UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

ANÁLISIS EVOLUTIVO DE LAS RUTAS METABÓLICAS ANCESTRALES

DERIVADAS DEL CATÁLOGO PROTEICO DEL

ÚLTIMO ANCESTRO COMÚN (LCA).

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS BIOLÓGICAS
(BIOLOGÍA EXPERIMENTAL)

P R E S E N T A

ANDRADE DÍAZ FERNANDO ABRAHAM

TUTOR PRINCIPAL DE TESIS: **DR. ARTURO CARLOS II BECERRA BRACHO**

COMITÉ TUTOR: **DR. ENRIQUE MERINO PÉREZ**
DR. ANTONIO LAZCANO ARAUJO

MÉXICO, D.F. **MAYO, 2012**

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS BIOLÓGICAS

FACULTAD DE CIENCIAS

ANÁLISIS EVOLUTIVO DE LAS RUTAS METABÓLICAS ANCESTRALES

DERIVADAS DEL CATÁLOGO PROTEICO DEL

ÚLTIMO ANCESTRO COMÚN (LCA).

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS BIOLÓGICAS
(BIOLOGÍA EXPERIMENTAL)

P R E S E N T A

ANDRADE DÍAZ FERNANDO ABRAHAM

TUTOR PRINCIPAL DE TESIS: **DR. ARTURO CARLOS II BECERRA BRACHO**

COMITÉ TUTOR: **DR. ENRIQUE MERINO PÉREZ**
DR. ANTONIO LAZCANO ARAUJO

MÉXICO, D.F. **MAYO, 2012**

VNIVERIDAD NACIONAL
AVTÓNOMA DE MEXICO

POSGRADO EN CIENCIAS BIOLÓGICAS
FACULTAD DE CIENCIAS
DIVISIÓN DE ESTUDIOS DE POSGRADO

OFICIO FCIE/DEP/112/12

ASUNTO: Oficio de Jurado

Dr. Isidro Ávila Martínez
Director General de Administración Escolar, UNAM
P r e s e n t e

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **16 de mayo de 2011** se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS (BIOLOGÍA EXPERIMENTAL)** del (la) alumno (a) **ANDRADE DIAZ FERNANDO ABRAHAM** con número de cuenta **99015163** con la tesis titulada **"Análisis evolutivo de las rutas metabólicas ancestrales derivado del catálogo proteico del último ancestro común (LCA)",** realizada bajo la dirección del (la**) DR. ARTURO CARLOS II BECERRA BRACHO:**

Presidente:	DR. VÍCTOR MANUEL VÁLDES LÓPEZ
Vocal:	DR. ENRIQUE MERINO PÉREZ
Secretario:	DR. LEÓN PATRICIO MARTÍNEZ CASTILLA
Suplente:	DR. RAFAEL CAMACHO CARRANZA
Suplente:	DR. LUIS FELIPE JIMÉNEZ GARCÍA

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

A t e n t a m e n t e
"POR MI RAZA HABLARA EL ESPÍRITU"
Cd. Universitaria, D.F., a 1ro. de marzo de 2012

Dra. María del Coro Arizmendi Arriaga
Coordinadora del Programa

FACULTAD DE CIENCIAS
DIVISION DE ESTUDIOS
DE POSGRADO

MCAA/MJFM/ASR/grf*

AGRADECIMIENTOS

Primeramente agradezco al programa de Maestría en Ciencias Biológicas (Biología experimental), programa en el cual participe.

Al apoyo económico que me brindo CONACYT. No. de becario: 220326.

A mi comité tutor conformado por el Dr. Enrique Merino y al Dr. Antonio Lazcano por sus comentarios y su infinita paciencia.

AGRADECIMIENTOS A TÍTULO PERSONAL

Primero que nada me permito agradecerle a mi tutor principal Dr. Arturo Becerra Bracho por haberme brindado la maravillosa oportunidad de trabajar a su lado. Por su guía y paciencia siempre estaré agradecido.

A todo el personal del laboratorio de Microbiología de la Facultad de Ciencias. Muy en especial a Ricardo, Daniel, Mario y Héctor, quienes fueron no sólo mis compañeros sino mis amigos. Siempre estuvieron ahí, tanto en tiempos buenos como malos.

A mis maravillosos profesores de las diversas materias que curse durante el programa. La preparación que me dieron todo el tiempo me resulta útil y fuente de conocimiento para mi vida

Por supuesto y no menos importante a mi más grande motor de vida: Berenice. Quién estuvo conmigo, no sólo a mi lado, sino brindándome inspiración, apoyo, regaños, ayuda incondicional, observaciones y en fin, todo lo necesario para que esta tesis fuera posible. Sin temor a equivocarme puedo decir que sin ella no hubiera logrado nada de esto. Esperando siga a mi lado mucho tiempo más.

ÍNDICE

Resumen.

El último ancestro común LCA (*Last common ancestor,* por sus siglas en inglés) puede ser definido como el conjunto de características comunes a todos los seres vivos. Para estudiar las propiedades del LCA diferentes estrategias se han ocupado, por ejemplo, la genómica comparada y la cladística molecular. Cada estudio ha tomado la información disponible en las bases de datos públicas con lo cual la cantidad de información con la que contó cada reconstrucción estuvo sesgada. La información genómica en las bases de datos públicas ha aumentado exponencialmente, aumentando la probabilidad de obtener sesgos en las reconstrucciones del LCA, debido al aumento de información contenida en dichas bases de datos. Para poder evitar esos sesgos definí una metodología que pudiera señalar el número mínimo de genomas a comparar para reconstruir al LCA. Hice comparaciones de genomas completos para realizar curvas de rarefacción para poder definir la muestra mínima que representara todo el listado de genes del LCA. Los resultados que obtuve muestran que al cabo de la muestra 31 de genomas comparados éstos cubren 85% de los genes altamente conservados, es decir, comparando 93 genomas de los tres dominios celulares se alcanza a cubrir la mayor parte de los genes conservados pertenecientes a diferentes procesos metabólicos. El metabolismo de nucleótidos, la traducción y el metabolismo de aminoácidos fueron los mejor representados. Si estos genes estuvieron en el LCA entonces éste debió de haber sido un organismo de alta complejidad no sólo genética, sino con diferentes funciones metabólicas.

Abstract

The last common ancestor can be defined as the group of common features to all living beings. For studying the LCA different strategies has been applied, as an example comparative genomics and molecular cladistics. Each study used the information available en the public databases. Since the information in the public databases has been growing up so is the probability for biases in the LCA reconstructions. To avoid this bias I defined a methodology able to point out the minimal number of genomes to be compared to reconstruct the LCA. I made samples of complete genomes to make rarefaction curves to define a gene list for the LCA. The results shows that by the 31th sample of the genome comparisons it is achieved 85% of the high conserved genes, we only need 93 complete genomes compared to reach most of the conserved genes. The nucleotide metabolism, translation and amino acids biosynthesis are the best represented, if those genes were already in the LCA, it means that it was a complex organism not only genetically but with different metabolic features.

Introducción

Último ancestro común.

El estudio del último ancestro común (LCA, por sus siglas en inglés *Last Common Ancestor*) es parte del estudio de la evolución temprana de la vida en la tierra. El LCA como entidad biológica es resultado de un proceso evolutivo que se llevó a cabo antes de la diversificación de los tres dominios celulares. Un resumen de estas etapas se encuentra en la figura 1.

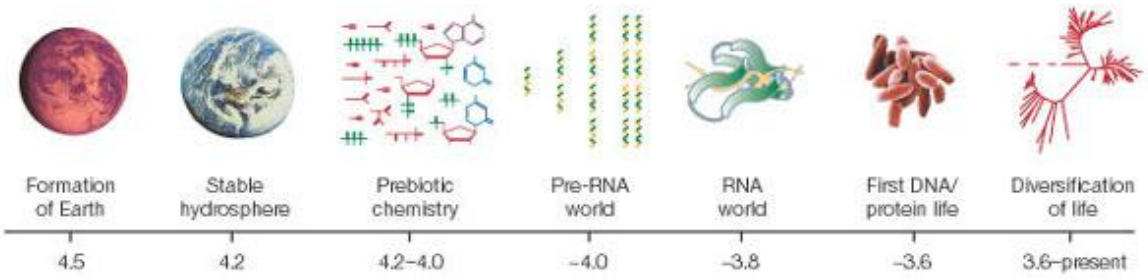


Figura 1. Eventos principales referentes a la aparición de la vida en la tierra, tomado de Joyce, 2002.

Los fósiles más antiguos corresponden a ensambles de procariontes de aproximadamente 3500 milloones de años (Schopf, 1993; Schopf 2006), formaciones que ya corresponden a comunidades de procariontes. Esta evidencia física está lejos del LCA. Por lo tanto evidencias indirectas, como genómica comparada y cladística molecular.

Las filogenias moleculares utilizando 16s/18s rRNA revelan que todos los organismos pueden estar agrupados en tres principales líneas monofiléticas, los dominios celulares

Bacteria, Archaea y *Eucarya* (Woese, 1990). Si bien esta filogenias fueron realizadas utilizando únicamente solo un marcador molecular existen también algunas moléculas comunes a los tres dominios celulares, y que al realizar una filogenia con estos marcadores se obtiene una topología con los mismos tres dominios (Cicarelli, 2006). De entre las diferentes características comunes a todos los seres vivos están el código genético, expresión y replicación génica, las mismas reacciones anabólicas, la forma de producción de energía dependiente de ATPasa, entre otras. El LCA se puede inferir a partir de las características compartidas a todos los seres vivos, localizado en la base de las filogenias moleculares. Variaciones menores en las características comunes pueden ser explicadas por cambios posteriores a que el LCA divergiese a lo largo del tiempo y se separaran en los tres dominios de la vida (Becerra, 2007).

Existe controversia con respecto a que dominios debió pertenecer el LCA. Siendo tres dominios en un árbol sin raíz, en principio la forma más sencilla de saber a que dominio perteneció es enraizando este árbol, así conocer cual es el grupo plesiomórfico. Di Gulio (2007) señala que el grupo ancestral debía de estar dentro de las *Archeas*, específicamente en la línea de nanoarchaeota, *Nanoarcheum equitans*, por poseer características únicas y ancestrales, por ejemplo, organización genómica carente de operonesLa postura de que el dominio más antiguo es el *Bacteria*, favorecido por inferencias filogenéticas de genes repetidos (Iwabe, 1989) y recientemente el uso de un conjunto de genes parálogos polarizando indeles ha hecho robusta la propuesta que *Bacteria* es el grupo más plesiomórfico (Valas, 2009). Pese a la evidencia incluso se ha señalado que el grupo más antiguo pudo haber sido la línea eucarionte (Forterre, 1999). Incluso con la falta de

consenso respecto a la naturaleza del LCA, la genómica comparada nos ha permitido acercarnos al contenido de genes de éste.

