ANÁLISIS ESTRUCTURAL DE PROTEÍNAS DE *Methanopyrus kandleri AV19,* MICROORGANISMO RELACIONADO CON LUCA.

UNIVERSIDAD COLEGIO MAYOR DE CUNDINAMARCA

FACULTAD DE CIENCIAS DE LA SALUD

PROGRAMA DE BACTERIOLOGIA Y LABORATORIO CLINICO

TRABAJO DE GRADO

Bogotá D.C., Mayo 2018

ANÁLISIS ESTRUCTURAL DE PROTEÍNAS DE Methanopyrus kandleri AV19*, MICROORGANISMO RELACIONADO CON LUCA.*

Autor

ANGIE MILENA BARRERA CORREA

Asesor externo

SILVIO ALEJANDRO LOPEZ PAZOS MSc. PhD.

Facultad de Ciencias, Universidad Antonio Nariño

Asesor interno

SANDRA MONICA ESTUPIÑAN M. Sc.

Facultad de Ciencias de la Salud, Universidad Colegio Mayor de Cundinamarca

UNIVERSIDAD COLEGIO MAYOR DE CUNDINAMARCA

FACULTAD DE CIENCIAS DE LA SALUD

PROGRAMA DE BACTERIOLOGIA Y LABORATORIO CLINICO

TRABAJO DE GRADO

Bogotá D.C., Mayo 2018

DEDICATORIA

Quiero dedicar esta tesis a mis padres Miguel Antonio Barrera Correa y Ana Yolanda Correa Quintero por su apoyo incondicional, moral y económico, sus consejos y su compañía. A mi esposo York Luis Gutiérrez Navia por respaldarme en todo momento y a mi hija Emily Gutiérrez Barrera por ser mi inspiración y fortaleza para culminar este proyecto.

Angie Milena Barrera Correa.

AGRADECIMIENTOS

Quiero agradecer en primer lugar, a mis padres por financiar mis estudios, por ser un apoyo incondicional durante el transcurso de mi carrera, alentarme a continuar y culminar de la mejor manera.

A mi asesor Dr. Silvio Alejandro López Pazos, por brindarme la oportunidad de elaborar un proyecto a su lado, brindarme su conocimiento y guiarme con paciencia y sabiduría. A mi asesora Msc. Sandra Mónica Estupiñan por acompañarme y corregirme durante todo este proceso.

A la Universidad Antonio Nariño, por abrirme las puertas para realizar este proyecto. A la Universidad Colegio Mayor de Cundinamarca, por formarme y brindarme los conocimientos necesarios.

A Colciencias por financiar este proyecto a través de la convocatoria en alianza SENA ¨Jóvenes Investigadores e Innovadores 2016-2017”

Contenido

INDICE DE TABLAS

UNIVERSIDAD COLEGIO MAYOR DE CUNDINAMARCA

FACULTAD DE CIENCIAS DE LA SALUD

PROGRAMA DE BACTERIOLOGIA Y LABORATORIO CLINICO

ANÁLISIS ESTRUCTURAL DE PROTEÍNAS DE *Methanopyrus kandleri* AV19, MICROORGANISMO RELACIONADO CON LUCA

RESUMEN

Se cree que el inicio de la vida estaría basado en el desarrollo de un metabolismo a partir del origen del código genético, o de la formación de moléculas informativas replicantes, que terminaron en el último ancestro común universal (LUCA del inglés Last Universal Common Ancestor). Los estudios filogenéticos basados en especies primitivas de Euryarchaeota, incluyendo a *Methanopyrus kandleri*, han permitido la construcción del posible genoma de LUCA, ya que probablemente comparte características metabólicas con la primera forma de vida, como la hipertermófilia y la metanógenesis. Por esto es relevante analizar bioquímica y estructuralmente el proteoma de la cepa de *M. kandleri* AV19, en base a la función de las proteínas que se han conservado en cuanto a metabolismo, estructura,

replicación, control de ciclo celular, entre otros. Se realizaron análisis estadísticos para establecer abundancias relativas de los aminoácidos y características bioquímicas de las proteínas de *M. kandleri* AV19 por grupo funcional. Además se estableció las cantidades relativas de estructuras secundarias y terciarias de este proteoma. Entre los resultados se resalta que hay diferencias para los aminoácidos arginina, tirosina y leucina. Además se encontró diferencias asociadas a aminoácidos que pertenecen a hélices transmembrana. Los resultados de este proyecto se relacionan principalmente con aquellas investigaciones enfocadas en el origen de la vida. Este proyecto fue favorecido en la convocatoria de Colciencias-SENA: “Jóvenes Investigadores e Innovadores 2016-2017”.

PALABRAS CLAVES: LUCA, Last Universal Common Ancestor, *Methanopyrus kandleri,* proteoma, modelamiento de proteínas.

Estudiante: Angie Milena Barrera Correa

Docentes: Silvio Alejandro López Pazos MSc. PhD., Universidad Antonio Nariño

Sandra Mónica Estupiñan M.Sc. UCMC.

Fecha: Marzo 2018

GLOSARIO DE TÉRMINOS

- **GENOTIPO:** Se refiere al conjunto de genes del organismo (1) .
- **FENOTIPO:** La expresión del genotipo, en función a un determinado ambiente se denomina fenotipo (1).
- **HOMOLOGIA:** se refiere a la similitud estadísticamente significativa entre secuencias de aminoácidos, ya que han evolucionado a partir de un gen ancestral común. Esta conservación estructural se usa para predecir actividades bioquímicas comunes y funciones biológicas de proteínas y secuencias no codificantes (1).
- **FILOGENIA:** Es el estudio de la historia de la evolución de una especie o grupo, las líneas de descendencia y las relaciones e interacciones entre diferentes grupos de organismos (2).
- **GENES PARALOGOS:** Son aquellos que se encuentran en un mismo genoma y tienen un alto grado de homología. Por lo tanto uno de ellos ha aparecido por duplicación del otro. Esto permite generar proteínas nuevas con una función similar, creando la posibilidad de la especialización (2).
- **GEN ORTOLOGO:** Se define así a un gen que se encuentra en diferentes especies y que es altamente similar debido a que se ha originado de los mismos loci del gen antecesor en un antepasado común después de los eventos de especiación. Las proteínas ortólogas generalmente tienen secuencias similares y realizan las mismas funciones biológicas (3) .

- **COG:** La base de datos de grupos de proteínas ortólogas (COG) tiene como finalidad ayudar en la clasificación filogenética de las proteínas codificadas en genomas completos de Bacterias, Arqueas y Eucariotas. Cada COG consiste en un grupo de proteínas que se encuentran ortólogas en al menos tres linajes y que probablemente corresponde a un antiguo dominio conservado (3).

INTRODUCCIÓN

Los microorganismos de ambientes extremos (ME) son diversos y distribuidos ampliamente, estos pueden adaptarse a ecosistemas con condiciones severas para otro tipo de organismos como aguas hidrotermales, lagos alcalinos, salinas, desiertos, nevados y glaciares, e incluso a plantas de energía nuclear y espacios contaminados con metales pesados (4).

Se cree que el inicio de la vida se dio con el desarrollo de un metabolismo a partir del origen del código genético, o de la formación de moléculas informativas replicantes, que terminaron en LUCA. Se han desarrollado estudios centrados en establecer la temperatura en la que esta se originó incluyendo los ambientes extremos como psicrófilos, mesófilos, termófilos y hipertermófilos y teniendo en cuenta las condiciones de la tierra primigenia, concluyendo que esta surgió en fuentes hidrotermales de profundidades marinas, estructuras geológicas porosas bajo reacciones químicas asociadas a roca sólida y agua (5).

El contenido de genes varía entre los organismos debido a la adaptación a sus nichos ecológicos. Sin embargo, la utilización del mismo código genético por todos los organismos existentes indica que todos descienden de un ancestro común que lo poseía (4). Se cree que *M kandleri,* es un microorganismo posiblemente cercano a LUCA ya que comparte características genéticas y metabólicas asociadas a la primera forma de vida, la hipertermófilia (crecimiento óptimo a

100°C) y metanógenesis (producción de metano a partir de H_2 y CO_2) (6). Por esto, en *M. kandleri* y otras especies primitivas de *Euryarchaeotas,* se han basado estudios filogenéticos para determinar su relación con LUCA, características metabólicas y funcionales y el posible genoma de esta.

Se ha enfatizado en entender la bioquímica que llevo al origen de la primera forma de vida y el ambiente en el que se pudo gestar, ya que una vez establecido, permitiría entender las relaciones evolutivas de los tres dominios *Bacteria, Archaea* y *Eukarya*. Esto abre paso a investigaciones que permitan sentar bases respecto a la relación de LUCA con otros organismos. Por lo expuesto, es relevante analizar estructural y bioquímicamente el proteoma de la cepa *M. kandleri* AV19 en cuanto a la función de las proteínas que se han conservado en cuanto a metabolismo, estructura, control de ciclo celular, adaptación, entre otros (7). Se ha establecido que el 73% de los productos de los genes de *M. kandleri* son proteínas conservadas, por lo que su caracterización, podría reflejar su relación con un ancestro común.

OBJETIVOS

- **Objetivo General**

Estudiar la estructura del proteoma de *Methanopyrus kandleri* AV19 *in silico* en relación al ancestro común universal.

- **Objetivos específicos:**

- Clasificar las proteínas codificadas por el genoma de *Methanopyrus kandleri* AV19.
- Establecer características funcionales de residuos en las proteínas de *Methanopyrus kandleri* AV19 importantes para el ancestro común universal.
- Comparar la estructura secundaria del proteoma de *Methanopyrus kandleri* AV19 con proteínas de referencia.

1. ANTECEDENTES

Slesarev et al., (2002) realizaron la secuenciación del genoma de *M. kandleri* AV19 (cromosoma circular de 1694969 pb), con 1692 genes codificantes de proteínas, y 39 genes codificantes de ARNs estructurales. Los autores encontraron que las proteínas de este microorganismo poseen un número elevado de aminoácidos cargados negativamente, y proteínas con punto isoeléctrico (pI) de 5. El análisis filogenético según la región 16S ARNr sugiere que *M. kandleri* estaría ubicado cerca de la base del árbol de Euryarchaeota, y estaría relacionado con metanógenos arqueales (*Methanococcus jannaschii* y *Methanothermobacter thermoautotrophicum*). *M. kandleri* tiene un numero bajo de proteínas relacionadas a la señalización y regulación de expresión génica. *M. kandleri* parece tener pocos genes adquiridos a través de transferencia lateral desde otras arqueas. El 73% de las proteínas de *M. kandleri* son ortólogos conservados. Cuando se comparó el genoma de *M. kandleri* con otros genomas de arqueas se logró determinar que sus proteínas se relacionan a sistemas funcionales conservados (6).

En su estudio, Brochier et al (2004), utilizaron 20 genomas en los que seleccionaron 14 proteínas implicadas en la transcripción y 53 proteínas ribosomales. Con estas, construyeron 15 conjuntos de datos correspondientes a 12 subunidades de ARN polimerasa y tres factores de transcripción para determinar la filogenia de las arqueas y las posibles transferencias laterales de

genes (LGT del inglés lateral gene transference), detectando un solo caso de LGT, la subunidad H de una ARN polimerasa de los Termoplasmatales (*Thermoplasma sp*, y *Acidiplasma sp*.) y *M. kandleri,* que comparten cinco o seis aminoácidos en las subunidades de ARN polimerasa. La cercanía de *M. kandleri* con Halobacteriales (*Halapricum sp*. y *Salarchaeum japonicum*) sugiere que este adquirió su subunidad H del gen de Termoplasmatales. Los árboles resultantes de los datos de transcripción ubican a *M. kandleri* como el primer descendiente justo antes de Termococcales, mientras que en el árbol de traducción los Termococcales se encuentran en la rama más basal, y *M. kandleri* agrupado parafiléticamente (agrupación que contiene algunos, pero no todos los descendientes del ancestro común) con Metanococcales y Metanobacteriales (7).

