaetes
■	Synergistetes
■	Tenericutes
■	Thermodesulfobacteria
■	Verrucomicrobia
■	Korarchaeota
■	Crenarchaeota
■	Bathyarchaeota
■	Thaumarchaeota
■	Lokiarchaeota
■	Euryarchaeota
■	Elusimicrobia
■	Thermotogae
■	Dictyoglomi
■	Fibrobacteres
■	Caldiserica

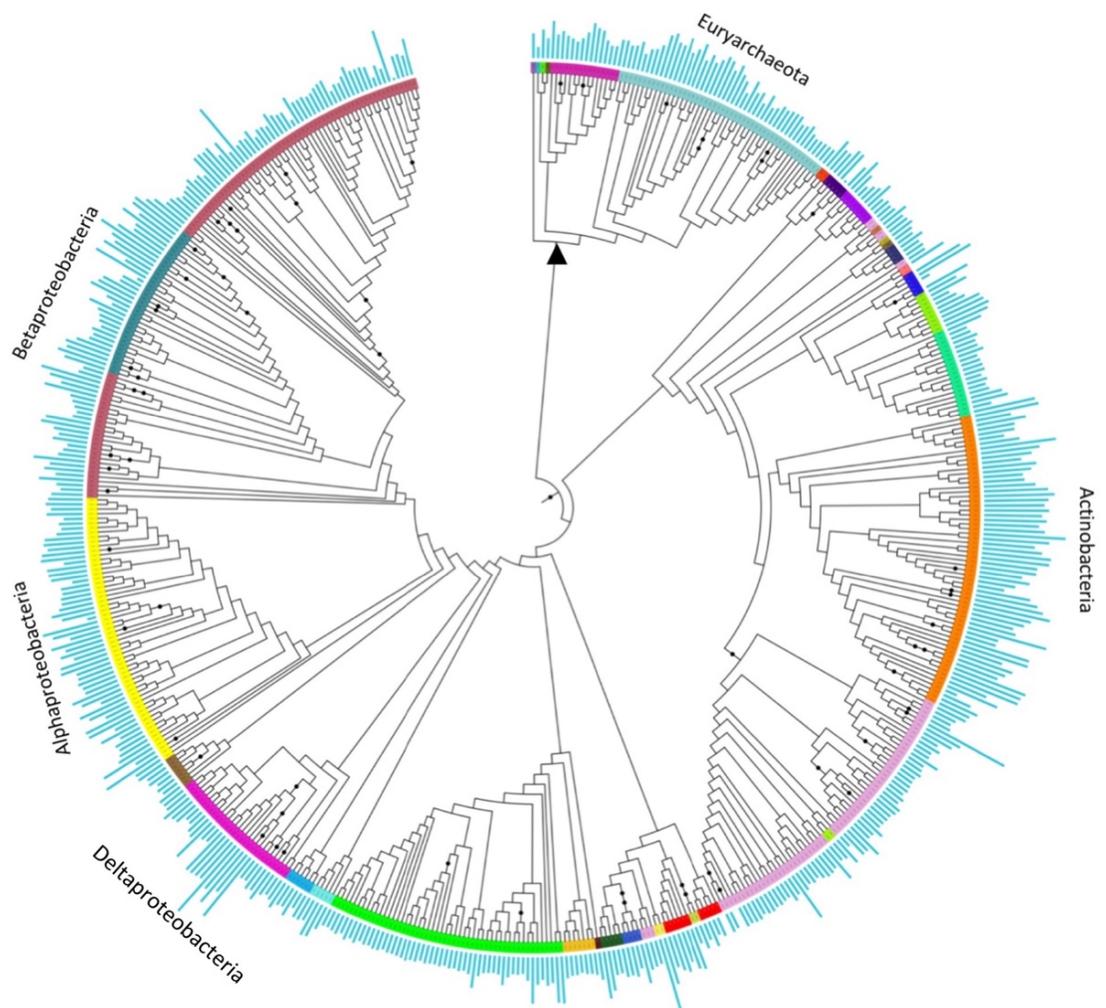


Figura 12: Árbol filogenético basado en 16S rRNA. Se muestran los porcentajes de proteínas con LCRs en barras de color azul claro. Los proteomas corresponden a bacterias y arqueas. El clado de Archaea está señalado con un triángulo negro. Algunos phyla con altos porcentajes de LCRs están indicados.

El organismo que presenta la mayor proporción de LCRs en su proteoma es la deltaproteobacteria *Anaeromyxobacter dehalogenans* (53.67 %). El procarionte que presenta menor porcentaje de secuencias simples es *Aerococcus viridans* (0.10 %) del phylum Firmicutes.

Contenido de GC en LCRs

El porcentaje de LCRs en los proteomas analizados parece estar relacionado con el contenido de GC. El coeficiente de Coeficiente de correlación de Pearson es positivo ($r=0.77$) (Fig. 13), es decir existe una asociación positiva entre ambos. La asociación del porcentaje de LCRs con el contenido de GC en Bacteria es $r=0.78$, mientras que para Archaea corresponde a $r=0.62$.

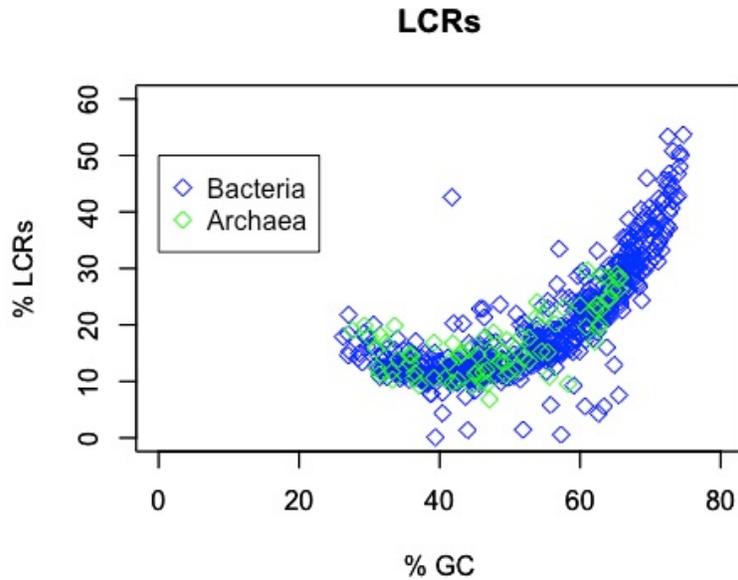


Figura 13: Asociación entre el contenido de GC y el porcentaje de LCRs en proteomas de Bacteria y Archaea.

La asociación positiva entre el porcentaje de LCRs y el contenido de GC, se aprecia también en la Figura 14, en donde a nivel de Phylum, aquellos con mayor porcentaje de GC (barras de color rojo) coinciden con aquellos Phyla con los porcentajes más altos de LCRs (barras de color azul), estos son Actinobacteria, Alphaproteobacteria, Deltaproteobacteria, Betaproteobacteria y Euryarchaeota.

Phyla	
■	Acidobacteria
■	Alphaproteobacteria
■	Actinobacteria
■	Armatimonadetes
■	Bacteroidetes
■	Betaproteobacteria
■	Aquificae
■	Chlorobi
■	Chloroflexi
■	Cyanobacteria
■	Deinococcus-Thermus
■	Deltaproteobacteria
■	Epsilonproteobacteria
■	Firmicutes
■	Fusobacteria
■	Gammaproteobacteria
■	Gemmatimonadetes
■	Nitrospirae
■	Planctomycetes
■	Spirochaetes
■	Synergistetes
■	Tenericutes
■	Thermodesulfobacteria
■	Verrucomicrobia
■	Korarchaeota
■	Crenarchaeota
■	Bathyarchaeota
■	Thaumarchaeota
■	Lokiarchaeota
■	Euryarchaeota
■	Elusimicrobia
■	Thermotogae
■	Dictyoglomi
■	Fibrobacteres
■	Caldiserica

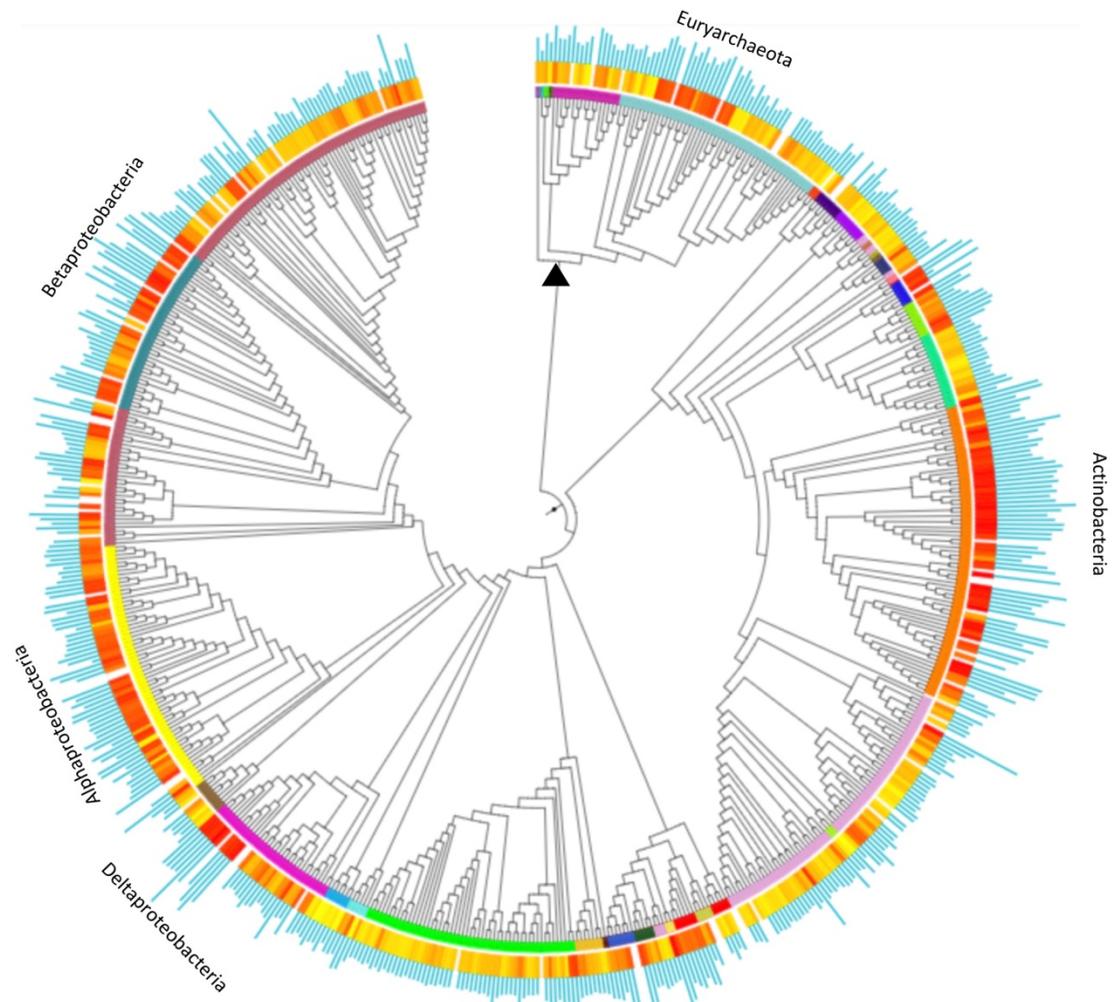


Figura 14: Árbol filogenético basado en 16S rRNA. Se muestra el contenido de GC en los proteomas en un gradiente de color, los valores de GC bajos corresponden a las barras color amarillo, los de valor medio de color naranja y los de valores altos de GC en color rojo. También se muestran los porcentajes de proteínas con LCRs en barras de color azul claro. Los proteomas corresponden a bacterias y arqueas. El clado de Archaea está señalado con un triángulo negro. Algunos phyla con valores altos de contenido de GC (barras de color rojo) y altos porcentajes de LCRs están indicados.

Composición de aminoácidos de LCRs

El sesgo composicional de los aminoácidos de las secuencias simples reveló que el aminoácido más abundante es alanina (A) en Bacteria y Archaea, también glicina (G), leucina (L) y valina (V) son frecuentes en ambos dominios celulares (Fig. 15). Sin embargo, es notable que, en las arqueas el número de los residuos Asp (D) y Glu (E) es más alto que en las bacterias. La composición de aminoácidos también refleja la ausencia de Cys (C), Phe (F), His (H), Ile (I), Met (M), Asn (N), Gln (Q), Thr (T), Trp (W) y Tyr (Y).

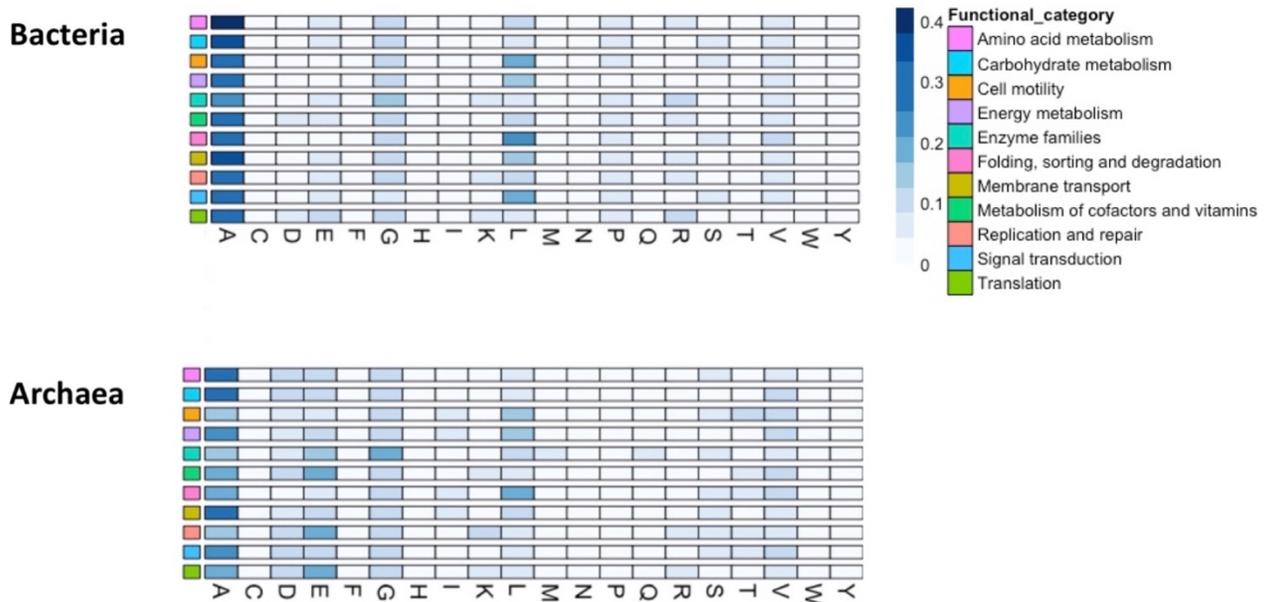


Figura 15: Composición de aminoácidos de las secuencias simples en en las clases funcionales más abundantes de proteínas con LCRs en Bacteria y Archaea.

Análisis de IDRs

A partir de la búsqueda de regiones intrínsecamente desordenadas en proteomas completos de 561 bacterias y 82 arqueas, se analizaron aquellas proteínas con IDRs en su secuencia. Utilizando el método de consenso en tres predictores de desorden (Fig. 16), se estudiaron aquellas proteínas con regiones intrínsecamente desordenadas de longitud igual o mayor a 30 aminoácidos. En total, fueron identificadas 76,607 proteínas con IDRs.

Es importante considerar que debido a que en este trabajo se realizó el análisis de aquellas IDRs con una longitud mayor o igual a 30 aminoácidos, existe la posibilidad de tener falsos negativos, ya que podrían haber IDRs de 29 residuos o menor longitud que no fueron considerados.

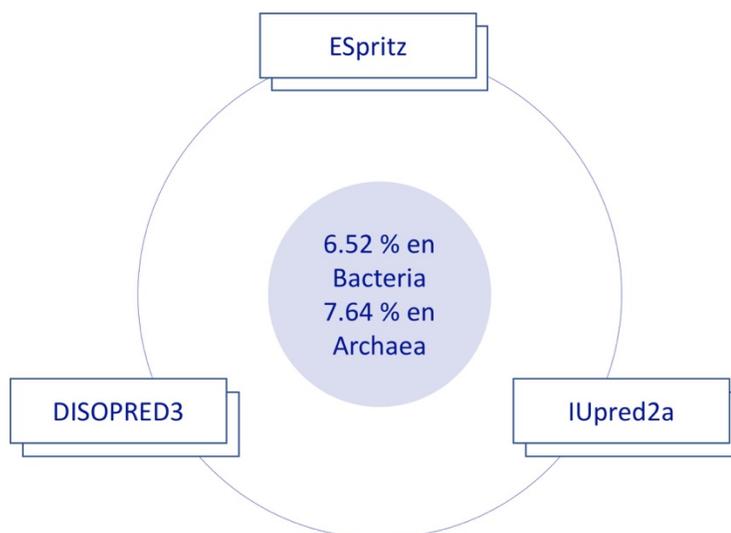


Figura 16: Resultados del método de consenso de tres predictores para la detección de regiones intrínsecamente desordenadas en los proteomas de Bacteria y Archaea. Se muestran los porcentajes de proteínas con IDRs respecto al total de las proteínas por dominio celular.

Distribución de IDRs en Bacteria

Al igual que las secuencias simples, las IDRs se encuentran en todos los proteomas de bacterias, y existe una diversidad de abundancia en todos los phyla (Fig. 17).

Los phyla bacterianos que presentan una mayor proporción de regiones intrínsecamente desordenadas en sus proteomas son Planctomycetes, Actinobacteria y Gemmatimonadetes (Fig. 17 y Anexo IV). En contraste, Dictyoglomi, Thermodesulfobacteria y Aquificae tienen porcentajes bajos de IDRs. Notablemente el phylum Deltaproteobacteria presenta la mayor variación de IDRs. Los grupos Tenericutes, Thermotogae y Aquificae presentan la menor variación.

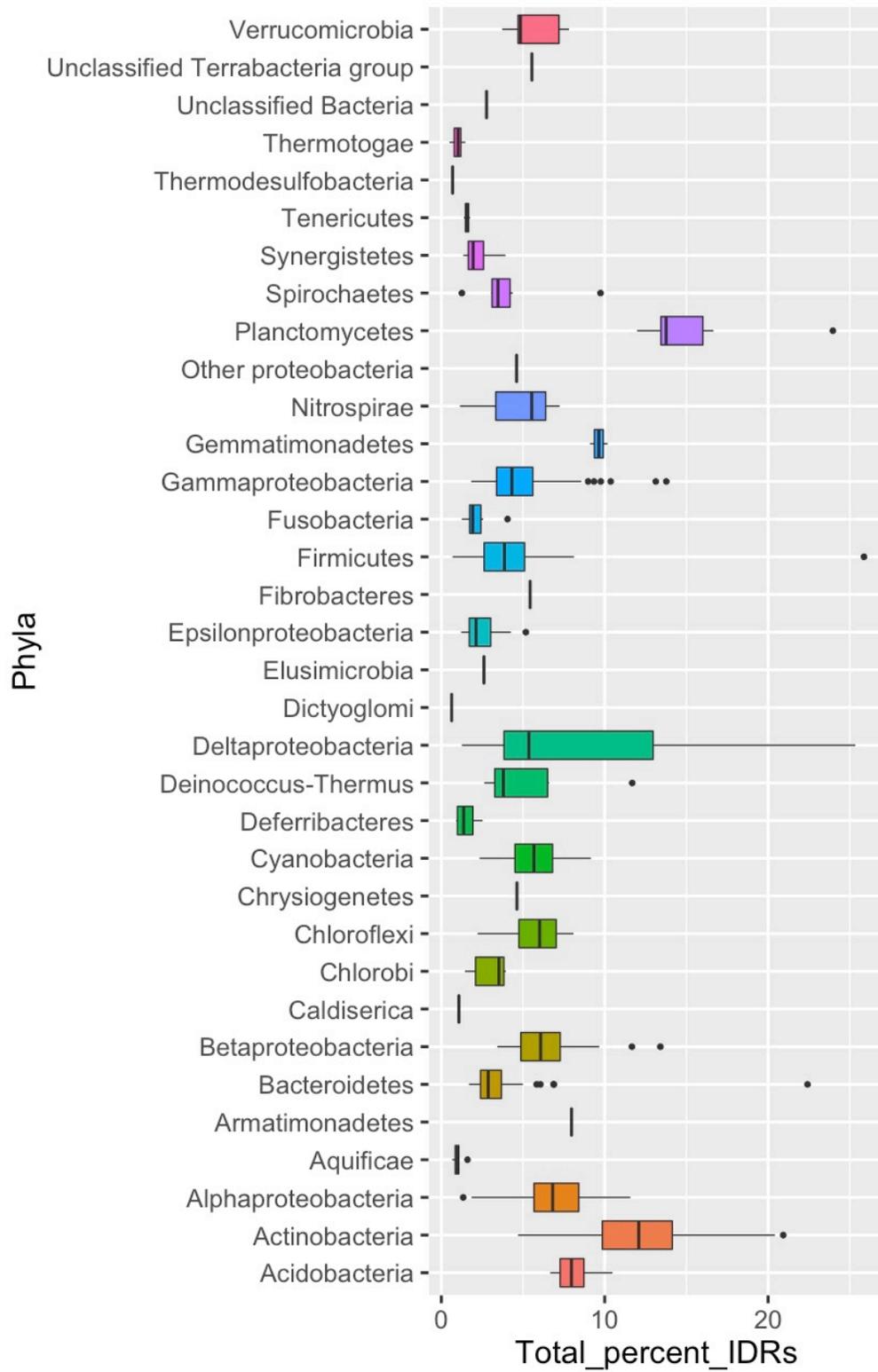


Figura 17: Porcentaje de regiones intrínsecamente desordenadas en bacterias. Los diagramas de caja y bigote muestran el valor mínimo, el primer cuartil, el segundo cuartil (la mediana), el tercer cuartil y el valor máximo.