El primer estudio que utilizó la comparación de genomas completos para la reconstrucción de caracteres ancestrales fue Mushegian y Koonin en 1996. En este estudio compararon los genomas de las dos únicas bacterias disponibles en el año antes mencionado, *Mycoplasma genitallium* y *Haemophillus influenza*, en busca de las características comunes mínimas a la vida. Ese estudio dio origen a la genómica comparada. Ésta ha sido ocupada para realizar inferencias evolutivas de cómo pudo haber sido el último ancestro común. Pero desde este comienzo de la genómica comparad sesgos han existido. El uso de organismos parásitos para realizar inferencias evolutivas puede resultar en fuertes sesgos por las pérdidas secundarias derivadas del estilo de vida de estos organismos (Becerra, 1997). Sin embargo las características compartidas entre genomas de organismos muy distantes implica que pudieron estar en el último ancestro común a ellos (Huynen, 1998; Koonin, 2003). Este principio de biología comparada aplicada a genomas completos ha servido como base para realizar inferencias sobre la naturaleza del LCA.

En adición a los trabajos de reconstrucción del LCA, existen otros que utilizando la genómica comparada intentan detallar otros aspectos del ancestro utilizando a organismos distantes. Kurland (2007) buscando a que súper familia de plegamiento (*fold superfamily*) se parecía más el LCA, con el fin de determinar a que dominio correspondía este. Analizando así organismos de los tres dominios; Peregrín-Álvarez (2007), comparó

proteomas completos de los tres dominios para observar la diversidad de secuencias únicas y homólogas a lo largo de todos los seres vivos.

Pese a las diversas metodologías existen categorías funcionales comunes a todas ellas, transcripción y traducción, replicación y reparación de DNA, y en menor medida, proteínas asociadas a membrana. Independientemente de la metodología de reconstrucción, esos mecanismos se pueden considerar presentes en el metabolismo del LCA y robustos desde el punto de vista metodológico. Sin embargo esta conjetura resulta *a priori* debido a que aun no existe ningún trabajo que señale lo robusto de las pruebas, entonces se puede concluir que independientemente de la metodología utilizada existirá un sesgos intrínsecos a la bases de datos, cada estudio estará limitado por el momento en el que se realice el análisis, el número de genomas completos secuenciados y la diversidad de organismos que posean genoma secuenciado. Un resumen de las metodologías ocupadas están enlistadas en la tabla 1.

Propiedad del Ancestro	**Metodología**	**Número de secuencias y categoría funcional.**
LCA		**80 COGs universalmente distribuidos**
Eficiente transcripción y estructura del ribosoma; funciones ligadas a membrana; capaz de sintetizar largas hebras de DNA (Harris, 2003)	Identificación universalmente distribuidos en los tres dominios.	Transcripción y traducción (63/80). Replicación y reparación de DNA (5/80) Proteínas asociadas a membrana (1/80) Metabolismo de aminoácidos)(1/80) Manejos de proteínas (2/80) Otros (2/80)
LUCA		**600 genes asignados a LUCA (COGs)**
Casi suficiente genes para mantener funcionalmente a un organismo (Mirkin, 2003)	Construcción de escenarios parsimoniosos para grupos individuales de COGs basados en	Transcripción y traducción (112/600) Replicación y reparación de DNA (30/600)

	árboles de especies.	Proteínas de membrana y metabolismo (287/600) Manejo de proteínas (25/600) Otras (94/600)
LUCA		**63 genes universales (proteínas)**
Simple con pocos genes; carente de un genoma de DNA y de un sistema de replicación. (Koonin, 2003).	Comparación de las secuencias de 100 genomas.	Transcripción y traducción (56/63) Replicación y reparación de DNA (3/63) Proteínas asociadas a membrana (1/49) Manejo de proteínas (1/63)
LCA		**49 plegamientos de proteínas universalmente distribuidos (superfamilias SCOP)**
Con una sofisticada maquinaria genética en estructura y equipo (Yang, 2005)	Distribuciones de las superfamilias de SCOP en 174 genomas completos.	Transcripción y traducción (39/42) Replicación y reparación de DNA (5/49) Metabolismo (5/49) Manejo de proteínas (1/49) Otros (5/49)
LCA		**115 dominios de proteínas (dominios de Pfam)**
Similares a las células actuales en complejidad genética (Delaye, 2005)	Comparación de secuencias de 20 genomas con BLAST e identifación de ortólogos utilizando la base de datos Pfam	Transcripción y traducción (56/115) Replicación y reparación de DNA (6/115) Proteínas asociadas a membrana (7/115) Metabolismo de nucleótidos y azúcares (33/115 Metabolismo de aminoácidos (12/115) Manejo de proteínas (1/115)
Propiedad del Ancestro	**Metodología**	**Número de secuencias y categoría funcional.**
El total de los octámeros deben de estar en el orden de magnitud de miles (Sobolevski, 2006)	Identificación de octámeros prácticamente omnipresentes en motivos de proteínas.	Transcripción y traducción; Replicación y reparación de DNA; manejo de proteínas; proteínas asociadas a membrana
LUCA		**1000 genes con un mínimo de 561 a 669 secuencias/ categorías funcionales (proteínas)**
Organismos complejos en sus genomas, similares a los procariontes de vida libre actuales (Ouzounis, 2006)	Identifación de secuencias homólogas entre 184 genomas, utilizando un método que corrige entre las pérdidas de genes.	Transcripción y traducción (34/659) Replicación y reparación de DNA (35/659) Proteínas asociadas a membrana (120/659) Metabolismo (309/659) Otras (161/659)
LUCA		**140 dominios de proteínas ancestraes (superfamilias CATH)**

Entidad genéticamente compleja, con prácticamente todos las características presentes en organismos actuales (Ranea, 2006)	Distribución de las superfamilias CATH en 114 genómas completos	Transcripción y traducción (52/140) Replicación y reparación de DNA (12/140) Proteínas asociadas a membrana (2/140) Metabolismo (46/140) Otros (28/140)

Tabla 1. Obtenida de Becerra, 2007, lista las reconstrucciones de último ancestro común, estimando el contenido de genes y la naturaleza de estos, así como listando las metodologías ocupadas para cada estudio.

Genómica e información.

Según la base de datos GOLD, *Genomes Online Data Base* (http://www.genomesonline.org/gold_statistics.htm), el número de genomas completamente secuenciados desde 1995 a la fecha ha ido aumentado exponencialmente (figura 2), sin embargo la representatividad de grupos ha sido desigual. Los organismos mejor representados son las Proteobacterias con 44% del total de las bases de datos, Firmicutes 28%, Actinobacteria 12% y el resto de los organismos 18%. Recientemente en las diferentes bases de datos públicas se alcanzo el genoma mil completamente secuenciado para *Bacteria* en diferentes bases de datos (Lagesen, 2010). Se confirmó que dentro de este dominio las Proteobacterias están sobre representadas. Esto puede permitir realizar cada vez comparaciones más completas y obtener un grupo de genes conservados para este dominio, pero aun no se tiene una completa representación de toda la biodiversidad en las bases de datos públicas ni siquiera para *Bacteria*. En base a esos datos la reconstrucción del LCA presenta problemas generados por sesgos en las bases de datos y, además, por factores biológicos como el transporte horizontal, la diferente taza de sustitución en

proteínas, el grado de pérdidas secundarias en adición a los sesgos de las bases de dato por si mismas (Mirkin, 2003)

Para evitar sesgos en la reconstrucción en este estudio se planteó una metodología ecológica aplicada a genómica comparada. Para evitar hacer comparaciones de los más de mil genomas completos disponibles en las bases de datos se construirán curvas de rarefacción que permitan otorgar un número mínimo de genomas a comparar.

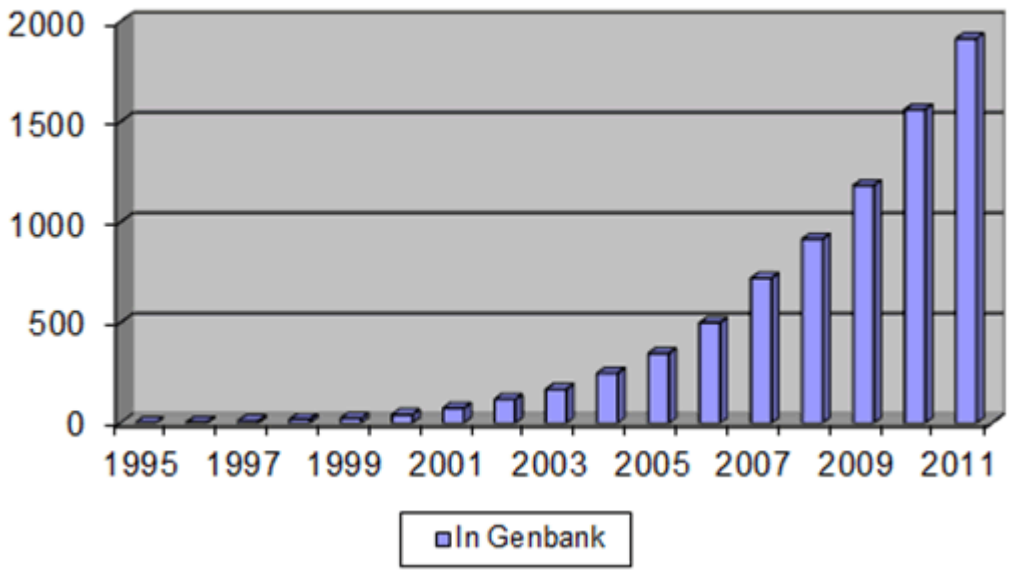


Figura 2. Genomas completamente secuenciados en Genbank, desde el primero completamente secuenciado en 1995 a la fecha. Figura obtenida de http://www.genomesonline.org/cgi-bin/GOLD/index.cgi?page_requested=Statistics.

Curvas de rarefacción

La riqueza de especies es una medida de la diversidad de un habitad. Comúnmente no se puede acceder a todas las especies de una región o localidad, se necesita hacer un muestreo para poder acceder a la riqueza. La cuantificación de la riqueza se ha utilizado para la comparación entre sitios y en conservación, para observar la saturación en la colonización de diferentes especies (Gotelli, 2001). La curva de acumulación de especies sirven para observar esta saturación de forma gráfica.