Greco et al. (2005) buscaban nuevas secuencias potencialmente compatibles con la estructura de un pseudodimero de las histonas H3-H4, para determinar si este es un gen primitivo, su papel biológico y si se ha conservado en otras especies. Para esto, realizaron un análisis *in silico* de dominios de histona a partir de virus 1 de *Heliothis zea* (HZV-1), la histona humana Sos, y la histona procariota de *M. kandleri*. Se determinó que la proteína viral de HZV-1 contenía el pliegue doble en la histona H3-H4, y la histona homóloga humana Sos, y la histona procariota de *M. kandleri* dos motivos plegables diferentes localizados a lo largo de la misma cadena polipeptídica. Se concluyó, que aunque tienen cierta homología, las histonas de *M. kandleri* y de proteínas Sos tienen papeles biológicos muy

diferentes: la histona pseudodimerica procariota está implicada en el ensamblaje de cromatina, mientras que Sos ejerce una acción inhibidora en la actividad de genes de transducción de señales (familia de genes Ras) (8).

El contenido de genes varía ampliamente entre los organismos, debido a sus adaptaciones. La utilización del mismo código genético por todos los organismos existentes indica que todos descienden de un ancestro común que lo poseía. En el estudio de Mat et al. (2008) se realizó un acercamiento al posible genoma de LUCA, basado en los genes comunes de ocho especies primitivas de *Euryarchaea* y *Crenarchaea* (*M. kandleri, M. thermautotrophicum, M. jannaschii, Pyrococcus. abyssi, P. furiosus, P. horikoshii, Aeropyrum pernix* y *P. aerophilum*. La evidencia basada en secuencias de ARN de transferencia y complementada por otras pruebas ha localizado cerca de LUCA al metanogeno hipertermofílico *M. kandleri,* lo que plantea que los primeros seres vivos heterótrofos que habitaban las zonas de temperaturas más frías desarrollaron la metanogénesis y un genoma de ADN para adaptarse a las fuentes hidrotermales profundas donde habían altas concentraciones de CO_2 e H_2, lo que les permitió producir grandes cantidades de metano por medio de este nuevo metabolismo. Los resultados de esta construcción genómica hacen pensar que LUCA contendría como mínimo ~463 genes y 39 genes de ARN estructural (5).

Wang et al. (2009) identificaron asociaciones genómicas para proteínas ribosomales (R-proteínas) conservadas, y discutieron las implicaciones de estas

asociaciones en la maquinaria de traducción. Entre los tres dominios de la vida, hay aproximadamente 102 familias de R-proteínas reconocidas, 36 son universales por lo que es probable que hayan aparecido antes de LUCA. El gran número de familias R-proteína compartidas por los dominios *Archaea* y *Eucaryota* sugieren que el sistema de traducción eucariota se originó a partir de una versión arqueal y que 12 de los grupos se encuentran en más de la mitad de los genomas analizados con representantes tanto en *Crenarqueota* como en *Euryarqueota*. Seis de los grupos se produjeron en más del 80% de las especies. El análisis de estas agrupaciones reveló asociaciones con genes implicados en la iniciación de la síntesis de proteínas, la transcripción y de otros procesos celulares. Entre los genes asociados con las R-proteínas no universales hay algunas que codifican las proteínas involucradas en el inicio de la traducción como son infB, eIF2, eIF2-GTPase, eIF2c y eIF6. La secuencia de eIF2 mostró homología con una proteína perteneciente a la familia del factor de elongación universal, EF-Tu en las bacterias y eEF1A en las arqueas. Debido a que los dos factores de iniciación arcaica son un complejo, es probable que aparecieran al mismo tiempo del ribosoma (9).

Dodd et al., (2017) hallaron fósiles biológicos, microfósiles de filamentos y túbulos de hemetita, forma mineral de óxido férrico u óxido, en Nuvvuagittuq Supracrustal Belt (NSB) (Quebec, Canadá). Estos restos son similares a los de bacterias oxidantes que viven formando parte de fuentes hidrotermales de alta mar ricos en minerales como hierro. La formación de estos filamentos pudo deberse a factores

no biológicos, como cambios de temperatura, presión de las rocas y sedimentación. Sin embargo, la presencia de carbonato, minerales como apatita y material carbonoso presentes en materia biológica, son evidencia de procesos de oxidación y actividad biológica. Las evidencias químicas y morfológicas de la formación de estos túbulos y filamentos de restos de bacterias que metabolizan hierro apoyan la hipótesis de que hace 3770 a 4300 millones de años la Tierra tuvo ambientes que proporcionaron las condiciones para el desarrollo las formas de vida. Este descubrimiento apoya la idea de que la vida surgió poco después de la formación del planeta en fuentes de altas temperaturas en el fondo marino (10) .

Friar et al. (2012) basados en la distribución Benford, ley que predice que en un conjunto determinado de números aquellos cuyo primer dígito es 1 aparecerán de forma más frecuente que los que empiezan por otros dígitos (11), plantean que los datos sobre el número de Marcos de Lectura Abierta (ORF del inglés Open Reading Frames) codificadas por los genomas de los tres dominios de la vida permiten describir algunas de las características generales y diferencias esenciales entre procariotas y eucariotas, donde el número de ORFs crecen linealmente con el tamaño total del genoma para las primeras, y solo de manera logarítmica para las segundas. Esto permite estimar algunos tamaños del genoma mínimo, dependiendo de las funciones biológicas requeridas. La ecuación utilizada predice que el tamaño mínimo del genoma de los procariotas está limitado a 167 + 400 ORF y es de máximo 8-12 Mpb y en eucariotas el genoma debe ser de un tamaño mínimo de 90-130 kpb y aproximadamente mayor a 4-5 Mbp (12).

2. MARCO REFERENCIAL

2.1. Extremófilos

Los organismos extremófilos son aquellos que sobreviven a condiciones ambientales extremas, y que suelen ser letales para otros. Estos ambientes extremos se agruparon en dos, geofísicos y geoquímicos extremos. En el primer grupo se incluyen condiciones como temperatura, presión, radiación electromagnética (radiación ionizante y no ionizante y radiación cósmica) y en los ambientes geoquímicos se encuentran pH, salinidad, desecación, desierto, tóxicos tales como especies reactivas de oxígeno y nitrógeno, o metales pesados (13). Los microorganismos de ambientes extremos (ME) pueden adaptarse diversos ecosistemas como aguas hidrotermales, lagos alcalinos, salinas, desiertos, nevados y glaciares, e incluso a plantas de energía nuclear y espacios contaminados con metales pesados (4). Una de las clasificaciones de los ME los agrupa en termófilos (45-80°C) como *Thermophilus aquaticus*, hipertermófilos (>80°C) como *P. furiosus*, psicrófilos (-5-20ºC) como *Methanogenium sp.*, barófilos (presión >1 atm) entre los que está *Shewanella benthica*, acidófilos (pH bajo) incluyendo a *Picrophilus oshimae*, alcalófilos (a pHs de incluso 11 o 12) como *Anaerobranca sp*., halófilos (alta salinidad) entre los que están *Halobacterium sp.* y *Halobacteroides sp.*, metanógenos (producen metano en su metabolismo) como *Methanobacterium sp.* y *Methanococcus sp.*, metalófilos (metales pesados) como

Ralstonia sp., y radiófilos (radiación ionizante) como *Deinococcus radiodurans* (14) (13).

Algunas evidencias como microfósiles, estromatolitos, isotopos de carbono sedimentario y azufre indican que los microorganismos del periodo arqueano, desarrollaron fuentes metabólicas que comparten cierta similitud con muchos de los microorganismos vivos actuales. Por lo que los extremófilos vivos actuales podrían servir para dilucidar las relaciones evolutivas y la transferencia a sus descendientes de este metabolismo y sus biomoléculas (15).

2.2. Origen de la vida

La idea de un organismo común a todos, se tiene presente desde que Charles Darwin afirmo que todos los seres orgánicos que han vivido en esta tierra han descendido probablemente de una forma primordial. Se cree que el inicio de la vida estaría basado en el desarrollo de un metabolismo a partir del origen del código genético, o de la formación de moléculas informativas replicantes, que terminaron con la aparición de la primera forma de vida o LUCA (16) (17).

Los organismos vivos son selectivos con los elementos químicos asimilados, el número de compuestos orgánicos y reacciones utilizadas en el metabolismo. Una minoría de compuestos desempeña papeles importantes en las reacciones autocatalíticas y son la base de múltiples reacciones derivadas. Ha sido un

desafío, comprender como estos compuestos actúan como catalizadores, y como pudieron haber surgido a partir de reacciones geoquímicas termodinámicas. La introducción efectiva de los compuestos ricos en energía a las reacciones endergónicas por los catalizadores primordiales orgánicos (compuestos metálicos catalíticos como clusters de hierro y hierro-azufre) fue un avance de la síntesis prebiótica. Un ejemplo, es el ciclo del ácido tricarboxílico inverso (ATCr), donde se asimila el carbono del CO_2 mediante una serie de etapas autocatalíticas y cada compuesto intermedio intensifica su producción. Este ciclo energético se puede asociar con un escenario primitivo de hierro-azufre o fijación fotosintética de CO_2 impulsada por sulfuro de zinc (ZnS) y ácidos carboxílicos. Es a partir del ciclo de ATCr y algunas derivaciones que se conduce a la síntesis de ácidos grasos, terpenoides y carbohidratos, luego el amoníaco (NH3) es captado por varios cetoácidos, para formar los aminoácidos comunes, que a su vez abrieron camino a otros metabolitos que contienen N y diversos compuestos aromáticos heterocíclicos implicados en reacciones catalíticas moleculares actuales. La primera síntesis orgánica catalizada probablemente ocurrió en minerales o en superficies planas y estas reacciones condujeron posiblemente a otros compuestos (18).

Se desea entender las relaciones evolutivas de los tres dominios: *Bacteria, Archaea* y *Eukarya* (19), por lo que se han establecido teorías con respecto a la temperatura en la que LUCA se desarrolló teniendo en cuenta las condiciones de

la Tierra poco después de que el planeta se formará. Varios estudios, han enfocado las posibilidades de surgimiento y desarrollo de LUCA a ambientes extremos como psicrófilos, mesófilos, termófilos y hipertermófilos, basados en los genes comunes de ocho especies primitivas de *Euryarchaea* y *Crenarchaea*: *M. kandleri, M. thermautotrophicum, M. jannaschii, P. abyssi, P. furiosus, P. horikoshii, A. pernix* y *Pyrobaculum aerophilum.* Se cree que en las temperaturas psicrofilas, la condensación de monómeros de ARN era posible, pero no era óptima para el ensamblaje de ácidos nucleicos, en el ambiente mesófilo, podría haberse presentado una crisis en la fijación de carbono antes de la invención de la fotosíntesis, la termófilia ofrecía un equilibrio en las reacciones quimiosintéticas y en temperaturas hipertermofilas, la energía geotérmica liberada podría proporcionar un ambiente propicio para el origen de la vida. Las fuentes hidrotermales están dotadas de abundante energía térmica, así como CO_2 e H_2 que podría ser capturado por litoautótrofos, sin embargo la inestabilidad de diversos compuestos esenciales para la vida, contradicen la hipertermófilia como temperatura óptima para las primeras formas de vida. Se cree, que la vida comenzó en zonas de temperatura mesófilas donde se facilitó la síntesis de plantillas de ARN, y a medida que las formas de vida heterótrofas se multiplicaron y se daba una crisis de fijación de carbono, estos organismos se trasladaron a fuentes hidrotermales. La biota emergente de temperaturas más bajas adaptada a la termófilia, desarrollaron un genoma de ADN, y la metanogénesis. Así, las formas de vida cruzaron progresivamente de las zonas de temperatura más bajas a la zona de hipertermofília, donde metanógenos sobrevivieron del CO_2 e H_2

abundante produciendo grandes cantidades de metano para protegerse de una era de hielo, y finalizaron en la invención del genoma de ADN y el código genético universal de 20 aminoácidos para dar lugar a LUCA (5). Se dice que las primeras células vivas surgieron en fuentes hidrotermales de profundidades marinas, estructuras geológicas porosas asociadas a roca sólida y agua bajo reacciones químicas permanentes (20). De acuerdo a un reciente estudio en el que se encontraron fósiles de origen biológico en la región de Nuvvuagittuq Supracrustal Belt (NSB), en Quebec, Canadá, zona que formó parte de un sistema de fuentes hidrotermales de alta mar, rico en minerales como el hierro, se concluyó que hace 3770 a 4280 millones de años estos ambientes proporcionarian las condiciones ideales para las primeras formas de vida en la Tierra (10), lo que reafirma las teorías de que LUCA surgió en ambientes hipertermófilos.