Distribución de IDRs en Archaea

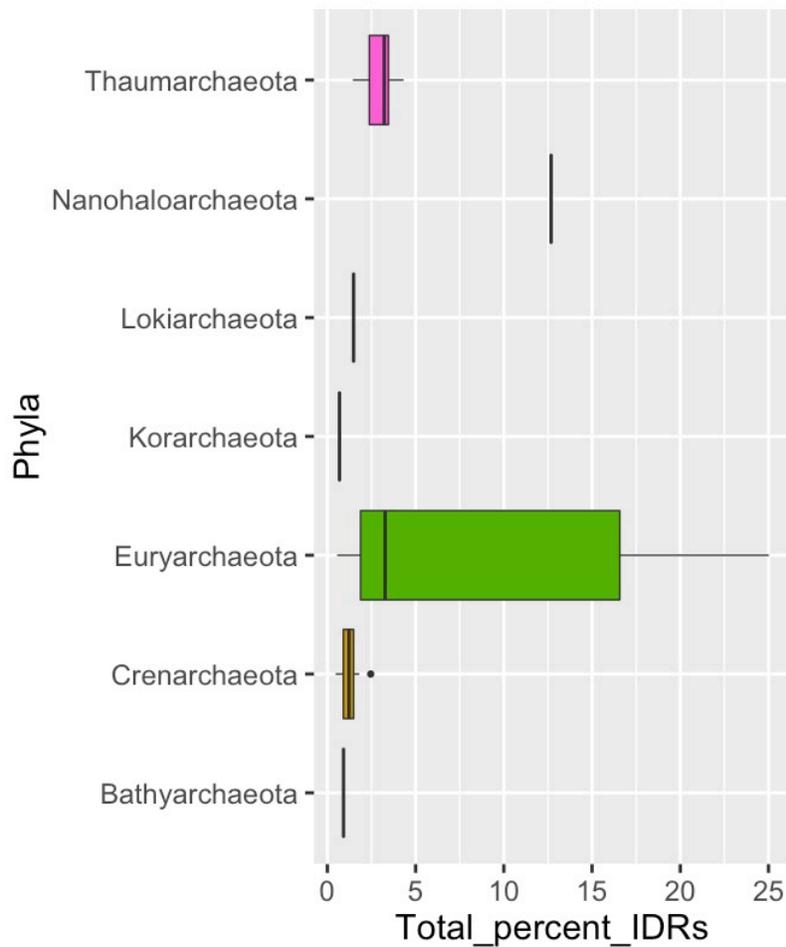


Figura 18: Porcentaje de regiones intrínsecamente desordenadas en arqueas. Los diagramas de caja y bigote muestran el valor mínimo, el primer cuartil, el segundo cuartil (la mediana), el tercer cuartil y el valor máximo.

Las arqueas que presentan la mayor proporción de regiones intrínsecamente desordenadas en sus proteomas son Nanohaloarchaeota y Euryarchaeota (Fig. 18 y Anexo IV). Aquellos phyla con menor proporción de IDRs son Korarchaeota y Bathyarchaeota. Al igual que con las LCRs, Euryarchaeota presenta la mayor variación en los porcentajes de IDRs.

Tamaño del proteoma vs el porcentaje de IDRs

Con el objetivo de evaluar la relación entre el tamaño del proteoma y el porcentaje de IDRs en Bacteria y Archaea, se utilizó el Coeficiente de correlación de Pearson (r). La asociación entre el tamaño del proteoma y el porcentaje de regiones intrínsecamente desordenadas en Bacteria y Archaea

es positiva ($r = 0.31$) (Fig. 19). Es importante mencionar que, aunque los valores de r impliquen una asociación positiva, es una relación lineal ascendente débil. La asociación del porcentaje de LCRs con el tamaño del proteoma en Bacteria es $r = 0.31$, mientras que para Archaea corresponde a $r = 0.64$.

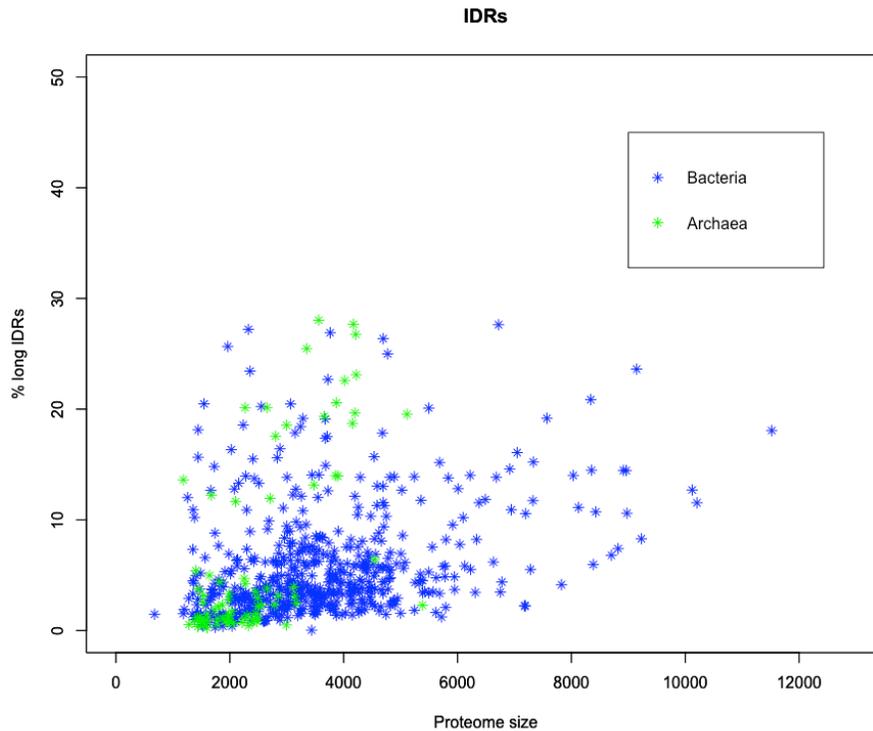


Figura 19: Asociación entre el tamaño del proteoma vs el porcentaje de regiones intrínsecamente desordenadas en Bacteria y Archaea

Clasificación funcional de las IDRs

El análisis funcional de las proteínas que presentan regiones intrínsecamente desordenadas demostró que son abundantes en procesos de *Traducción*, *Transporte de membrana* y *Replicación y Reparación* (Anexo V y Fig. 20) en los dos dominios celulares. Es importante mencionar que de las 76,607 proteínas con IDRs, únicamente 28,959 se encuentran anotadas funcionalmente.

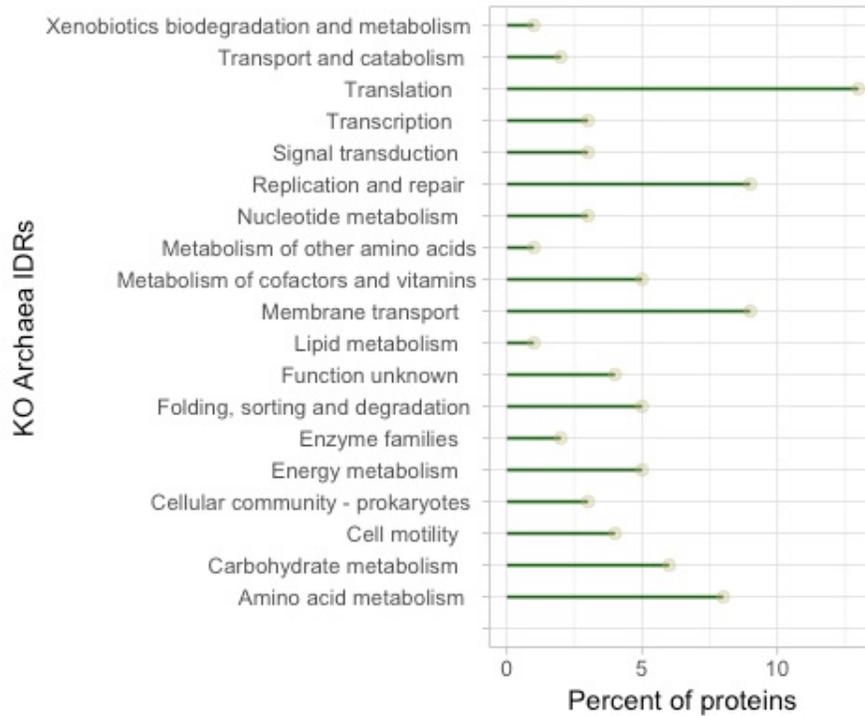
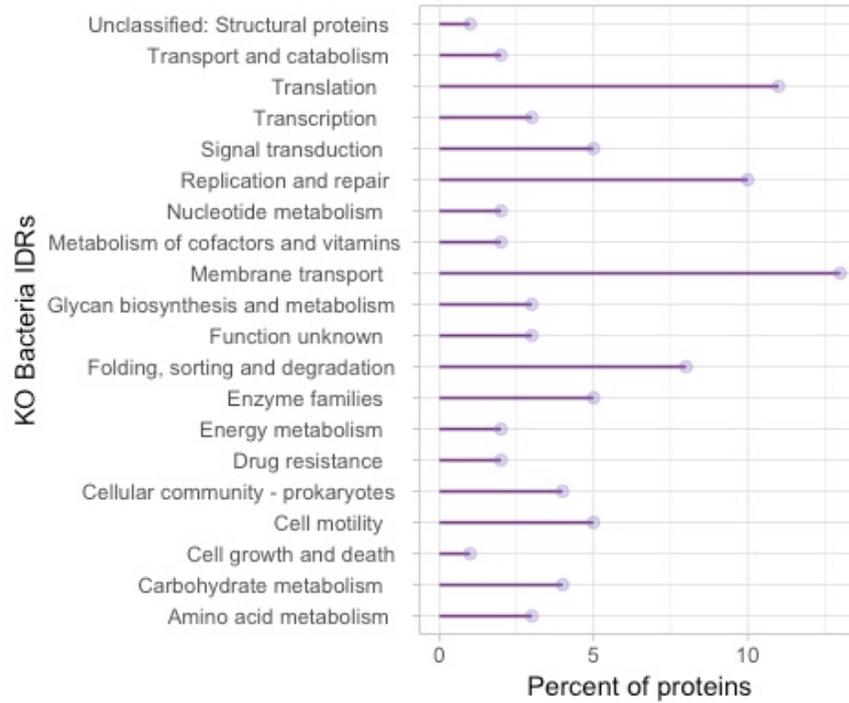


Figura 20: Categorías funcionales del segundo nivel de los grupos *KEGG orthology* (KO) de mayor abundancia en las proteínas con IDRs. En la parte superior se muestran los valores porcentuales pertenecientes a las categorías funcionales más abundantes en Bacteria y en la inferior, de Archaea.

A nivel de phylum de Bacteria y Archaea, las proteínas con IDRs presentan distintas proporciones de desorden (Fig. 21) en las categorías funcionales más abundantes (Fig. 20), a diferencia de las LCRs (Fig. 11) que presentan valores similares en sus proporciones.

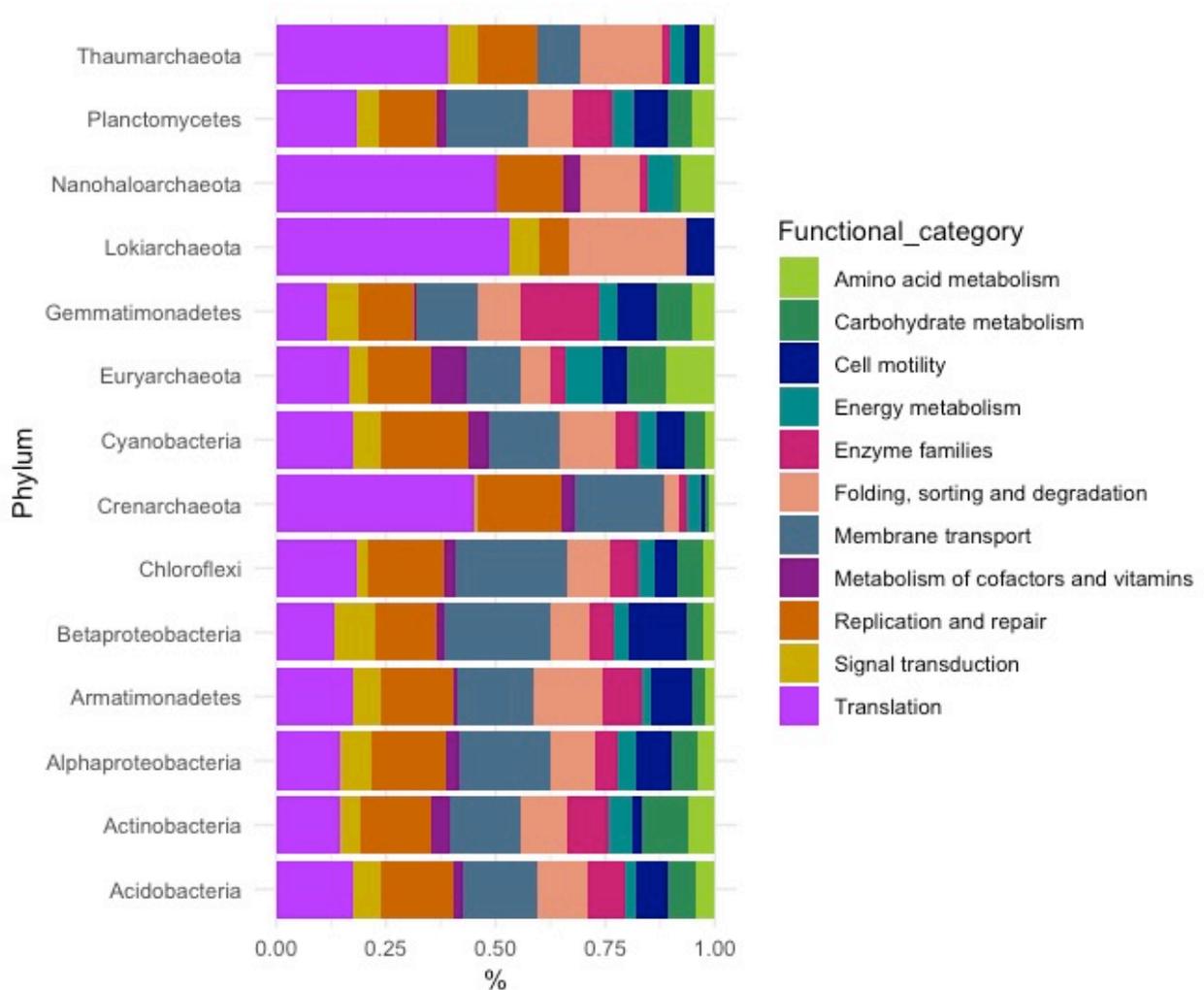


Figura 21: Funciones más abundantes en los phyla con mayor proporción de IDRs en sus proteomas. Se muestran los valores relativos.

Análisis evolutivo de las regiones intrínsecamente desordenadas

Como se aprecia en el árbol filogenético (Fig. 22), las proteínas con regiones intrínsecamente desordenadas forman parte de todos los proteomas de bacterias y arqueas analizados. Las barras de color azul rey representan los porcentajes de IDRs por organismo.

Phyla	
■	Acidobacteria
■	Alphaproteobacteria
■	Actinobacteria
■	Armatimonadetes
■	Bacteroidetes
■	Betaproteobacteria
■	Aquificae
■	Chlorobi
■	Chloroflexi
■	Cyanobacteria
■	Deinococcus-Thermus
■	Deltaproteobacteria
■	Epsilonproteobacteria
■	Firmicutes
■	Fusobacteria
■	Gammaaproteobacteria
■	Gemmatimonadetes
■	Nitrospirae
■	Planctomycetes
■	Spirochaetes
■	Synergistetes
■	Tenericutes
■	Thermodesulfobacteria
■	Verrucomicrobia
■	Korarchaeota
■	Crenarchaeota
■	Bathyarchaeota
■	Thaumarchaeota
■	Lokiarchaeota
■	Euryarchaeota
■	Elusimicrobia
■	Thermotogae
■	Dictyoglomi
■	Fibrobacteres
■	Caldiserica

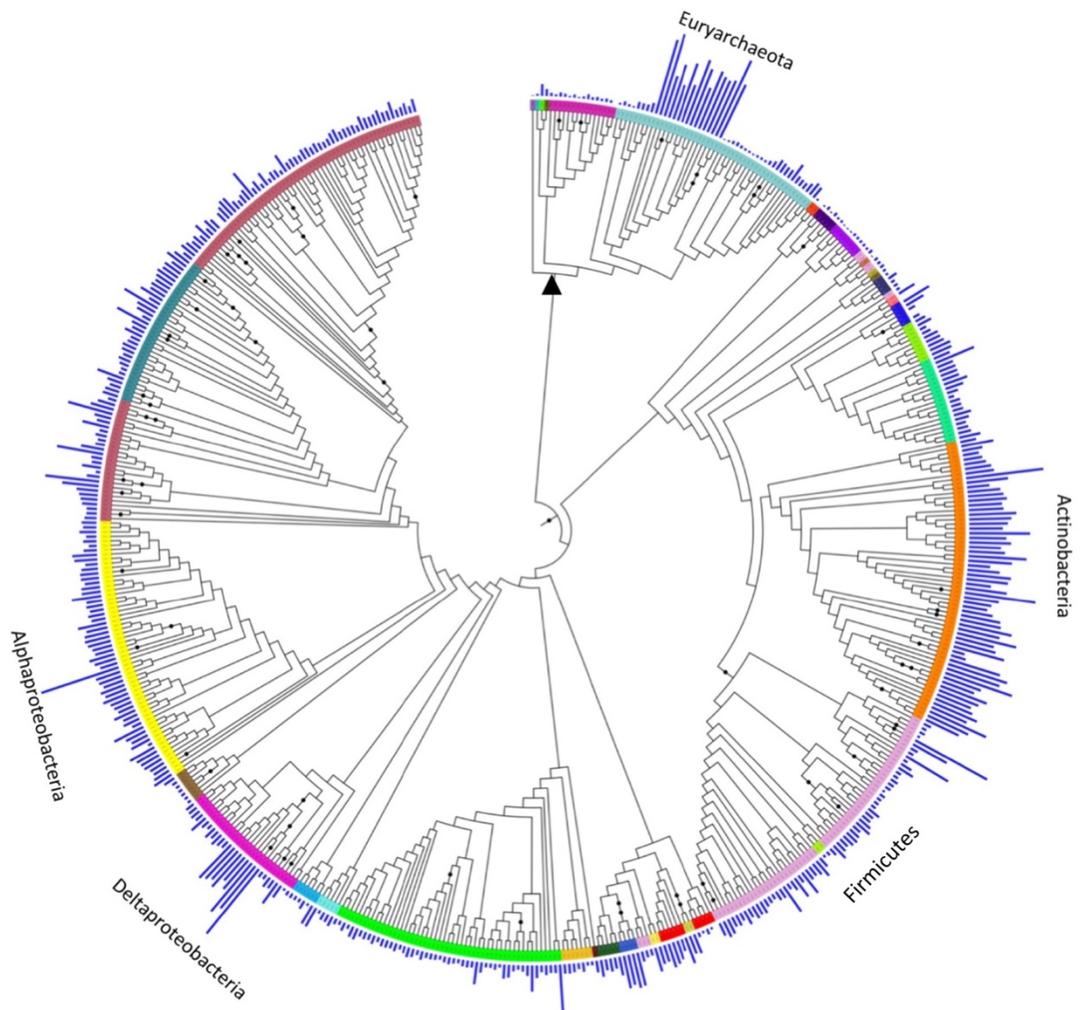


Figura 22: Árbol filogenético basado en 16S rRNA. Los porcentajes de proteínas con IDRs de 30 o más residuos de longitud, se muestran en barras color azul rey. Los proteomas corresponden a bacterias y arqueas. El clado de Archaea esta señalado con un triángulo negro. Algunos phyla con altos porcentajes de LCRs están indicados.

El organismo que presenta el porcentaje más alto de IDRs en su proteoma es *Thermaerobacter marianensis* (25.87 %) del phylum Firmicutes. La crenarqueota *Sulfolobus*

solfataricus contiene la menor proporción de desorden (0.46 %). Es importante mencionar que al igual que las LCRs (Fig. 12), las IDRs (Fig. 22) presentan altos porcentajes en los phyla Actinobacteria, Alphaproteobacteria y Euryarchaeota.

Contenido de GC en IDRs

El porcentaje de IDRs en los proteomas analizados parece estar relacionado con el contenido de GC en los proteomas (Fig. 23). Aunque el coeficiente de Coeficiente de correlación de Pearson es positivo ($r= 0.67$), la asociación es débil, a diferencia de lo observado en las LCRs (Fig. 13). La asociación del porcentaje de IDRs con el contenido de GC en Bacteria es $r= 0.69$, mientras que para Archaea corresponde a $r= 0.70$.

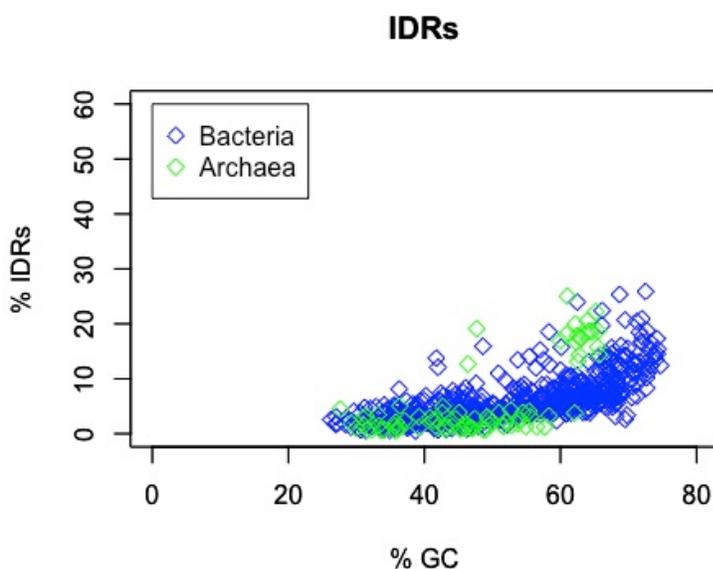


Figura 23: Asociación entre el contenido de GC y el porcentaje de IDRs en proteomas de Bacteria y Archaea.

La asociación positiva entre el porcentaje de IDRs y el contenido de GC en los proteomas, se observa en la Figura 24, en donde aquellos Phyla con los mayores porcentajes de IDRs (barras de color azul rey) también presentan los valores más altos de contenido de GC (barras de color rojo), estos son Actinobacteria, Alphaproteobacteria, Deltaproteobacteria y Euryarchaeota.