La curva de acumulación de especies es la forma gráfica de observar las especies a través del esfuerzo de muestreo (Colwell, 2004). En Ecología ha sido utilizada para medir el ensamble de especies, para estimar el número de especies dado un muestreo y para planear un muestreo mínimo con una riqueza representativa (Mao, 2005), y recientemente se ha utilizado para estimar porcentajes de cobertura de secuencias en análisis de metagenómica (Qin, *et al* 2010). Para su construcción se necesita graficar secuencialmente los individuos en una misma muestra o bien la acumulación de muestras. Mientras que las curvas de individuos suelen ser discontinuas, la curva de muestras es continua (Gotelli, 2001) y es llamada rarefacción basada en muestras.

Para realizar la rarefacción basada en muestras es posible trabajar con dos diferentes tipos de matrices, la matriz de abundancia y la de presencia-ausencia. La primera refiere a la abundancia relativa en el muestreo, indicando el número de individuos para cada muestra. La de presencia-ausencia es una matriz binaria que, como su nombre lo indica, únicamente

hace referencia con 0 y 1 a si una especie está o no en el análisis. Siempre es posible hacer la transformación de una matriz de abundancia en una de presencia-ausencia al transformar los valores diferentes de 0 por 1 (Colwell, 2004). El uso de estas matrices para el ensamble de una curva de rarefacción es rutinario en Ecología y la formalización estadística de este modelo ha sido sujeto de estudio. Como consecuencia se ha desarrollado las formulas analíticas para la realización de las curvas de rarefacción y para calcular los intervalos de confianza como una medida de dispersión de datos. Los parámetros antes mencionados están implementados en el programa *EstimateS* (Colwell, 2004).

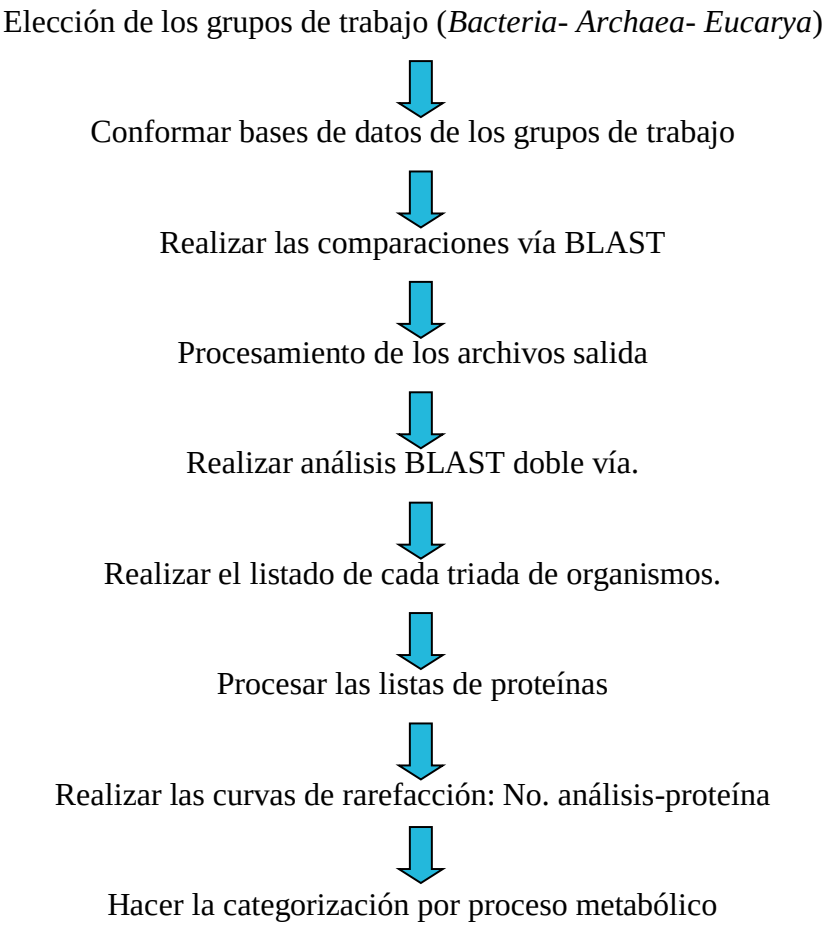
Objetivos

- Obtener las proteínas homólogas a los tres dominios, usando una metodología que incluya el uso creciente de bases de datos. A partir de dicho listado realizar un listado de proteínas del LCA.
- Establecer si las bases de datos actuales son representativas para reconstruir al LCA completamente. Conocer el sesgo de las bases de datos en la reconstrucción del LCA.
- Reconstruir posibles rutas metabólicas ancestrales que se pudieron haber encontrado en el último ancestro común.

Metodología

Filogenia de cada dominio celular.

El resumen de la metodología se muestra en el siguiente diagrama de flujo:



Antes de la selección de organismos de trabajo requerí de una filogenia que incluyera a todos los dominios celulares. Para *Bacteria* y *Archaea* existe una filogenia resuelta dentro del *All species living tree*, el cual ocupa como marcador 16S rRNA (Yarza, 2008).

Realicé la filogenia para los eucariontes con genomas completamente secuenciados del KEGG (*Kyoto Encyclopedia of Genes and Genomes*). Obtuve las secuencias 18s rRNA del

European Bioinformatics Institute (http://www.ebi.ac.uk/) y utilicé el programa MEGA 4 para realizar la filogenia (Tamura, 2007).

Reconstrucción del LCA, realización del catálogo proteínico del LCA

Para realizar la reconstrucción del LCA conformé 46 grupos. Éste número representa el número máximo de grupos que se pudieron conformar al momento de realizar el análisis, debido a que sólo existían 46 Archaeas con genoma completamente secuenciado. Cada grupo incluye un organismo de cada dominio celular, *Bacteria*, *Archaea* y *Eucarya*. Para conformar los grupos primero elegí un organismo del clado más plesiomófico de la filogenia *All species living tree* (Yarza, 2008) para *Bacteria* y *Archaea*, y elegí un organismo del clado más plesiomófico de la filogenia 18s RNA para *Eucarya.* Así cada grupo está agrupado con un representante de cada dominio. De cada grupo elegido, comprobé si existía un organismo con proteoma liberado para el KEGG. De ser así, elegía un organismo de ese grupo para el análisis. El proceso se repitió para cada clado que tuviese un representante en la base KEGG, agrupando organismos plesiomórficos con otros organismos plesiomórficos pero de otro dominio hasta tener representantes de toda la filogenia y al mismo tiempo se pudo obtener tríos de organismos, uno de cada dominio, con representantes a lo largo de toda la filogenia del dominio. La lista de organismos tomados se encuentra en la tabla del Apéndice A. A este grupo se le llamó organismos seleccionados con sesgo biológico, debido a que se utilizó una lógica filogenética para su selección.

También hice un análisis control tomando organismos dependiendo del año en que se completó su genoma. Para ello, ordene la lista de organismos de la base de datos KEGG en

la sección que muestra el año de liberación de cada genoma, lo ordené del más antiguo al más reciente para conformar los grupos. De manera similar al primer grupo de organismos elegidos, sesgo biológico, se conformó tríos de organismos, uno de cada dominio. Este segundo grupo puede ser considerado como control, debido a que fue el resultado de selección por conveniencia y fue llamado sesgo cronológico.

Obtuve las secuencias de proteínas de los organismos seleccionados del ftp KEGG (ftp://ftp.genome.ad.jp/pub/Kegg). Para cada grupo agrupé estas secuencias de tal manera que al final obtuve bases de datos para los 46 grupos que conforme. Realicé esto para poder comparar las secuencias posteriormente.

Comparé las secuencias utilizando el programa BLAST (*Basic Local Alignment Search Tool*). Efectué la alineación del concatenado contra el proteoma de cada organismo del grupo. Realicé el proceso inverso, e hice la alineación del proteoma de cada organismo del grupo contra el concatenado. Con las alineaciones que generé realicé el análisis de BLAST doble vía. Los algoritmos para realizar el doble vía fueron obtenidos del *System Biology Research and Resources* de Harvard (http://sysbio.harvard.edu/CSB/resources/computational/scriptome/UNIX/Protocols/Sequences.html).

Realicé programas con el lenguaje PERL para el manejo de los archivos resultados. Eliminé la redundancia de los archivos de las alineaciones. Eliminé del análisis cada una de las proteínas que encontraban un hit con ellas mismas del 100% de identidad. Para extraer los

resultados hice un programa que extrajera aquellas proteínas que encontrarán al menos un homólogo en cada uno de los otros dos dominios.

El principio del BLAST doble vía es buscar los mejores hits de la alineación organismo contra la base de datos y corroborar si son el mejor hit del análisis inverso de la base de datos contra el organismo. El resultado fue un listado de los mejores hits. Estos se pueden considerar ortólogos y fueron ocupados para la realización del listado de proteínas del LCA.

Realización de las curvas de rarefacción y categorización funcional.

Para llevar a cabo el procesamiento de las listas y realizar la curva de rarefacción obtuve el número de ortólogo KEGG (KO *number*). Comparé los KO de cada grupo buscando aquellos que estuvieran en cada uno de los tres análisis. De estos KOs encontré todos los que aparecían al menos una vez en todos los 46 grupos. Este listado es reportado como el listado de proteínas del LCA (Apéndice B).

Para construir la curva de rarefacción general para todos los grupos, convertí los listados de KOs en matrices de presencia/ausencia en binario, ocupando números 0 y 1 respectivamente. Cada matriz realizada toma en cuenta el listado de proteínas del LCA en número de KO como referencia a la diversidad total de secuencias. Todos los parámetros de diversidad fueron calculados utilizando el programa *EstimateS* (Colwell, 2005). Utilicé el parámetro de Mao Tao Sobs y para la asíntota el parámetro de Michaelis Menten al no

existir una diferencia entre los diferentes parámetros de extrapolación de curvas (Shaw, 2008).

Para cada número de KO existe al menos un número de *path*. El número de *path* representa a que ruta metabólica corresponde la proteína de KO. Con esta asociación fue posible categorizar cada proteína del listado de proteínas del LCA. Debido a que la anotación KO/*path* está dada para cada organismo, las diferentes listas para los 46 grupos fueron realizadas a partir de los organismos del mismo grupo, pero para el listado general del LCA ocupé el organismo modelo *Escherichia coli* K-12 para categorizarlo de manera independiente a los organismos utilizados.

Estas categorías a su vez pueden ser englobadas en súper-categorías del KEGG, ampliamente usadas para análisis globales de metabolismo (Peregrín-Álvarez, 2009). Cada número de *path* fue asociado a la súper-categoría correspondiente para poder analizar el conjunto de resultados de manera más práctica y para poder analizar lo completo de cada súper-categoría metabólica. Para cada grupo se categorizaron sus números KO en sus *paths* y a su vez en sus respectivas súper-categorías. Para cada súper-categoría se realizó una curva de rarefacción siguiendo los parámetros antes mencionados. Cada súper-categoría sirvió como guía de riqueza total de secuencias conservadas para la realización de la matriz presencia/ausencia. Además cada súper-categoría sirvió para categorizar todos los KOs por grupo y de manera global para todo el análisis.