La existencia de un antepasado común se ratifica por las propiedades comunes en los diferentes organismos vivos (código genético universal, 20 aminoácidos, ADN con T, C, A, G, ARN con U, C, A, G, ATP como principal fuente de energía y las coenzimas NAD y CoA). Los métodos para identificar la naturaleza de LUCA en la base de la vida dependen de la construcción de árboles filogenéticos basados en genes parálogos, o con análisis combinado de secuencias y estructuras. Sin embargo, debido a que los genes parálogos proteicos contienen artefactos que afecta el enraizamiento, otros biopolímeros se han analizados para tal fin como el ARN de transferencia (ARNt). Los organismos vivos libres contienen 20 familias de ARNt afines a los 20 aminoácidos. El análisis de las distancias genéticas entre las

secuencias de ARNt de cada familia localiza a LUCA en el dominio *Archaea* en la proximidad de *M. kandleri* y a medida que evolucionaban, las secuencias se dispersaron y distanciaron entre sí, derivando en las diferencias genómicas de los organismos actuales. LUCA, por ser una especie extinta, su origen es complejo y controvertido, estudios han proporcionado evidencias que favorecen una LUCA bacteriana, eucariótica o arcaica, pero las que soportan una LUCA arcaica cercana a *M. kandleri* son mayores. La transición de una atmósfera en la Tierra pobre en oxígeno a una rica en oxígeno obligó a adaptaciones extensas por la mayoría de los linajes vivos. Sólo un pequeño número de organismos, incluyendo *M. kandleri*, escaparon a adaptaciones masivas. Los respiraderos hidrotermales, donde habita *M. kandleri*, representan uno de los nichos ecológicos más conservados en la Tierra. Por lo tanto, *M. kandleri* y otros habitantes de la ventilación anaeróbica podrían sobrevivir con mínimas variaciones genómicas. Es esto lo que permite que *M. kandleri* proporcione la mejor semejanza de LUCA (21).

Los citocromos participan en el transporte de electrones en las mitocondrias, cloroplastos y organismos que reducen sulfatos. Los genes codificantes de citocromos están ausentes de los genomas de seis especies antiguas conservadas en la evolución molecular: *M. kandleri*, *M. thermautotrophicum*, *M. jannaschii*, *P. furiosus*, *P. abyssi* y *P. horikoshii*. Puesto que *M. kandleri*, *M. thermautotrophicum* y *M. jannaschii* consumen metabólicamente H_2 y CO_2, mientras que *P. furiosus*, *P. abyssi* y *P. horikoshii* producen H_2 y CO_2, la

deficiencia de citocromos en estas especies no proviene de similitud metabólica, sino de su proximidad filogenética a una LUCA posiblemente sin citocromo. Por esto, se usaron los genes comunes de estas especies para definir el genoma de LUCA. Como LUCA no podría sobrevivir sin ningún metabolismo energético debido a la falta de citocromos, el conjunto de genes de LUCA debe ser complementado con los genes metabólicos de algún grupo. Dado que LUCA se ubica cercanamente a un metanógeno hipertermófilo, los genes de metanogénesis comunes a *M. kandleri*, *M. thermautotrophicum* y *M. jannaschii*, *s*e añadieron al genoma plausible de LUCA, sabiendo que genes comunes podrían estra originalmente ausentes de LUCA. Para contrarrestar esta posibilidad, sólo se tomaron genes comunes a estos antiguos seis. Sobre esta base, el genoma mínimo LUCA se estima que contiene esencialmente 463 genes, incluyendo un proteoma de 424 Clúster de Grupos Ortólogos de proteínas (COGs) y 39 genes ARNs estructurales del genoma de *M. kandleri*. El proteoma mínimo de LUCA contendría 24 COGs para la replicación, recombinación y reparación del ADN, 26 COGs para la transcripción, más otros COG relacionados con el ADN. Los 463 genes de LUCA, exceden el número estimado de genes de 150-340 para organismos mínimos, lo que sugiere que LUCA no fue un organismo mínimo, sino un organismo modestamente complejo (5) (21).

Desde el descubrimiento de comunidades vivas en los respiraderos hidrotermales submarinos, estos han sido considerados posibles sitios de origen de la vida, donde la energía geotérmica liberada en las aberturas proporciona un ambiente

idóneo para la evolución prebiótica, apoyando las síntesis autotróficas de una gama de compuestos orgánicos (5). Sin embargo, existen contradicciones de este escenario como la Inestabilidad del ARN a temperaturas elevadas; metabolitos lábiles e intermedios que requieren catalización metabólica por biocatalizadores evolucionados; y ARNs aleatorios, incluso en una temperatura termofílica se presentarían estos inconvenientes. Esto deja una primera célula mesófila que dio lugar a una vida pre-Lucana, y no psicrófila, ya que las velocidades de crecimiento celular son lentas a bajas temperaturas. Así esta se trasladó a una zona de mayor temperatura, respiraderos hidrotermales, que con abundantes suministros de CO_2 e H_2 y fijación exergónica de carbono, permitió la síntesis prebióticas de compuestos orgánicos. En el proceso de traslado a las porosidades de rocas en las fuentes hidrotermales, los pre-Lucanos, necesitaban proteínas fueran resistentes al calor, así que la resistencia al calor del ADN sobre el ARN se convirtió en una ventaja selectiva, y los pre-Lucanos que desarrollaron una maquinaria informativa de ácidos nucleicos que podía funcionar eficientemente de ADN, perfeccionaron la metanogénesis y dieron lugar a LUCA. Esta produjo descendientes que se extendieron a zonas más frías, eliminando aquellos que no tenian un genoma de ADN y/o código genético de 20 aminoácidos (21).

Los genomas de las especies de organismos en los tres dominios *Bacteria*, *Archaea* y *Eukarya*, se distinguen fácilmente entre sí por sus secuencias de ARN ribosomal (ARNr). Sin embargo, algunas diferencias fenotípicas entre los dominios podrían dilucidar los incentivos evolutivos que provocarían las disgregaciones que

formaron estos dominios: Estrategias anticodón, lípidos de membrana y membrana nuclear. Una diferencia clave de *Archaea* de *Bacteria* y *Eukaria* es la conservación de la complementaridad codón-anti codón. Esta diferencia se detectó con la proteína LepA (o EF4-factor de alargamiento durante la fase de elongación que mejora la tolerancia ribosomal a los cambios en las concentraciones iónicas) una de las más conservadas, presente en todas las bacterias y en la mayoría de eucariotas, pero no en arqueas. Así, bacterias y eucariotas con LepA podrían ser más tolerantes a variaciones del medio interno, y por lo tanto adaptables a entornos cambiantes en comparación con las arqueas. Debido a los bajos nutrientes prebióticos, los pre-Lucanos se trasladaron a aberturas geológicas ricas en elementos orgánicos donde fue indispensable la adquisición de lípidos de éter-glicerol, ya que, proporcionaban mayor estabilidad térmica, lo que sería necesario para la supervivencia y la posterior aparición de LUCA, por lo que esta adquirió lípidos de éter en su membrana, y así sus descendientes. Cuando los descendientes de LUCA adoptaron el nuevo código genético y se trasladaron a zonas mesófilas, no eran necesarios estos lípidos en su membrana por lo que fueron eliminados, esto explicaría la diferencia en la composición de la membrana de arqueas de las bacterias y eucariotas. La membrana nuclear de *Eukarya* difiere de los genomas procarioticos que contienen hasta 50 nucleótidos, mientras que eucariotas contienen más de 50 nucleótidos, lo que sugiere que la protección del núcleo por una membrana permite un mayor genoma y por lo tanto más genes. La membrana nuclear, al facilitar el aumento del tamaño del genoma y la expansión del proteoma, podría facilitar nuevas formas de

vida de otro modo inalcanzable y, al hacerlo, se habrían alejado de *Archaea* y *Bacteria* para formar su propio dominio (22).

2.3. *Methanopyrus kandleri*

Una de las especies de *Euryarchaeota* primitivas en las que se han basado los estudios filogenéticos y la construcción del posible genoma de LUCA es *M. kandleri.* Este es el microorganismo vivo considerado posiblemente el más cercano a la primera forma de vida, esto por sus características térmicas y metabólicas. *M. kandleri* es un bacilo que fue aislado del fondo del mar en la base de una chimenea volcánica de 2,000m de profundidad "black smoker" en el Golfo de California (6), que comparte características con los hipertermófilos con una temperatura de crecimiento óptima de 100°C y con los metanógenos que producen metano a partir de H_2 y CO_2 (9). Se estima que el rango de temperaturas compatible con la vida es de -25°C, y hasta 122°C en respiraderos hidrotermales donde habita *M. kandleri*. Sin embargo, un hallazgo reciente, indica que *M. kandleri* puede soportar temperaturas de 130°C, por lo que se considera esta temperatura como el límite para la vida (23). Mediante el estudio de diferentes cepas de *M. kandleri* comparando las proteinas valyl-ARNt sintetasa (ValSR) y la isoleucil- ARNt sintetasa (IleRS), se estableció que los linajes vivos más primitivos de *M. kandleri* se encontrarían en la cordillera central del Océano Indico (24).

Para la mayoría de genomas procarioticos secuenciados, un tercio de los genes codificantes de proteínas anotadas son proteínas huérfanas, ya que carecen de homología con proteínas conocidas. Se ha encontrado que gran parte de los genes que codifican estas proteínas huérfanas en el genoma de *M. kandleri* AV19 se produce dentro de dos grandes regiones. Estos genes no tienen homólogos conocidos excepto de otros genes de *M. kandleri*. Aunque estas regiones podrían ser consecuencia de TLG, estos también podrían ser resultado de la integración de plásmidos (25).

Se identificó a partir de células de *M. kandleri*, un nuevo lípido de éter 2,3-di-O-geranilgeranil-sn-glicerol y un geranylgeraniol. Este lípido terpenoide, es una característica primitiva y se asemeja con la idea de un linaje de M.kandleri corto y en la base del árbol filogenético basado en la subunidad 16S ARNr (26). Se ha descrito una girasa inversa en *M. kandleri* como único ejemplo conocido de una topoisomerasa heterodimérica de tipo I. La enzima está formada por una subunidad de 50 kDa (RgyA) que interactúa covalentemente con el ADN y una subunidad de 150 kDa (RgyB) involucrada en la hidrolisis de ATP. Esta girasa inversa, difiere notablemente en sus subunidades y secuencias de *Sulfolobus acidocaldarius*. Como estos dos organismos representan dos ramas filogenéticas de Archaea, Euryarchaeota y Crenarchaeota, es posible que estas diferencias se mantengan en otros miembros de estos grupos (27).