Phyla	
■	Acidobacteria
■	Alphaproteobacteria
■	Actinobacteria
■	Armatimonadetes
■	Bacteroidetes
■	Betaproteobacteria
■	Aquificae
■	Chlorobi
■	Chloroflexi
■	Cyanobacteria
■	Deinococcus-Thermus
■	Deltaproteobacteria
■	Epsilonproteobacteria
■	Firmicutes
■	Fusobacteria
■	Gammaproteobacteria
■	Gemmatimonadetes
■	Nitrospirae
■	Planctomycetes
■	Spirochaetes
■	Synergistetes
■	Tenericutes
■	Thermodesulfobacteria
■	Verrucomicrobia
■	Korarchaeota
■	Crenarchaeota
■	Bathyarchaeota
■	Thaumarchaeota
■	Lokiarchaeota
■	Euryarchaeota
■	Elusimicrobia
■	Thermotogae
■	Dictyoglomi
■	Fibrobacteres
■	Caldiserica

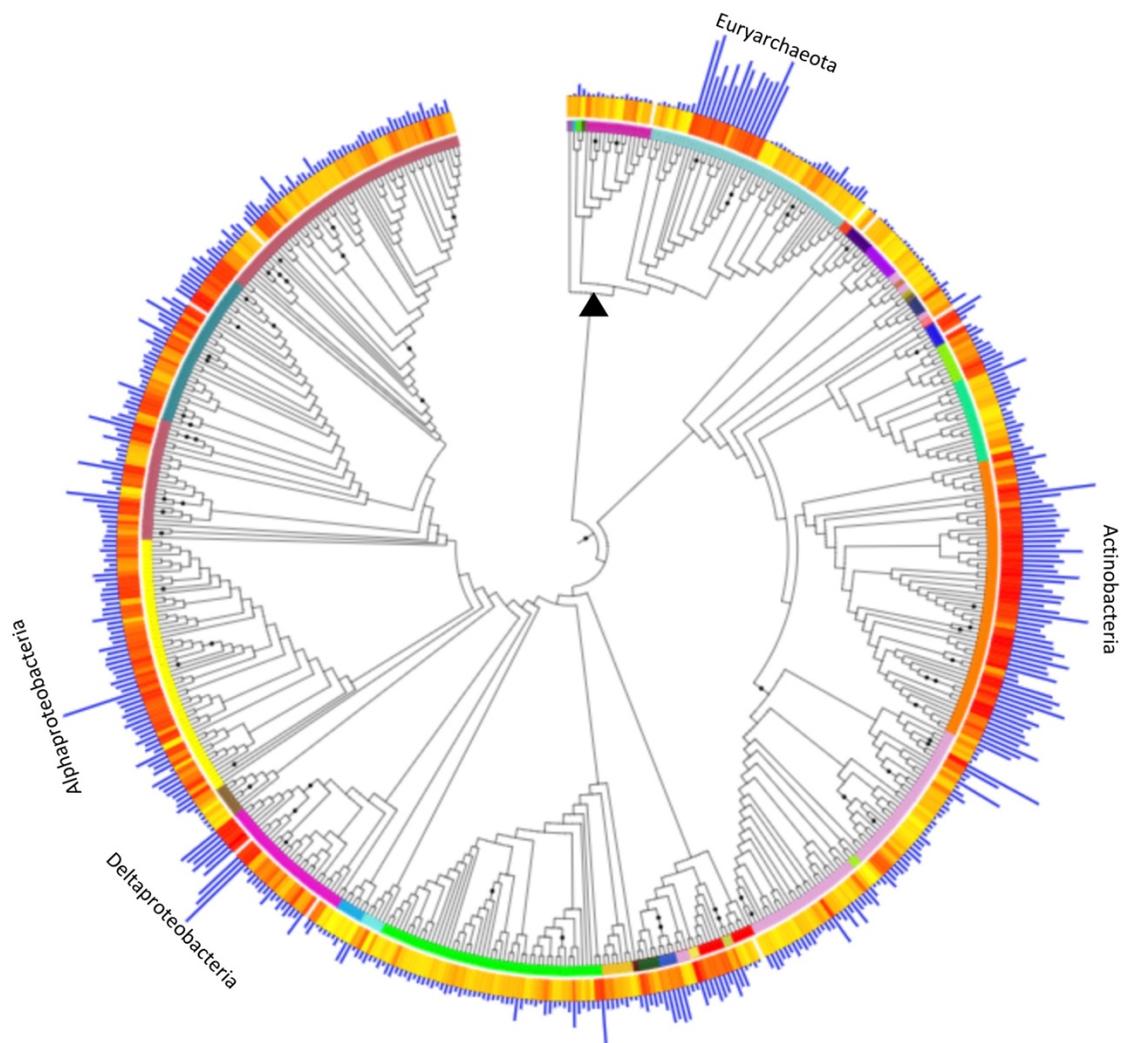


Figura 24: Árbol filogenético basado en 16S rRNA. Se muestra el contenido de GC en los proteomas en un gradiente de color, los valores de GC bajos corresponden a las barras color amarillo, los de valor medio de color naranja y los de valores altos de GC en color rojo. También se muestran los porcentajes de proteínas con IDRs en barras de color azul rey. Los proteomas corresponden a bacterias y arqueas. El clado de Archaea esta señalado con un triángulo negro. Algunos phyla con valores altos de contenido de GC (barras de color rojo) y altos porcentajes de IDRs están indicados.

Composición de aminoácidos de IDRs

La composición de aminoácidos en las regiones intrínsecamente desordenadas reveló su abundancia en Ala (A), Asp (D), Glu (E), Gly (G), Leu (L), Pro (P), Arg (R), Ser (S), Thr (T) y Val (V) en bacterias y arqueas (Fig. 25). La Lys (K) es frecuente en la Traducción, mientras que la Gln (Q) se encuentra en más categorías funcionales de Bacteria, en comparación con Archaea. Es notable la ausencia de Cys (C), Phe (F), His (H), Ile (I), Met (M), Asn (N), Trp (W) y Tyr (Y).

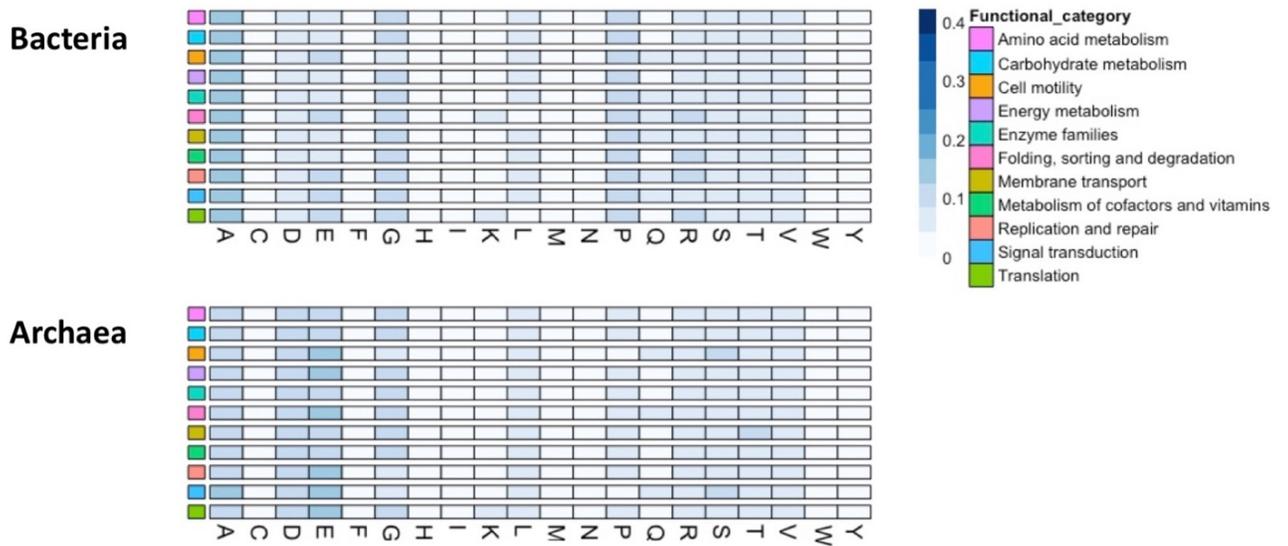
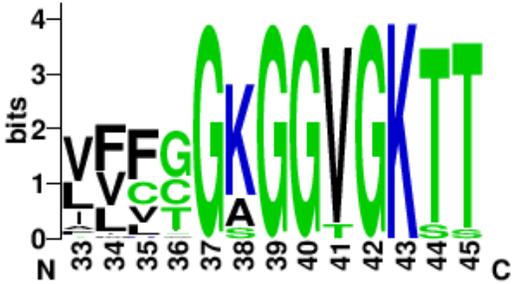
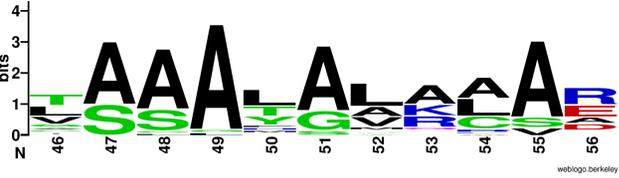
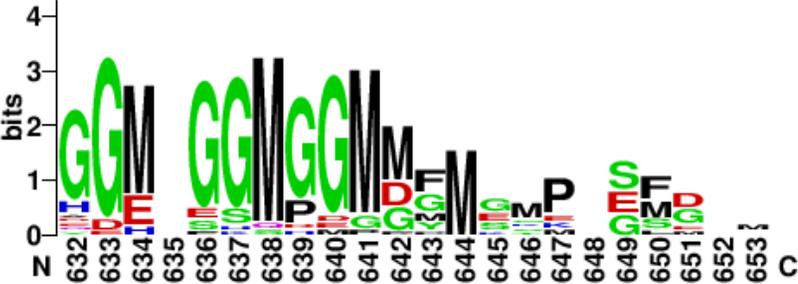


Figura 25: Composición de aminoácidos en regiones intrínsecamente desordenadas en las clases funcionales más abundantes de proteínas con IDRs en Bacteria y Archaea.

Secuencias simples con amplia distribución filogenética

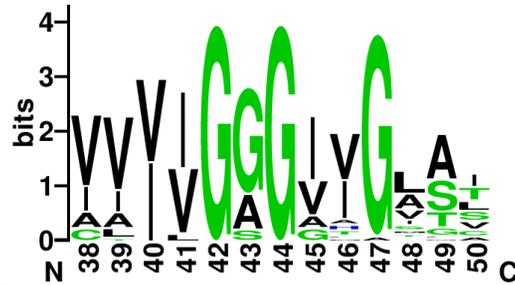
La búsqueda de LCRs en proteomas completos de bacterias y arqueas reveló su presencia en todos estos (Fig. 12). Sin embargo, únicamente 42 presentan una amplia distribución filogenética (Anexo VI), en la Tabla IV se muestran algunos ejemplos.

Tabla IV: LCRs con amplia distribución filogenética

Proteína	LCR	Posición en la secuencia	Dominio de localización de la LCR	Categoría funcional
ATPasa transportadora de arseniato (EC: 3.6.3.16)	<p>1)</p> 	N-terminal	ArsA_ATPase	Procesos celulares y de señalización
	<p>2)</p> 	C-terminal	ninguno	Plegamiento, modificación y degradación
Chaperona HSP60		C-terminal	ninguno	Plegamiento, modificación y degradación

D-aminoácido

**deshidrogenasa (EC:
1.4.5.1)**

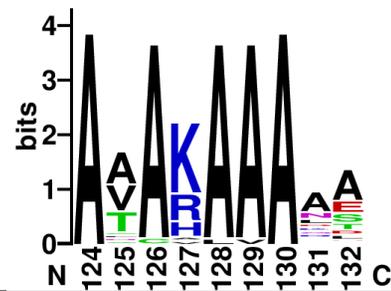


N-terminal

DAO

Metabolismo de
aminoácidos

Enolasa (EC: 4.2.1.11)



Central

Enolase_N

Metabolismo de
carbohidratos

**Metiltioribosa-1-fosfato
isomerasa (EC: 5.3.1.23)**



Central

IF-2B

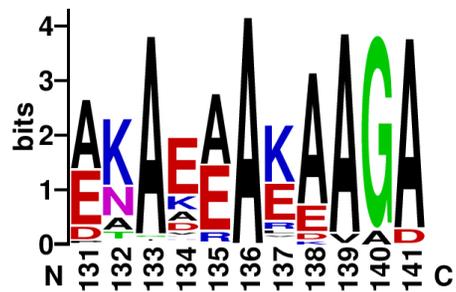
Metabolismo de
aminoácidos

Proteína ribosomal L1

Central

Ribosomal_L1

Traducción



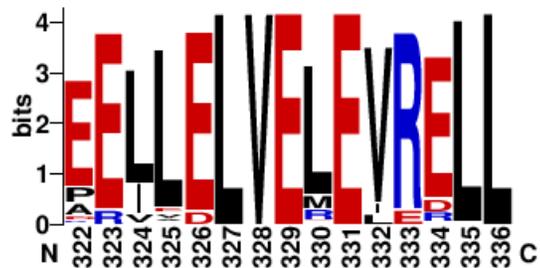
Factor de elongación

Central

GTP_EFTU

Traducción

EF-Tu



Conservación de las LCRs con altos niveles de distribución

Los alineamientos múltiples de las proteínas con LCRs de amplia distribución filogenética permitieron analizar la conservación de las secuencias simples en Bacteria y Archaea. Adicionalmente, es posible identificar aquellos motivos de repeticiones de aminoácidos que caracterizan a las secuencias simples y determinan su sesgo composicional. A continuación, se muestran algunos alineamientos representativos de LCRs.

ATPasa transportadora de arseniato

	LCR 1	LCR 2
Anaeromyxobacter dehalogenans	CVGAGGVGKTTAAAAALALSRALE	
Hoyosella subflava	CCGAGGVGKTTTAAALALRAADM	
Gordonia bronchialis	CCGAGGVGKTTTAAAMAMYAAEH	
Acidothermus cellulolyticus	CCGSGGVGKTTLAAALGLRAAEA	
Stackebrandtia nassauensis	CCGSGGVGKTTTAAALALRAAEV	
Salinispora tropica	CCGAGGVGKTTTAAALALRAAEE	
Luteipulveratus mongoliensis	CCGSGGVGKTTTAAAVAVRAAEA	
Corallococcus coralloides	LCGAGGVGKTTTAAALGVAAARS	
Myxococcus xanthus	LCGAGGVGKTTTAAALGVAAARA	
Stigmatella aurantiaca	LCGAGGVGKTTTAAALGVAAARA	
Actinosynnema mirum	FTGKGGVGKTTLAAATAARLAES	
Saccharothrix espanaensis	FTGKGGVGKTTLAAATSAALAAG	
Kutzneria albida DSM 43870	FTGKGGVGKTTLAAATAAALVEG	
Saccharopolyspora erythraea	FTGKGGVGKTTLAAATAARLAAR	
Amycolatopsis mediterranei	CTGKGGVGKTTLAAATGAALAL	
Nakamurella multipartita	HTGKGGVGKTTISAATAIACAAG	
Saprospira grandis	FTGKGGVGKTTSSAAATALAAAQ	
Prosthecochloris aestuarii	FTGKGGVGKTTIAAATALKAAGM	
Halogeometricum borinquense	FGGKGGVGKTTLASAYALKCARE	
Haloferax volcanii	FGGKGGVGKTTMSSAYAVKCARD	
Natronobacterium gregoryi	FGGKGGVGKTTVSSAHALQCEG	
Natronococcus occultus SP4	FGGKGGVGKTTVSSAYALNCVDA	
Natrialba magadii	FGGKGGVGKTTVSSAYALECATV	
Natrinema pellirubrum	FGGKGGVGKTTVSSAYGLKCARD	
Halanaeroarchaeum sulfurireducens	FGGKGGVGKTTVSSAYGLKCARD	
Alcanivorax borkumensis	VGGKGGVGKTTTASALAVRAADQ	
Truepera radiovictrix	VGGKGGVGKTTTAAALALQWARR	
Aminobacterium colombiense	FGGKGGTGKTTCAAAYAYALSRL	
Acetomicrobium mobile	FGGKGGTGKTTCAAFLSKASRM	
Halobacillus halophilus	VGGKGGVGKSTSSAAMAAFARE	
Oceanobacillus iheyensis	VGGKGGVGKSSSAAAIAWKLAKE	
Aquifex aeolicus	FGGKGGVGKTTASSAFVKLSEQ	
Thermovibrio ammonificans	LSGKGGVGKSTLSAALAAALSRK	
Brevibacterium linens	VGGKGGTGKTTVSSGLAMARADD	
Thermobifida fusca	FGGKGGVGKTTLAAHALALADS	
Thermosphaera aggregans	VIGKGGVGKTTVSLLLGSALARL	

Figura 26: Alineamiento múltiple de las LCRs de la proteína ATPasa transportadora de arseniato. Ambas LCRs forman parte del dominio ArsA_ATPase y se localizan en la región N-terminal de la secuencia. Se observa que los motivos de repeticiones AA, LL, GG, FF, CC y AA son frecuentes. Las arqueas se encuentran señaladas en negritas.

Chaperona HSP60

Acidipropionibacterium acidipropionici	GG---DMDGM-GGMGGMM-----
Koribacter versatilis	AG-G-HGGGM-GGM----Y-----
Coraliomargarita akajimensis	AM-P-DMGGM-GGMGGMM-----
Asticcacaulis excentricus	-M-P--GGGM-GGMGGMDF-----
Methylocella silvestris	---P-GGGGM-GGMGGMDF-----
Comamonas testosteroni	-M-P-DMGGM-GGMGGMGM-----
Gallionella capsiferriformans	MG---GGDM-GGMGGMM-----
Pseudomonas aeruginosa	GM-P-DMGGM-GGMGGMM-----
Bacteroidales bacterium	GA-P----GM-GGMGGMM-----
Zunongwangia profunda	-M-P---QGMGGGMPGMM-----
Chloroherpeton thalassium	-M-P-PGGGM-GGMGGMGGM-----Y
Anaeromyxobacter dehalogenans	AG----GAGMGGGMDMDY-----
Alkaliphilus metalliredigens	GG----MGGM-GGMGGGMPM-----M
Oceanobacillus iheyensis	GGMP-DMGGM-GGMGGMM-----
Salinispira pacifica	AA----GGGM-GGMGGMPMM-----
Truepera radiovictrix	AGAG-AGGGM-GGMGGMDF-----
Arcobacter butzleri	SM-P-DMGGM--GMPGMM-----
Pseudothertotoga lettingae	PAMP-PEY-----
Fimbriimonas ginsengisoli	GGGG-HSHG--GGMGDMDF-----
Isosphaera pallida	HH-H-HHHDH-GGMGMM-----
Arthrospira platensis	DM-D-GMGGM-GGMGGMGGMGGMGGM
Verrucosispora maris	AGHG-HGHGH-SHQHGPGF-----
Aciduliprofundum boonei	GGMPGGMGGM-GGMPGMM-----
Archaeoglobus fulgidus	GGMP-----EMP-EF---
Candidatus Bathyarchaeota archaeon	GLEEEEGGGE-----EF---
Lokiarchaeum sp. GC14_75	GGMP-PGGGM-GGMGGGMG-GMPPGMM--
Candidatus Nitrosopelagicus brevis	GGMP-DMGGM-GGMPGMGGMPPGMMG-M
Desulfurococcus amylolyticus	GKKE----GE-ES-----KMP-SFD--
Sulfolobus solfataricus	GSKE-ESGGE-GG-----STP-SLG-D
Vulcanisaeta distributa	GKSKTESGGE-ES-----ESK-SSE--
Caldisphaera lagunensis	GPSSKSSEEE-SS-----SKE-SSD--
Korarchaeum cryptofilum	GEEEEEGGGE-FKS-----EFD--

Figura 27: Alineamiento múltiple de la LCR de la proteína Chaperona HSP60. La secuencia simple no forma parte de ningún dominio proteínico y se localiza en el extremo C-terminal de la secuencia. Los motivos de repeticiones GG, MM y GGM son abundantes. Las arqueas se encuentran señaladas en negritas.

D-aminoácido deshidrogenasa

Actinosynnema mirum	VVVVGGGAMGSAA
Conexibacter woesei	VVIIIGGGIGASA
Pyrococcus horikoshii	IVVIGGGIVGVTI.
Palaeococcus pacificus DY20341	ITIIGGGIIGATL
Kosmotoga olearia	VVIIIGGGIIGTAL.
Thermofilum pendens	VVIVGGGIVGVSL
Sphaerobacter thermophilus	VVVVGGGVIGCAS.
Caldivirga maquilingensis	VVVVGGGVVGLAT.
alpha proteobacterium HIMB59	AIIGGGGHGLGT.
Serratia sp. FGI94	VIIIGGGHGLGT.
Agrobacterium fabrum	VVVIGGGIVGTST.
Bordetella pertussis	VAIIGAGAAGVAT.
Collimonas pratensis	VVVIGGGIIGIFT.
Brevibacillus brevis	IAVIGGGIIGAAI.
Acidiphilium multivorum	VAIIGGGIIGLAL
Chloroherpeton thalassium	VVIIIGGGIIGLSL
Crinalium epipsammum PCC 9333	ILIIIGSGIIGLSL
Trichodesmium erythraeum	IIIIGGGIIGISI.
Atelocyanobacterium thalassa	VIIIGGGIIGMSI.
Thermocrinis albus	VLIVGGGILGLSS
Jeotgalibacillus malaysiensis	AIVVGGGVIGGAI.
Halogeometricum borinquense	AVVVGGGIVGSSV.
Natrialba magadii	AVVVGGGIVGSSI.
Dokdonia donghaensis DSW-1	IIIIGGGIVGVSC.
Leadbetterella byssofila	VVIVGGGIVGLSS.
Rubinisphaera brasiliensis	VVIVGGGIIGIAS.
Planctopirus limnophila	VVVVGGGVVGAMC.
Paraburkholderia xenovorans	VVVLGSGVVGVT.
Cedecea neteri	VVVLGSGVVGVAS.
Atelocyanobacterium thalassa	IVIIIGAGIIGATI.
Chloroflexus aurantiacus	VVVVGAGVVGAAAT.
Desulfovibrio vulgaris	IAIIGGGGTAAL
Haloarcula hispanica	VLVVGGGATGAGV.
Methylacidiphilum infernorum	CIIIIGAGIAGATA
Azospirillum lipoferum	CVVVGAGVVGLAV.
Alicyclophilus denitrificans	CVVVGAGVVGLAV.
Delftia acidovorans	CVVVGAGVVGLAV.
Accumulibacter phosphatis	AAVIGAGVVGLAC.
Caulobacter vibrioides	VVVVGAGAVGLAC.

Figura 28: Alineamiento múltiple de la LCR de la proteína D-aminoácido deshidrogenasa. La secuencia simple se encuentra en el dominio DAO y se localiza en el extremo N-terminal de la secuencia. Los motivos de repeticiones **VV**, **GGG** e **II** son frecuentes. Las arqueas se encuentran señaladas en negritas.

Enolasa

Acidipropionibacterium acidipropionici	AAARAAAAS
Saccharothrix espanaensis	AVAKAAAAS
Acholeplasma laidlawii	AAAKAAADL
Acidithiobacillus ferrooxidans	ATAHAAAHA
Laribacter hongkongensis	AVAKAAALE
Candidatus Tenderia electrophaga	AAAKAAAEE
Carboxydotherrnus hydrogenoformans	AVAKAAANY
Flexistipes sinusarabici	ACAKAAAADA
Blastochloris viridis	ATAKAAAAA
Methyloceanibacter caenitepidi	AAAKAAAANA
Akkermansia muciniphila	ALAKAAAQ
Coraliomargarita akajimensis	ATAKAAAALA
Flavobacteriaceae bacterium	AAAKAAAEE
Sphaerobacter thermophilus	ATARAAAAA
Thermomicrobium roseum	AVARAAAAA
Truepera radiovictrix	AAARAVAQT
Solibacter usitatus	ATARAAAAA
Saccharophagus degradans	AAAKAAAQD
Salmonella typhi	ANAKAAAAA
Rubinisphaera brasiliensis	AAAHAAART
Planctopirus limnophila	AAAHAAARA
Stanieria cyanosphaera	ATAKAAAEE
Mesoplasma florum	AAAHAAASE
Methanoculleus marisnigri	AVARAAAAA
Methylococcus capsulatus	AVAAAAAKS
Fibrobacter succinogenes	AVCVAAAKD
Pyrolobus fumarii	AVAKAAAAT
Conexibacter woesei	ATARAAAALA
Lokiarchaeum sp. GC14_75	AVAKLAAVL
Ferroglobus placidus	AAAKAAANS
Geoglobus acetivorans	ASAKAAAANA

Figura 29: Alineamiento múltiple de la LCR de la proteína Enolasa. La secuencia simple se encuentra en el dominio Enolase_N. Los motivos de repeticiones de AAA, son altamente frecuentes. Las arqueas se encuentran señaladas en negritas.