Resultados

Análisis con sesgo biológico

El listado de proteínas del último ancestro común aparece en el apéndice B. Cada proteína está junto con su categorización metabólica y su número de acceso KO. Este listado fue realizado a partir de las proteínas de *E. coli* (ver metodología).

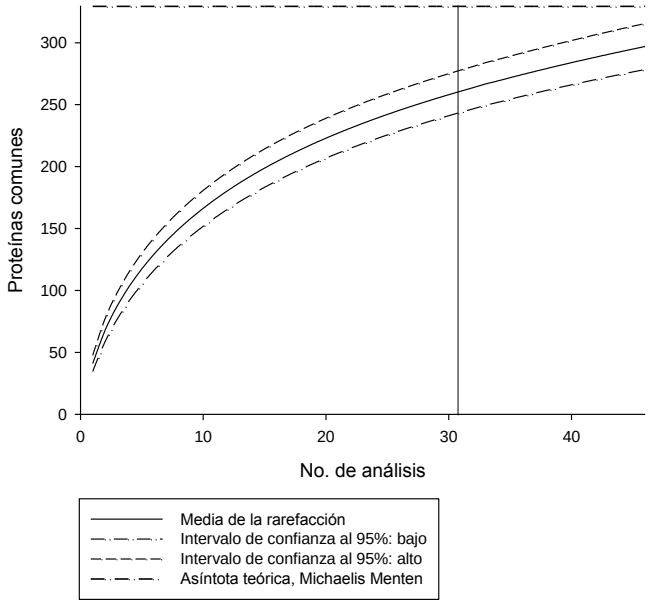
A partir del listado de la tabla 2 construí las curvas de rarefacción totales para todos los análisis. En la figura 3 se incluyen las curvas que incluyen todas las proteínas comunes, sin importar su categoría funcional. Debido a que las matrices para la construcción de las curvas fueron realizadas con los números de KO, todos los marcos de lectura abiertos y proteínas hipotéticas no fueron incluidas de primera instancia. Para tener en cuenta todas las secuencias altamente conservadas las proteínas hipotéticas también fueron incluidas, pero fueron analizadas en curvas separadas.

En cada curva se incluyen 4 parámetros que son la media de la rarefacción: la media de la rarefacción, los intervalos de confianza al 95% alto y bajo, y la asíntota teórica calculada con el parámetro de Michaelis Menten (Colwell, 2005). Estos parámetros permiten un análisis por coberturas. Para poder determinar si la media de la rarefacción está suficientemente próxima a la asíntota necesita alcanzar al menos un 85% de cobertura del valor total de la asíntota, mientras que el intervalo de confianza al 95% alto debe corresponder al 90% del valor total de la asíntota. Si ambas condiciones se cumplen,

podemos inferir que la media de la rarefacción ha alcanzado a la asíntota y por lo tanto se alcanza el muestreo mínimo.

Los porcentajes de cobertura son diferentes para los análisis generales con y sin hipotéticas. Cuando no se incluyen proteínas hipotéticas en el análisis 31 se alcanzan los valores de cobertura suficientes para el muestreo mínimo y para el análisis 46, el último, se alcanza un 90.16% de cobertura total con la media de la rarefacción y un 95.80% para el intervalo de confianza alto. El análisis que incluyó proteínas hipotéticas no alcanzó los valores para el muestreo mínimo. Para el análisis 46 se alcanzo una cobertura de 86% para la media de la rarefacción y del 89.14 para el intervalo de confianza alto. La diferencia en las medias de la rarefacción es clara, también gráficamente, en la cual la riqueza total es de 297 en el análisis sin hipotéticas y de 361 en el análisis con hipotéticas (figura 4).

Curva de acumulación de secuencias conservadas:
análisis sin proteínas hipotéticas



Curva de acumulación de secuencias conservadas:
análisis utilizando proteínas hipotéticas

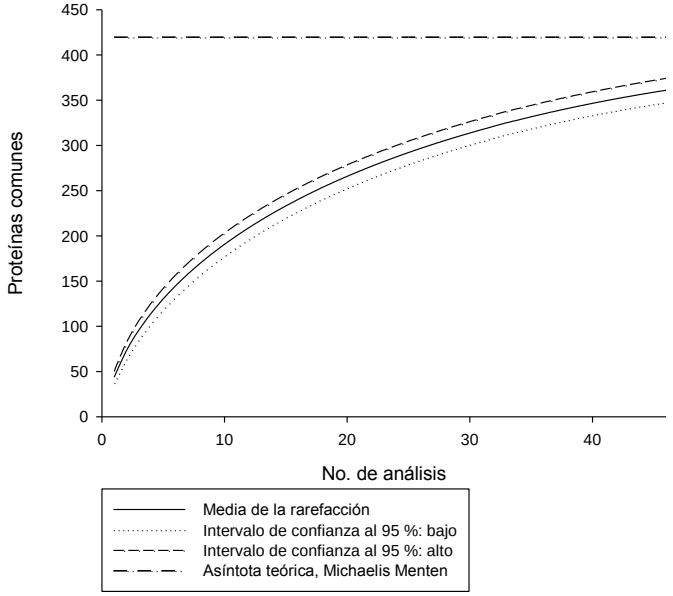


Figura 3. Curvas de rarefacción general. En la parte superior de la figura se muestra el análisis que no incluyó proteínas hipotéticas. En la parte inferior de la figura se muestra la curva que incluyó proteínas hipotéticas.

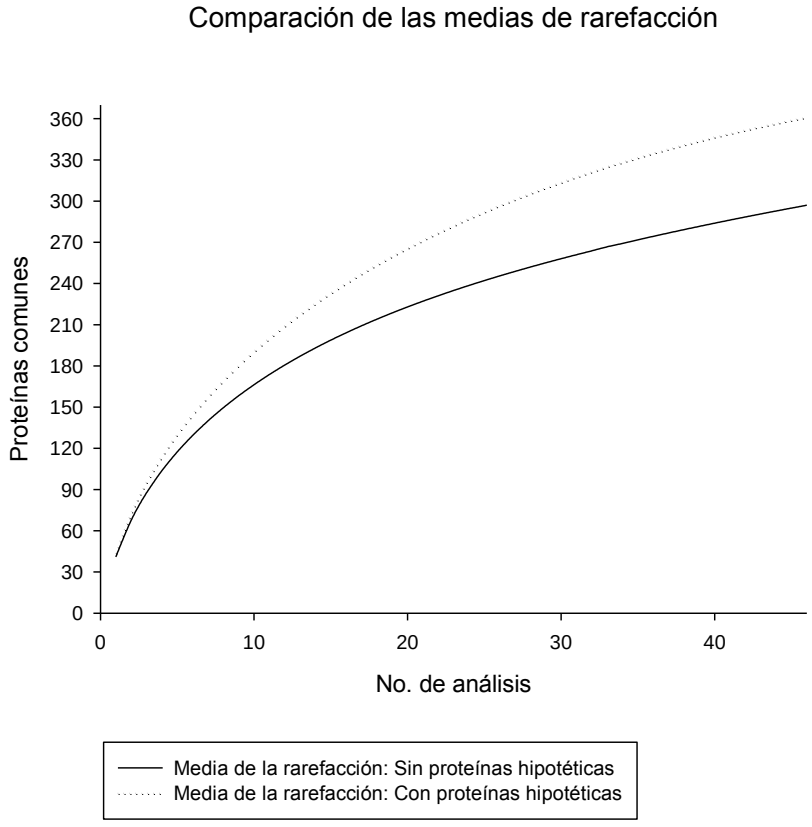


Figura 4. Se comparan las medias de la rarefacción de los análisis generales de la figura 2 para poder comparar diferencias entre ambos.

Los análisis de rarefacción fueron realizados sobre la categorización del listado del LCA, pero realizando un análisis por tipo de metabolismo. El resumen de estas curvas está en la figura 5. El metabolismo de aminoácidos alcanza los porcentajes requeridos para la muestra mínima después del análisis 22. El metabolismo de carbohidratos lo alcanza después del

análisis 44. El metabolismo de cofactores y vitaminas no alcanza la cobertura necesaria. El valor máximo que alcanza para la media de la rarefacción en el último análisis es de 79.51%. El metabolismo energético cumple con ambos parámetros para el análisis 44, mientras que su intervalo de confianza incluso separa la asíntota. El metabolismo de lípidos no llega a la cobertura para muestreo mínimo. El metabolismo de metabolitos secundarios tampoco cumple con ambos parámetros para el muestreo mínimo. El metabolismo de nucleótidos para el análisis 27 ya alcanza el los parámetros de muestreo mínimo. La traducción y proteínas asociadas alcanzan el muestreo mínimo en el análisis 23.

Para poder determinar el estado de cada ruta metabólica elaboré un mapa de calor. Estos presentan cada tipo de metabolismo y a que porcentaje del análisis corresponde cada categoría. Realicé dos gráficas, la primera corresponde a todas las proteínas que encontré en el estudio y el segundo las mismas proteínas pero tomando en cuenta todas las proteínas anotadas para cada categoría metabólica. Estas gráficas se encuentran en la figura 6. Los procesos mejor representados son el metabolismo de aminoácidos y la transcripción. Generalmente presentan siempre los porcentajes mayores y son constantes para cada análisis. El metabolismo de carbohidratos, pese a que no es constante como los anteriores, es el tercer mejor representado. Cuando este mismo conjunto de proteínas se evalúa con respecto a toda la ruta metabólica, la transcripción es el que representa un porcentaje mayor de toda la ruta, pero éste apenas alcanza a representar el 6.37% de todas las proteínas referentes a la transcripción. Estos porcentajes se pueden observar en la tabla 3.

Curva de acumulación de proteínas conservadas.