Se ha caracterizado una histona de *M. kandleri*, que a diferencia de histonas procariotas, tiene una estructura que consta de dos motivos de pliegues repetidos en un único polipéptido. Los análisis indican que la repetición N-terminal está estrechamente relacionada con las histonas eucarióticas H2A y H4, mientras que la repetición C-terminal se asemeja a la que se encuentra en las histonas procariotas. Estos resultados implican un *M. kandleri* antes de la divergencia de la familia de genes de histonas de procariotas y eucariotas (28).

El genoma de *M. kandleri* fue secuenciado directamente a partir de ADN genómico. De esta secuenciación se determinó que su genoma consta de un único cromosoma circular de 1.694.969 pares de bases, 1.691 genes codificadores de proteínas y 39 genes de ARNs estructurales. Al igual que en otras arqueas y bacterias, el 73% de los productos de los genes de *M. kandleri* son proteínas conservadas que pertenecen a los COG, lo que les permite habitar en condiciones extremas y demuestra su limitada capacidad de intercambiar genes con otros organismos (6).

2.4. Parámetros estructurales y bioquímicos de proteínas

Los aminoácidos, son moléculas que poseen un grupo ácido carboxílico y un grupo amino, unidos a un átomo denominado carbono alfa y su variedad química

proviene de la cadena lateral que también está unida a este. Se han establecido 20 tipos de aminoácidos que se encuentran comúnmente en las proteínas, cada uno con diferente cadena lateral unida al átomo de carbono alfa. La importancia de estos aminoácidos es la formación de proteínas necesarias para cumplir las funciones necesarias en una célula (29).

Cada aminoácido posee características bioquímicas como el peso molecular (MW), que se calcula mediante la suma de masas isotópicas medias de los aminoácidos presentes en la proteína y la masa isotópica media de una molécula de agua y se da en Dalton (Da). El punto isoeléctrico, también llamado pI, se define como el pH en el que una molécula tiene igual número de cargas positivas y negativas y, por tanto, es eléctricamente neutra. A este valor de pH la solubilidad de la sustancia es casi nula (30).

Hay aminoácidos que tienen un radical (-R) con carga debido a la presencia de un grupo adicional ácido o base en la molécula. Los aminoácidos aspartato y glutamato (Asp + Glu) son ácidos y tienen grupos - R cargados negativamente a pH 7.0 por la presencia del grupo -COOH en el radical. Los aminoácidos cargados positivamente a pH 7.0 son los aminoácidos básicos, arginina y lisina (Arg + Lys), que contienen uno o más –NH2 en el radical. El Índice de inestabilidad, es una estimación de la estabilidad de la proteína en un tubo de ensayo. Una proteína cuyo índice es menor que 40 se predice como estable; un valor superior a 40 sugiere que la proteína es inestable. El Índice alifático de una proteína se define

como el volumen relativo ocupado por las cadenas laterales alifáticas (alanina (Ala), valina (Val), isoleucina (Ile) y leucina (Leu)). La gran media de hidropatícidad (GRAVY) para un péptido o proteína se calcula como la suma de los valores de hidropatícidad de todos los aminoácidos, dividido por el número de estos en la secuencia (31).

Se han establecido cuatro niveles de organización en la estructura de una proteína. La secuencia de aminoácidos se conoce como la estructura primaria. Los plegamientos de la cadena polipeptídica que forman alfa hélices y beta láminas constituyen la estructura secundaria de la proteína. La organización tridimensional completa de una cadena polipeptídica se denomina estructura terciaria, y si una molécula proteica particular se forma como un complejo de más de una cadena polipeptídica, la estructura completa se conoce como estructura cuaternaria. Una proteína está constituida por cadenas laterales de aminoácidos, unidos por enlaces peptídicos covalentes (el carbono del grupo carboxilo de un aminoácido comparte electrones con el átomo de nitrógeno del grupo amino de un segundo aminoácido) formando una cadena principal polipeptídica. Cada proteína difiere de otras por su secuencia y la cantidad de aminoácidos que la conforman. Los dos extremos de la cadena polipeptídica son químicamente diferentes, lo que hace que cada proteína sea distinta. Un extremo lleva el grupo amino libre o N-terminal (NH_3^+, también escrito NH_2) y el otro lleva el grupo carboxilo libre o C-terminal (COO^-, también escrito COOH). El plegamiento de las proteínas está determinado por las interacciones atómicas de los aminoácidos

presentes en ellas. Estas interacciones entre los átomos de la cadena principal y de las cadenas laterales, dan lugar a enlaces covalentes y no covalentes débiles (enlaces de hidrógeno, enlaces iónicos y fuerzas de Van der Waals), y como resultado la formación de una estructura tridimensional única de cada proteína. La estructura plegada final, adoptada por cualquier cadena polipeptídica es generalmente aquella en la que la conformación es estable y se minimiza la energía libre (32).

Las proteínas se pliegan en una amplia variedad de formas, y generalmente tienen entre 50 y 2000 aminoácidos de largo. Las moléculas de proteína más pequeñas contienen un dominio (unidades estructurales que se pliegan independientemente entre sí) único, mientras que las proteínas grandes pueden contener hasta docenas de dominios, conectados entre sí por cadenas cortas de polipéptidos. Cuando se comparan las estructuras tridimensionales de muchas moléculas de proteínas diferentes, se encuentran dos patrones de plegado regulares en ellas: El patrón hélice alfa y hoja beta. Estos dos patrones son comunes porque resultan del enlace de hidrógeno entre los grupos N - H y C = O en la cadena principal del polipéptido, sin involucrar las cadenas laterales de los aminoácidos. Por lo tanto, pueden formarse por muchas secuencias de aminoácidos diferentes (32).

- Alfa-hélices: Una hélice α se genera cuando una única cadena polipeptídica se dobla sobre sí misma para formar un cilindro rígido. El N-H de cada enlace peptídico está unido por enlaces de hidrógeno al C=O de un enlace peptídico vecino localizado a cuatro enlaces peptídicos en la misma cadena. Las regiones cortas de alfa-hélice son abundantes en las proteínas localizadas en las membranas celulares, como las proteínas de transporte y los receptores (32).
- Beta-laminas: Se forman a partir de cadenas polipeptídicas vecinas que se encuentran en una misma orientación (paralelas) o de una cadena polipeptídica que se pliega sobre sí misma, con cada sección de la cadena ejecutándose en dirección opuesta a la de sus vecinos inmediatos (anti paralelas). Ambos tipos de beta-láminas producen una estructura muy rígida, unida por enlaces de hidrógeno que conectan los enlaces peptídicos de las cadenas vecinas (32).
- Hélices transmembrana: Son aquellas estructuras que cruzan la bicapa lipídica como una hélice alfa compuesta principalmente de aminoácidos con cadenas laterales no polares (32). Las hélices transmembrana están predominantemente compuestas de residuos hidrófobos y están implicadas en los procesos bioquímicos que se dan en la bicapa lipídica (78.5%) (33).

3. DISEÑO METODOLOGICO

3.1. Población: proteínas de *M. kandleri*.

Muestra: 597 proteínas del genoma de *M. kandleri* AV19 curadas en la base de datos de SwissProt.

3.2. Hipótesis: La estructura bioquímica de las proteínas de *M. kandleri* está relacionada con la función.

Variables:

- **Independientes:** Función de las proteínas de *M. kandleri AV19.*
- **Dependientes:** Estructura de las proteínas de *M. kandleri AV19.*

Indicadores:

- **Estructurales**
    - Alfa-hélices
    - Beta-laminas
    - Proteínas transmembrana con el programa

- **Parámetros bioquímicos**
    - Tipo de aminoácido
    - Número total de aminoácidos
    - Peso molecular

- Punto isoeléctrico
- Residuos cargados negativamente (Asp + Glu),
- Residuos cargados positivamente (Arg + Lys),
- Índice de inestabilidad
- Índice alifático
- Media de hidropatícidad (GRAVY)

3.3. Técnicas y procedimientos:

Clasificación funcional

Los 561 genes que posiblemente pertenecieron a LUCA, se clasificaron de acuerdo a las diferentes funciones que cumple cada proteína codificada por estos (597 proteínas) (5):

- (A) Traslación, estructura ribosomal y biogénesis.
- (B) Modificación y procesamiento de ARN.
- (C) Transcripción
- (D) Replicación, recombinación y reparación.
- (E) Estructura y dinámica de cromatina
- (F) Control del ciclo celular, división celular, y partición cromosómica
- (G) Mecanismos de transducción de señales.
- (H) Pared celular, membrana y biogénesis.

- (I) Tráfico intracelular, secreción y transporte vesicular
- (J) Modificación postraduccional, recambio proteico, chaperones
- (K) Producción y conversión de energía.
- (L) Transporte y metabolismo de carbohidratos
- (M) Transporte y metabolismo de aminoácidos
- (N) Transporte y metabolismo de nucleótidos
- (O) Transporte y metabolismo de coenzimas
- (P) Transporte y metabolismo de lípidos
- (Q) Transporte y metabolismo de iones inorgánicos.
- (R) Biosíntesis, transporte y catabolismo de metabolitos secundarios.
- (S) Predicción de funciones generales

Basados en los COGs de los 1614 genes de *M. kandleri*, se obtuvo la secuencia primaria de cada una de las 597 proteínas. Esta revisión se hizo usando las bases de datos curada SwissProt/UniProt (http://www.uniprot.org/). UniProt es una base de datos de secuencias de proteínas curadas que ofrece anotaciones que incluyen función, familia, análisis de dominios, modificaciones posttraduccionales entre otras. Está conformada por tres componentes, bases de datos de proteínas como SwissProt, UniRef, que contiene clusters de secuencia para búsquedas de similitud de secuencia rápida, y UniParc que contiene archivos de secuencias para el seguimiento e identificación (34). SwissProt es una base de datos de secuencias de proteínas curadas manualmente, que proporciona un alto nivel de

anotación (descripción de la función de una proteína, su estructura de dominio, modificaciones post-traduccionales, etc.), con un nivel mínimo de redundancia y un alto nivel de integración con otras bases de datos (35).

Análisis estructural y bioquímico

Se recolectaron los datos de los parámetros bioquímicos de cada proteína: número y tipo de aminoácidos, peso molecular, punto isoeléctrico, residuos cargados negativamente (Asp + Glu), residuos cargados positivamente (Arg + Lys), índice de inestabilidad, índice alifático, media de hidropatícidad (GRAVY), con el programa ProtParam de la suite ExPasy (https://www.expasy.org/). ExPasy-ProtParam es una herramienta que permite el cálculo de parámetros físicos y bioquímicos para una proteína almacenada en Swiss-prot, para una secuencia introducida por un usuario (36).

Modelamiento proteico

La estructura secundaria (número de alfa-hélices, beta-laminas y hélices transmembrana) de la totalidad de las secuencias de aminoácidos (597 proteínas), y la estructura terciaria de 546 secuencias de aminoácidos, asociadas a los COGs de *M. kandleri*, se calculó con el programa Phyre2 (http://www.sbg.bio.ic.ac.uk/~phyre2/) (37). Phyre2 es un servidor conformado por varias herramientas que predice y analiza la estructura de proteínas, su función y

las mutaciones que estas puedan presentar. Esta interfaz utiliza métodos avanzados de detección de homología para construir modelos tridimensionales (3D), predecir sitios de unión a ligando, y analizar el efecto de variantes de aminoácidos para una secuencia de proteínas dada por un usuario. En comparación con otros métodos (I-TASSER, Swiss-Model, HHpred, PSI-Pred, Robetta y Raptor), Phyre2 ocupa el sexto lugar de 55. Los cinco superiores tienen una mejora en la calidad del modelo de 2.8%, mientras I-TASSER muestra una mejora del 5%. Este sistema presenta algunas limitaciones, una de ellas es que no predice los efectos estructurales de mutaciones puntuales, y que aún no existen métodos confiables para predecir una estructura de proteína a partir de la secuencia sola, sin referenciar con estructuras conocidas (37). Phyre2 usa en su método de cálculo de estructura de proteínas la base de datos de macromoléculas biológicas (proteínas, ADN, ARN) determinadas experimentalmente llamada Protein Data Bank (RCSB PDB). Esta contiene herramientas para depositar secuencias, anotación, consulta, análisis y visualización, y recursos educativos para su uso con el archivo PDB, que contiene las coordenadas atómicas 3D y datos experimentales de proteínas, ácidos nucleicos y conjuntos complejos (38).