Metiltioribosa-1-fosfato isomerasa

<i>Alicyclobacillus acidocaldarius</i>	GAPAI GAAAAFGLALE
<i>Desulfobacca acetoxidans</i>	GAPAI GVAAAMAAALG
<i>Desulfuromonas soudanensis</i>	GAPAI GVAAAFGAAFG
<i>Geoalkalibacter subterraneus</i>	GAPAI GVAAAYGAAFG
<i>Pelobacter carbinolicus</i>	GAPAI GVAAAYGAALG
<i>Desulfotomaculum reducens</i>	GAPAI GAAAAYGLVVG
<i>Flavonifractor plautii</i>	GAPAI GVAAAYAYCLA
<i>Caldanaerobacter subterraneus</i>	GAPAI GAAAAYGVVLA
<i>Oceanithermus profundus</i>	GAPAI GAAAAFGVVLA
<i>Fimbriimonas ginsengisoli</i>	GAPAI GVAAAYGLALA
<i>Leptospirillum ferrooxidans</i>	GAPAI GIAAAYGIAIG
<i>Halorhodospira halophila</i>	GAPAI GVAAGYGAALA
<i>Eggerthella lenta</i>	GAPAI GVAGAAVALW
<i>Sphaerobacter thermophilus</i>	GAPAI GIAAAGMAIA
<i>Palaeococcus pacificus</i>	GAPAI GAAAAYGLALL
<i>Thermococcus kodakarensis</i>	GAPAI GAAAAFGLALY
<i>Amycolatopsis mediterranei</i>	GAPALGGAGALGVALS
<i>Cronobacter sakazakii</i>	GAPLIGLSASLLLALL
<i>Klebsiella pneumoniae</i>	GAPLIGLSASLLLALL
<i>Dickeya dadantii</i>	GAPLIGLSASLLLALL
<i>Thermofilum pendens</i>	GAPAI GVAAAYAVALF
<i>Methanohalophilus mahii</i>	GAPALAAAGAYGIALA
<i>Hyperthermus butylicus</i>	GAPAI GVAAAYGLALA
<i>Streptomyces coelicolor</i>	GAPLLGIAGGYGVALA
<i>Blastochloris viridis</i>	GAPLIGAAAAYGLALG
<i>Opitutus terrae</i>	GAPLIGATAAWGLWLA
<i>Salinivirga cyanobacteriivorans</i>	GAGAIGGAAAFAMAQA
<i>Natrialba magadii</i>	GAAAIADAAAAALATQ
<i>Natrinema pellirubrum</i>	GAATIADAAAAALATQ
<i>Pyrodictium delaneyi</i>	GAGRIARAAARALMIA

Figura 30: Alineamiento múltiple de la LCR de la proteína Metiltioribosa-1-fosfato isomerasa. La secuencia simple se encuentra en el dominio IF-2B. Los motivos de repeticiones de AA y LL, son frecuentes. Las arqueas se encuentran señaladas en negritas.

Regiones intrínsecamente desordenadas con amplia distribución filogenética

Las regiones intrínsecamente desordenadas forman parte de los proteomas de todos los phyla de Bacteria y Archaea (Fig. 22) pero sólo 22 presentan una amplia distribución filogenética (Anexo VII); en la Tabla V se muestran algunos ejemplos.

Tabla V: IDRs con amplia distribución filogenética

Proteína	Clasificación de IDR	Posición en la secuencia	Dominio de localización de la IDR	Categoría funcional
Proteína chaperona DnaK (HSP70)	1) IDR conservada en secuencia 2) IDR conservada	1) Central 2) C-terminal	1) HSP70 2) HSP70	Plegamiento, modificación y degradación
Proteína ribosomal L2	1) IDR conservada 2) IDR conservada en secuencia	1) N-terminal 2) C-terminal	1) Ribosomal_L2 2) Ribosomal_L2_C	Traducción
Proteína ribosomal L3	IDR conservada en secuencia	Central	Ribosomal_L3	Traducción
Proteína ribosomal L4	IDR conservada en secuencia	Central	Ribosomal_L4	Traducción
Proteína ribosomal S3	IDR conservada	C-terminal	ninguno	Traducción
Proteína ribosomal S4	IDR conservada	N-terminal	Ribosomal_S4	Traducción

Proteína ribosomal S9	IDR conservada	C-terminal	Ribosomal_S9	Traducción
Proteína ribosomal S13	IDR conservada en secuencia	C-terminal	Ribosomal_S13	Traducción
Partícula de reconocimiento de señal (EC:3.6.5.4)	IDR conservada	C-terminal	SRP_SPB	Plegamiento, modificación y degradación
Receptor de la Partícula de reconocimiento de señal	IDR conservada	N-terminal	ninguno	Plegamiento, modificación y degradación
Helicasa de RNA DeaD (EC:3.6.4.13)	IDR conservada	Central	ninguno	Plegamiento, modificación y degradación

Conservación de IDRs con altos niveles de distribución

Los alineamientos múltiples de las regiones intrínsecamente desordenadas de amplia distribución filogenética permitieron analizar su conservación en las proteínas de bacterias y arqueas. A continuación, se muestran alineamientos representativos de IDRs.

Proteína chaperona DnaK (HSP70)

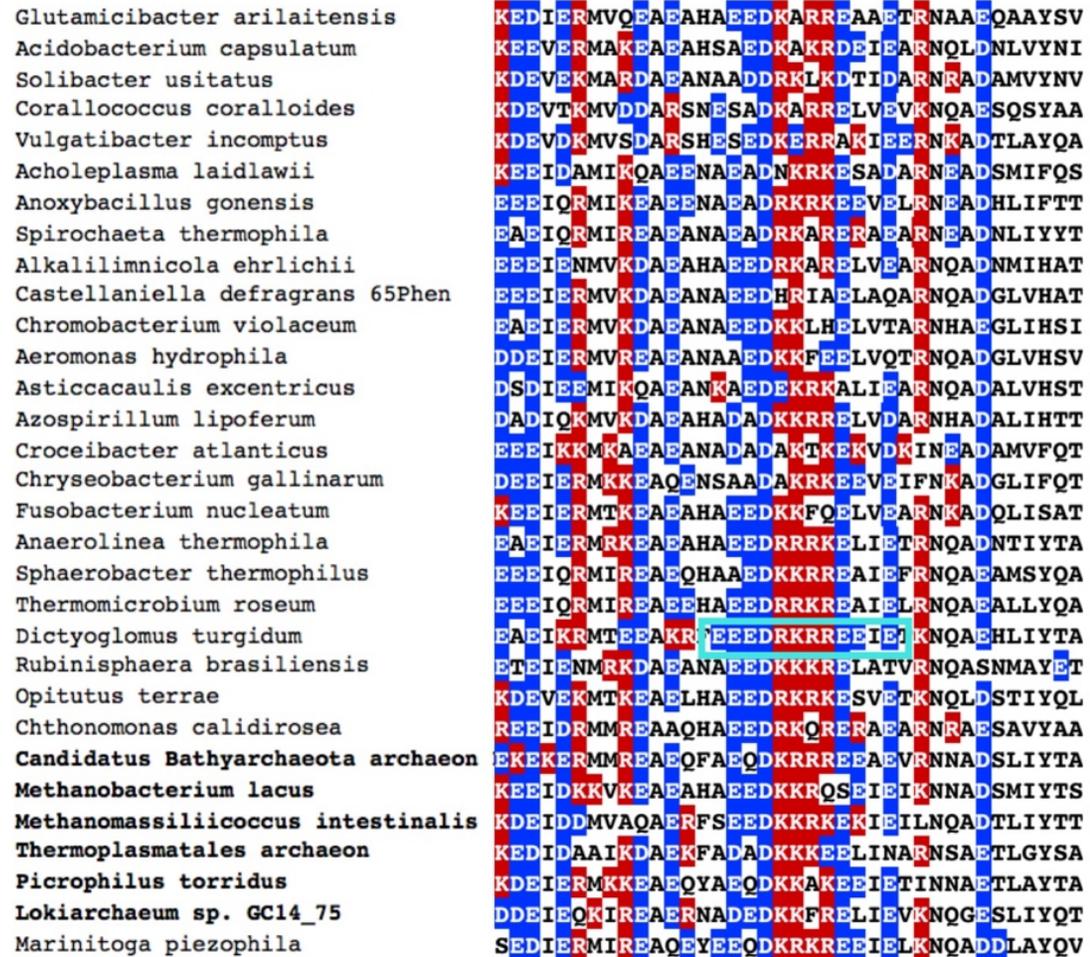


Figura 31: Alineamiento múltiple de la IDR (1) conservada en secuencia de la proteína chaperona DnaK (HSP70). La región intrínsecamente desordenada se encuentra en el dominio HSP70. Se identificó una LCR señalada con un recuadro azul claro. Las arqueas se encuentran señaladas en negritas. Los aminoácidos cargados se resaltan en color azul rey y rojo.

Proteína chaperona DnaK (HSP70)

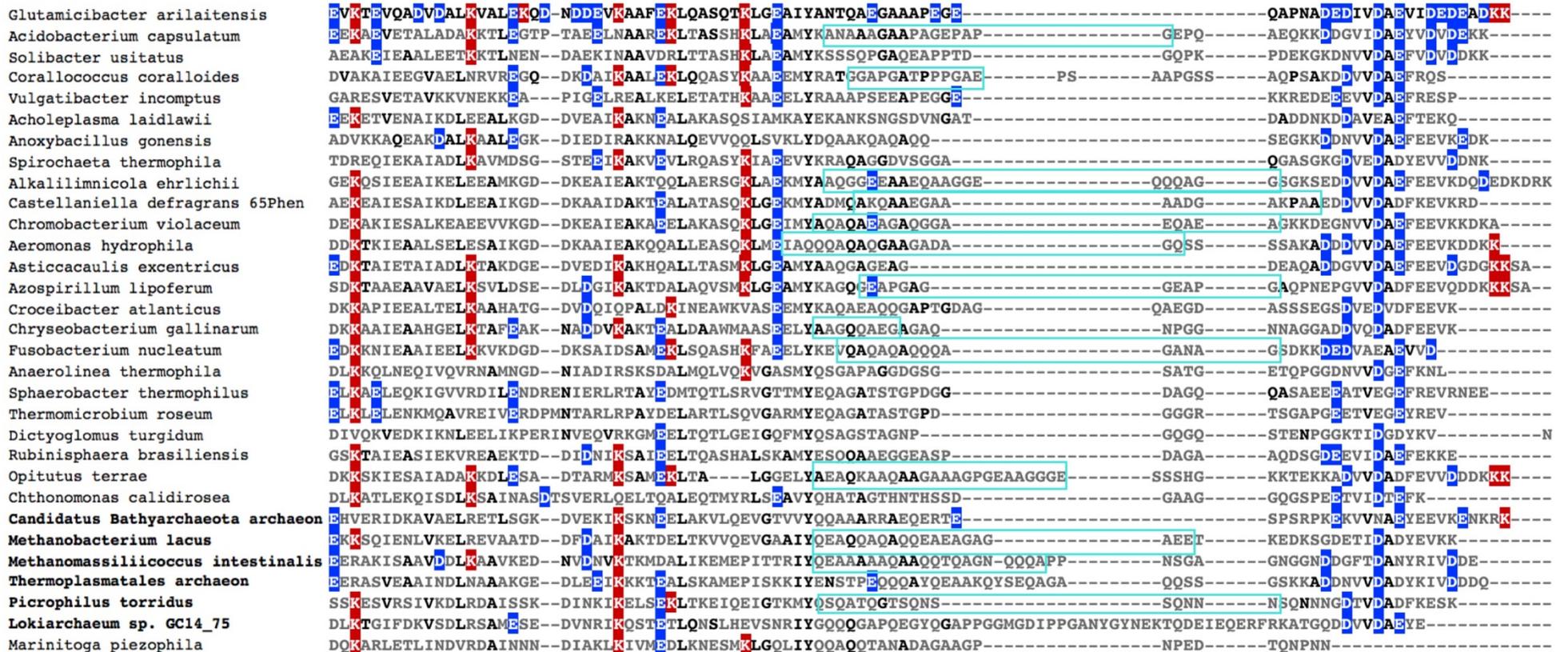


Figura 32: Alineamiento múltiple de la IDR (2) conservada de la proteína chaperona DnaK (HSP70). La región intrínsecamente desordenada se encuentra en el dominio HSP70 y también forma parte de la región C-terminal en la secuencia. Se identificaron algunas LCRs, señaladas con recuadros azul claro, que contienen motivos de repeticiones de AA y QQ. Las arqueas se encuentran señaladas en negritas. Los aminoácidos cargados se resaltan en color azul rey y rojo.

Proteína ribosomal L4

Lokiarchaeum sp. GC14_75	NNKSIQGRDPSAGLKNISEGWGT-GFGMSRAPRRKGS GFPTS RHVGRVPFATGG
Haloarcula hispanica	NRKQDYGSDEYAGLRTPAESFGS-GRG--QAHVPKQDGRAR-----RVPQAVKGG
Natronobacterium gregoryi	NRKQDYGADEFAGLRTPAESFGS-GRG--MAHVPRQDGRAR-----RVPQSIKGG
Granulibacter bethesdensis	CRRAG-----THKVKGMGEVSGT-TK---KPYRQKGTGNARQGS L-RAPQFR TGG
Beijerinckia indica	KRRAG-----THQSLGRAD I HRT-GK---KMYKQKGTGSARHGSA-RAPQFR TGG
Chthonomonas calidirosea	NQRQG-----THDTKQRSEVRGG-GR---KPWRQKGTGRARQGS I-RAPHWRGG
Lactococcus lactis	SLRQG-----THAHKNRS AVSGG-GK---KPWRQKGTGRARQGS T-RSPQWRGG
Halobacillus halophilus	NLRQG-----THKVKGRSEVSGG-GR---KPWRQKGTGRARQGS I-RAPQVWGG
Acholeplasma laidlawii	ALRQG-----TAKTKTRAEVRGG-GK---KPWRQKGTGRARQGS I-RSPQWRGG
Aminobacterium colombiense	NLRQG-----THSCKGRGEVRGG-GR---KPWRQKHTGRARHGST-RSPIWVGG
Chloroflexus aurantiacus	NARLG-----THNTRGRGEVKGS-TR---KLYRQKGTGRARQGS I-RAPHHKGG
Thermomicrobium roseum	NARAG-----THDTKTRGEVRGG-GR---KPWRQKGTGRARQGS I-RAPHWKGG
Ilyobacter polytropus	AARQG-----TAATKTRAMVRGG-GR---KPFKQKGTGRARQGS T-RAPHMVGG
Kosmotoga olearia	NRRAG-----TAKAKTRSEVSGG-GR---KPWPQKHTGRARTGS I-RNPLWRHG
Dictyoglomus turgidum	NGRQG-----THSTKRRSEVNRS-GR---KVWPQKGTGHARQGD R-KATHWVGG
Chroococcidiopsis thermalis	NARQG-----TANTKTRAEVRGG-GR---KPWRQKGTGRARAGS I-RSPLWRGG
Trichodesmium erythraeum	NNRQG-----TASTKTRSEVRGG-GR---KPWRQKGTGRARAGS I-RSPLWRGG
Dehalogenimonas lykanthroporepellens	NARQG-----TSSTRTRSEVSGT-TQ---KMFROKGTGEARAGSR-KSGLRPGG
Desulfocapsa sulfexigens	AKRAG-----NASTKTRREVRGG-GA---KPWKQKGTGRARAGTR-NSPIWRGG
Desulfuromonas soudanensis	ARRQG-----TASTKTRSEVSGG-GK---KPYKQKGTGNARQGC I-RAPHYVGG
Denitrovibrio acetiphilus	NRRAG-----THSTLNRAKMKGGRGA---KPWRQKGTGRARSGSR-KSPIWRGG
Pirellula staleyi	NLRQG-----THRTKGRGEVAGS-TK---KMYRQKGTGNARAG SR-RSGVRRGG
Rubinisphaera brasiliensis	NRRVG-----TFATKSRADVAGS-KK---KMYRQKGTGNARAGGK-RSPIRRGG
Gemmatimonas aurantiaca	NRRQG-----TAKTKTRGEVTGG-NQ---KPWKQKGTGRARQGS T-RAPNWPGG
Gemmatirosa kalamazoonesis	NQRQG-----TAATKIRKYVTGG-NQ---KPWRQKGTGRARQGS T-RAPHVWGG
Odoribacter splanchnicus	NNRQG-----THKSKQRNEISGS-TK---KLKKQKGTGGARAGS I-KNPEFRGG
Rufibacter tibetensis	NQRQG-----THKSKERA EVAGS-TK---KIKRQKGTGGARAGSM-KSPVFKGG
Deinococcus radiodurans	SRRRG-----TASTRTRAQVSKT-GR---KMYGQKGTGNARHGDR-SVPTFVGG
Leptospira interrogans	NLRSG-----NHATKTRSMVSGG-GK---KPWSQKGTGRARQGS T-RAPHVWGG
Methylotenera mobilis	NARTA-----TRAQKGRD T VAHT-TH---KPYAQKGTGNARSGMS-SSPIWRGG
Chromobacterium violaceum	NARSG-----NRAQLTRA EVKHS-TK---KPFROKGTGNARAGMT-STPNRRGG

Figura 33: Alineamiento múltiple de la IDR conservada en secuencia de la proteína ribosomal L4. La región intrínsecamente desordenada se encuentra en el dominio Ribosomal_L4. Se identificó una LCR, señalada con el recuadro azul claro. Aunque el algoritmo SEG (Wootton & Federhen, 1993) únicamente detecto una LCR, se observan varios motivos de repeticiones GG. Las arqueas se encuentran señaladas en negritas. Los aminoácidos cargados se resaltan en color azul rey y rojo.

Proteína ribosomal S13

Sulfolobus solfataricus	HSLGLKVRGQRTTRTTGRTGMTI ---GVAR KKAAQF Q SQQSSSQQQKSS ---
Metallosphaera sedula	HSLGLKVRGQRTTRTTGRTGTI ---G VKR SKAA QPSGGQSSQQQK ---
Methanobrevibacter smithii	HEAGLPVRGQRTKSTFRNSSSV---G VKR S-----
Ferroplasma acidarmanus	HE QGH KVRGQRT RSNCRHGLSM ---G VI R KR Q EQ KK-----
Halomicrobium mukohataei	HKR G Q KVRGQRT KSTGRTEGTIGV -N VEA I KEEQAE --E AAAAE DEGGEE
Natronomonas pharaonis	HKR G Q KVRGQRT KSTGRTEGTIGV -N VEA I KEEQAE --D GGDEE -----
Granulicella tundricola	HRR SLPVRGQRT H T NAR TRKGPRKGT VAG KK KATK -----
Koribacter versatilis	HRR SLPVRGQRT H T NAR TRKGPRKGT VAN KK KATAK -----
Pseudarthrobacter chlorophenolicus	HR KGLPVRGQRT K T NAR TRKGPKR-T VAG KK KAR -----
Catenulispora acidiphila	HR KGLP VHG QRT H T NAR TRKGPRK-A IAG KK KAGRK -----
Chthonomonas calidirosea	HRR GLPVRGQRT K T NAR TRKGKRR-T VAG KK KAKK -----
Deinococcus radiodurans	HRR GLPVRGQRT K T NAR TRKGPKK-T VAG KK KATRK -----
Desulfosporosinus orientis	HRR GLPVRGQRT K T NAR TRKGPAK-T VGA KK RRK -----
Thermovirga lienii	HKL GLPVRGQ K T KTNAR TRKGPRR-A VAG KK KPTGKK -----
Thermobaculum terrenum	HRR NLP VHG QRT R T NAR QRRGPRK-T VGA RR SKR -----
Marinithermus hydrothermalis	HRR GLPVRGQRT R T NAR TRKGPRK-T VAG KK KAPRK -----
Caldilinea aerophila	HRR NLP TRG QRT R T NAR TRRGARK-T VAG KK KAPRK -----
Carnobacterium maltaromaticum	HRR GLPVRGQ N T KNNAR TRKGPAR-T VAG KK K -----
Solibacillus silvestris	HRR GLPVRGQ N T KNNAR TRKGPRK-T VAN KK K -----
Acholeplasma laidlawii	HR KGLP VNG Q N T RNNAR TRKGPK-A VTG KK QAGK -----
Gemmatimonas aurantiaca	HRR GLPVRGQRT H T NAR T KKG PRR-A IAG KK KVTK -----
Corallococcus coralloides	HR KGLPVRGQRT H T NAR TRKGPKR-G IVRA K PAAPAR -----
Aquifex aeolicus	HAR GLPVRGQ Q T R T NAR TRKGKRR-T VGG T KKAKAK -----
Hyphomicrobium denitrificans	HR KGLPVRGQRT H T NAR TRKGKAV-P IAG KK KATK -----
Endomicrobium proavitum	HRR NLPVRGQRT K T NAR TRRGKRR-T V G AGK AA APG KK G ---
Micavibrio aeruginosavorus	HRR NLPVRGQRT K T NAR TRKGPAK-P IAG KK QATK -----
Fibrella aestuarina	HR KGLPVRGQ K T KNNSR TRKGKRR-T VAN KK KATK -----
Nitrosospira multiformis	HRR GLPVRGQRT R T NAR TRKGPRK-A VRAS SA KAGR -----
Turneriella parva	HR KGLP ARG QRT K T NAR TRKGRR-L-T VAG KK SAP SK G -----
Caldisericum exile	HKR NLPVRGQRT R T NAR TRKGPKK-T VGS VR KS A-----
Isosphaera pallida	HR KKLPVRGQRT R T NAR TRKGPRR-T VAG KK GVK DMR-----
Pirellula staleyii	HRL GLPVRGQRT K T NAR TRKGPKK-T VAG KK GVK DMR-----
Mesotoga prima	HR NGLPVRGQ K T HSNCR TRKGSRA-S KIR KK S -----
Akkermansia muciniphila	HRR GLPVRGQRT R T NAR TRKGKKK-T VGA Q A KK K -----

Figura 34: Alineamiento múltiple de la IDR conservada en secuencia de la proteína ribosomal S13. La región intrínsecamente desordenada se encuentra en el dominio Ribosomal_S13. Se identificaron algunas LCRs, señaladas con recuadros azul claro, que contienen motivos de repeticiones de AA y QQ. Las arqueas se encuentran señaladas en negritas. Los aminoácidos cargados se resaltan en color azul rey y rojo.