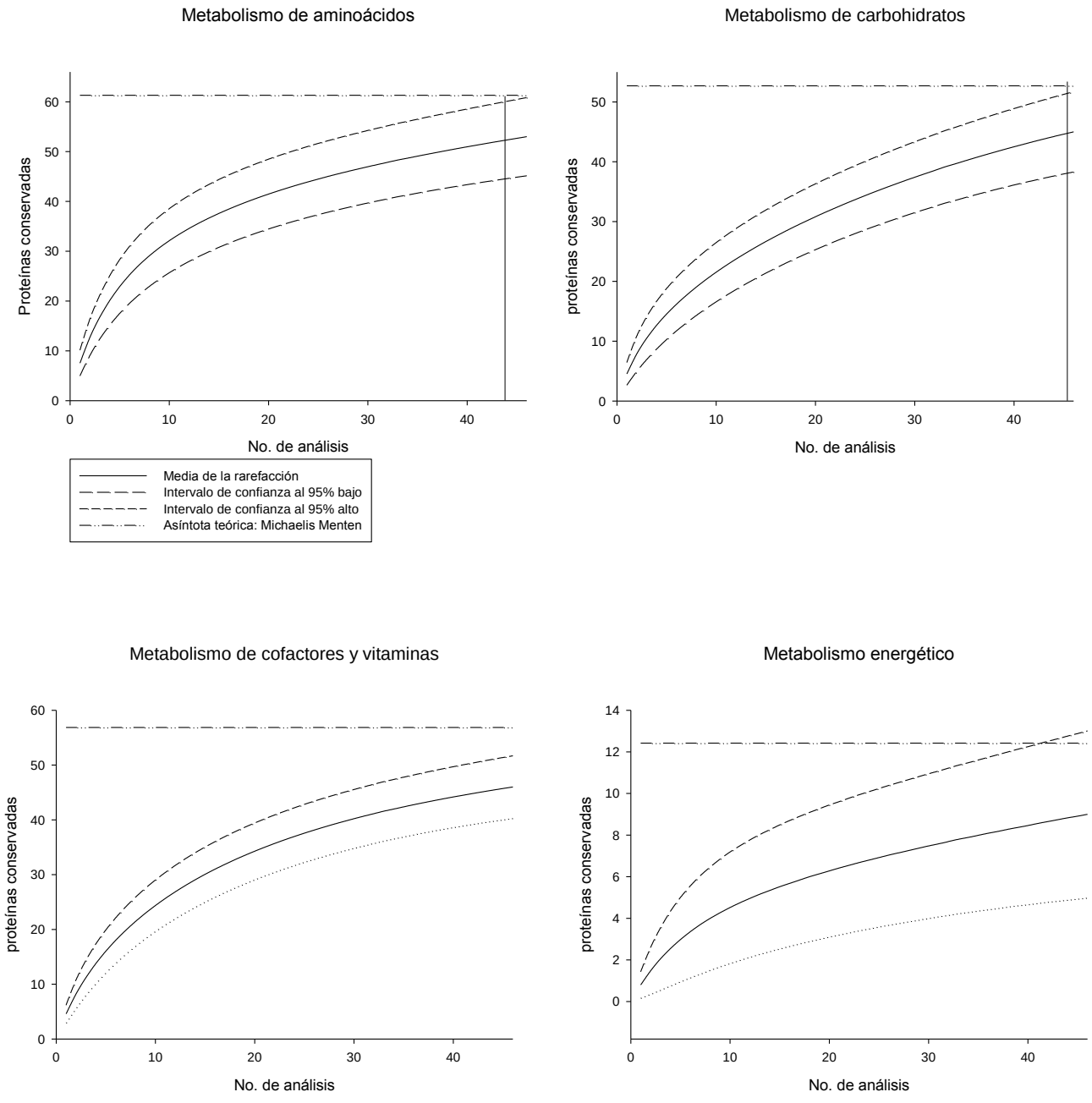


Figura 5. Curvas de rarefacción para cada súper-categoría metabólica. Continúa en la siguiente página

Curva de acumulación de proteínas conservadas.

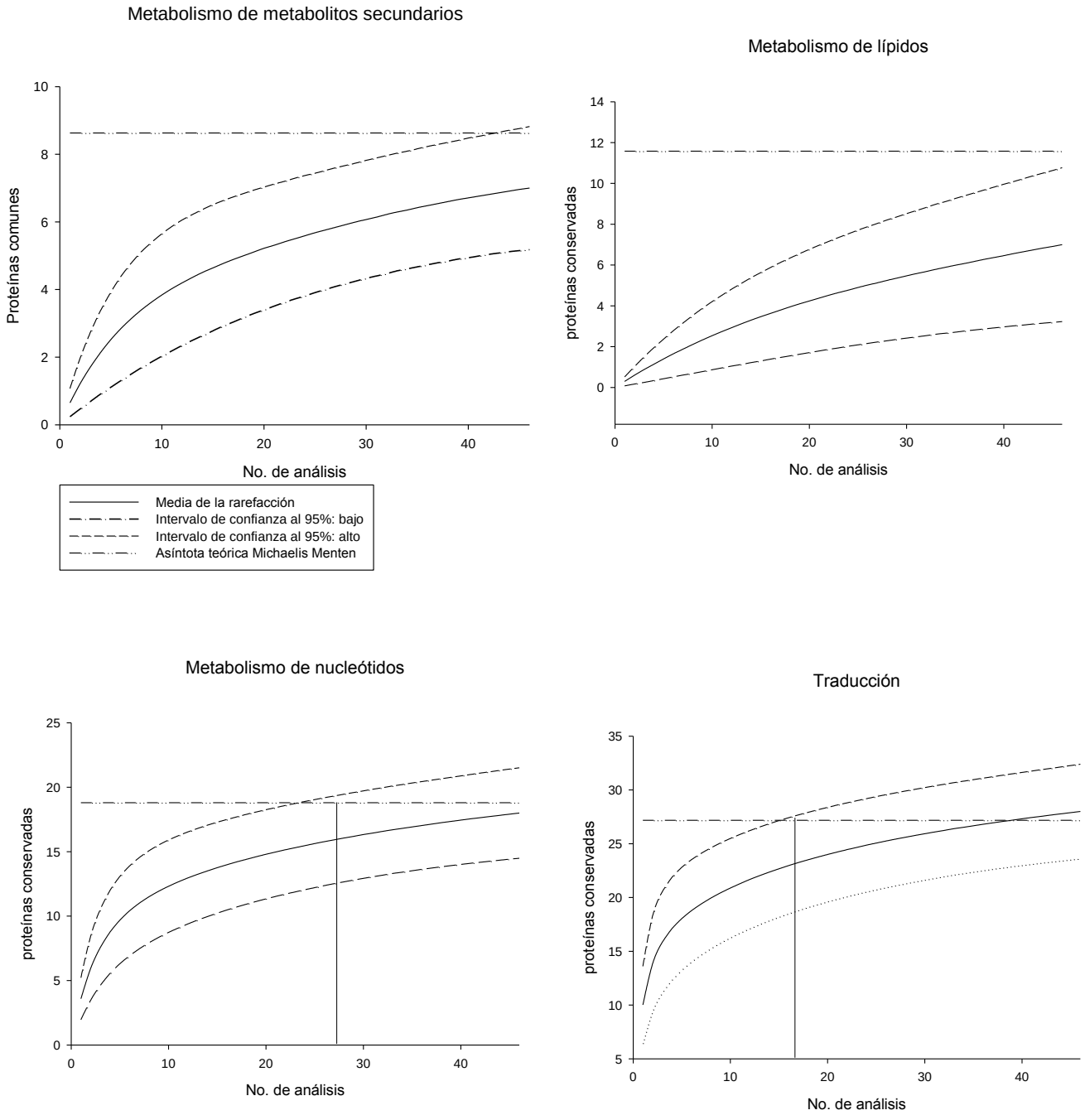


Figura 5. Curvas de rarefacción para cada súper-categoría metabólica.

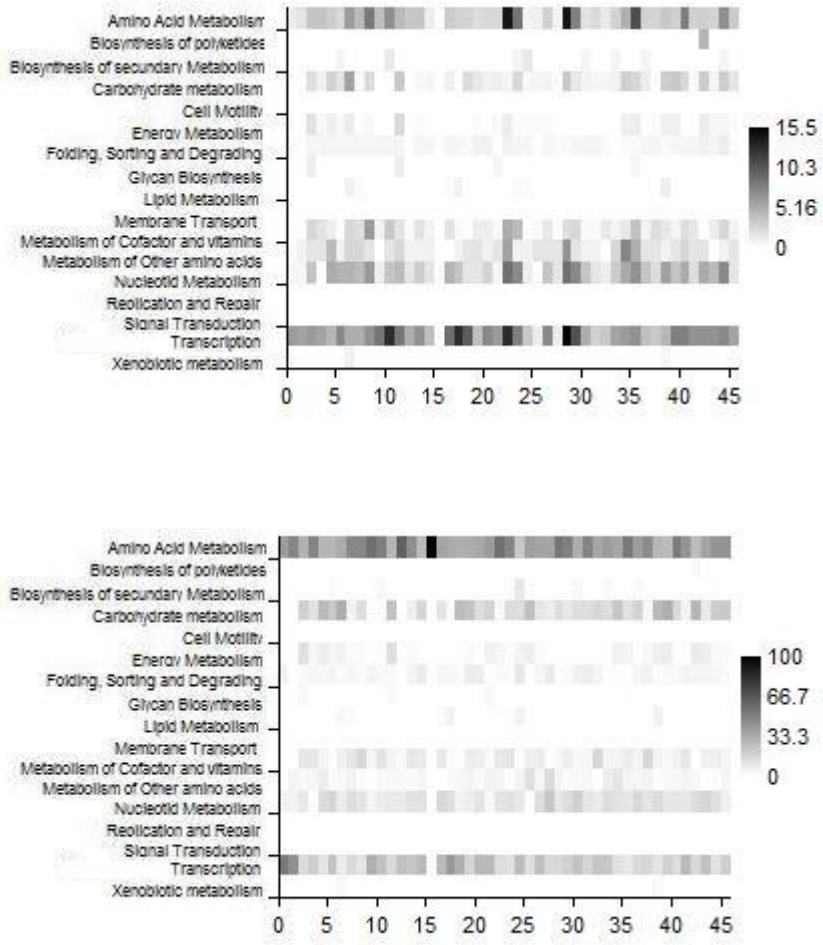


Figura 6. Gráficas de matriz de las proteínas comunes encontradas en el estudio. El panel superior muestra las proteínas tomando como el 100% todas las proteínas anotadas para cada categoría. El panel inferior muestra únicamente las proteínas encontradas en el estudio. En el eje de las x se muestra el número de análisis y cada parte del eje y muestra cada súper-categoría metabólica.