Obtenido el resultado de cada proteína, seleccionamos los modelos que tuvieran un porcentaje de cobertura mayor a 70% y de confianza mayor al 90%, aquellos modelos que no cumplían con estos valores se enviaron al programa I-Tasser (https://zhanglab.ccmb.med.umich.edu/I-TASSER/). I-Tasser (Iterative Threading ASSEmbly Refinement) es una plataforma en línea que realiza la predicción

funcional y el modelamiento 3D de proteínas a partir de su secuencia de aminoácidos mediante el enfoque de enhebrado múltiple. En recientes evaluaciones criticas de técnicas para la predicción de la estructura proteica, I-Tasser clasificó como el servidor número 1 en CASP7, CASP8, CASP9 y CASP10 (39).

Análisis estadístico

Para el análisis estadístico, se realizó la prueba de Shapiro Test que permitió identificar los grupos de datos que están distribuidos normalmente (paramétricos) de aquellos que no siguen una distribución normal (no paramétricos), con un umbral de p=0.01. A los grupos de datos que seguían una distribución normal se les realizó la prueba de ANOVA de una vía, con un umbral de p=0.01, y a los no paramétricos la prueba de Kruskal Wallis, con un umbral de p=0.01, con el fin de determinar si habían diferencias significativas. Para los grupos de datos que presentaron diferencias significativas y que no seguían una distribución normal se les aplico la prueba de Dunn con un umbral de p=0.01.

4. RESULTADOS

Se obtuvieron las secuencias del genoma de *M. kandleri* AV19 depositadas en la base de datos UniProt/SwissProt con un total de 1719 entradas, 397 revisadas (Swissprot), y 1322 no curadas. A Los 561 genes probables de LUCA (5), se les asignó una estructura secundaria basada en los COGs de 1614 genes de *M. kandleri* AV19 depositados en la base de datos curada Swiss-prot, para un proteoma anotado teóricamente de LUCA de 597 secuencias de aminoácidos para este trabajo (182997 aminoácidos).

Se calculó la media por cada tipo de aminoácido de las secuencias proteicas relacionadas según su función, para determinar si existían diferencias significativas. Se determinó que Arg y tirosina (Tyr) diferían significativamente del resto de aminoácidos en alguno de los grupos funcionales. Para Arg y Tyr, el valor de p en la prueba de Kruskal Wallis fue de 0.000011 y 0.000293 respectivamente, lo que indica el rechazo de la hipótesis nula (todos los grupos funcionales tienen la misma distribución). Para determinar el grupo funcional que específicamente difiere del resto de grupos, para Arg y Tyr, se realizó la prueba de Dunn. Las proteínas que cumplen función de replicación, recombinación y reparación (Grupo D) contienen una cantidad significativamente mayor de Arg que las que pertenecen al grupo de traducción, estructura ribosomal y biogénesis (Grupo A), y a las que tienen función de transcripción (Grupo C). Las proteínas del Grupo D también contienen una cantidad significativamente mayor de Arg que las proteínas

encargadas de modificación posttraduccional, recambio proteico y chaperonas (Grupo J), transporte y metabolismo de carbohidratos (Grupo L) y de metabolismo y transporte de nucleótidos (Grupo N). Para el aminoácido Tyr se encontró que el Grupo D tiene diferencia significativa en abundancia, respecto a las proteínas del Grupo C (Tablas 1, 2 y 3).

Se determinaron diferentes parámetros bioquímicos: peso molecular, punto isoeléctrico, número total de residuos cargados negativamente (Asp y Glu), número total de residuos cargados positivamente (Arg y Lys), índice de inestabilidad, índice alifático, media de hidropatícidad (GRAVY) (Tabla 4). De estos datos se encontró que únicamente para la cantidad de residuos cargados positivamente (Arg + Lys) un grupo es significativamente diferente de los demás grupos funcionales, con un p de 0.000014 en la prueba de Kruskal Wallis. Para determinar el grupo funcional que difiere del resto de grupos, en cuanto a los residuos cargados positivamente, se realizó la prueba de Dunn (Tabla 5) donde se pudo concluir que el grupo proteico con función de replicación, recombinación y reparación (Grupo D) contiene una cantidad significativamente mayor de estos aminoácidos frente al Grupo C. Asimismo, las proteínas del Grupo D poseen una cantidad mayor de aminoácidos cargados positivamente respecto a los Grupos J (modificación posttraduccional, recambio proteico y chaperonas), K (producción y conversión de energía), N (transporte y metabolismo de nucleótidos), O (transporte y metabolismo de coenzimas) y P (transporte y metabolismo de lípidos).

Tabla 1. Media de aminoácidos en cada grupo funcional. Aln: Alanina, Arg: Arginina, Asn: Asparagina, Asp: Aspartato, Cys: Cisteína, Gln: Glutamina, Glu, Ácido glutámico, Gly: Glicina, His: Histidina, Ile: Isoleucina, Leu: Leucina, Lys: Lisina, Met: Metionina, Phe: Fenilalanina, Pro: Prolina, Ser: Serina, Thr: Treonina, Trp: Triptófano, Tyr: Tirosina, Val: Valina. Azul: Grupo de datos no paramétricos.

Numero de secuencias	Nombre del grupo funcional	Aln	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
128	(A) Traducción, estructura ribosomal y biogénesis.	19	24,6	4,9	16	3,1	4,4	32	20	5,5	13,4	22,3	14,7	5,23	7,7	15	9	11	3,3	7,2	25,1
2	(B) Modificación y procesamiento de ARN.	22	23,5	4,5	13	2,5	5	28	29	7	13,5	28,5	11	4,5	6	17	12	12	2	6,5	25,5
31	(C)Transcripción	16	23,8	4,8	14	4	3,6	32	16	5,1	13,1	22,7	14,8	5	5,6	12	10	9,7	1,8	7,2	21,1
36	(D) Replicación, recombinación y reparación.	34	45,9	8,5	29	4,3	7,4	54	32	8,4	20,4	42,4	24,2	7,36	13	22	19	17	4,1	12	44,6
2	(E) Estructura y dinámica de cromatina	23	17,5	3	17	1,5	4	23	21	7	9,5	18,5	4,5	5	5	13	11	9,5	1,5	9,5	20
14	(F) Control del ciclo celular, división celular, y partición cromosómica	24	29,7	6,3	19	4,9	3,7	32	23	4,4	15,1	29,9	13,6	5,93	8,9	17	13	13	2,7	7,1	28,5
5	(G) Mecanismos de transducción de señales.	14	15,2	2,8	12	1,2	1,6	24	13	4,8	11	20,4	7,6	4	4,2	10	9	6,8	1,2	4,4	19,2
8	(H) Pared celular, membrana y biogénesis.	36	34	10	20	5,3	6,6	40	31	9,5	23,6	38,6	16,6	8,63	13	21	18	19	5,3	15	44,8
12	(I) Tráfico intracelular, secreción y	28	22	6,8	15	1,5	5,7	27	26	3,3	24,5	38,2	17	10	9,6	15	15	16	2,8	8,6	31

	transporte vesicular																				
28	(J) Modificación postraduccional, recambio proteico, chaperones	28	24,6	5,8	20	5,2	7,1	35	25	5,4	16,8	29,7	15,1	6,82	7,7	16	12	14	2,5	7,7	30,9
92	(K) Producción y conversión de energía.	31	24,1	7,2	21	8,2	5,4	38	29	6,6	19,5	30,6	14,3	7,56	9,4	19	13	16	3,1	9,1	34,6
22	(L) Transporte y metabolismo de carbohidratos	26	21,8	6,5	22	3,3	3,5	36	28	7	16	28,5	11,9	5,45	8,4	16	14	15	2	7	34,6
31	(M) Transporte y metabolismo de aminoácidos	32	25,7	7,8	22	4,2	4,5	39	29	7,9	17	30,2	11,9	7,16	9,5	17	15	15	2,5	7,9	40,5
36	(N) Transporte y metabolismo de nucleótidos	24	22,8	6,1	20	3,8	3,7	34	24	6,4	15,4	26,6	11,3	5,61	7,8	15	10	12	1,7	7,8	30,9
31	(O) Transporte y metabolismo de coenzimas	27	24,7	4,5	18	4,4	3,1	30	25	4,8	12,4	28,2	9,84	6,03	8,4	16	12	12	2,9	5,9	33,7
12	(P) Transporte y metabolismo de lípidos	28	17,8	4,8	12	2,6	3,1	22	25	5,6	13,3	25,1	10,7	5,08	9,3	11	12	14	4,2	7,3	30,5
18	(Q) Transporte y metabolismo de iones inorgánicos	37	23,2	6,4	15	3,3	4,3	22	28	6,3	17,3	43,2	10,8	6,39	11	15	18	18	3,8	8,8	39,6
6	(R) Biosíntesis, transporte y catabolismo de metabolitos secundarios.	16	21,3	5,2	17	3,2	1,7	24	16	5,7	10,8	25	7,17	4,67	5,7	12	10	8,7	2	8	25,8
83	(S) Predicción de funciones generales	23	26,6	5,2	18	4	4	33	24	7	14,6	28,6	11,9	5,67	8,6	17	13	14	2,7	8,1	32,3
597	Valor P	0,398	0,000011	0,32	0,337	0,495	0,30100	0,000002	0,331	0,24	0,306	0,374	0,362	0,281	0,301	0,27	0,42	0,3	0,3	0,000293	0,4

Tabla 2. Prueba de Dunn del aminoácido arginina. Los círculos azules representan los grupos funcionales que tienen una diferencia significativa respecto a los otros grupos.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
B	1.000000																	
C	1.000000	1.0																
D	0.000004	1.0	0.000075															
E	1.000000	1.0	1.000000	1.000000														
F	1.000000	1.0	0.815487	1.000000	1.0													
G	1.000000	1.0	1.000000	0.093019	1.0	1.000000												
H	1.000000	1.0	1.000000	1.000000	1.0	1.000000	1.0											
I	1.000000	1.0	1.000000	0.188483	1.0	1.000000	1.0	1.0										
J	1.000000	1.0	1.000000	0.014806	1.0	1.000000	1.0	1.0	1.0									
K	1.000000	1.0	1.000000	0.000103	1.0	1.000000	1.0	1.0	1.0	1.0								
L	1.000000	1.0	1.000000	0.010127	1.0	1.000000	1.0	1.0	1.0	1.0	1.0							
M	1.000000	1.0	1.000000	0.565135	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0						
N	1.000000	1.0	1.000000	0.000733	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0					
O	1.000000	1.0	1.000000	0.047534	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0				
P	1.000000	1.0	1.000000	0.003083	1.0	0.966028	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
Q	1.000000	1.0	1.000000	0.108590	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
R	1.000000	1.0	1.000000	1.000000	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
S	1.000000	1.0	1.000000	0.006107	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 3. Prueba de Dunn del aminoácido tirosina. Los círculos azules representan los grupos funcionales que tienen una diferencia significativa respecto a los otros grupos.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
B	1.000000																	
C	1.000000	1.0																
D	0.000331	1.0	0.006509															
E	1.000000	1.0	1.000000	1.000000														
F	1.000000	1.0	1.000000	1.000000	1.0													
G	1.000000	1.0	1.000000	1.000000	1.0	1.0												
H	0.110184	1.0	0.100682	1.000000	1.0	1.0	0.643951											
I	1.000000	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000										
J	1.000000	1.0	1.000000	0.486799	1.0	1.0	1.000000	1.000000	1.0									
K	0.338377	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000	1.0	1.0								
L	1.000000	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0							
M	1.000000	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0	1.0						
N	1.000000	1.0	1.000000	0.876072	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0					
O	1.000000	1.0	1.000000	0.011551	1.0	1.0	1.000000	0.146442	1.0	1.0	1.0	1.0	1.0	1.0				
P	1.000000	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
Q	1.000000	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
R	1.000000	1.0	1.000000	1.000000	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
S	1.000000	1.0	1.000000	0.453929	1.0	1.0	1.000000	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Tabla 4. Parámetros bioquímicos.