Helicasa de RNA Dead

```

Paenarthrobacter aurescens      GGQPL--LVKELPAA-----PEYQKRE-----RSKDGF
Pseudarthrobacter chlorophenicus GGQPL--LVKELAAA-----PDFQKRE-----RSKDGF
Acaryochloris marina           KDKPL--V---PPKED-LSSDPGHAVQTSNWA-----ADLGK-PSH-----
Dechlorosoma suillum           KSRPL--LM---PEKEV-RSFEADSAE---GR-----PERKERF-ERPERGE-RFERKERPERGE
Thauera humireducens           RERPL--QIEQ---SG-RGWLATAEPGSGGR-----APREEHF---TTGA-PRERTPRPNRDE
Pelodictyon phaeoclathratiforme GETPL--LLSNKPEKS--RSYEEREGS---GR-----DRGSDRE-RSTDRGR-SSDRDRGPVK--
Hahella chejuensis            VEQPLFPILKDLPKLD-----KDFRGRDRDRDGRGRNRDRDRRGRE--
Planctopirus limnophila        GDAPL--LVKD-DIRDF-NSIETRNREPAP-GR-----GNPRLARA---O-DAERSMRPRGRE
Thioflavicoccus mobilis       RERPL--VVAEVPQPR-----IDARRGRR---DRPD-RAGRDPRAGHGP
Rhodopirellula baltica        NGRPF--LLKERPKKQR-ERSDRDNSR---GR-----DSFDSND-RSGNRGS-FGDDRRR---
Chloroflexus aurantiacus       NED-----QTD--QVDTITRLDPPP-----AERPQRE---RSE-RSARSPR---
Coralloccoccus coralloides     EKED-----EGKDTADQEI PQVSAPAERKFPRTERS GPPGRFADRGAE RPSRGPRTADRGE-RPERGERSERGE
Stigmatella aurantiaca        EARD-----EGHAQNEEEI PVVPPPSEKR-----GR-----TERPSRGAPPERRG-KAE-APP SERPR
Gemmatimonas aurantiaca       ARS-----EGE---EAEI PAVTPRPDRGG---ERSFDRGDRGDRGGERGSRREGARDLRGK-PIGRESRGATGE
Methanocorpusculum labreanum   DSG-----KNAA-QSEEIKPLQSPSSS---ETFTP GTVENS GAE
Methanosphaera stadtmanae     DDKRY-----KEEF-----
Rufibacter tibetensis         KDT-----KAKE-KSAEAGEAKGGPK-----
Parvimonas micra              ENNKL-----N-----NHEELIDVDIKKNKK---KGNDKTSISNKGKSNVTE
Methanococcus maripaludis     KDE-----L-SESNYKRIG-KP-----ANREGRD---RGE-GRSEGRR---
Acholeplasma laidlawii       PTK-----KTYETIEI IPEK-----ASRRSQDERGNDKGR-KDSRSSR---
Nonlabens dokdonensis        NSIY-----KRNSITDLN-----DRSKGRE---DSGE-YRERK---
Akkermansia muciniphila       EARTH-----REVQIPEDK-----PARRARR---MPQTGE-EAESP---
Ketogulonicigenium vulgare    EGRP-----APEALTVVNTPT-----PGAAPAPRE-----
Novosphingobium aromaticivorans AALP-----QPEELLS-----GTDEGRK-----

```

Figura 36: Alineamiento múltiple de la IDR conservada de la proteína Helicasa de RNA Dead. La región intrínsecamente desordenada no se localiza en ningún dominio proteínico. Se identificaron algunas LCRs, señaladas con recuadros azul claro, que contienen motivos de repeticiones de **RR** y **RD**. Las arqueas se encuentran señaladas en negritas. Los aminoácidos cargados se resaltan en color azul rey y rojo.

Discusión

Secuencias simples en Bacteria y Archaea

La asociación entre el tamaño del proteoma y el porcentaje de secuencias simples en Bacteria y Archaea (Fig. 9) es débil, lo que sugiere que existe algún factor que determina el sesgo en el contenido de LCRs en los procariontes.

Las secuencias simples tienen presencia en todos los proteomas. Gemmatimonadetes es el phylum bacteriano con mayor proporción de LCRs en su proteoma, 37.73% de sus proteínas contienen secuencias simples, mientras que la bacteria de la categoría “Unclassified Bacteria” contiene 11.15% y el phylum Deferribacteres con 12.26% presentan los menores porcentajes (Fig. 7 y Anexo II).

Los genes que contienen secuencias simples tienden a presentar un alto contenido de GC (Albà & Guigó, 2004). Esto también se observó a nivel de proteomas (Fig. 13). Las bacterias de Gemmatimonadetes poseen un alto porcentaje de GC: 72.6%; lo que coincide con el segundo phylum con mayor cantidad de LCRs, Actinobacteria, el cual se caracteriza por presentar un alto contenido de GC en su genoma (Ventura *et al.*, 2007). Las actinobacterias presentan un rango variable de porcentajes de secuencias simples, como puede observarse en el valor de desviación estándar (Fig. 7 y Anexo II); sin embargo, las especies con mayor porcentaje de LCRs presentan altos contenidos de GC. *Kitasatospora setae* tiene el mayor porcentaje de LCRs y un contenido de GC de 74.2%, en contraste, *Atopobium parvulum*, presenta un 45.69% de GC y es una de las actinobacterias con menores porcentajes de LCRs (Fig. 14).

Los altos niveles de redundancia de genes que presenta *Deinococcus radiodurans* (White *et al.*, 1999) podrían relacionarse con el alto porcentaje de LCRs en su proteoma, además del alto contenido de GC, 67%.

En Archaea se mantiene la relación entre altos contenidos de GC y una gran proporción de secuencias simples (Fig. 13 y Fig. 14). Euryarchaeota (Fig. 12 y Anexo II), contiene a la arquea con el mayor porcentaje de LCRs, *Natrialba magadii* cuyo contenido de GC es de 61%, mientras que *Methanobus psychrophilus*, 44.6% de GC, la euriarqueota con menos secuencias simples.

Las arqueas de hábitat halófilo presentan los mayores porcentajes de secuencias simples y altos contenidos de GC (60- 70%). No obstante, esta última característica podría estar relacionada con presiones selectivas de adaptación al ambiente halófilo y en particular, a la estabilidad del DNA y proteínas en respuesta a las condiciones ambientales extremas (Paul *et al.*, 2008).

Las secuencias simples no están restringidas a una categoría funcional, aunque la mayoría se puede encontrar en la categoría *Transporte de membrana* (Fig. 10 y Anexo III). Nuestros hallazgos concuerdan con lo encontrado en eucariontes (Sim & Creamer, 2002), donde algunas de las LCRs pueden encontrarse en proteínas de señalización asociadas a la membrana y otras forman parte de dominios integrales de membrana.

En general, el sesgo composicional de las LCRs (Fig. 15) reveló la ausencia de los aminoácidos His y Cys, los dos más abundantes en los sitios activos de las proteínas (Holliday *et al.*, 2011), así como de aquellos aromáticos Phe, Tyr y Trp, lo cual tiene sentido biológico, ya que una secuencia repetitiva de estos aminoácidos sería aberrante en la estructura de una proteína.

En algunos casos, la composición de aminoácidos de las secuencias simples tiende a relacionarse con su temporalidad. Se ha reportado que las LCRs en las proteínas más recientes de vertebrados son ricas en Ala y Gly, mientras que las más antiguas poseen altos niveles de aminoácidos con carga positiva Lys y Arg, lo que implica el efecto de la selección purificadora en su preservación (Radó-Trilla & Albà, 2012).

Regiones intrínsecamente desordenadas en Bacteria y Archaea

Al igual que ocurre con las secuencias simples, el porcentaje de IDRs no se relaciona con el tamaño del proteoma (Fig. 19), lo que indica que tienen alguna importancia biológica. El porcentaje de IDRs en los proteomas parece estar relacionado con el contenido de GC (Fig. 23 y Fig. 24), aquellos Phyla con los mayores porcentajes de IDRs también presentan los valores más altos de contenido de GC, esta asociación es similar a la observada en las LCRs (Fig. 14).

Los porcentajes de desorden en el proteoma de los procariontes es variable, el rango varía de 0.63% en Dictyoglomi a 15.6% en Planctomycetes (Fig. 17 y Anexo IV). Resulta interesante que los phyla de hábitats termófilos presentan los menores porcentajes de desorden, Dictyoglomi, Thermodesulfobacteria, Aquificae, Thermotogae y Caldiserica. El bajo contenido de IDRs en estos phyla puede deberse a una adaptación a condiciones extremas de las altas temperaturas (Burra *et al.*,

2010). Sin embargo, existe una excepción, ya que la bacteria con mayor porcentaje de IDRs en su proteoma es *Thermaerobacter marianensis*, un organismo hipertermófilo del phylum Firmicutes.

Como se mencionó anteriormente, el phylum que posee el mayor porcentaje de IDRs es Planctomycetes (Fig. 17), lo que coincide con reportes anteriores (Pavlović-Lažetić *et al.*, 2011; Peng *et al.*, 2015; Bordin *et al.*, 2018). Se ha sugerido que los altos niveles de proteínas desordenadas en Planctomycetes, se relacionan con el desarrollo de sus membranas, lo cual desde una perspectiva evolutiva resulta fascinante, ya que los Planctomycetes, que pertenecen al grupo PVC, presentan características que no se observan frecuentemente en las bacterias: algunas especies presentan una extensa organización de la membrana interna, un sistema de transporte de macromoléculas inusual (Bordin *et al.*, 2018), y recientemente se ha descrito la presencia de poros nucleares en la especie *Gemmata obscuriglobus*, los cuales tienen elementos estructurales similares a los poros nucleares de eucariontes (Sagulenko *et al.*, 2017).

En Archaea, el phylum con mayor porcentaje de IDRs es Nanohaloarchaeota, con 12.6%, mientras que Korarchaeota con 0.68% presenta el menor porcentaje de proteínas desordenadas (Fig. 15 y Anexo IV). Al igual que ocurre en las bacterias, al parecer los bajos contenidos de IDRs en arqueas se relacionan con los hábitats termófilos. Korarchaeota es un phylum de arqueas termófilas; Crenarchaeota es el segundo phylum con menor proporción de IDRs en sus proteomas, no obstante, en este trabajo solo se analizaron especies termófilas de este último. Lo anterior confirma que uno de los factores que originan la baja proporción de IDRs en los organismos es la adaptación a distintos ambientes extremos (Xue *et al.*, 2010; Burra *et al.*, 2010).

Nanohaloarchaeota, el phylum con mayor porcentaje de proteínas desordenadas, es un grupo de arqueas hiperhalófilas. El hábitat halófilo de *Nanohaloarchaea archaeon* parece ser otra característica de las arqueas con mayores porcentajes de IDRs. Euryarchaeota es el segundo phylum con mayores niveles de IDRs, aunque presenta una gran variación, lo que se puede ver con la Desviación estándar (Fig. 18 y Anexo IV). Este grupo contiene especies de diversos hábitats, sin embargo, las especies con mayores porcentajes de IDRs son halófilas, algunos ejemplos son: *Natrialba magadii*, *Halopiger xanaduensis* y *Halorubrum lacusprofundi*. La dependencia del contenido de proteínas desordenadas en organismos de ambientes halófilos puede explicarse debido a que estas proteínas pueden ayudarles en la supervivencia en ambientes con altas concentraciones de iones en el ambiente (Xue *et al.*, 2010).

Por otro lado, las funciones en las que participan las proteínas con regiones intrínsecamente desordenadas son principalmente *Traducción*, *Transporte de membrana*, así como *Replicación y reparación* (Fig. 20, Fig. 21 y Anexo V). Lo que sugieren dos de las categorías más abundantes, *Traducción* y la *Replicación y reparación*, es que en la interacción con ácidos nucleicos, las regiones y proteínas intrínsecamente desordenadas tienen alguna importancia. Se ha reportado que las regiones intrínsecamente desordenadas en proteínas de unión a ácidos nucleicos frecuentemente requieren el desorden para sus funciones y regulación (Wang *et al.*, 2016). Las proteínas de unión a RNA son ricas en desorden, y la mayor fracción de los residuos desordenados, se localizan en la interfaz de unión en contacto directo con el RNA (Varadi *et al.*, 2015). La importancia del desorden también subyace en que la clase de proteínas con mayor frecuencia de desorden es la de chaperonas de RNA (Tompa & Csermely, 2004). Las proteínas que interactúan con DNA también presentan altos contenidos de regiones desordenadas, aunque los dominios proteínicos desordenados que se unen al DNA son abundantes en eucariontes, en comparación con los procariontes.

El análisis de la composición de las IDRs (Fig. 25), demostró que son abundantes en los aminoácidos promotores de desorden Glu, Pro, Ser y Gln (Olfield & Dunker, 2014); la Lys que también pertenece a esa categoría solo es frecuente en *Traducción*. Los residuos Ala, Asp, Gly, Arg y Thr que también presentan una alta frecuencia, son aminoácidos neutrales que pueden estar presentes en regiones ordenadas o desordenadas. Cabe destacar que al igual que en las LCRs, los aminoácidos más abundantes en los sitios activos enzimáticos (Holliday *et al.*, 2011), la His y la Cys están ausentes, estos también se caracterizan por ser promotores de orden en las proteínas.

LCRs con amplia distribución filogenética

Los proteomas de las bacterias y arqueas analizadas comparten LCRs, que han preservado su sesgo composicional a lo largo de la evolución. En la Tabla IV y Anexo VI se muestran LCRs conservadas en Bacteria y Archaea presentes en genes que posiblemente formaban parte del complemento génico del LCA, el ancestro de todos los seres vivos.

Lo que define a las secuencias simples, es un sesgo composicional, el cual se refleja en la presencia de motivos repetidos señalados en los alineamientos múltiples; no obstante, la divergencia a nivel de secuencia primaria y la longitud corta de las secuencias simples, dificultan la detección de los motivos repetidos (Andrade *et al.*, 2001).

A continuación, se muestra un listado de LCRs conservadas en bacterias y arqueas y la importancia que presentan en las proteínas en las que se encuentran.

ATPasa transportadora de arseniato

Las LCRs forman parte de dominios funcionales, un ejemplo son las que se identificaron en la proteína ArsA celulares (Tabla IV y Fig. 26), una ATPasa de la membrana interna de las bacterias. La conservación de las secuencias simples en esta proteína demuestra que cumplen funciones importantes, ArsA participa en la expulsión del arseniato de las células. Las LCRs conservadas forman parte del *P-loop*, un sitio de unión a Mg^{2+} (Zhou *et al.*, 2001).

Chaperona HSP60

Se ha demostrado que las LCRs conservadas en los dos dominios celulares (Tabla IV y Fig. 27) presentes en el extremo C-terminal de la familia de chaperonas moleculares HSP60 son indispensables para su función, ya que la eliminación del extremo C-terminal en GroEL puede producir consecuencias desfavorables: en la proteína más abundante de la biósfera, la Rubisco, se produce una disminución en la velocidad de plegamiento (Weaver & Rye, 2014); mientras que en la proteína verde fluorescente (o GFP, por sus siglas en inglés, Green Fluorescent Protein), el extremo C-terminal es indispensables para retener a la GFP dentro del complejo GroEL-GroES (Ishino *et al.*, 2015).

El extremo C-terminal donde se localiza la LCR, presenta altos niveles de inserciones y remociones, y un alto contenido de aminoácidos hidrofóbicos casi ubicuo, caracterizado por muchas repeticiones de Gly-Gly-Met.

D-aminoácido deshidrogenasa

En la flavoproteína D-aminoácido deshidrogenasa, que cataliza la desaminación oxidativa de los D-aminoácidos, también se identificó una LCR conservada en el dominio proteico DAO (Fig. 28), lo que indica que una vez que se originan las secuencias simples, pueden asociarse a funciones. La enzima cataliza la desaminación oxidativa de los D-aminoácidos y tiene amplia especificidad de sustrato. Esta enzima es principalmente activa en D-prolina y, en menor grado, en varios otros D-aminoácidos como D-alanina, D-fenilalanina y D-serina (He *et al.*; 2011).

Enolasa

Las secuencias simples pueden presentar cualquier tipo de estructura secundaria (Kumari *et al.*, 2014). La LCR conservada en la Enolasa (Tabla IV y Fig. 29) conforma una alfa hélice dentro del dominio Enolase_N (Kühnel & Luisi, 2001). La estructura secundaria de la secuencia simple, es predecible tomando en consideración la alta frecuencia de alaninas en su composición. El residuo Lys 119 es un contacto clave en la unión a la RNasa E (Chandran & Luisi, 2006). La enolasa es una enzima universal que, además de ser clave en los últimos pasos de la glucólisis, forma parte del degradosoma de RNA, un complejo multienzimático que interviene en el procesamiento y degradación del mRNA (Carpousis, 2007). El procesamiento del mRNA mediante el degradosoma es un antiguo mecanismo que posiblemente se estableció en épocas evolutivas anteriores al LCA, cuando “las moléculas de RNA desempeñaban un papel más conspicuo en los procesos celulares” (Delaye *et al.*, 2005).

IDRs con amplia distribución filogenética

Las regiones intrínsecamente desordenadas forman parte de todos los proteomas de bacterias y arqueas (Fig. 22). Con el objetivo de conocer aquellas IDRs conservadas y sus posibles funciones en la evolución temprana de la vida, se analizaron las IDRs con amplia distribución filogenética en Bacteria y Archaea (Tabla V y Anexo VII). A continuación, se muestra un listado de proteínas con regiones intrínsecamente desordenadas conservadas en los dos dominios celulares y la importancia funcional que presentan.

Proteína chaperona DnaK (HSP70)

La chaperona Hsp70, codificada por el gen DnaK, contiene dos IDRs con amplia distribución filogenética (Tabla V) que, al igual que las LCRs, forman parte de dominios funcionales. Ambas IDRs se localizan en el dominio HSP70; sin embargo, difieren en su clasificación y posición. La IDR 1 está conservada en secuencia (Fig. 31) y se encuentra dentro del dominio HSP70, mientras que la IDR 2, cuya secuencia primaria no se conserva sino solo el desorden, contiene una parte dentro del dominio y también abarca el extremo C-terminal (Fig. 32).

La familia de chaperonas Hsp70 se conoce por ser asistentes de plegamiento de proteínas, sin embargo, se ha demostrado que la chaperona Hsp70 también presenta actividad chaperona de RNA (Zimmer *et al.*, 2001). La proteína DnaK interactúa con el 16S rRNA y facilita el ensamblaje de la subunidad 30S ribosomal y participa en la *Biogénesis del Ribosoma* (Maki *et al.*, 2002). Los residuos de la IDR 1 (Fig. 31) se encuentran conservados en la Hsp70 de mamíferos y corresponden a los sitios

de interacción con RNA, lo que sugiere que la IDR 1 es sitio de unión al ácido nucleico, lo que se demuestra con la conservación a nivel de secuencia de la región desordenada, ya que aquellas IDRs que preservan su secuencia primaria tienden a interactuar con RNA y presentar actividad chaperona (Bellay *et al.*, 2011).

La IDR 2 (Fig. 32) que forma parte del extremo C-terminal de la DnaK, contiene residuos que presentan un sitio de unión débil, auxiliar para proteínas desnaturalizadas. La asociación débil de la región intrínsecamente desordenada promueve ciclos subsecuentes de unión y liberación del substrato, esta actividad corresponde a una función frecuente en las IDRs, la de actuar como *flexible tether* y presentan motivos de aminoácidos que les confieren características de reconocimiento molecular; la flexibilidad resultante permite a las regiones participar en múltiples modos de unión (Smock *et al.*, 2011), la conservación de la IDR 2 en Bacteria y Archaea, sugiere la antigüedad de este fenómeno.

Proteínas ribosomales de la subunidad grande

La estructura del ribosoma es una ventana hacia el pasado, la universalidad de varios de sus componentes, implica que ya estaban establecidos antes de la divergencia de los tres linajes celulares (Becerra *et al.*, 2007). En las proteínas ribosomales que corresponden L4, L3y L2, se encontraron IDRs conservadas en secuencia en Bacteria y Archaea (Tabla V y Fig. 28).

El orden de ensamblaje del ribosoma, permite asignar una temporalidad relativa a las proteínas ribosomales, lo que permite hacer una inferencia sobre la evolución de las IDRs que presentan; se han establecido cuatro grupos de antigüedad (Fox, 2010) del más antiguo al más reciente: I): L2, L3, L4; II): L22, L23, L24; III): L5, L6, L10, L11, L12, L15 y IV): L1, L13, L14, L16, L18, L29, L30.

Es probable que la preservación de IDRs conservadas en secuencia de las proteínas L2, L3 y L4, se deba a factores que se relacionen con su temporalidad, diversos autores han propuesto que son muy antiguas: i) son las más próximas al sitio de la peptidil transferasa (PTC, por sus siglas en inglés) (Worbs *et al.*, 2000; Fox, 2010; Hsiao *et al.*, 2013), uno de los componentes ribonucleotídicos más antiguas del ribosoma. ii) las proteínas L3 y L4 son imprescindibles en las primeras etapas del ensamblaje de la subunidad grande del ribosoma bacteriano (Timsit *et al.*, 2009; Fox, 2010). iii) poseen extensiones no globulares, que cubren la mayor área de la superficie del rRNA en la subunidad grande del ribosoma de *Haloarcula marismortui* (Klein *et al.*, 2004). iv) se unen de manera normal

al 23S rRNA en presencia de Mg²⁺ (Hsiao *et al.*, 2013) y v) tienen, en conjunto con la L15, las extensiones más largas y complejas de todas las proteínas ribosomales (Klein *et al.*, 2004).

Un fósil molecular presente en todas las células es el ribosoma, el cual tiene grabado una cronología molecular del origen y evolución de las proteínas; los segmentos de las proteínas ribosomales revelan la historia del plegamiento de las proteínas a nivel atómico (Hsiao *et al.*, 2013; Kovacs *et al.*, 2017), en otras palabras, “las proteínas ribosomales ofrecen una ventana en el tiempo cuando las proteínas adquirieron la habilidad de plegarse” (Lupas & Alva, 2017); resulta intrigante que la etapa más temprana de las proteínas ribosomales está representada por protopéptidos cortos “random coil”, es decir están intrínsecamente desordenados, lo que sugiere que uno de los fenómenos más antiguos de las proteínas son los polipéptidos desordenados.

Es posible que las primeras proteínas ribosomales hayan sido pequeños péptidos básicos, cuyos vestigios posiblemente forman parte de dominios funcionales de proteínas ribosomales actuales (Lazcano *et al.*, 1992). Por lo que podemos inferir, las regiones intrínsecamente desordenadas conservadas en Bacteria y Archaea podrían ser vestigios de las proteínas más antiguas. Los primeros péptidos adquirieron la habilidad de plegarse resultado de una propiedad emergente de la coevolución péptido-RNA (Kovacs *et al.*, 2017; Lupas & Alva, 2017).