Categoría metabólica	**Porcentaje de aparición**	**Total de proteínas anotadas**
Amino Acid Metabolism	3.96	689
Biosynthesis of Polyketides	0.15	44
Biosynthesis of Secondary Metabolites	0.14	150
Carbohydrate Metabolism	1.34	720
Cell motility	0	191
Energy Metabolism	0.36	533
Folding, Sorting and Degradation	0.48	361
Glycan Biosynthesis and Metabolism	0.07	199
Lipid Metabolism	0.08	372
Membrane Transport	0.03	86
Metabolism of Cofactors and Vitamins	1.4	295
Metabolism of Other Amino Acids	1.71	123
Nucleotide Metabolism	3.41	196
Replication and Repair	0.13	172
Signal Transduction	0	689
Translation	6.37	181
Xenobiotics Biodegradation and Metabolism	0.03	370

Tabla 3. Se muestra cada categoría metabólica. El porcentaje de aparición representa que el porcentaje con respecto al total de proteínas anotadas.

Discusión.

La diversidad de organismos no está representada en las bases de datos actuales. Los intereses médicos y económicos han dominado los proyectos de secuenciación (Peregrín-Álvarez, 2007), pero no así las características propias de los organismos, especialmente en procariontes. Eso supone un importante sesgo en las bases de datos en lo que a distribución filogenética se refiere, dejando dos opciones para estudios de genómica a larga escala, tomar secuencias de organismos al azar o bien tomando en cuenta una perspectiva filogenética. Si se toma en cuenta la selección de organismos tratando de tener una cobertura máxima de todos los organismos se pueden obtener mejores resultados en diversidad de secuencias y propiedades biológicas (Wu, 2009).

La selección de los organismos fue tomada en cuenta tratando de maximizar la cobertura filogenética de árboles 16-18s RNA. Esto permite obtener el mayor número de secuencias nuevas y abate el sesgo de la falta de la representación de la biodiversidad en las bases de datos. Asimismo tomando en cuenta la taza de descubrimiento creciente con el número de genomas, a partir de 56 genomas se puede tener una importante cobertura de prácticamente todas las funciones biológicas y muchas de ellas pueden coincidir con al menos nuevas familias de proteínas (Wu, 2009). Por lo tanto, todas las funciones biológicas se pueden evaluar en el contexto de secuencias altamente conservadas y encontrar homólogos distantes. Hecho que se puede observar en los análisis de rarefacción utilizando proteínas hipotéticas.

Las curvas de rarefacción indican riqueza a través del esfuerzo de muestreo (Colwell, 2004). Si bien en estudios de genómica comparada nunca se utilizado muestreos de comparación, puede ser comparable con las curvas de rarefacción y por lo tanto se pueden inferir los mismos aspectos en Ecología. De las curvas de la figura 3 la riqueza de secuencias conservadas es de 297 en el análisis sin hipotéticas y de 361 en el análisis con hipotéticas. Es un número muy reducido comparado con la totalidad de secuencias en bases de datos y de los más de 7 millones secuencias encontradas en análisis del *Global Ocean Sampling* (Yoeseph, 2007). Tomando en cuenta esta proporción, pocas secuencias pueden ser trazadas hasta el LCA, pero las proteínas que se muestran se puede estar seguro con un importante nivel de confianza 90.16% para el análisis número 46 de su aparición en el LCA.

En las curvas de la figura 3 aquella sin incluir proteínas hipotéticas alcanzó la cobertura necesaria para considerar que se alcanzó una muestra significativa como para considerar que pese al esfuerzo de muestreo no se incluirán un número importante de proteínas conservadas. A partir de eso se puede concluir que el universo de secuencias conservadas está completo y disponible en las bases de datos actuales. Pese al fuerte sesgo en biodiversidad de organismos incluidos en las bases de datos, las proteínas conservadas ya están ahí. De manera opuesta a la creciente diversidad de organismos que se pueden incluir para abatir el sesgo en biodiversidad, aquellos nuevos organismos que lleguen a ser secuenciados deberán tener varias o todas las proteínas incluidas en el análisis.

Así como en los muestreos ecológicos, donde existen especies raras o que aparecen poco representadas en el estudio, existirán algunas proteínas que al aparecer poco representados en el análisis, habrá poca probabilidad de encontrarlas en las diferentes organismos, pero aquellas que aparecen continuamente en el análisis será más probable encontrarlas en organismos nuevos secuenciados. Gráficamente y por categoría, la figura 5 puede servir de guía de a que categorías podrán pertenecer estas proteínas y si alcanzan o no la asíntota teórica indicará a que proteína pertenecerá.

Mientras que la secuenciación se convierte en un proceso más rutinario, especialmente utilizando herramientas informáticas (Reeves, 2009). La anotación sigue presentado problemas con respecto al conocimiento bioquímico (Schnoes, 2009). Ésto es claro también en este estudio. Al incluir proteínas hipotéticas a los análisis de rarefacción, la riqueza de secuencias conservadas aumenta, pero no es posible reconocer a que función corresponden todas aquellas proteínas anotadas como indefinidas o pobremente caracterizadas. Como se señalado en otros trabajos (Becerra, 2007), uno de los mayores retos actuales es la anotación de estas proteínas poco caracterizadas y que además están bien conservadas en los tres dominios. Esto se puede reconocer debido a que al incluirlas a la matriz de presencia y ausencia y compararlas utilizando BLAST obtuve más de 50 de estas proteínas individuales carentes de una anotación formal en KEGG.

Tomando en cuenta estas proteínas hipotéticas no se puede alcanzar un número mínimo de genomas a comparar para poder alcanzar la asíntota. Lo que significaría que aun pueden encontrarse algunas proteínas que estarán bien conservadas en organismos de reciente

secuenciación. Pero todavía no se puede conocer la función bioquímica de estas. Pese a la falta de caracterización bioquímica, la representación de proteínas hipotéticas con respecto a todo el genoma de un organismo suele ser poca. En organismos como *Drosophila melanogaster* todas las proteínas hipotéticas están asociadas a un número de KO, en contraste con *Nanoarquaeum equitans* en la cual 272 de sus 546 proteínas están anotadas como hipotéticas y no están asociadas a ningún KO, de está la fracción de proteínas hipotéticas no anotadas y además conservadas es bajo. (Tabla del apéndice C.)

Lo encontrado utilizando la curva de rarefacción sin proteínas hipotéticas indica que se cuenta con un muestreo mínimo para el LCA, pese al resultado de la curva utilizando proteínas hipotéticas. La presencia de hipotéticas en los genomas seguirá siendo algo común, en especial para aquellas bases de datos que tengan anotación automatizada (Reeves, 2009), pero la aparición de proteínas hipotéticas conservadas irá en declive (figura 2). Esto permitirá alcanzar los parámetros de muestreo mínimo una vez que estas proteínas se encuentren anotadas.

Idealmente la extrapolación de la riqueza al alcanzar la asíntota significaría que se cuenta con toda la diversidad posible, hecho que en la naturaleza suele no ocurrir. Sin embargo dentro del universo de secuencias conservadas podría ser posible alcanzar la riqueza total. Esto es claro para el caso de las curvas de rarefacción de la figura 5 como el metabolismo de nucleótidos y traducción siendo estos los casos más claros al haber obtenido una riqueza total de las secuencias que están conservadas. Para estos casos las curvas de rarefacción pierden una fuente de sesgo debido a que se cuenta con la riqueza total de proteínas

conservadas, convirtiéndolas así en una técnica de análisis precisa y poco sesgada (Hughes, 2005).

Las categorías funcionales mayor representadas son metabolismo de aminoácidos y traducción. Buena parte de las proteínas relacionadas con metabolismo de aminoácidos están también relacionadas con traducción. Ejemplo de estos son las aminoacil t-RNA sintetazas, en las cuales encontré: treonil-tRNA sintetasa,fenilalanil-tRNA sintetasa cadenas alfa y beta, alanil-tRNA sintetasa, seril-tRNA sintetasa, arginil-tRNA sintetasa, valil-tRNA sintetasa, cisteinil-tRNA sintetasa, histidil-tRNA sintetasa, isoleucil-tRNA sintetasa, metionil-tRNA sintetasa, aspartil-tRNA sintetasa, glycil-tRNA sintetasa clase II, leucil-tRNA sintetasa, glutamil-tRNA sintetasa, prolil-tRNA sintetasa. Estas representan al menos 15 de los 20 aminoácidos esenciales, hecho que es similar a otras reconstrucciones (Ouzounis, 2006; Ranea, 2006).

El metabolismo energético alcanza los parámetros necesarios para muestreo mínimo. Esto señala que existe una gran variedad de proteínas conservadas. En particular aquellas que están relacionadas con el amonio y que a su vez están ligadas a síntesis de aminoácidos como el ácido glutámico están bien representadas. Entre estas proteínas se encuentran la glutamina sintetasa, glutamato sintetasa (NADPH) cadenas grande y chica, glutamato deshidrogenasa (NAD(P)+). Aun cuando se alcanza el muestreo mínimo hay una alta cantidad de proteínas anotadas para este tipo de metabolismo. Esto refleja la complejidad de mecanismos para obtener y manejar la energía. Hecho que es consistente con la irregularidad de las filogenias de las enzimas involucradas en este metabolismo

(Castresana, 2001). Así mismo, no es posible concluir cual fue la fuente de energía del LCA, ni como éste se relacionaba con su medio abiótico.

Pese a que obtengo resultados similares al de reconstrucciones anteriores, existe una discrepancia mayor. El metabolismo de aminoácidos debería de estar pobremente representado (Koonin, 2003; Harris 2003; Delaye 2005). Sin embargo en Mirkin 2003, al aumentar la penalidad de ganancia (*g*) se puede asignar un número mayor de genes al LCA, si se asume que el mecanismo más importante para la adquisición de genes del LCA es la pérdida secundaria. Bajo la metodología realizada en este trabajo la pérdida secundaria en principio puede ser un problema, debido a que en lo organismos incluidos existen parásitos obligados (Becerra, 1997). Pero al ser una metodología de uso creciente de bases de datos, se pueden ir incluyendo proteínas que no son tomadas en cuenta en tríos en que alguno de ellos sea parásito obligado. Entonces el metabolismo de aminoácidos puede estar representado por su distribución en diferentes grupos debido al diseño de este estudio.

La biosíntesis de aminoácidos es un proceso ampliamente distribuido en los tres dominios y algunas enzimas que participan en este metabolismo pueden ser trazadas al LCA. Entre aquellos que encontré están parte de la síntesis de arginina procedente de la vía de la ornitina (*ornithine carbamoyltransferase, ornithine decarboxylase*). La cadena alfa y beta de la triptofano sintetaza aparecen también en el análisis. Una enzima altamente distribuida, *anthranilate phosphoribosyltransferase*, esta conservada y ampliamente distribuida y participa no solo en la síntesis de triptofano si no también de histidina que se encuentra en el presente estudio. Así mismo están presentes en este estudio partes del metabolismo del

glutamato, aquellas que lo metabolizan: *glutamate dehydrogenase (NAD(P)+), glutamate 5-kinase* y *glutamate-5-semialdehyde dehydrogenase.* Estos resultados son consistentes con reportes previos de enzimas ampliamente distribuidas (Hernández-Montes, 2008).