Nombre Grupo funcional	Aminoácidos	Peso molecular	Punto isoelectrico	Residuo cargados negativamente (Asp + Glu)	Residuos cargados positivamente (Arg + Lys) **	Índice de inestabilidad	Índice alifático.	Indice de hidropaticidad GRAVY
(A) Traslación, estructura ribosomal y biogénesis.	262,8	29810	7,1050	40,5000	34,5000	41,3000	96,2850	-0,2325
(B) Modificación y procesamiento de ARN.	272	29817	6,9342	48,2344	39,1719	46,5202	84,8095	-0,5371
(C) Transcripción	242,14	27633	6,0579	45,8276	40,7241	46,7890	93,4503	-0,4499
(D) Replicación, recombinación y reparación.	454,4	50896	6,3272	82,9722	70,1111	43,9692	91,0897	-0,4238
(E) Estructura y dinámica de cromatina	221,5	24380	6,8500	39,0000	22,0000	39,1800	92,0650	-0,2750
(F) Control del ciclo celular, división celular, y partición cromosómica	301,21	33743	6,5529	50,3571	43,5000	45,2657	93,6579	-0,2959
(G) Mecanismos de transducción de señales.	186,2	20862	4,8860	36,2000	22,8000	46,1180	102,5440	-0,2010
(H) Pared celular, membrana y biogénesis.	413	46166	6,4700	59,3750	50,6250	39,7775	99,1488	-0,0879
(I) Tráfico intracelular, secreción y transporte vesicular	321,92	35506	7,8292	41,1667	39,0000	38,8017	114,1175	0,1569
(J) Modificación posttraduccional, recambio proteico, chaperones	314,86	34934	5,1582	55,5714	39,7500	41,4582	95,4400	-0,2277
(K) Producción y conversión de energía.	347,58	38274	5,2395	58,9474	38,3579	41,4308	92,5461	-0,1624
(L) Transporte y metabolismo de carbohidratos	312,59	34273	4,9595	58,6818	33,7273	36,7391	96,2786	-0,1846
(M) Transporte y metabolismo de aminoácidos	347,03	38060	4,7748	61,4516	37,6129	37,9065	96,4926	-0,1445
(N) Transporte y metabolismo de nucleótidos	288,61	32024	4,9336	53,4167	34,0556	37,2861	97,1533	-0,2251
(O) Transporte y metabolismo de coenzimas	288,23	31618	5,6797	47,1613	34,5161	39,1026	98,2484	-0,0741
(P) Transporte y metabolismo de lípidos	263,25	28684	7,3050	34,3333	28,4167	32,0242	106,5375	0,2093
(Q) Transporte y metabolismo de iones inorgánicos	336,44	36413	7,5439	37,0556	34,0000	31,8717	114,5422	0,3266
(R) Biosíntesis, transporte y catabolismo de metabolitos secundarios.	228,17	25775	5,6767	40,3333	28,5000	43,4667	100,8550	-0,1917
(S)Predicción de funciones generales	301,37	33510	5,7880	50,6747	38,4819	42,0047	95,7841	-0,2262
P valué <0,01	0,02	0,336	0,1175	0,36	0,000014	0,35	0,42	0,52

** Uno de los grupos presenta diferencias significativas respecto a los demás.

Tabla 5. Prueba de Dunn de los aminoacidos cargados positivamente (Arg + Lys). Los círculos azules representan los grupos funcionales que tienen una diferencia significativa respecto a los otros grupos.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
B	1.000000																	
C	1.000000	1.0																
D	0.000063	1.0	0.001612															
E	1.000000	1.0	1.000000	1.000000														
F	1.000000	1.0	1.000000	1.000000	1.0													
G	1.000000	1.0	1.000000	0.028635	1.0	0.592470												
H	1.000000	1.0	1.000000	1.000000	1.0	1.000000	1.0											
I	1.000000	1.0	1.000000	1.000000	1.0	1.000000	1.0	1.0										
J	1.000000	1.0	1.000000	0.007904	1.0	1.000000	1.0	1.0	1.0									
K	1.000000	1.0	1.000000	0.001645	1.0	1.000000	1.0	1.0	1.0	1.0								
L	1.000000	1.0	1.000000	0.005197	1.0	1.000000	1.0	1.0	1.0	1.0	1.0							
M	1.000000	1.0	1.000000	0.156613	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0						
N	1.000000	1.0	1.000000	0.000190	1.0	0.912600	1.0	1.0	1.0	1.0	1.0	1.0	1.0					
O	1.000000	1.0	1.000000	0.003677	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0				
P	1.000000	1.0	1.000000	0.005049	1.0	0.750638	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
Q	1.000000	1.0	1.000000	0.032598	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0		
R	1.000000	1.0	1.000000	0.237832	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
S	1.000000	1.0	1.000000	0.000430	1.0	1.000000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Se obtuvieron los datos de la composición estructural de cada proteína: número de aminoácidos que forman alfa-hélices, beta-laminas y hélices transmembrana, según su grupo funcional. La media de las 597 proteínas de *M. kandleri* AV19, que se analizaron en este trabajo, contienen 134.46 hélices, 59.63 láminas y 12.18 hélices transmembrana (Tabla 6). A estos datos se les realizo análisis estadístico para determinar si algún grupo de datos tenía una diferencia significativa respecto

a los demás. Con este análisis se concluye que hay diferencias significativas en la composición estructural por grupos funcionales para el caso de las hélices transmembrana (Tabla 6 y 7). Se encontró que los grupos funcionales que tienen mayor contenido de aminoácidos asociados a hélices transmembrana son el Grupo H (pared celular, membrana y biogénesis), el Grupo I (tráfico intracelular, secreción y transporte vesicular), y el Grupo Q (transporte y metabolismo de iones inorgánicos).

Tabla 6. Composición estructural por grupos funcionales. El F calculado es menor que el F de tabla para todos los grupos en el ANOVA. El asterisco indica el valor p significativo en el análisis de Kruskal-Wallis ($p<0.01$)

Grupo funcional	Media de aminoácidos en alfa-hélices	Media de aminoácidos en beta-láminas	Media de aminoácidos en hélices transmembrana
(A) Traslación, estructura ribosomal y biogénesis	102,29	64,71	2,12
(B) Modificación y procesamiento de ARN	76,48	95,52	8,8
(C) Transcripción	99,54	42,86	0
(D) Replicación, recombinación y reparación	209,33	76,99	2,9
(E) Estructura y dinámica de cromatina	111,23	22,88	0
(F) Control del ciclo celular, división celular, y partición cromosómica	149,69	39,32	3,4
(G) Mecanismos de transducción de señales	92,31	44,34	6,44
(H) Pared celular, membrana y biogénesis	195,29	76,1	68,14
(I) Tráfico intracelular, secreción y transporte vesicular	206,15	27,61	49,87

(J) Modificación postraduccional, recambio proteico, chaperones	142,44	55,61	11,8
(K) Producción y conversión de energía.	151,53	57,94	15,8
(L) Transporte y metabolismo de carbohidratos	120,47	68,14	1,69
(M) Transporte y metabolismo de aminoácidos	139,97	69,59	3,23
(N) Transporte y metabolismo de nucleótidos	111,44	65,69	2,58
(O) Transporte y metabolismo de coenzimas	133,99	50,73	18,88
(P) Transporte y metabolismo de lípidos	141,06	37,22	45,48
(Q) Transporte y metabolismo de iones inorgánicos	228,83	25,36	101,96
(R) Biosíntesis, transporte y catabolismo de metabolitos secundarios.	79,44	57,41	2,58
(S) Predicción de funciones generales	123.1	67.16	7.94
F de tabla=1.607	F=0.4291	F=0.3179	p=3.718478e-13*

* Valor p significativo en el análisis de Kruskal-Wallis (p<0.01)

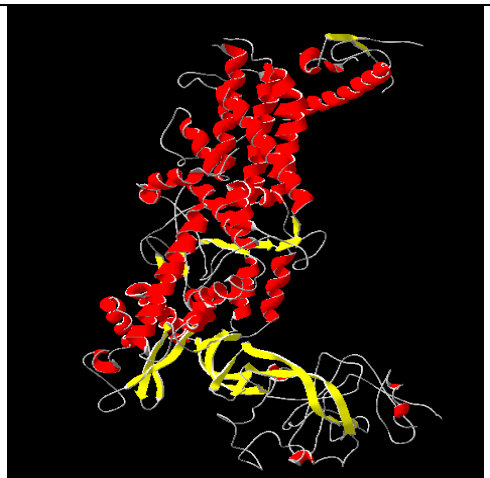
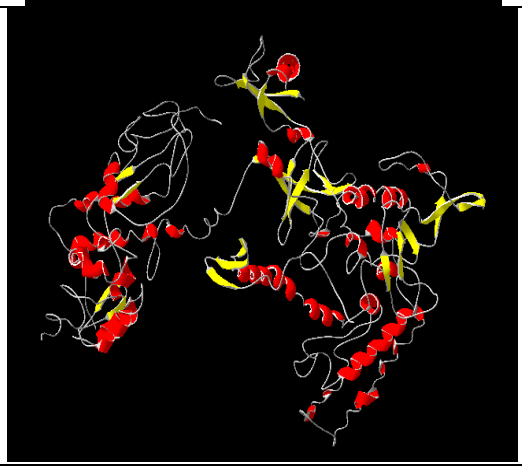
Tabla 7. Prueba de Dunn correspondiente a hélices transmembrana. En rojo se indican los resultados que corresponde a diferencias significativas (p<0.01).