Algunos autores han realizado estudios de diversos péptidos de unión a RNA, definidos como cadenas de ≤ 30 aminoácidos que carecen de una estructura estable en ausencia del Ácido ribonucleico y no obstante, se unen a este de manera específica (Frankel, 2000); estas transiciones de desorden a orden están presentes en otros péptidos de unión a RNA y muchos de ellos presentan simplicidad en la secuencia (Das & Franklin, 2003).

El análisis de las IDRs conservadas en Bacteria y Archaea (Tabla V), revela una preservación a nivel estructural de componentes peptídicos antiguos del ribosoma. Asimismo, la simplicidad representada por los motivos de repeticiones presentes en la secuencia de aminoácidos (Fig. 33), sugiere el papel del deslizamiento de la polimerasa en etapas evolutivas tempranas.

Proteínas ribosomales de la subunidad pequeña

Todas las proteínas de la subunidad pequeña del ribosoma presentan extensiones no globulares o IDRs, con excepción de la S4 y S15 (Peng *et al.*, 2015). Algunas IDRs fueron

identificadas en las proteínas ribosomales S4, S3, S9 y S13 (Tabla V y Fig. 34), cuya distribución universal permite inferir que formaban parte del proteoma del LCA.

Se ha establecido una temporalidad de las proteínas ribosomales de la subunidad pequeña (Shajani *et al.*, 2011) que van del más antiguo al más reciente: **I**: S4, S7, S8, S15, S17; **II**: S9, S13, S19 y **III**: S2, S3, S5, S10, S11, S12, S14. En la subunidad pequeña del ribosoma, la distribución y conservación de las IDRs, sugiere que no se relacionan con la temporalidad de las riboproteínas y probablemente la evolución de estas es de un carácter más flexible, a diferencia de las proteínas en la subunidad grande.

Las proteínas S9 y S13 (Fig. 34) incrementan la estabilidad del ribosoma, sin embargo, no son indispensables para que la traducción funcione (Peng *et al.*, 2015). Las riboproteínas S9 y S13 están en contacto con el tRNA (Yusupov *et al.*, 2001). La IDR en la proteína S4, pertenece al tipo de desorden conservado (Tabla V) y la secuencia es conservada solo por dominio celular, esto sugiere que son divergencias y su origen es posterior al LCA. La proteína ribosomal S3 contiene una IDR conservada y de manera extraordinaria, presenta una alta abundancia de secuencias simples (Anexo VI y VII) y una de las LCRs más largas, de 162 aminoácidos de longitud en la bacteria *Conexibacter woesei*. Es posible que la presencia de LCRs en la proteína S3 se deba a que es una proteína ribosomal perteneciente a la etapa III de la temporalidad relativa, es decir, es más reciente que las demás riboproteínas, lo cual coincide con que las secuencias simples tienden a disminuir con el tiempo (Toll-Riera *et al.*, 2012).

Partícula de reconocimiento de señal

La partícula de reconocimiento de señal (SRP) es un complejo ribonucleoproteínico que transporta las proteínas membranales y de secreción a las membranas celulares en procariontes y hacia el retículo endoplásmico rugoso en eucariontes. Se identificaron IDRs conservadas en dos proteínas del complejo, una IDR en la Partícula de reconocimiento de señal y otra en el Receptor de la Partícula de reconocimiento de señal.

Todas las células transportan proteínas con la partícula de reconocimiento de señal, pero los únicos componentes universales son las proteínas SRP, llamada Ffh en bacterias y SRP54, su homóloga en arqueas, unidas a las hélices 5 y 8 del SRP RNA. Estos llevan a cabo el reconocimiento de la secuencia señal, la interacción con el ribosoma y con el receptor de SRP (Egea *et al.*, 2008;

Hainzl & Sauer-Eriksson, 2015), lo que sugiere que estos procesos son los más antiguos de todo el complejo SRP.

En la IDR conservada (Fig. 35) en el extremo C-terminal de las proteínas Ffh y SRP54 se identificaron LCRs, lo que permite proponer que ambas estaban presentes en el LCA. El dominio M y los residuos de su extremo C-terminal son esenciales en la interacción con la secuencia señal. En su conformación libre, es decir no unido a la secuencia señal, el dominio M de la SRP54 se define solo por sus hélices $\alpha M1$ - $\alpha M5$, el finger loop, el GM linker y los residuos del extremo C-terminal están desordenados. Las regiones desordenadas del dominio M en su conformación libre corresponden al GM linker: Met 303 - Ile 318, finger loop: Lys 346 - His 364 y el extremo C-terminal: Lys 431- Gly 451 (Hainzl & Sauer-Eriksson, 2015). En el momento que la SRP se une a la secuencia señal, es comunicado al dominio NG, lo que produce la interacción con el SR; esta unión ocasiona cambios conformacionales en el dominio M, causando que el GM linker, el finger loop y los residuos del extremo C-terminal se plieguen. Este cambio estructural de desorden a orden es evolutivamente relevante, ya que provee una capacidad estructural en los componentes que reconocen la secuencia señal; estas IDRs son promiscuas y podrían reconocer diversas secuencias señal debido a su plasticidad estructural (Keenan *et al.*, 1998; Hainzl & Sauer-Eriksson, 2015).

Las LCRs pertenecen al extremo C-terminal del dominio M de las proteínas Ffh y SRP54. Este es importante en la unión del SRP con la secuencia señal (Hainzl & Sauer-Eriksson, 2015) (Fig. 35) y varía mucho en longitud (Keenan *et al.*, 1998), pero presenta muchas repeticiones de motivos MM y GGM. La metionina es muy abundante en estas secuencias simples, un aminoácido hidrofóbico cuya cadena lateral flexible confiere plasticidad en el sitio de unión a la secuencia señal (Keenan *et al.*, 1998); otro aminoácido abundante es lisina, lo que sugiere que sea una región de unión a RNA, que se comprueba por la interacción de los residuos del extremo C-terminal del dominio M con el rRNA. Los motivos de repeticiones GGM, también son característicos en la IDR de las chaperonas moleculares HSP60 (Fig. 27), lo cual indica que este aporta una flexibilidad inherente a estas regiones terminales.

La IDR conservada del Receptor de la Partícula de reconocimiento de señal (Tabla V) se localiza en el extremo N-terminal. La partícula de reconocimiento de señal (SRP) primero reconoce la secuencia de señal N-terminal de proteínas nacientes y luego interactúa con el receptor de SRP.

Debido a que las IDR del dominio M están conservadas en los dos grupos celulares es muy probable que el cambio conformacional de desorden a orden que ocurre en la partícula de reconocimiento de señal, sea un mecanismo universal, lo que implica que: 1) este cambio estructural es antiguo y estaba presente en LCA, 2) la presencia y conservación de la LCR (Fig. 35) dentro de la IDR, sugiere la importancia del deslizamiento de la polimerasa en la generación de partes funcionales en las proteínas.

DEAD-box RNA helicasas

Se identificó una IDR conservada en el extremo C- terminal de la helicasa DeaD (Tabla V y Fig. 36), la cual interviene en la biogénesis de la subunidad grande del ribosoma, en la iniciación de la traducción y en el ensamblaje del degradosoma “cold shock” (Redder *et al.*, 2015). Aunque la IDR no presenta ningún motivo conservado, la alta proporción de Lys y Arg, aminoácidos con carga positiva en las extensiones de las DEAD-box helicasas analizadas, sugiere que se unen al RNA (López-Ramírez *et al.*, 2011), de manera no específica (Rudolph & Klostermeier, 2009).

Las proteínas DEAD-box son helicasas de RNA, ATP dependientes que forman parte del proteoma de todos los seres vivos e intervienen en varios procesos centrales en el metabolismo del RNA (López-Ramírez *et al.*, 2011). La relación directa de las DEAD helicasas con el RNA, sugiere que tenían importancia durante el mundo de RNA/proteínas y su temporalidad antecede a la divergencia de los dominios celulares (Delaye *et al.*, 2005).

Relación entre LCRs e IDRs

Se encontraron diversos motivos de repeticiones y LCRs en las IDRs, lo que sugiere que posiblemente hayan evolucionado a partir de expansión de repeticiones. La inestabilidad genética de las regiones genómicas de repeticiones, en combinación con la naturaleza permisiva estructural de las IDRs, tiende a incrementar la cantidad de desorden en la evolución (Tompa, 2003).

La mayoría de los homopolímeros, secuencias simples compuestas por un solo aminoácido, se localizan en regiones intrínsecamente desordenadas (Albà *et al.*, 2007). Es importante mencionar que las regiones desordenadas no siempre presentan una baja complejidad en la secuencia primaria (Romero *et al.*, 2001; Tompa & Kovacs, 2010).

Las diferencias en longitud de LCRs e IDR de los extremos terminales en chaperonas HSP60, DEAD-box helicasas de RNA, la partícula de reconocimiento de señal y algunas riboproteínas, caracterizadas por una elevada cantidad de repeticiones, indica que las secuencias simples tienden a extender la secuencia de las proteínas (Ellegren, 2004; Radó-Trilla & Albà, 2012), en general se ha observado una tendencia a originarse en esas regiones terminales (Toll-Riera & Albà, 2013).

Aquellas IDRs que interaccionan con ácidos nucleicos presentan una frecuencia alta de argininas y lisinas. La elevada proporción de glicinas en IDRs de unión a RNA es frecuentemente favorecida por su flexibilidad conformacional (Fournier *et al.*, 2010). Se ha propuesto que ciertas repeticiones de aminoácidos, como Arg-Gly-Gly y Arg-Ser se encuentran en proteínas de unión a RNA y su estructura desordenada y baja complejidad en la secuencia, posiblemente fueron fundamentales en proteínas muy antiguas (Fornerod, 2012).

La contribución de las secuencias simples en el origen de genes

En la comprensión del origen de genes durante la evolución temprana de la vida siempre habrá incertidumbre. Es probable que las polimerasas primitivas, al igual que las actuales, inevitablemente tuvieran como característica el deslizamiento de la polimerasa, proceso que puede generar secuencias simples en el material genético y consecuentemente, en proteínas (Becerra *et al.*, 2002). Es importante mencionar que el deslizamiento de la polimerasa no siempre genera baja complejidad o repeticiones en la secuencia (V. Valdés, comunicación personal).

Aún si el deslizamiento de la polimerasa careciera de cualquier implicación en la formación de los primeros genes, es un mecanismo que ocurre constantemente en las células durante la replicación del DNA.

En el surgimiento de una secuencia simple, pueden ocurrir dos posibles eventos: que las repeticiones por sí mismas, no presenten alguna función inicial, se produzcan mutaciones puntuales y la huella de lo simple desaparezca con el tiempo, entonces la región repetitiva actuará como materia prima para nuevas funciones; si por el contrario, el sesgo composicional resulta beneficioso, probablemente la selección purificadora actuará y mantendrá la secuencia simple, reteniendo la huella en sus motivos de repeticiones (Sim & Creamer, 2004; Radó-Trilla & Albà, 2012; Radó-Trilla, 2013).

Las proteínas de origen reciente contienen una mayor proporción de secuencias simples que las proteínas antiguas, lo que sugiere que las LCRs contribuyen al origen de nuevos genes (Toll-Riera *et al.*, 2012). Además de la baja complejidad en la secuencia (Toll-Riera *et al.*, 2012), los genes nuevos tienden a ser estructuralmente desordenados (Kersting *et al.*, 2012; Bornberg-Bauer *et al.*, 2015; Basile *et al.*, 2017; Wilson *et al.*, 2017), lo que les confiere promiscuidad y la capacidad de unión a diversos sustratos; estas regiones actúan como un buffer, presentando altas tasas de mutación (Tokuriki & Tawfik, 2009).

La simplicidad o baja complejidad en la secuencia también se presenta en los nuevos dominios o nuevas regiones en las proteínas y existe una propensión a encontrarse en el dominio N-terminal de la proteína (Toll-Riera & Alba, 2013).

Se ha considerado que los genes nuevos han sido cruciales en las innovaciones evolutivas adaptativas. Las LCRs son fuente de variación genética (Radó-Trilla, 2013), al igual que las IDRs; resulta interesante que las clases de proteínas cuyas funciones y mecanismos moleculares pueden ser acoplados a procesos adaptativos, tienden a presentar un contenido de porcentaje más alto de IDRs, entre las clases principales se encuentran los factores de *transcripción*, *membrana nuclear*, *dominios de unión al DNA* (coordinación del Zn), *factores de defensa*, *virulencia y enfermedades* y *unión a RNA* (Nilsson *et al.*, 2011).

El papel de las LCRs e IDRs en la evolución temprana de la vida

La preservación de algunas secuencias simples y regiones intrínsecamente desordenadas en Bacteria y Archaea (Tabla IV y Tabla V), sugiere su presencia en el proteoma del último ancestro común (LCA). La presencia de secuencias simples en proteínas ancestrales, implica que el deslizamiento de la polimerasa es un mecanismo antiguo.

Algunas LCRs e IDRs conservadas tienen importancia en la estructura y función de las proteínas donde se encuentran: en dominios funcionales, en segmentos que unen a los dominios proteicos y en extremos terminales, sugiere que pudieron tener un papel importante en la evolución temprana de la vida, contribuyendo al aumento del tamaño del genoma, origen de genes y a la variación genética y materia prima (Becerra *et al.*, 2002).

Las proteínas tienen una temporalidad relativa, esto permite establecer una polaridad de las LCRs dentro de las proteínas; aquellas conservadas en los dos dominios celulares que interactúan con

RNA, son más antiguas que las que lo hacen con DNA (Delaye & Lazcano, 2000). En el mundo de RNA/proteínas, aquellas secuencias simples que presentaban una composición de aminoácidos básica, pudieron haber contribuido en la estabilización de las conformaciones del RNA necesarias para funciones catalíticas, replicativas o estructurales. La conservación de LCRs en la Proteína ribosomal L1 y de IDRs en la Chaperona HSP70, en proteínas ribosomales y en DEAD-box helicasas de RNA, sugieren que tuvieron un aporte durante esa etapa temprana; una evidencia adicional, es el incremento de la actividad ribozímica mediante la asociación con un péptido corto rico en lisinas y de composición simple (AAKK) (Bergstrom *et al.*, 2001).

El nacimiento de las proteínas donde el RNA era el protagonista, sugiere que sus primeras funciones eran estructurales y se relacionaban en mayor medida con la estabilización del RNA, actuando como chaperonas de RNA (Csermely, 1997; Poole *et al.*, 1998; Delaye & Lazcano, 2000; Tompa & Csermely, 2004). “Es altamente improbable que las primeras proteínas fueran enzimas complejas con una actividad catalítica exquisita y finamente establecida” (Delaye & Lazcano, 2000). Posteriormente, una vez establecido un “andamio” comenzaron a tener actividad catalítica. Se han descrito asociaciones entre proteínas que no requieren de un plegamiento específico ni de la interacción específica de residuos, éstas solo se originan mediante la atracción entre cargas opuestas y pueden formar complejos estables y dinámicos (Borgia *et al.*, 2018); es probable que este tipo de asociación sea antiguo.

Es posible que los primeros genes codificantes, fueran altamente repetitivos, como resultado de la expansión de repeticiones. Conforme fueron adquiriendo periodicidades más largas, adoptaron una estructura secundaria de hélices alfa y hojas beta (Ohno & Epplen, 1983), sin embargo, la presencia de LCRs dentro de IDRs demuestra que también dan lugar a regiones desordenadas.

Las secuencias simples, demuestran que solo se requiere un pequeño número de aminoácidos para hacer proteínas con funciones importantes, lo que sostiene la teoría de que las primeras proteínas codificadas genéticamente tenían LCRs (Poole *et al.*, 1998). Posiblemente, los aminoácidos de esos polipéptidos estaban conformados por aminoácidos tempranos, de acuerdo con las teorías de la temporalidad de los aminoácidos, se ha descrito que los polipéptidos más antiguos tienden a ser más desordenados (Trifonov, 2009; Di Mauro *et al.*; 2012).

Uno de los métodos para el estudio del mundo de RNA/proteínas, es el análisis de virus con genomas de RNA, utilizando la naturaleza de su material genético como modelo (Jácome *et al.*, 2015;

Campillo-Balderas *et al.*, 2015). Las IDR's se encuentran en la nucleocápside de virus de RNA pertenecientes a las familias *Flaviviridae* y *Coronaviridae*, donde interactúan directamente con el RNA y le ayudan a adoptar su conformación funcional, además de actuar como chaperonas de RNA, incrementar la actividad de ribozimas y habilitar el apareamiento de las cadenas del RNA (Zuñiga *et al.*, 2007; Nagy *et al.*, 2008); otras IDR's son abundantes en la maquinaria de replicación de los paramyxovirus (Communie *et al.*, 2014), al igual que ocurre con las LCR's en la proteína gp120 del virus VIH que proveen variabilidad y materia prima para generar nuevas funciones (Velasco *et al.*, 2013) lo que indica que el surgimiento de secuencias simples pudo haber incrementado el tamaño de los genomas primitivos de RNA (Poole *et al.*, 1998; Velasco *et al.*, 2013).

Las secuencias simples también pudieron haber tenido importancia en la generación de proteínas de unión a membrana. La abundancia de LCR's en la categoría funcional de transporte de membrana (Fig. 10) y la conservación de ATPasas en los dos dominios celulares indica que es probable su presencia en etapas evolutivas tempranas. Más aun, se ha reportado que pequeños oligopéptidos sintéticos de baja complejidad en su composición (Leu-Ser-Ser-Leu-Leu-Ser-Leu) (Lear *et al.*, 1988), los cuales pudieron haberse producido en condiciones primitivas y son lo suficientemente largas para abarcar la fase hidrocarbonada de las bicapas lipídicas, tienen permeabilidades y tiempos de vida similares a los de los canales selectivos de protones y al receptor de acetilcolina.

El deslizamiento de la polimerasa parece tener una tendencia que acumula el número de secuencias simples a lo largo de la evolución, ya que son abundantes en todos los seres vivos.

El estudio de secuencias simples y regiones intrínsecamente desordenadas conservadas en Bacteria y Archaea, permite indagar acerca del papel que tuvieron en etapas tempranas de la evolución biológica; debido a que la causa de su conservación es por motivos funcionales y estructurales, quizá en la actualidad, no se aprecie la contribución del deslizamiento de la polimerasa y la huella de repeticiones en las proteínas se difumine con el transcurso del tiempo.

Conclusiones

Las secuencias simples y regiones intrínsecamente desordenadas forman parte de los proteomas de Bacteria y Archaea. La proporción de secuencias simples en los procariontes parece estar relacionada con el contenido de GC en los genomas y es mayor su prevalencia en comparación con las IDRs en los phyla de Bacteria y Archaea. Las IDRs son abundantes en los organismos halófilos y presentan bajos porcentajes en termófilos. El análisis de la secuencia primaria en LCRs e IDRs, reveló un sesgo composicional no azaroso y la frecuente localización de secuencias simples dentro de regiones intrínsecamente desordenadas, sugiere alguna relación evolutiva entre ambos. La conservación de aquellas que presentan una alta distribución filogenética, sugiere su posible presencia en el complemento génico del último ancestro común (LCA por sus siglas en inglés). En el mundo de RNA/proteínas, aquellas secuencias simples y regiones intrínsecamente desordenadas que presentaban una composición de aminoácidos básica, pudieron haber contribuido en la estabilización de las conformaciones de RNA catalíticas, replicativas o estructurales, lo cual es sugerido por la conservación de LCRs en la riboproteína L1 y las IDRs conservadas en proteínas ribosomales antiguas y en DEAD-box helicasas de RNA, por lo que es posible inferir ambos fenómenos han estado presentes en los proteomas de los seres vivos desde la evolución temprana de la vida.

Literatura citada

Albà M, Tompa P, Veitia. (2007). Amino Acid Repeats and the Structure and Evolution of Proteins. En Volff J-N (ed): Gene and Protein Evolution. *Genome Dyn.* Basel, Karger. **3**, pp 119-130.

Albà M, Guigó R. (2004). Comparative analysis of amino acid repeats in rodents and humans. *Genome Res.* **14**:549–554.

Altschul S, Gish W, Miller W, Myers E, Lipman D. (1990). Basic local alignment search tool. *J Mol Biol.* **215**(3):403-10.

Andrade M, Perez-Iratxeta C, Ponting C. (2001). Protein Repeats: Structures, Functions, and Evolution. *Journal of Structural Biology.* **134**: 117–131.

Atkins J, Boateng S, Sorensen T, McGuffin L. (2015). Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies. *Int. J. Mol. Sci.* **16**:19040-19054.

Ball, K, Wemmer, D, Head-Gordon, T. (2014). Comparison of structure determination methods for intrinsically disordered amyloid- β peptides. *The journal of physical chemistry. B.* **118**(24), 6405-16.

Basile W, Sachenkova O, Light S, Elofsson A. (2017), High GC content causes orphan proteins to be intrinsically disordered. *PLOS Computational Biology.* **13**(3): e1005375.

Basile W, Elofsson A. (2018). Why do eukaryotic proteins contain more intrinsically disordered regions?. *bioRxiv.* 270694.

Bebenek K, Kunkel T. (1990). Frameshift errors initiated by nucleotide misincorporation. *Proc. Natl. Acad. Sci.* Vol. **87**: 4946-4950.

Becerra A, Islas S, Leguina J, Silva E, Lazcano A. (1997). Polyphyletic gene losses can bias backtrack characterizations of the cenancestor. *J Mol Evol.* 45:115–118.

Becerra A, Cocho G, Delaye L, Lazcano A. (2002). Simple sequences It is something you have whether you like it or not. *Origins Of Life & Evolution Of The Biosphere.* **32**(5-6): 485.

Becerra A, Delaye L, Islas S, Lazcano A. (2007). The Very Early Stages of Biological Evolution and the Nature of the Last Common Ancestor of the Three Major Cell Domains. *Annu. Rev. Ecol. Evol. Syst.* **38**:361–79.

Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews B, Boone C, Bader G, Myers C, Kim P. (2011). Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biology.* **12**:R14.

Bergstrom R, Mayfield L, Corey D. (2001) A bridge between the RNA and protein worlds? Accelerating delivery of chemical reactivity to RNA and DNA by a specific short peptide (AAKK). *Chem Biol.* **2**:199-205.

Britten R, Kohne D. (1968). Repeated Sequences in DNA. *Science.* **161**: 529-540.