Otro proceso metabólico ampliamente distribuido y conservado es el metabolismo de nucleótidos. Una parte del metabolismo de pirimidinas presenta pasos sucesivos conservados casi en su totalidad. Las siguientes reacciones indican los reactantes y el producto y escrito entre paréntesis la enzima y el número KO. La reacción de L-glutamina a carbamoil fosfato (K01956, *carbamoyl-phosphate synthase small chain*) está conservada; la reacción de carbamoil fosfato a N-carbamoil L-aspartato (K00609, *aspartate carbamoyltransferase catalytic subunit*) está conservada. El paso de N-carbamoil L-aspartato a dihidroorato no está presente en este estudio debido a que es una enzima representada únicamente en eucariontes. Pero la reacción de dihidroorotato a orotato (K00226, *dihydroorotate oxidase*) si está conservada. La reacción reversible de orotato a orotidina 5 fosfato con la liberación de PRPP (K00762, *orotate phosphoribosyltransferase*) está conservada. La oritidina 5 fosfato es precursor para uridin monofosfato UMP base nitrogenada del RNA que junto con otros procesos conservados relacionados al RNA dan soporte al mundo del RNA (Delaye, 2005). Mientras que el PRPP (*5-Phospho-alpha-D-ribose 1-diphosphate*) es intermediario central en metabolismo de nucleótidos.

No todos los tipos de metabolismo están completamente representados como en los anteriores mencionados sino que algunos no cumplen con los requisitos de muestreo mínimo. El resumen de estos procesos metabólicos están en la figura 4. El metabolismo de

lípidos además de estar pobremente representado en este estudio no alcanza los parámetros de muestreo mínimo. Esto se debe a que, a pesar que los lípidos son fundamentales para todos los seres vivos, existen diferentes formas de metabolizarlos. Por ejemplo la síntesis de isoprenoides es realizada por dos rutas independientes no homólogas: la ruta del mevalonato en *Eucarya* y *Archaea* y la ruta del metileritritol fosfato en *Bacteria* (Lombard, 2010). Incluso la estructura de la membrana varía entre dominios. Las *Archaeas* tienen una membrana formada por fosfolípidos compuestos por grupos isoprenilos ligados por grupos eter al glicerol, mientras que *Bacteria* y *Eucarya* utilizan ácidos grasos ligados al glicerol utilizando enlaces ester. No solo en estructura difieren los lípidos sino también en las enzimas que los sintetizan (Calo, 2010).

El metabolismo de metabolitos secundarios es otro proceso que está pobremente representado en el estudio debido a que dentro de la clasificación funcional de la base de datos KEGG presenta características contemporáneas como manejo antibióticos. Por lo tanto las enzimas conservadas que encontré corresponden a algunas enzimas aisladas que tienen características compartidas a otros metabolismos, por ejemplo la *acetyl-CoA C-acetyltransferase* que participa a su vez en metabolismo de lípidos, carbohidratos y aminoácidos, en adición de metabolitos secundarios.

El hecho que que no todas las rutas estén conservadas no implica necesariamente que no pudieron estar presentes en el LCA. Por ejemplo ninguna célula contemporánea se puede concebir sin la presencia de membrana plasmática, por lo tanto es fácil suponer que el LCA

tuvo una membrana. Sin embargo, con esta metodología solo se pueden apreciar aquellas rutas metabólicas que estén conservadas.

Se puede considerar que ésta es una reconstrucción menos estricta debido a que no existen correcciones por dichas pérdidas secundarias (Ouzounis, 2006) y a su vez pueden existir genes anotados en el Apéndice B que pudieran ser resultado de transporte horizontal. Pero si se toman en cuenta como el número mínimo de secuencias conservadas posiblemente encontradas en el LCA pueda considerarse como un escenario más realista. Así mismo contesta al sesgo de patrones filéticos (Mirkin, 2003). No se requiere de un gran número de genomas completos a comparar para obtener una reconstrucción estadísticamente confiable pero no es así para cada categoría metabólica.

Conclusiones

La metodología utilizada permitió realizar un listado de proteínas conservadas que pueden ser trazadas hasta el LCA. Por lo tanto es posible obtener un listado de proteínas del LCA con el contenido de información de las bases de datos actuales. Las curvas de rarefacción mostraron que esta lista es representativa y estadísticamente confiable.

Así mismo se pudo observar que, pese a la falta de representatividad de la biodiversidad en las bases de datos, es posible tener una reconstrucción utilizando un número relativamente limitado de genomas secuenciados en su totalidad.

Esta reconstrucción incluye muchas rutas conservadas como transcripción y metabolismo de nucleótidos, que a través de la curva de rarefacción se puede suponer que son de las rutas de las cuales se conocen prácticamente todas las proteínas conservadas.

Perspectivas

Este estudio puede ser continuado en alguna de las siguientes vertientes: profundizando en esta línea. Por ejemplo comprobar si durante el procesamiento de datos no fue eliminada alguna proteína con 100% de identidad, como pudieran ser las de resultado de duplicación.
Es posible que la continuación del estudio sea en realizar controles con diferentes técnicas de selección de organismos. Realizando esto con la finalidad de darle solidez al estudio.

Literatura citada

Becerra, A., Islas S., Leguina J. I., Silva E., Lazcano A. 1997. *Polyphyletic gene losses can bias backtrack characterization of the cenancestor*. J. Mol. Evo. 45: 115-118

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*. Ann. Rev. Ecol. Evol. Syst. 38: 261-279.

Calo D., Eichler J. 2010. *Crossing the membrane in Archaea, the third domain of life*. Biochimica e Biophysica Acta. En prensa

Castresana J. 2001. *Comparative genomics and bionergetics*. Biochemica et Biophysica Acta. 1506: 147-162

Cicarelli F., Doerks T., von Mering C., Creevey C., Snel B., Bork P. 2006. *Toward automatic reconstruction of a highly resolved tree of life.* Science. 311: 1283-1287.

Colwell, R. K., Mao., C. X., Chang J. 2004. *Interpolatin, extrapolating and comparing incidence-based speces accumulation curves*. Ecology. 85:2717-2727

Di Gulio, M. 2007. *The tree of life might be rooted in the branch leading to Nanoarchaeota*. Gene. 401: 108-113.

Forterre P., Phillippe H. 1999. *Where is the root of the universal tree of life?*. Bioessays. 21: 871-879.

Gotelli N. J., Collwell R. 2001. *Quantifying biodiversity: procedures and pitfalls in the measurements and comparison of species richness*. Ecology Letters. 4: 379-391.

Harris J. K., Kelly S. T., Spigelman G. B., Pace N. R. 2003. *The genetic core of the universal ancestor*. Genome Res. 13: 407-412.

Hernández-Montes G., Díaz-Mejía J., Pérez-Rueda E., Segovia L. 2008. The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. Genome Biology. 9: R95.

Hughes J., Hellman J. 2005. *The application of rarefaction techniques to molecular inventories of microbial diversity*. Methoods in Enzymology. 397: 292-308.

Huynen M. A., Bork P. 1998. *Mesuring genome evolution*. PNAS. 95: 5849-5856.

Iwabe N., Kuma K., Hasewaga M., Osawa S., Miyata T. 1989. *Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes*. PNAS. 86: 9355-9359.

Joyce G. F. 2002. *The antiquity of RNA-based evolution*. Nature. 418: 214:221.

Koonin E. V. 2003. *Comparative genomics, minimal gene-set and the last common ancestor*. Nature Reviews. Microbiology. 1: 127, 136.

Kurland C. G., Canbak B., Berg O. G. 2007. *The origin of modern proteomes*. Biochimie. 89: 1454-1463.

Lagesen K., Ussery D., Wassenaar T. 2010. G*enome update: the 1000th genome – a cautionary tale*. Microbiology. 156: 603-608.

Mao C., Colwell R., Chang J. 2005. *Estimating the species accumulation curves using mixtures*. Biometrics. 61: 433-441.

Mirkin B. G., Fenner T. I., Galperin G., Koonin E. V. 2003. *Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes*. BMC Evol Biol. 6;3: 2.

Lombard J., Moreira D. 2010. *Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life*. Molecular Biology and Evolution. Jul 22. En prensa.

Mushegian A R., Koonin E. V . 1996. *A minimal gene set for cellular life derivied by comparision of complete bacterial genomes*. PNAS. 93: 10268-10273.

Ouzounis A. C., Kunin V., Darzentas N., Goldovsky L. 2006. *A minimal stimate for gene content of the last common ancestor – exobiology from a extraterrestrial perspective*. Res. Microbiol. 157: 57-68.

Peregrín-Álvarez. J., Parkinson J. 2007. *The global landscape of sequence diversity. Genome Biology*. 8: R238.

Peregrín-Álvarez. J., Sandford C., Parkinson J. 2009. *The conservation and evolutionary modularity of metabolism*. Genome Biology 10: R63.

Qin J., Li R., Raes J., Arumugan M., *et al*. 2010. *A human gut microbial gene catalogue established by metagenomics sequencing*. Nature. 464: 59-67.

Ranea A. G., Sillero A., Thorton M. G., Orengon A. C. 2006. *Protein superfamily evolution and the last universal common ancestor (LUCA)*. J. Mol. Evol. 63:513-525.

Reeves G. A., Talavera D., Thornton J. M. 2009. *Genome and proteome annotation: organization, interpretation and integration*. J. R. Soc. Interfase. 6: 129-147.

Schnoes A. M., Brown S. D., Dodevsky I., Babbit P. C. 2009. *Annotation error in public databases: Misannotation in molecular function in enzyme superfamilies*. PloS Computational Biology. 5: e1000605.

Schopf J. W. 1993. *Microfossils of early Archaean Apex Chert: new evidence of antiquity life*. Science. 30, 260: 640-646.

Schopf J. W., 2006. *Fossils evidence to Archaean life*. Phil. Trans. R. Soc. B. 361: 869-885.

Shaw A., Halpern A., Beeson K., Tran B., Venter J., Martiny J. 2008. *It's all relative: ranking the diversity of aquatic bacterial communities*. Enviromental microbiology. 10: 2200-2210.

Sobolevsky Y., Trifonov E. N. 2006. *Protein modules conserved since LUCA*. J. Mol. Evol. 63: 622-634.

Tamura K, Dudley J, Nei M., Kumar S. 2007. *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0*. Molecular Biology and Evolution. 24:1596-1599.

Valas R. E., Bourne P. E. 2009. *Structural analysis of polarazing indels: an emergency consensus on the root of the tree of life*. Biol. Direct. 25: 4-30.