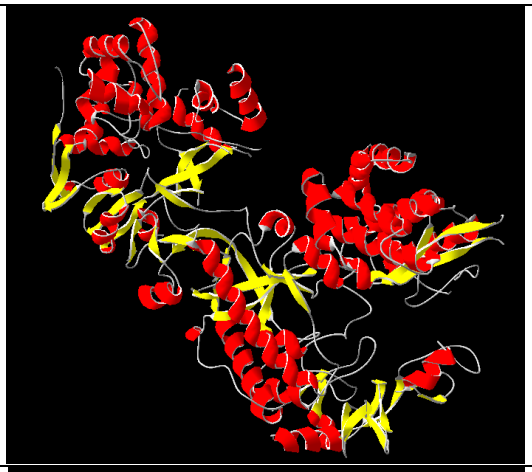
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
B	1.0																	
C	1.0	1.0																
D	1.0	1.0	1.0															
E	1.0	1.0	1.0	1.0														
F	1.0	1.0	1.0	1.0	1.0													
G	1.0	1.0	1.0	1.0	1.0	1.0												
H	3.418 06 9e-10	1.0	3.4810 30e-10	7.8076 20e-08	0.0630 63	0.000 34	0.159 01											
I	2.018 06 0e-01	1.0	6.9773 13e-02	1.0	1.0	1.0	1.0	6.8196 31e-02										
J	1.0	1.0	1.0	1.0	1.0	1.0	1.0	9.7576 09e-06	1.0									
K	1.0	1.0	8.2117 05e	1.0	1.0	1.0	1.0	2.9298 40e	1.0	1.0								

			-01					-07										
L	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0 146 56e -08	0.4 464 12	1.0	1.0							
M	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.4 621 16e -07	1.0	1.0	1.0	1.0						
N	1.0	1.0	1.0	1.0	1.0	1.0	1.0	6.9 561 14e -08	1.0	1.0	1.0	1.0	1.0					
O	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.3 250 54e -05	1.0	1.0	1.0	1.0	1.0	1.0				
P	1.5 22 56 2e- 01	1.0	5.3 615 89e -02	1.0	1.0	1.0	1.0	8.2 847 49e -02	1.0	1.0	1.0	0.3 564 34	1.0	1.0	1.0			
Q	7.5 76 78 0e- 05	1.0	7.5 767 80e -05	1.2 280 30e -02	1.0	0.6 745 45	1.0	3.2 795 45e -01	1.0	0.5 187	0.0 52 87 3	0.0 028 28	0.0 308 13	0.0 110 98	0.6 954 58	1.0		
R	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.2 755 72e -03	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
S	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.7 263 29e -08	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.00 381 6	1.0

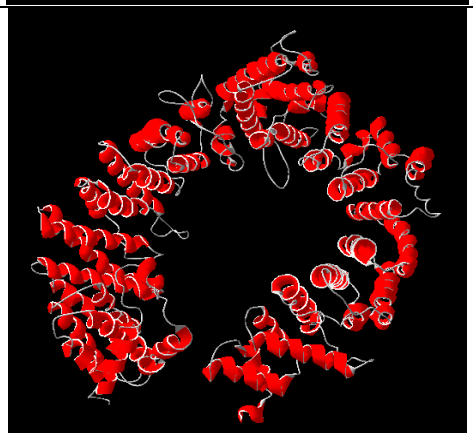
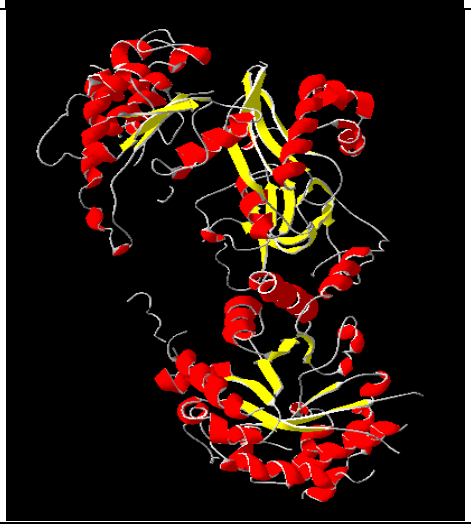
La totalidad de las secuencias obtenidas en la revisión por medio de los COGs de *M. kandleri* AV19 (597 secuencias de aminoácidos) fueron modeladas 3D en formato PDB, para determinar una aproximación al modelo estándar de las proteínas relacionadas a cada grupo funcional. En la Tabla 8 se presenta la estructura 3D de mayor numero de aminoácidos por grupo funcional, representando la estructura de hélices y láminas. Se puede observar que aquellas proteínas asociadas a grupos funcionales con presencia de mayor número de aminoácidos relacionados a hélices transmembrana poseen estructura de hélices acompañadas de láminas beta (Grupos H, I), o solo hélices (Grupo Q).

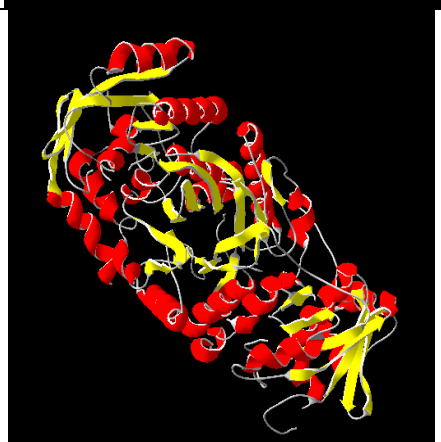
Tabla 8. Estructura 3D de la proteína de mayor número de aminoácidos por función.

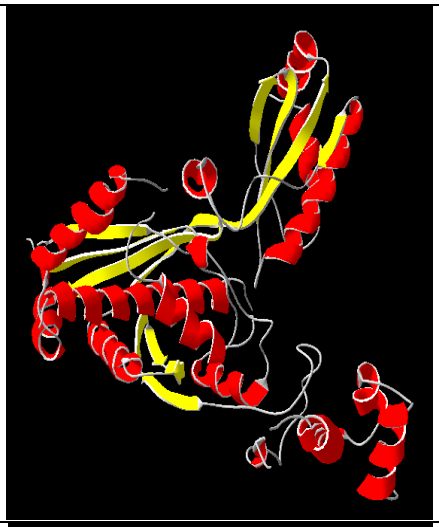
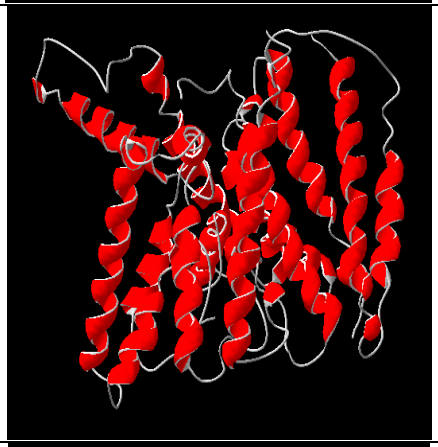
Nombre Grupo funcional	Estructura 3D
(A) Traslación, estructura ribosomal y biogénesis.	
(B) Modificación y procesamiento de ARN.	
(C) Transcripción	

(D) Replicación, recombinación y reparación.			
(E) Estructura y dinámica de cromatina			
(F) Control del ciclo celular, división celular, y partición cromosómica			

(G) Mecanismos de transducción de señales.			
(H) Pared celular, membrana y biogénesis.			
(I) Tráfico intracelular, secreción y transporte vesicular			

(J) Modificación postraduccional, recambio proteico, chaperones			
(K) Producción y conversión de energía.			
(L) Transporte y metabolismo de carbohidratos			

(M) Transporte y metabolismo de aminoácidos			
(N) Transporte y metabolismo de nucleótidos			
(O) Transporte y metabolismo de coenzimas			

(P) Transporte y metabolismo de lípidos			
(Q) Transporte y metabolismo de iones inorgánicos			
(R) Biosíntesis, transporte y catabolismo de metabolitos secundarios.			

(S) Predicción de funciones generales			

5. DISCUSIÓN

M. kandleri es un microorganismo primitivo que habita en fuentes hidrotermales, y resiste a temperaturas de hasta 130°C (23), su metabolismo está basado en la metanogénesis, y se ha considerado por algunos como la Euryarchaeota viva posiblemente más cercana a LUCA (5) (6). Es por esto que ha sido utilizado en estudios filogenéticos para determinar el origen y el genoma de LUCA. Se han realizado varios análisis de la relación de este organismo con LUCA, en la que se discute su proximidad en el árbol de la vida. Son varias las afirmaciones de que debe ser ubicado en la rama más basal de este, basados en construcciones filogenéticas a partir de la subunidad 16s ARNr, y proteínas implicadas en diferentes mecanismos como la transcripción (por ejemplo topoisomerasa I) (6) (27). Sin embargo, otras construcciones basadas en proteínas como de traducción difieren, ubicando a *M. kandleri* como un grupo uniforme con Metanococcales y Metanobacteriales, y no cercano al primer organismo primitivo (6) (7). Las diversas teorías del origen de un organismo común a todos, han surgido de evaluaciones sobre las diferentes temperaturas en las que se pudo originar LUCA, y coinciden en un origen hipertermofilo cerca a fuentes hidrotermales y conos volcánicos, y por lo tanto, un metabolismo metanogénico que les permitió sobrevivir a concentraciones de H2 y CO2 abundantes en la Tierra primigenia (5) (10) (20). La invención del código genético, le permitió a las especies que lo adoptó la supervivencia en la tierra primitiva. Este se desarrolló durante millones de años, y para explicar esta evolución, se han definido diversas etapas: el desarrollo de un

gen peptidasa, el mundo de ARN donde se originó el ARN mensajero, ARN de transferencia, intrones, codones, y luego el código genético que dio lugar a LUCA y a la vida sintética (40).

La Arg (C6 H12 N4 O) es un aminoácido polar, cargado positivamente, por lo que es altamente hidrofilico, posee en su cadena lateral un grupo guanido y está implicado en varios procesos biológicos, como la regulación de la conformación o los potenciales redox, transporte de iones H+, y translocación de péptidos. También participa en las interfaces proteína-proteína, en sitios activos enzimáticos, y en de canales de transporte (41) (42). En nuestro estudio, encontramos que este aminoácido es significativamente mayor en las proteínas implicadas en la función de replicación, recombinación y reparación (Grupo D).

La Tyr es un aminoácido polar cargado positivamente a pH neutro, por lo que es hidrófilo, posee en su cadena lateral como grupo ionizable a un grupo amino primario (41). Para el aminoácido Tyr encontramos que el Grupo D tiene diferencia significativa en abundancia respecto a las proteínas implicadas en la transcripción del Grupo C.

Las proteínas de *M. kandleri* muestran un contenido alto de aminoácidos cargados negativamente, y proteínas con pI de aproximadamente 5, para la adaptación a alta salinidad intracelular (>3 M K+) (6). En nuestro estudio, la media de pI es de 6.1. En cuanto a residuos cargados positivamente se determinó una

media de aminoácidos de 37.4 para esta característica. Estos resultados pueden deberse al número de proteínas usadas en este análisis (35.3%).

Una característica distintiva de *M. kandleri* AV19 es la escasez de proteínas implicadas en la señalización y regulación de la expresión génica (6). En nuestro estudio, encontramos un aproximado de 135 de secuencias proteicas (22.6%) que cumplen esta función, lo que podría concordar con dicha afirmación.

También determinamos, que la conformación de aminoácidos presentes en hélices transmembrana, difería significativamente entre varios grupos funcionales. Se encontró, que los grupos funcionales que tienen mayor contenido de aminoácidos asociados a hélices transmembrana son el Grupo H (pared celular, membrana y biogénesis), el Grupo I (tráfico intracelular, secreción y transporte vesicular), y el Grupo Q (transporte y metabolismo de iones inorgánicos), lo que está relacionado directamente con la función que cumplen, ya que las hélices transmembrana están implicadas en los procesos bioquímicos que tienen lugar en las bicapas lipídicas (33), la señalización de células y el transporte de iones y solutos a través de la membrana (43). Se ha predicho que el proteoma primitivo de LUCA podría haber sido hidrófobico, y que esta característica en la composición proteica individual fue disminuyendo durante la evolución (44).

Dado que la función de una proteína está determinada en parte por su estructura, predecir la estructura de una proteína a partir de su secuencia de aminoácidos

puede ser útil para comprender las funciones moleculares y su papel en las vías biológicas. Para ello se han desarrollado múltiples mecanismos computacionales basados básicamente en el modelado de homología (comparativo), enhebrado (reconocimiento de pliegue) y modelado libre, que difieren solo en un pequeño porcentaje de rendimiento de acuerdo con los puntos de referencia recientes de CASP. El enfoque computacional más utilizado para la predicción de la estructura de proteínas se basa en la detección de una relación homóloga con una proteína de estructura conocida y el uso de esta proteína como plantilla para modelar la estructura de la proteína de consulta en él (45), (46).