- Bordin N, González-Sánchez J, Devos D. (2018). PVCbase: an integrated web resource for the PVC bacterial proteomes. *Database*. bay042.
- Borgia M, Bugge K, Kissling V, Heidarsson P, Fernandes C, Sottini A, Soranno A, Buholzer K, Nettels D, Kragelund B, Best R, Schuler B. (2018). Extreme disorder in an ultrahigh-affinity protein complex. *Nature*. **555**(7694):61-66.
- Bornberg-Bauer E, Schmitz J, Heberlein M. (2015). Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult'. *Biochem Soc Trans*. **43**(5):867-73.
- Burra P, Kalmar L, Tompa P. (2010). Reduction in Structural Disorder and Functional Complexity in the Thermal Adaptation of Prokaryotes. *PLoS ONE*. **5**(8): e12069.
- Campillo-Balderas J, Lazcano A, Becerra A. (2015). Viral Genome Size Distribution Does not correlate with the Antiquity of the Host Lineages. *Front. Ecol. Evol*. **3**. Art. 143.
- Carpousis A. (2007). The RNA Degradosome of Escherichia coli: An mRNA-Degrading Machine Assembled on RNase E. *Annu. Rev. Microbiol*. **61**:71–87.
- Chandran, Luisi B. (2006). Recognition of Enolase in the Escherichia coli RNA Degradosome. *J. Mol. Biol*. **358**: 8–15.
- Chaudhry, S. R., Lwin, N., Phelan, D., Escalante, A. A., & Battistuzzi, F. U. (2018). Comparative analysis of low complexity regions in Plasmodia. *Scientific reports*. **8**(1), 335.
- Csermely P. (1997). Proteins, RNAs and chaperones in enzyme evolution: a folding perspective. *TIBS*. **22**.
- Coletta A, Pinney J, Weiss Solís D, Marsh J, Pettifer S, Attwood T. (2010). Low-complexity regions within protein sequences have position-dependent roles. *BMC Systems Biology*. **4**:43.
- Communie G, Ruigrok R, Jensen M, Blackledge M. (2014). Intrinsically disordered proteins implicated in paramyxoviral replication machinery. *Curr Opin Virol*. **5**:72-81.
- Das C, Franklin A. (2003). Sequence and Structure Space of RNA-Binding Peptides. *Biopolymers*. **70**: 80–85.
- Delaye, L, Lazcano, A. (2000) RNA-binding peptides as molecular fossils In J.Chela-Flores, G. Lemerchand, J. Oró (eds) *Astrobiology: Origins from the Big-Bang to Civilization*. Proceedings of the First Ibero-American School of Astrobiology (Kluwer Academic Publishers, Dordrecht), pp. 285-288.
- Delaye L, Becerra A, Lazcano A. (2005). The last common ancestor: What's in a name?. *Orig. Life Evol. Biosph*. **35**:537–54.
- Delaye L, Becerra A. (2012). Cenancestor, the Last Universal Common Ancestor. *Evo Edu Outreach*. **5**:382–388.
- Di Mauro E, Dunker A. K, Trifonov E.N. (2012), “Disorder to Order, Nonlife to Life: In the Beginning There Was a Mistake”. En: Seckbach J. (Eds) *Genesis - In The Beginning*. Cellular Origin, Life in Extreme Habitats and Astrobiology, vol 22. Springer, Dordrecht.

- Dunker K, Brown C, Lawson J, Iakoucheva M, Obradovic Z. (2002). Intrinsic Disorder and Protein Function. *Biochemistry*. Vol. 41, **21**:6573-6582.
- Dunker K, Obradovic Z, Romero P, Garner E, Brown C. (2010). Intrinsic protein disorder in complete genomes. *Genome Informatics*. **11**:161-71.
- Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature reviews, Genetics*. **5**(6):435-445.
- Egea P, Tsuruta H, P. de Leon G, Napetschnig J, Walter P, Stroud R. (2008). Structures of the Signal Recognition Particle Receptor from the Archaeon *Pyrococcus furiosus*: Implications for the Targeting Step at the Membrane. *PLoS one*. **3**, Issue 11.
- Fornerod M. (2012). RS and RGG repeats as primitive proteins at the transition between the RNA and RNP worlds. *Nucleus*. **3**(1):4-5.
- Fournier G, Neumann J, Gogarten P (2010). Inferring the ancient history of the translation machinery and genetic code via recapitulation of ribosomal subunit assembly orders. *PLoS ONE*. **5**(3): e9437.
- Fox G. (2010). Origin and Evolution of the Ribosome. In Deamer D & Szostak J (ed): *Cold Spring Harb Perspect Biol*. **2**:a003483.
- Frankel D. (2000). Fitting peptides into the RNA world. *Current Opinion in Structural Biology*. **10**:332- 340.
- Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M, Appel R, Bairoch A (2005). Protein Identification and Analysis Tools on the ExPASy Server. In Walker J (ed): *The Proteomics Protocols Handbook*, Humana Press. pp. 571-607.
- Hainzl T, Sauer-Eriksson A. (2015). Signal-sequence induced conformational changes in the signal recognition particle. *Nature Communications*. **6**:7163.
- Hancock J. (1995). The Contribution of slippage-Like Processes to Genome Evolution. *J Mol Evol*. **41**:1038-1047.
- Hancock J. (1996). Simple sequences and the expanding genome. *BioEssays*. **18**: 421-425.
- Harris J, Kelley S, Spiegelman G, Pace N. (2003). The genetic core of the universal ancestor. *Genome Res*. **13**:407-12.
- He W, Li C, Lu C. (2011). Regulation and characterization of the dadRAX locus for D-amino acid catabolism in *Pseudomonas aeruginosa* PAO1. *J Bacteriol*. **193**(9):2107-15.
- Holliday G, Fischer J, Mitchell J, Thornton J. (2011). Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis. *The FEBS journal*. **278**(20), 3835-45.
- Hsiao C, Lenz T, Peters J, Fang P, Schneider D, Anderson E, Preeprem T, Bowman J, O'Neili N, Lie L, Athavale S, Gossett J, Trippe C, Murray J, Petrov A, Wartell R, Harvey S, Hud N, Williams L.

- (2013). Molecular paleontology: a biochemical model of the ancestral ribosome. *Nucleic Acids Research*. Vol. **41**, No. 5: 3373–3385.
- Ishino S, Kawata Y, Taguchi H, Kajimura N, Matsuzaki K, Hoshino M. (2015). Effects of C-terminal truncation of chaperonin GroEL on the yield of in-cage folding of the green fluorescent protein. *J Biol Chem*. **290**(24):15042-51.
- Ivanyi-Nagy R, Lavergne J, Gabus C, Ficheux D, Darlix J. (2008). RNA chaperoning and intrinsic disorder in the core proteins of Flaviviridae. *Nucleic Acids Research*. Vol. **36**, No. 3: 712–725.
- Jácome R, Becerra A, Ponce de León S, Lazcano A. (2015). Structural Analysis of Monomeric RNA-Dependent Polymerases: Evolutionary and Therapeutic Implications. *PLoS ONE*. **10**(9):e0139001.
- Jones, D, Cozzetto D. (2014). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics (Oxford, England)*. **31**(6), 857-63.
- Joyce G. (2002). The antiquity of RNA-based evolution. *Nature* **418**: 214-221.
- Kanehisa, M, Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28**: 27-30.
- Katti M, Ranjekar P, Gupta V. (2001). Differential distribution of simple sequences repeats in eukaryotic genome sequences. *Mol Biol Evol*. **18**: 1161-1167.
- Katoh K, Standley D. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol*. **30**(4):772–780.
- Keenan R, Freymann D, Walter P, Stroud R. (1998). Crystal Structure of the Signal Sequence Binding Subunit of the Signal Recognition Particle. *Cell*. **94**:181–191.
- Kersting A, Bornberg-Bauer E, Moore A, Grath S. (2012). Dynamics and Adaptive Benefits of Protein Domain Emergence and Arrangements during Plant Genome Evolution. *Genome Biol. Evol*. **4**(3):316– 329.
- Klein D, Moore P, Steitz A. (2004). The Roles of Ribosomal Proteins in the Structure Assembly, and Evolution of the Large Ribosomal Subunit. *J. Mol. Biol*. **340**:141–177.
- Kovacs N, Petrov A, Lanier K, Williams L. (2017). Frozen in Time: The History of Proteins. *Molecular biology and evolution*. **34**(5), 1252-1260.
- Kühnel K, Luisi B (2001). Crystal Structure of the Escherichia coli RNA Degradosome Component Enolase. *J. Mol. Biol*. **313**: 583-592.
- Kumari B, Kum R, Kumar M (2014). Low complexity and disordered regions of proteins have different structural and amino acid preferences. *The Royal Society of Chemistry*. DOI: 10.1039/c4mb00425f.
- Lane D, Scott D, Hebl M, Guerra R, Osherson D, Zimmer H. (2013). Introduction to Statistics. Rice University. pp: 170.

- Lazcano A, Guerrero R, Margulis L, Oro J. (1988). The Evolutionary Transition from RNA to DNA in early cells. *J Mol Evol.* **27**:283-290.
- Lazcano A, Fox G, Oró J. (1992). Life before DNA: the origin and early evolution of early archaean cells. In R. P. Mortlock: (ed), *The Evolution of Metabolic Function*: CRC Press, Boca Raton.
- Lear J, Wasserman Z, DeGrado W. (1988). Synthetic amphiphilic peptide models for protein ion channels. *Science.* **240**:1177-1181.
- Lechner M, Findei S, Steiner L, Marz M, Stadler P, Prohaska S. (2011). Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics.* **12**:124.
- Levinson G, Gutman G. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol.* **4**(3):203-21.
- López-Ramírez V, Alcaraz L, Moreno-Hagelsieb G, Olmedo-Álvarez G. (2011). Phylogenetic Distribution and Evolutionary History of Bacterial DEAD-Box Proteins. *J Mol Evol.* **72**:413–431.
- Lozada-Chávez I. (2004). El papel de las secuencias simples en los proteomas virales. (Tesis de Licenciatura), Facultad de Ciencias, UNAM, México.
- Lupas A, Alva V. (2017). Ribosomal proteins as documents of the transition from unstructured (poly)peptides to folded proteins. *Journal of Structural Biology.* **198**: 74–81.
- Maki J, Schnobrich D, Culver G. (2002). The DnaK chaperone system facilitates 30S ribosomal subunit assembly. *Mol Cell.* **10**(1):129-38.
- Meng F, Uversky V, Kurgan L. (2017). Computational prediction of intrinsic disorder in proteins. *Current Protocols in Protein Science.* **88**, 2.16.1–2.16.14.
- Mészáros B, Erdos G, Dosztányi Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**(W1):W329-W337.
- Mirkin B, Fenner T, Galperin M, Koonin E. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last common ancestor, and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**:2.
- Necci, M, Piovesan, D, Tosatto, S. (2016). Large-scale analysis of intrinsic disorder flavors and associated functions in the protein sequence universe. *Protein science.* **25**(12), 2164-2174.
- Nilsson J, Grahn M, Wright A. (2011). Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome biology.* **12**(7), R65.
- Ohno S, Epplen J. (1983). The primitive code and repeats of base oligomers as the primordial-encoding sequences. *Prod. Natl. Acad. Sci.* **80**: 3391-3395.
- Olfield C, Dunker K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **83**:553–84.

- Patel A, Louder R, Greber B, Grünberg S, Luo J, Fang J, Liu Y, Ranish J, Hahn S, Nogales E. (2018). Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science*. **21**:362(6421).
- Paul S, Bag S, Das S, Harvill E, Dutta C. (2008). Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biology*. **9**:R70.
- Pavlović-Lažetić, G, Mitić, N, Kovačević, J, Obradović, Z, Malkov, S, Beljanski, M. (2011). Bioinformatics analysis of disordered proteins in prokaryotes. *BMC bioinformatics*. **12**, 66.
- Pearson, K. (1948). *Early Statistical Papers*. Cambridge, England: University Press.
- Peng Z, Yan J, Fan X, Mizianty M, Xue B, Wang K, Hu G, Uversky V, Kurgan L. (2015). Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci*. **72**(1):137-51.
- Poole A, Jeffares D, Penny D. (1998). The Path from the RNA World. *J Mol Evol*. **46**:1–17.
- Radó-Trilla N, Albà M. (2012). Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol Biol*. **12**: 155.
- Radó-Trilla N. (2013). Low-complexity regions in proteins as a source of evolutionary innovation. (Tesis Doctoral), Universtitat Pompeu Fabra, Barcelona, España.
- Ranea A, Sillero A, Thornton M, Orengo A. (2006). Protein superfamily evolution and the last universal common ancestor (LUCA). *J. Mol. Evol*. **63**:513–25.
- Raymann K, Brochier-Armanet C, Gribaldo S. (2015). The two-domain tree of life is linked to a new root for the Archaea. *PNAS*. **21**; 6670–6675.
- Redder P, Hausmann S, Khemici V, Yasrebi H, Linder P. (2015). Bacterial versatility requires DEAD- box RNA helicases. *FEMS Microbiology Reviews*. **39**: 392–412.
- Romero P, Obradovic Z, Li X, Garner E, Brown C, Dunker K. (2001). Sequence Complexity of Disordered Protein. *PROTEINS: Structure, Function, and Genetics* .42:38–48.
- Sagulenko E, Nouwens A, Webb I, Green K, Yee B, Morgan G, Leis A, Lee K, Butler M, Chia N, Pham U, Lindgreen S, Catchpole R, Poole A, Fuerst J. (2017). Nuclear Pore-Like Structures in a Compartmentalized Bacterium, *PLoS One*. **12**(2): e0169432.
- Saqi M. (1995). An analysis of structural instances of low complexity sequence segments. *Protein Eng*. **8**:1069–1073.
- Shajani Z, Sykes M, Williamson J. (2011). Assembly of bacterial ribosomes. *Annu Rev Biochem*. **80**:501–526.
- Siliqi D, Foadi J, Mazzorana M, Altamura D, Méndez-Godoy A, Sánchez-Puig N. (2018). Conformational Flexibility of Proteins Involved in Ribosome Biogenesis: Investigations via Small Angle X-ray Scattering (SAXS). *Crystals*. **8**(3), 109.

- Sim K, Creamer T. (2002). Abundance and distributions of eukaryote protein simple sequences. *Mol Cell Proteomics*. **1**(12):983-95.
- Smock R, Blackburn M, Gierasch M. (2011). Conserved, Disordered C Terminus of DnaK Enhances Cellular Survival upon Stress and DnaK *in Vitro* Chaperone Activity. *The Journal of Biological Chemistry*. **286**:36, pp.31821–31829.
- Solano L, Rojas C. (2005). Estadística descriptiva y distribuciones de probabilidad. Universidad del Norte. pp: 76-77.
- Tatusova T, Ciuffo S, Fedorov B, O'Neill K, Tolstoy I. (2014). RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res*. **42**: D553–D559.
- Tautz D, Renz M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res*.**12**:4127-4138.
- Tautz D, Trick M, Dover G. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature*. **322**:652–656.
- The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources, *Nucleic Acids Research*. **45**, Issue D1: D331–D338.
- The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*. 45, Issue: D158–D169.
- Timsit Y, Acosta Z, Allemand F, Chiaruttini C, Springer M. (2009). The Role of Disordered Ribosomal Protein Extensions in the Early Steps of Eubacterial 50 S Ribosomal Subunit Assembly. *Int. J. Mol. Sci*. **10**: 817-834.
- Toll-Riera M, Radó-Trilla N, Martys F, Albà M. (2012). Role of low-complexity sequences in the formation of novel protein coding sequences. *Mol Biol Evol*. **29**(3):883-6.
- Toll-Riera M, Albà M. (2013). Emergence of novel domains in proteins. *BMC evolutionary biology*. **13**, 47.
- Tokuriki N, Tawfik D. (2009). Protein dynamism and evolvability. *Science*. **324**(5924):203-7. doi: 10.1126/science.1169375.
- Tompa P, Csermely P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *FASEB*. 0892-6638/04/0018-1169.
- Tompa P, Fersht A. (2009). Structure and Function of Intrinsically Disordered Proteins. CRC press. p 106.
- Tompa P. (2010). Intrinsically disordered proteins: a 10-year recap. *Trends in Biochemical Sciences December*. Vol. 37, No. **12**. 50916.
- Tompa P, Kovacs D. (2010). Intrinsically disordered chaperones in plants and animals. *Biochemistry and cell biology*. **88**(2):167–174.

- Trifonov E. (2009). The origin of the genetic code and of the earliest oligopeptides. *Research in Microbiology*. **160**:481- 486.
- Uversky V (2014). Introduction to Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **114**: 6557–6560 .
- van der Lee R, Buljan M, Lang B, Weatheritt R, Daughdrill G, Dunker A, Fuxreiter M, Gough J, Gsponer J, Jones D, Kim P, Kriwacki, Oldfield C, Pappu R, Tompa P, Uversky V, Wright P, Babu M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114**: 6589–6631.
- Varadi M, Zsolyomi F, Guharoy M, Tompa P (2015) Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS ONE*. **10**(10): e0139731.
- Vázquez-Cuevas C, Covarrubias-Robles A. 2011. Las proteínas desordenadas y su función: una nueva forma de ver la estructura de las proteínas y la respuesta de las plantas al estrés. *Revista Especializada en Ciencias Químico-Biológicas*. **14**:2.
- Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D. (2007). Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev.* **71**(3):495-548.
- Velasco A, Becerra A, Hernández-Morales R, Delaye L, Jiménez-Corona M, Ponce-de-León S, Lazcano A. (2013). Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein. *Journal of Theoretical Biology*. **338**: 80–86.
- Viguera E, Canceill D, Ehrlich S. (2001). Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.* **20**: 2587–2595.
- Walsh I, Martin A, Domenico T, Tosatto S. (2012). ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*. **28**(4):503-9.
- Wang C, Uversky V, Kurgan L. (2016). Disordered nucleome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics*. **16**(10):1486-98.
- Ward J, Sodhi J, McGuffin L, Buxton B, Jones D. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol.* **26**. 337(3):635-45.
- Weaver J, Rye H. (2014). The C-terminal Tails of the Bacterial Chaperonin GroEL Stimulate Protein Folding by Directly Altering the Conformation of a Substrate Protein. *THE JOURNAL OF BIOLOGICAL CHEMISTRY*. **289**, 33: 23219–23232.
- White O et al. (1999). Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science*. **286**:1571-7.
- Williams T, Foster P, Cox C, Embley T. (2013). An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. **504**: 231-236.
- Wilson B, Foy S, Neme R, Masel J. (2017). Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nat Ecol Evol.* **1**(6):0146-146.

Worbs M, Huber R, Wahl M. (2000) Crystal structure of ribosomal protein L4 shows RNA-binding sites for ribosome incorporation and feedback control of the S10 operon. *EMBO J* **19**: 807-818.

Wootton J. & Federhen S. (1993). Statistics of local complexity in amino acid sequence and sequences database. *Compt Chem.* **17**:149-163.

Xue B, Williams R, Oldfield C, Dunker K, Uversky V. (2010). Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst Biol.* 4(Suppl 1): S1.

Yang, S., Doolittle, R, Bourne, P. (2005). Phylogeny determined by protein domain content. *Proc Natl Acad Sci. USA.* **102**: 373-378.

Yusupov M, Yusupova G, Baucom A, Lieberman K, Earnest T, Cate J, Noller H. (2001). Crystal structure of the ribosome at 5.5 Å resolution. *Science.* **292**: 883-896.

Zhou T, Radaev S, Rosen B, Gatti D. (2001). Conformational changes in four regions of the Escherichia coli ArsA ATPase link ATP hydrolysis to ion translocation. *J Biol Chem.* **276**(32):30414-22.

Zimmer C, von Gabain A, Henics T. (2001). Analysis of sequence-specific binding of RNA to Hsp70 and its various homologs indicates the involvement of N- and C-terminal interactions. *RNA.* **7**(11), 1628-37.

Zuckerkindl E, Pauling L. (1965). Molecules as documents of evolutionary history. *J Theor Biol.* **8**(2):357-66.

Anexos

Anexo I) Algoritmo SEG

Descripción

SEG es un algoritmo que divide a la secuencia en dos tipos: segmentos de baja y alta complejidad. Las regiones de baja complejidad representan “regiones con un sesgo composicional”. El sesgo se basa en la composición de los residuos, no necesariamente en las repeticiones o periodicidades en la secuencia. Permite localizar secuencias simples desde homopolímeros a regiones no globulares de las proteínas.

Formato de entrada

En formato FASTA, pueden ser una secuencia o varias.

Sinópsis

```
seg sequence [ W ] [ K(1) ] [ K(2) ] [ -x ] [ options ]
```

Parámetros

La búsqueda de segmentos de secuencias simples utiliza tres parámetros, los cuales pueden ser especificados en la línea de comandos seguidos del nombre del archivo y antes de las opciones; van en el siguiente orden:

- 1) Longitud de la ventana del disparador [W]:** Representa a la longitud de la ventana inicial de búsqueda y debe representar un entero mayor a cero. El valor default es 12
- 2) Complejidad del disparador [K1]:** Es la complejidad máxima de una ventana inicial en unidades de bits; debe ser de un número entre cero (para la búsqueda de las secuencias más simples) a 4.322 bits ($\log [\text{base}2] 20$) para secuencias de aminoácidos. El valor default es 2.2 bits
- 3) Complejidad de la extensión [K2]:** La complejidad máxima de la extensión de una ventana en unidades de bits. Solamente valores mayores que K1 son efectivos en extender ventanas adicionales. El intervalo posible de valores es similar al de K1. El valor default es 2.5 bits.

El algoritmo SEG tiene dos estados:

i) La primera etapa identifica segmentos aproximadas de baja complejidad en bruto. La búsqueda se basa en los parámetros W, K2 (1) y K2 (2). Las ventanas disparadoras son definidas, incluyendo a las ventanas sobrelapantes, de longitud W y complejidad menores a K1. La “Complejidad” se define como la ecuación (3) de Wootton y Federhen (1993). Cada ventana disparadora es extendida en un contig (segmento en bruto) en ambas direcciones, por la fusión con las ventanas de extensión, que son ventanas sobrelapantes de longitud W y complejidad menor o igual a K2.

ii) La segunda etapa es la optimización local de cada segmento identificado en la etapa anterior. Cada segmento es reducido en un segmento de baja complejidad, el cual puede ser el segmento en bruto completo, aunque generalmente es una subsecuencia. La subsecuencia óptima tiene el valor menor de la probabilidad P (0) de la ecuación 5 de Wootton y Federhen (1993).

Opciones

Se colocan después de los parámetros en la línea de comandos

Opción	Descripción
-a	Salida para los segmentos de alta y baja complejidad en un archivo con formato FASTA, como un sistema de entradas separadas con líneas principales.
-c	[caracteres por línea] Número de caracteres por línea de salida. El valor estándar es 60. Otros caracteres, tales como el número de residuos, son adicionales.
-h	Salida solamente a los segmentos de alta complejidad en un archivo con formato FASTA, como un sistema de entradas separadas con líneas principales.
-l	Salida solamente a los segmentos de baja complejidad en un archivo con formato FASTA, como un sistema de entradas separadas con líneas principales.
-m	[longitud] Longitud mínima en residuos para un segmento de alta complejidad. El valor estándar es 0. Segmentos cortos son fusionados con segmentos de baja complejidad adyacentes.
-o	Muestra todos los sobrelapamientos, segmentos de baja complejidad disparados independientemente. Éstos son fusionados por estandarización.
-q	Produce un formato de salida con la secuencia en bloques numerados con marcas que ayudan a contar los residuos. Los segmentos de baja y alta complejidad están en minúsculas y mayúsculas, respectivamente.
-t	[longitud] Parámetro de la “longitud máxima de ajuste”. El valor estándar es 100. Éste controla el espacio de la búsqueda (y el tiempo de la búsqueda) durante la optimización de segmentos crudos. Por default, las subsecuencias de 100 residuos o más, de menor longitud que el segmento crudo son

	omitidas de la búsqueda. Este parámetro se puede aumentar para dar una búsqueda más extensa si los segmentos crudos son más largos que 100 residuos.
-x	Opción que enmascara las secuencias de amino. Cada secuencia de entrada es representada por una sola secuencia de la salida en formato FASTA con las regiones de baja complejidad, sustituidas por cadenas con caracteres “x”.