6. CONCLUSIONES

- En este trabajo se determinó que existían diferencias significativas para los aminoácidos Arg en las proteínas implicadas en la función de replicación, recombinación y reparación (Grupo D) y Tyr en el Grupo D respecto a las proteínas implicadas en la transcripción (Grupo C).

- Se determinó que algunos Grupos funcionales tienen alta abundancia de aminoácidos asociados a hélices transmembrana, los cuales son el Grupo H (pared celular, membrana y biogénesis), el Grupo I (tráfico intracelular, secreción y transporte vesicular), y el Grupo Q (transporte y metabolismo de iones inorgánicos).

- Se estableció las estructuras 3D de las proteínas de *M. kandleri* AV19 por cada grupo funcional, indicando una probable estructura estándar para cada función.

- Este trabajo es una aproximación teórica a un análisis bioquímico del proteoma de *M. kandleri* AV19, para inferir su posible relación con LUCA.

7. BIBLIOGRAFIA

1. Alberts B, Johnson A, Lewis J, et al. The Shape and Structure of Proteins. *Molecular Biology of the Cell.* 4th edition. New York : Garland Science, 2002.

2. J., Gittleman. Phylogeny - biology. [En línea] Encyclopedia Britannica, 2016. https://www.britannica.com/science/phylogeny#ref281267 .

3. Koonin EV, Galperin MY.. Chapter 2, Evolutionary Concept in Genetics and Genomics. *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics.* Boston : Kluwer Academic, 2003.

4. *Biotechnological applications of extremophiles, extremozymes and extremolytes. 99:.* Raddadi, N., Cherif, A., Daffonchio, D., Neifar, M. & Fava, F. 2015, Appl. Microbiol. Biotechnol., Vol. 99, págs. 7907–7913.

5. *The genomics of LUCA.* W., Mat. 2008, Frontiers in Bioscience, Vol. 13, pág. 5605.

6. *The complete genome of hyperthermophile Methanopyrus kandleri AV19 and monophyly of archaeal methanogens.* Slesarev A, Mezhevaya K, Makarova K, Polushin N, Shcherbinina O, Shakhova V et al. 2002, Proceedings of the National Academy of Sciences, Vol. 7, págs. 46, 99.

7. *Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. .* Brochier C, Forterre P, Gribaldo S. 3, 2004, Genome Biology, Vol. 5, pág. R17.

8. *In silico functional characterization of a double histone fold domain from the Heliothis zea virus 1. .* Greco C, Fantucci P, De Gioia L. 2005, Italian Society of Bioinformatics (BITS): Annual Meeting , Vol. 6(Suppl 4), pág. S15.

9. *Many non-universal archaeal ribosomal proteins are found in conserved gene clusters. .* Wang J, Dasgupta I, Fox G. 4, 2009, Archaea, Vol. 2, págs. 241-251.

10. *Evidence for early life in Earth's oldest hydrothermal vent precipitates. .* Dodd M, Papineau D, Grenne T, Slack J, Rittner M, Pirajno F et al. 543, 2017, Nature, Vol. 7643, págs. 60-64.

11. LEY DE BENFORD. [En línea] Estadistica para todos, 2008. http://www.estadisticaparatodos.es/taller/benford/benford.html.

12. *Genome Sizes and the Benford Distribution.* Friar J, Goldman T, Pérez–Mercader J. 5, 2012, PLoS ONE, Vol. 7, pág. e36624.

13. Madigan M, Martinko J, Parker J, Gacto Fernández M. *Biología de los microorganismos. Brock. .* 10th ed. Madrid, España : Pearson, 2004.

14. Casanova V. Clasificación de los extremófilos. (Curso de Astrobiología, capítulo 2). . [En línea] Astrofisicayfisica.com, 2017. http://www.astrofisicayfisica.com/2013/05/clasificacion-de-los-extremofilos-curso.html .

15. *Extremophiles: Sustainable Resource of Natural Compounds-Extremolytes.* Kumas R, Patel D, Bansal D, Mishra S, Mohammed A, Arora R et al,. 2009, Sustainable Biotechnology, págs. 279-294.

16. Darwin, C. *El origen de las especies por medio de la selección natural.* s.l. : Editorial CSIC-CSIC Press., 2009. pág. 464.

17. *Looking for the Last Universal Common Ancestor (LUCA).* Koskela M, Annila A. 4, 2012, Genes, Vol. 3, págs. 81-87.

18. Egel, Richard. Integrative Perspectives: In Quest of a Coherent Framework for Origins of Life on Earth. [aut. libro] Dirk-Henner Lankenau, Armen Y. Mulkidjanian Richard Egel. *Origins of Life: The Primal Self-Organization.* s.l. : Springer Science & Business Media, 2011.

19. *The universal tree of life: an update. .* P, Forterre. 2015, Frontiers in Microbiology, Vol. 6, pág. 717.

20. Arunas L. . We've been wrong about the origins of life for 90 years. [En línea] EDT. The Conversation US, Inc., 2016. http://theconversation.com/weve-been-wrong-about-the-origins-of-life-for-90-years-63744.

21. *Emergence of life: from functional RNA selection to natural selection and beyond. .* J., Tze-Fei Wong. 7, 2014, Frontiers in Bioscience. , Vol. 19, pág. 1117.

22. *Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic.* Wong J, Ng S, Mat W, Hu T, Xue H. 1, 2016, Life. MDPI-Life, Vol. 6, pág. 12.

23. *Protein folding at extreme temperatures: Current issues.* Feller, Georges. 2017, Seminars in Cell & Developmental Biology, pág. 9.

24. *Search for Primitive Methanopyrus Based on Genetic Distance Between Val- and Ile-tRNA Synthetases.* Yu, Z., Takai, K., Slesarev, A. et al. J Mol Evol. 2009, Journal of Molecular Evolution, Vol. 69, págs. 386–394 .

25. *Analysis of two large functionally uncharacterized regions in the Methanopyrus kandleri AV19 genome.* Jensen LJ, Skovgaard M, Sicheritz-Pontén T, et al. 12, 2003, BMC Genomics , Vol. 4.

26. *A Novel Unsaturated Archaeal Ether Core Lipid from the Hyperthermophile Methanopyrus kandleri.* Hafenbradl D., Keller M., Thierick R., Stetter k. 2, 1993, Elsevier, Vol. 16, págs. 165-169.

27. *A two-subunit type I DNA topoisomerase (reverse gyrase) from an extreme hyperthermophile. .* Krah R, Kozyavkin SA, Slesarev AI, Gellert M. 1, 1996, Proceedings of the National Academy of Sciences of the United States of America, Vol. 93, págs. 106-110.

28. *Evidence for an early prokaryotic origin of histones H2A and H4 prior to the emergence of eukaryotes.* Slesarev AI, Belova GI, Kozyavkin SA, Lake JA. 2, 1998, Nucleic Acids Research, Vol. 26, págs. 427-430.

29. The Chemical Components of a Cell. [aut. libro] Johnson A, Lewis J, et al. Alberts B. *Molecular Biology of the Cell.* 4th edition. . New York : Garland Science., 2002.

30. Rodwell V, Murray R, Granner D, Mayes P. 26th ed. United States of America : McGraw-Hill Publishing, 2003.

31. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins M, Appel R et al. Protein Identification and Analysis Tools on the ExPASy Server . [En línea] Expasy, 2018. https://web.expasy.org/docs/expasy_tools05.pd.

32. Alberts B, Johnson A, Lewis J, et al. The Shape and Structure of Proteins. *Molecular Biology of the Cell.* 4th edition. New York : Garland Science, 2002.

33. *Structural features of transmembrane helices.* Werner P., Preissner R., Frömmel C. 2002, Elsevier, Vol. 559, págs. 145-151.

34. *Protein Databases on the Internet.* . Xu D, Xu Y. 2004, Current Protocols in Molecular Biology, págs. Unit–19.4.

35. *The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999.* Bairoch A, Apweiler R. 1, 1999, Nucleic Acids Res, Vol. 27, págs. 49-54.

36. Home, ExPASy: SIB Bioinformatics Resource Portal -. ExPASy. [En línea] 2017. https://www.expasy.org/.

37. *The Phyre2 web portal for protein modeling, prediction and analysis.* Kelley L, Mezulis S, Yates C, Wass M, Sternberg M. 6, 2015, Nature Protocols, Vol. 10, págs. 845-858.

38. *The RCSB Protein Data Bank: views of structural biology for basic and applied research and education.* Rose P, Prli A, Bi C, Bluhm W, Christie C, Dutta S et al. 1, 2014, Nucleic Acids Research, Vol. 43, págs. 345 - 356.

39. *The I-TASSER Suite: Protein structure and function prediction. .* J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang. 7 - 8, 2015, Nature Methods, Vol. 12.

40. *Coevolution Theory of the Genetic Code at Age Forty: Pathway to Translation and Synthetic Life.* Wong, J. T.-F., Ng, S.-K., Mat, W.-K., Hu, T., & Xue, H. 1, 2016, LIFE, Vol. 6, pág. 12.

41. Scwald N, Dieter H. *Peptides: Chemistry and Biology.* s.l. : Wilye VCH, 2002.

42. *Arginine residues at internal positions in a protein are always charged. .* Harms MJ, Schlessman JL, Sue GR, García-Moreno E. B. 47, 2011, Proceedings of the National Academy of Sciences of the United States of America., Vol. 108, págs. 18954-18959.

43. *Transmembrane helix predictions revisited.* Chen CP, Kernytsky A, Rost B. 12, 2002, Protein Science : A Publication of the Protein Society, Vol. 11, págs. 2774-2791.

44. *A Universal Trend among Proteomes Indicates an Oily Last Common Ancestor.* Ranjan V. Mannige, Charles L. Brooks, Eugene I. Shakhnovich. 12, 2012, PLoS Comput Biology, Vol. 8.

45. *Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling.* Meier A, Söding J. 10, 2015, PLOS Computational Biology, Vol. 11, pág. e1004343.

46. *Mass Spectrometry Coupled Experiments and Protein Structure Modeling Methods.* Pi J, Sael L. 12, 2013, International Journal of Molecular Sciences., Vol. 14, págs. 20635-20657.

47. *Extremophiles and their application to veterinary medicine.* Irwin J, Baird A. 6, 2004, Irish Veterinary Journal., Vol. 57, pág. 348.

48. *The COG database: a tool for genome-scale analysis of protein functions and evolution. .* Tatusov R, Galperin M, Natale D, Koonin E. 1, 2000, Nucleic Acids Research., Vol. 28, págs. 33-36.

49. Koonin EV, Galperin MY. Chapter 2, Evolutionary Concept in Genetics and Genomics. . *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics.* Boston : Kluwer Academic , 2003. .

50. *Emergence of life: from functional RNA selection to natural selection and beyond.* J., Tze-Fei Wong. 7, 2014, Frontiers in Bioscience., Vol. 19, pág. 1117.

51. *Database resources of the National Center for Biotechnology Information.* Coordinators, NCBI Resource. 1, 2015, Nucleic Acids Research, Vol. 44, págs. 7 - 19.

52. Tool, NCBI N. BLAST: Basic Local Alignment Search. [En línea] 2017. https://blast.ncbi.nlm.nih.gov/Blast.cgi.

53. *BLAST: a more efficient report with usability improvements.* Boratyn G, Camacho C, Cooper P, Coulouris G, Fong A, Ma N et al. 1, 2013, Nucleic Acids Research, Vol. 41, págs. 29 - 33.

54. *The Phyre2 web portal for protein modeling, prediction and analysis.* al., Kelley LA et. 2015, Nature Protocols, Vol. 10, págs. 845-858 .