Anexo II) LCRs en phyla de Bacteria y Archaea

Phylum	N	Media	Desviación estándar	Mediana	Mínimo	Máximo
Bacteria						
Gemmatimonadetes	2	37.74	11.17	37.74	29.83	45.64
Actinobacteria	75	32.61	9.11	33.20	7.56	50.42
Deinococcus-Thermus	7	27.60	11.97	31.10	5.56	37.04
Chloroflexi	8	22.90	7.22	25.76	10.95	30.39
Betaproteobacteria	41	24.27	8.75	24.90	4.30	42.37
Acidobacteria	6	23.10	2.90	23.95	19.47	26.39
Armatimonadetes	1	23.30	NA	23.30	23.30	23.30
Alphaproteobacteria	65	22	7.09	22.12	0.57	37.56
Planctomycetes	6	26.70	13.48	21.98	14.61	50.79
Thermodesulfobacteria	1	20.05	NA	20.05	20.05	20.05
Unclassified Terrabacteria group	1	18.54	NA	18.54	18.54	18.54
Verrucomicrobia	5	20.62	6.23	18.15	15.09	29.29
Dictyoglomi	1	17.57	NA	17.57	17.57	17.57
Other proteobacteria	1	17	NA	17	17	17
Fibrobacteres	1	17	NA	17	17	17
Deltaproteobacteria	32	22.32	11.25	16.59	9.88	53.68
Chrysiogenetes	1	15.56	NA	15.56	15.56	15.56
Aquificae	5	14.57	7.78	15.51	1.37	20.41
Tenericutes	2	15.46	4.13	15.46	12.54	18.38
Spirochaetes	6	16.81	4.42	15.32	11.73	22.70
Elusimicrobia	1	15.06	NA	15.06	15.06	15.06
Fusobacteria	6	14.81	2.80	14.72	10.52	17.89

Gammaproteobacteria	96	16.21	6.94	13.92	1.45	42.61
Epsilonproteobacteria	8	13.75	2.49	13.77	10.40	18.06
Cyanobacteria	19	15.58	6.65	13.03	10.19	38.45
Firmicutes	80	13.68	5.66	12.69	0.10	53.33
Caldiserica	1	12.58	NA	12.58	12.58	12.58
Chlorobi	7	12.60	0.89	12.47	11.19	13.80
Nitrospirae	3	12.73	2.29	12.28	10.70	15.21
Thermotogae	8	13.58	4.96	12.22	7.79	22.87
Synergistetes	4	12.66	2.30	11.81	11.01	16.01
Bacteroidetes	56	13.13	4.30	11.60	9.55	29.09
Deferribacteres	4	12.26	2.78	11.44	9.95	16.22
Unclassified Bacteria	1	11.15	NA	11.15	11.15	11.15
Archaea						
Nanohaloarchaeota	1	17.67	NA	17.67	17.67	17.67
Euryarchaeota	57	17.26	6.14	15.52	9.19	29.68
Korarchaeota	1	15.29	NA	15.29	15.29	15.29
Crenarchaeota	15	16.81	4.29	14.98	11.61	24.02
Lokiarchaeota	1	12.39	NA	12.39	12.39	12.39
Thaumarchaeota	5	13.58	3.97	11.70	9.32	17.92
Bathyarchaeota	1	6.82	NA	6.82	6.82	6.82

Anexo III) Clasificación funcional de LCRs

Functional categories of LCRs in Bacteria	
Function	%
Membrane transport	20.618
Translation	6.451
Amino acid metabolism	6.290
Carbohydrate metabolism	5.853
Replication and repair	5.497
Energy metabolism	4.797
Signal transduction	4.149
Cellular community - prokaryotes	4.117
Folding, sorting and degradation	4.091
Metabolism of cofactors and vitamins	3.978
Enzyme families	3.776

Cell motility	3.084
Function unknown	3.078
Glycan biosynthesis and metabolism	2.712
Lipid metabolism	2.208
Nucleotide metabolism	2.047
Transcription	1.807
Transport and catabolism	1.619
Metabolism of other amino acids	1.493
Metabolism of terpenoids and polyketides	1.154
Xenobiotics biodegradation and metabolism	1.130
Cell growth and death	1.112
Unclassified: Energy metabolism	1.056
Biosynthesis of other secondary metabolites	1.038
Unclassified: Transport	0.950
Prokaryotic defense system	0.751
Unclassified: Structural proteins	0.742
Unclassified: Protein processing	0.713
Unclassified: Replication and repair	0.502
Unclassified: Others	0.495
Environmental adaptation	0.406
Unclassified: Carbohydrate metabolism	0.303
Unclassified: Cell growth	0.287
Unclassified: Cofactor metabolism	0.283
Unclassified: Signaling proteins	0.277
Membrane trafficking	0.198
Unclassified: Amino acid metabolism	0.182
Signaling molecules and interaction	0.148
Unclassified: Glycan metabolism	0.137
Unclassified: Lipid metabolism	0.116
Unclassified: Cell motility	0.105
Unclassified: Secondary metabolism	0.104
Unclassified: Transcription	0.076
Unclassified: Translation	0.040
Unclassified: Nucleotide metabolism	0.026
Unclassified viral proteins	0.002

Functional categories of LCRs in Archaea	
Function	%
Membrane transport	16.804

Translation	9.072
Amino acid metabolism	8.380
Energy metabolism	7.479
Carbohydrate metabolism	5.686
Replication and repair	5.512
Metabolism of cofactors and vitamins	5.318
Function unknown	5.243
Folding, sorting and degradation	4.725
Cellular community - prokaryotes	3.555
Cell motility	2.908
Nucleotide metabolism	2.549
Signal transduction	2.490
Enzyme families	2.330
Transport and catabolism	2.156
Transcription	1.773
Unclassified: Energy metabolism	1.753
General function prediction only	1.743
Lipid metabolism	1.733
Prokaryotic defense system	1.409
Metabolism of terpenoids and polyketides	1.285
Metabolism of other amino acids	1.280
Unclassified: Transport	1.135
Xenobiotics biodegradation and metabolism	1.061
Biosynthesis of other secondary metabolites	1.051
Unclassified: Replication and repair	0.936
Glycan biosynthesis and metabolism	0.931
Cell growth and death	0.876
Unclassified: Protein processing	0.533
Unclassified: Cofactor metabolism	0.408
Drug resistance	0.388
Unclassified: Signaling proteins	0.363
Unclassified: Others	0.339
Environmental adaptation	0.334
Unclassified: Transcription	0.304
Membrane trafficking	0.294
Unclassified: Carbohydrate metabolism	0.194
Unclassified: Translation	0.194
Unclassified: Lipid metabolism	0.144
Unclassified: Cell motility	0.139

Unclassified: Secondary metabolism	0.110
Signaling molecules and interaction	0.095
Unclassified: Amino acid metabolism	0.085
Unclassified: Cell growth	0.075
Unclassified: Structural proteins	0.055
Unclassified: Glycan metabolism	0.010
Unclassified: Nucleotide metabolism	0.010

Anexo IV) IDR en phyla de Bacteria y Archaea

Phylum	N	Media	Desviación estándar	Mediana	Mínimo	Máximo
Bacteria						
Planctomycetes	6	15.59	4.38	13.76	11.98	23.97
Actinobacteria	75	12.05	3.53	12.08	4.69	20.93
Gemmatimonadetes	2	9.64	0.77	9.64	9.10	10.18
Armatimonadetes	1	7.97	NA	7.97	7.97	7.97
Acidobacteria	6	8.18	1.38	7.96	6.66	10.48
Alphaproteobacteria	65	6.96	2.07	6.82	1.34	11.57
Betaproteobacteria	41	6.36	2.07	6.08	3.42	13.41
Chloroflexi	8	5.80	1.92	6.02	2.22	8.09
Cyanobacteria	19	5.55	1.69	5.67	2.33	9.16
Unclassified Terrabacteria group	1	5.54	NA	5.54	5.54	5.54
Nitrospirae	3	4.64	3.15	5.53	1.13	7.24
Fibrobacteres	1	5.43	NA	5.43	5.43	5.43
Deltaproteobacteria	32	8.40	6.40	5.36	1.25	25.34
Verrucomicrobia	5	5.65	1.76	4.84	3.72	7.81
Chrysiogenetes	1	4.63	NA	4.63	4.63	4.63
Other proteobacteria	1	4.61	NA	4.61	4.61	4.61
Gammaproteobacteria	96	4.89	2.20	4.32	1.83	13.77
Firmicutes	80	4.23	2.97	3.86	0.69	25.87
Deinococcus-Thermus	7	5.38	3.20	3.80	2.62	11.68
Chlorobi	7	2.97	1.07	3.52	1.44	3.95
Spirochaetes	6	4.23	2.89	3.47	1.26	9.74

Bacteroidetes	56	3.53	2.84	2.87	1.69	22.41
Unclassified Bacteria	1	2.77	NA	2.77	2.77	2.77
Elusimicrobia	1	2.61	NA	2.61	2.61	2.61
Epsilonproteobacteria	8	2.61	1.38	2.13	1.21	5.17
Synergistetes	4	2.29	1.14	1.95	1.34	3.93
Fusobacteria	6	2.23	0.99	1.92	1.25	4.05
Tenericutes	2	1.57	0.27	1.57	1.38	1.76
Deferribacteres	4	1.54	0.75	1.37	0.90	2.53
Caldiserica	1	1.07	NA	1.07	1.07	1.07
Thermotogae	8	0.99	0.34	1.02	0.49	1.48
Aquificae	5	1.03	0.35	0.95	0.66	1.60
Thermodesulfobacteria	1	0.69	NA	0.69	0.69	0.69
Dictyoglomi	1	0.63	NA	0.63	0.63	0.63
Archaea						
Nanohaloarchaeota	1	12.68	NA	12.68	12.68	12.68
Euryarchaeota	57	7.63	7.68	3.27	0.56	25.02
Thaumarchaeota	5	2.96	1.10	3.22	1.44	4.32
Lokiarchaeota	1	1.49	NA	1.49	1.49	1.49
Crenarchaeota	15	1.24	0.50	1.23	0.47	2.46
Bathyarchaeota	1	0.92	NA	0.92	0.92	0.92
Korarchaeota	1	0.69	NA	0.69	0.69	0.69

Anexo V) Clasificación funcional de IDRs

Functional categories of IDRs in Bacteria	
Function	%
Membrane transport	13.271
Translation	11.413
Replication and repair	10.255
Folding, sorting and degradation	8.123
Cell motility	5.443
Enzyme families	5.406
Signal transduction	4.639
Carbohydrate metabolism	4.310
Cellular community - prokaryotes	4.260

Transcription	2.963
Amino acid metabolism	2.790
Function unknown	2.581
Glycan biosynthesis and metabolism	2.504
Energy metabolism	2.315
Nucleotide metabolism	2.021
Metabolism of cofactors and vitamins	1.894
Drug resistance	1.737
Transport and catabolism	1.683
Cell growth and death	1.334
Unclassified: Structural proteins	1.042
Prokaryotic defense system	1.024
Lipid metabolism	0.961
General function prediction only	0.942
Unclassified: Replication and repair	0.860
Unclassified: Protein processing	0.756
Metabolism of other amino acids	0.616
Unclassified: Energy metabolism	0.579
Unclassified: Others	0.501
Unclassified: Cell growth	0.444
Xenobiotics biodegradation and metabolism	0.442
Biosynthesis of other secondary metabolites	0.400
Metabolism of terpenoids and polyketides	0.318
Unclassified: Signaling proteins	0.284
Unclassified: Carbohydrate metabolism	0.275
Unclassified: Transcription	0.269
Unclassified: Cell motility	0.183
Signaling molecules and interaction	0.169
Unclassified: Glycan metabolism	0.158
Environmental adaptation	0.152
Unclassified: Cofactor metabolism	0.142
Unclassified: Transport	0.134
Membrane trafficking	0.114
Unclassified: Secondary metabolism	0.097
Unclassified: Amino acid metabolism	0.070
Unclassified: Translation	0.070
Unclassified: Lipid metabolism	0.042
Unclassified: Nucleotide metabolism	0.011
Unclassified viral proteins	0.003

Functional categories of IDRs in Archaea	
Function	%
Translation	12.507
Replication and repair	9.167
Membrane transport	9.057
Amino acid metabolism	7.546
Carbohydrate metabolism	5.815
Metabolism of cofactors and vitamins	5.410
Energy metabolism	4.950
Folding, sorting and degradation	4.928
Function unknown	4.271
Cell motility	3.745
Cellular community - prokaryotes	3.406
Nucleotide metabolism	3.329
Transcription	3.209
Signal transduction	2.847
Enzyme families	2.322
Transport and catabolism	2.300
Xenobiotics biodegradation and metabolism	1.369
Lipid metabolism	1.358
Metabolism of other amino acids	1.325
Unclassified: Replication and repair	1.194
General function prediction only	0.964
Unclassified: Energy metabolism	0.964
Biosynthesis of other secondary metabolites	0.920
Metabolism of terpenoids and polyketides	0.909
Prokaryotic defense system	0.909
Cell growth and death	0.690
Glycan biosynthesis and metabolism	0.548
Unclassified: Protein processing	0.548
Unclassified: Others	0.526
Membrane trafficking	0.504
Drug resistance	0.361
Unclassified: Transcription	0.340
Unclassified: Cofactor metabolism	0.296
Environmental adaptation	0.285
Unclassified: Signaling proteins	0.252
Unclassified: Translation	0.208

Unclassified: Carbohydrate metabolism	0.186
Unclassified: Amino acid metabolism	0.175
Unclassified: Transport	0.131
Unclassified: Cell motility	0.088
Unclassified: Cell growth	0.077
Unclassified: Structural proteins	0.033
Unclassified: Glycan metabolism	0.022
Unclassified: Lipid metabolism	0.011

Anexo VI) LCRs con amplia distribución filogenética

Proteína	LCR	Posición en la secuencia	Dominio de localización de la LCR	Categoría funcional
Galactoquinasa (EC 2.7.1.6)	ggkllgagg g	Central	GHMP_kinases_ C	Biosíntesis y metabolismo del glicanos
Cisteína sintasa (EC 2.5.1.47)	gigtggtimgt g	Central	PALP	Metabolismo de aminoácidos
S-adenosilhomocisteína hidrolasa (EC 3.3.1.1)	aaaiaaaa	Central	AdoHcyase	Metabolismo de aminoácidos
Aldehído deshidrogenasa	aavaaraaq pa	N-terminal	Aldedh	Metabolismo de aminoácidos
D-aminoácido deshidrogenasa	vvvigggivg	N-terminal	DAO	Metabolismo de aminoácidos
Homoserina quinasa (EC 2.7.1.39)	aaaivaayaaa dallpa	Central	GHMP_kinases_ N	Metabolismo de aminoácidos
Metiltioribosa-1-fosfato isomerasa	gapaigaaaaf g	Central	IF-2B	Metabolismo de aminoácidos
Glicina deshidrogenasa (EC 1.4.4.2)	ggggpgagp vgv	Central	ninguno	Metabolismo de aminoácidos
Shikimato deshidrogenasa (EC 1.1.1.25)	gaggaaraaa wa	Central	Shikimate_DH	Metabolismo de aminoácidos

Treonina deshidratasa	ggglaagiaia	Central	PALP	Metabolismo de aminoácidos
Treonina sintasa (EC 4.2.3.1)	sasaaayaara	Central	PALP	Metabolismo de aminoácidos
Aconitato hidratasa (EC 4.2.1.3)	glgvvgwgv gg	Central	Aconitase	Metabolismo de carbohidratos
Biotin carboxilasa	kaaagggkg mkk	Central	CPSase_L_D2	Metabolismo de carbohidratos
Enolasa (EC 4.2.1.11)	aaaraaasa	Central	Enolase_N	Metabolismo de carbohidratos
Oxidorreductasa piruvato ferredoxina (flavodoxina), subunidad beta	ggdgdgfgig lg	Central	TPP_enzyme_C	Metabolismo de carbohidratos
Cobalto-precorrin-5B metiltransferasa (EC 2.1.1.195)	tgacataatka a	N-terminal	CbiD	Metabolismo de cofactores y vitaminas
Proteína bifuncional CoaBC de biosíntesis de la Coenzima A (EC 4.1.1.36; EC 6.3.2.5)	aalaadadarg ad	Central	DFP	Metabolismo de cofactores y vitaminas
Magnesio quelatasa (EC 6.6.1.1)	liddhlvdvll d	Central	Mg_chelatase	Metabolismo de cofactores y vitaminas
Nicotinato fosforibosil transferasa (EC. 6.3.4.21)	eellevelevr ell	Central	GTP_EFTU	Metabolismo de cofactores y vitaminas
Proteína de biosíntesis de tiamina ThiC (EC 4.1.99.17)	iagaiggalaayga	Central	ThiC_Rad_SAM	Metabolismo de cofactores y vitaminas

Descarboxilasa de la familia UbiD	vvfdddvdvq dv	Central	UbiD	Metabolismo de cofactores y vitaminas
Subunidad A de la glicerol-3-fosfato deshidrogenasa	iiigggatgagi a	N-terminal	DAO	Metabolismo de lípidos
Chaperona HSP60	ggdmdgmg gmggmm	C-terminal	ninguno	Plegamiento, modificación y degradación
Chaperona DnaJ	aagggfgdag fgggfdf	Central	Ninguno	Plegamiento, modificación y degradación
Chaperona DnaK (HSP70)	1) qaiyaaaqaa qqsapa 2) anaaagaapa gepapg	1) Central 2) C-terminal	ninguno	Plegamiento, modificación y degradación
Partícula de reconocimiento de señal	qmsqgmgg gmmgsm	Central	ninguno	Plegamiento, modificación y degradación
ATPasa transportadora de arseniato	1)vggkggvg kttt 2)laaalglraa ea	N-terminal	ArsA_ATPase	Procesos celulares y de señalización
Proteína de división celular FtsZ	tagegggtgtg ga	Central	Tubulin	Replicación y reparación
Proteína ribosomal S3	gsgrgrgnrr grgdrpdrgr rr	C-terminal	ninguno	Traducción
Proteína ribosomal L1	anaeaakaag a	Central	Ribosomal_L1	Traducción

Alanina - tRNA ligasa (EC 6.1.1.7)	aeaaggrggg rea	Central	DHHA1	Traducción
Factor de elongación Tu (EF-Tu)	eellevelevr ell	Central	GTP_EFTU	Traducción
Subunidad A de la glutamil-tRNA (Gln) amidotransferasa A (EC 6.3.5.7)	ggssggsaas	Central	Amidase	Traducción
Histidina quinasa CheA de transducción de señales	adagcgaagd aaaa	Central	ninguno	Transducción de señales
Glutamato metilesterasa (EC 3.1.1.61)	aakatpvtata a	Central	ninguno	Transducción de señales
Bomba ABC, subunidad de membrana interna	ggilgvllggil	Central	FtsX	Transporte de membrana
Transportador ABC	llleapdlill	Central	ABC_tran	Transporte de membrana
Proteína de los canales mecanosensibles	ggavvgvglgl gl	Central	MS_channel	Transporte de membrana
Transportador de membrana	gflgltgagg g	N-terminal	TauE	Transporte de membrana
Transportador de membrana	gggagigagl gagia	Central	ninguno	Transporte de membrana
CoA-disulfuro reductasa (EC 1.8.1.14)	vggvaggasv aa	N-terminal	Pyr_redox_2	
Proteína de fusión multifuncional [H-hidrato deshidratasa ADP- dependiente ; NAD (P) H-hidrato epimerasa] (EC 4.2.1.136; 5.1.99.6)	aavlaveaavr a	Central	Carb_kinase	

Anexo V) IDRs con amplia distribución filogenética

Proteína	Clasificación de IDR	Posición en la secuencia	Dominio de localización de la IDR	Categoría funcional
Fumarato liasa	IDR conservada en secuencia	C-terminal	ninguno	Metabolismo de aminoácidos
Formato deshidrogenasa subunidad alfa (EC 1.2.1.2)	IDR conservada	C-terminal	ninguno	Metabolismo de carbohidratos
Nitrito reductasa (EC 1.7.2.1)	IDR conservada en secuencia	Central	ninguno	Metabolismo de energía
RNA polimerasa, subunidad beta (EC 2.7.7.6)	IDR conservada	C-terminal	ninguno	Transcripción
Proteína chaperona DnaK (HSP70)	1) IDR conservada en secuencia 2) IDR conservada	1) Central 2) C-terminal	1) HSP70 2) HSP70	Plegamiento, modificación y degradación
Helicasa de RNA DeaD (EC:3.6.4.13)	IDR conservada	Central	ninguno	Plegamiento, modificación y degradación
Partícula de reconocimiento de señal	IDR conservada	C-terminal	SRP_SPB	Plegamiento, modificación y degradación
Proteína TatA translocasa sec-independiente	IDR conservada	C-terminal	MttA_Hcf106	Plegamiento, modificación y degradación
Receptor de la Partícula de reconocimiento de señal	IDR conservada	N-terminal	ninguno	Plegamiento, modificación y degradación
Helicasa de DNA ATP-dependiente (EC 3.6.4.12)	IDR conservada	Central	ninguno	Replicación y reparación

Proteína RecA (Recombinasa A)	IDR conservada	C-terminal	ninguno	Replicación y reparación
Proteína ribosomal L2	1) IDR conservada 2) IDR conservada en secuencia	1) N- terminal 2) C- terminal	1) Ribosomal_L2 2) Ribosomal_L2_ C	Traducción
Proteína ribosomal S13	IDR conservada en secuencia	C-terminal	Ribosomal_S13	Traducción
Proteína ribosomal S3	IDR conservada	C-terminal	ninguno	Traducción
Proteína ribosomal L3	IDR conservada en secuencia	Central	Ribosomal_L3	Traducción
Proteína ribosomal L4	IDR conservada en secuencia	Central	Ribosomal_L4	Traducción
Proteína ribosomal S4	IDR conservada	N-terminal	Ribosomal_S4	Traducción
Proteína ribosomal S9	IDR conservada	C-terminal	Ribosomal_S9	Traducción
Transportador ABC de hierro / manganeso / zinc, proteína de unión al sustrato	IDR conservada	Central	ZnuA	Transporte de membrana
Subunidad YbbK de la proteasa de membrana de la familia estomatina / prohibitina	IDR conservada	C-terminal	ninguno	Transporte de membrana
Transportador de cationes	IDR conservada	C-terminal	ninguno	Transporte de membrana
Deshidrogenasa/ reductasa SDR de cadena corta	IDR conservada	C-terminal	ninguno	