Wilde S. A., Valley J. W., Graham C.M. 2001. *Evidence from detrital zircons for the existence of continental crustand oceans on the Earth 4.4 Gyr ago*. Nature 409: 175-178.

Woese C. R., Kandler O., Wheelis M. L. 1990. *Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya*. PNAS. 877: 4756-4759.

Yang S., DoolitleR. F., Bourne P. E. 2005. *Phylogeny determined by protein domain content*. PNAS. 102: 373-378.

Yarza, P., Richter, M., Peplies J., Euzeby, J., Amann R., Schleifer, K., Ludwig W., Glöckner, F. O., Rossello-Mora, R. 2008. *The All-Species Living Tree Project: a 16S rRNA-based phylogenetic tree of all sequenced type strains.* System. Appl. Microbiol. 31: 241-250.

Yoseph S., Sutton G., Rush D. B., Halpern A. L., et al. 2007. T*he Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families*. PloS Biology. 5: e6.

Apéndice A. Grupos utilizados en el presente estudio

Grupo	**Bacteria**	**Archaea**	**Eucarya**
Core 1	*Thermotoga petrophila*	*Nanoarcheaum equitans*	*Encephalitozoon cuniculi*
Core 2	*Thermus thermophilus HB27*	*Thermofilum pendens*	*Trichomonas vaginalis*
Core 3	*Deinococcus radiudurans*	*Aeropyrum pernix*	*Dictyostelium discoideum*
Core 4	*Roseiflexus castenholzii*	*Hyperthermus butylicus*	*Rattus norvergicus*
Core 5	*Rubrobacter xylanophilus*	*Sulfolobus solfatacarius*	*Bombyx mori*
Core 6	*Propionobacter acnes*	*Thermoplasma volcanicum*	*Theileria parva*
Core 7	*Nocardioides spjs614*	*Picrophilus torridus*	*Nematostella vectensis*
Core 8	*Corynebacterium diphteriae*	*Thermococcus kodakaraensis*	*Sacharomices cerevisae*
Core 9	*Mycobacterium vanbaalenii*	*Methanococcus vannielii*	*Ostreococcus tauri*
Core 10	*Bifidobacterium longum ncc2705*	*Archaeoglobus fulgidus*	*Psychomitrella patens*
Core 11	*Artrobacter aurescens*	*Methanobacteria thermoautotrophicus*	*Oryza sativa*
Core 12	*Streptomyces griseus*	*Methanosarcina mazei*	*Plasmodium yoelii*
Core 13	*Thermoanaerobacter tengcongensis*	*Methanocuellus marisnigri*	*Mus musculus*
Core 14	*Clostridium acetobutylicum*	*Methanocorpusculum labreanum*	*Appis mellifera*
Core 15	*Mycoplasma pneumoniae*	*Natronomonas pharaonis*	*Crypstosporidium*

			parvum
Core 16	*Staphylococcus haemolyticus*	*Pycobaculum calidifontis*	*Kluveromyces waltii*
Core 17	*Lactobacillus sakei*	*Ignococcus hospitalis*	*Ostreococcus lucimarinus*
Core 18	*Enterococcus faecalis*	*Sulfolobus acidocaldarius*	*Arabidopsis thailiana*
Core 19	*Streptococcus mutans*	*Thermoplasma acidophilum*	*Entamoeba histolytica*
Core 20	*Prochlorococcus marinus 9515*	*Thermococcus onnurineus*	*Equus caballus*
Core 21	*Fusobacterium nucleatum*	*Methanococcus aeolicus*	*Drosophila pseudoscura*
Core 22	*Leptospira biflexa (ames)*	*Methanobrevibacter smithii*	*Crypstosporidium hominis*
Core 23	*Rodhopirellula baltica*	*Methanosarcina acetivorans*	*Candida glabrata*
Core 24	*Akkermansia muciniphila*	*Holoarcula marismortui*	*Cyanidioschyzon merolae*
Core 25	*Chloroherpeton thalassium*	*Pyrobaculum arsenaticum*	*Trypanosoma cruzi*
Core 26	*Cytophaga hutchinsonii*	*Staphylothermos marinus*	*Sus scrofa*
Core 27	*Porphyromonas gingivalis W83*	*Sulfolobus tokodaii*	*Drosophila melanogaster*
Core 28	*Flavobacterium psychrophilum*	*Pyrococcus horikoshii*	*Paramecium tetraurelia*
Core 29	*Geobacter sulfurreducens*	*Methanococcus jannaschii*	*Candida albicans*
Core 30	*Myxococcus xanthus*	*Methanospharae stadmanae*	*Chlamydomonas reindhardtii*

Core 31	*Helicobacter hepaticus*	*Methanosarcina barkeri*	*Trypanosoma brucei*
Core 32	*Rhodospirillum rubrum*	*Methanopyrus kandleri*	*Ornitorhynchus anatius*
Core 33	*Orientia tsutsugamushi Ikeda*	*Halobacterium NCR1*	*Anopheles gambiae*
Core 34	*Paracoccus denitrificans*	*Pyrobaculum islandicum*	*Tetrahymenia thermophila*
Core 35	*Rhizobium etli CFN 42*	*Methalosphaera sedula*	*Monosiga brevicollis*
Core 36	*Brucella suis 1330*	*Mehtanococcoides burtonii*	*Yarrowia lipolitica*
Core 37	*Methylobacterium radiotolerans*	*Methanosaeta thermophila*	*Leishmania mayor*
Core 38	*Xanthomonas oryzae KACC10331*	*Candidatus Methanosphaerela palustris*	*Monodeplhis domestica*
Core 39	*Acinetobacter baumannii ATCC 17978*	*Holoquadratum walbsyi*	*Aedes aegypti*
Core 40	*Chromobacterium violaceum*	*Pyrobaculum aerophilum*	*Theileria anmulata*
Core 41	*Burkholderia thailandensis*	*Methanococcus maripaludis s2*	*Nuerospora crassa*
Core 42	*Pseudomonas mendocina*	*Methanospirillum hungatei*	*Gallus galllus*
Core 43	*Alteromonas macleodii*	*Methanococcus maripaludis c5*	*Caenorhabditis elegans*
Core 44	*Shewanella denitrificans*	*Candidatus Methanoregula boonei*	*Plasmodium falsiparum 3d7*
Core 45	*Photobacterium profundum*	*Methanococcus maripaludis c6*	*Aspergillus fumigatus*
Core 46	*Serratia proteamaculans*	*Candidatus methanogenic archaeon RC-1*	*Xenopus laevis*

En la tabla se muestran todos los organismos utilizados en el presente estudio. Se obtuvieron las secuencias de proteínas de ftp del KEGG (ver Metodología). Todos los archivos *.pep fueron descargado de la base de datos en agosto de 2008, fecha cuando inicié el estudio. Cada trío de organismos *core* está integrado por un organismo de cada dominio celular.

Apéndice B. Lista de proteínas obtenidas que pudieron estar en el LCA.

No. de K	**Nombre de la proteína**	**Proceso metabólico donde participa**
K01868	*threonyl-tRNA synthetase*	*Translation*
K01889	*phenylalanyl-tRNA synthetase alpha chain*	*Translation*
K01872	*alanyl-tRNA synthetase*	*Translation*
K01875	*seryl-tRNA synthetase*	*Translation*
K01887	*arginyl-tRNA synthetase*	*Translation*
K01873	*valyl-tRNA synthetase*	*Amino Acid Metabolism, Translation*
K01883	*cysteinyl-tRNA synthetase*	*Translation*
K01892	*histidyl-tRNA synthetase*	*Translation*
K01870	*isoleucyl-tRNA synthetase*	*Amino Acid Metabolism, Translation*
K01874	*methionyl-tRNA synthetase*	*Metabolism of Other Amino Acids, Translation*
K00384	*thioredoxin reductase (NADPH)*	*Nucleotide Metabolism*
K01409	*O-sialoglycoprotein endopeptidase*	*Metabolism; Enzyme Families; Peptidases*
K01876	*aspartyl-tRNA synthetase*	*Translation*
K01736	*chorismate synthase*	*Amino Acid Metabolism*
K01880	*glycyl-tRNA synthetase, class II*	*Translation*
K01687	*dihydroxy-acid dehydratase*	*Amino Acid Metabolism, Metabolism of Cofactors and Vitamins*
K00773	*queuine tRNA-ribosyltransferase*	*Translation*
K06173	*tRNA pseudouridine synthase A*	*Translation*
K03306	*inorganic phosphate transporter, PiT family*	*Cellular Processes and Signaling*
K00930	*acetylglutamate kinase*	*Amino Acid Metabolism*
K00766	*anthranilate phosphoribosyltransferase*	*Amino Acid Metabolism*
K01693	*imidazoleglycerol-phosphate dehydratase*	*Amino Acid Metabolism*
K02434	*aspartyl-tRNA(Asn)/glutamyl-tRNA (Gln) amidotransferase subunit B*	*Translation*
K01940	*argininosuccinate synthase*	*Amino Acid Metabolism*
K03168	*DNA topoisomerase I*	*Replication and Repair*
K03106	*signal recognition particle, subunit SRP54*	*Folding, Sorting and Degradation*
K00620	*amino-acid N-acetyltransferase*	*Amino Acid Metabolism*
K00642	*glutamate N-acetyltransferase*	*Amino Acid Metabolism*
K02469	*DNA gyrase subunit A*	*Replication and Repair*
K01952	*phosphoribosylformylglycinamidine synthase*	*Nucleotide Metabolism*
K01695	*tryptophan synthase alpha chain*	*Amino Acid Metabolism*
K01951	*GMP synthase (glutamine-hydrolysing)*	*Nucleotide Metabolism*

K00767	*nicotinate-nucleotide pyrophosphorylase (carboxylating)*	*Metabolism of Cofactors and Vitamins*
K01755	*argininosuccinate lyase*	*Amino Acid Metabolism*
K01869	*leucyl-tRNA synthetase*	*Amino Acid Metabolism, Translation*
K03177	*tRNA pseudouridine synthase B*	*Translation*
K03110	*signal recognition particle receptor*	*Folding, Sorting and Degradation*
K01885	*glutamyl-tRNA synthetase*	*Metabolism of Cofactors and Vitamins, Translation*
K02994	*small subunit ribosomal protein S8*	*Translation*
K02874	*large subunit ribosomal protein L14*	*Translation*
K02965	*small subunit ribosomal protein S19*	*Translation*
K02886	*large subunit ribosomal protein L2*	*Translation*
K00088	*IMP dehydrogenase*	*Nucleotide Metabolism*
K00013	*histidinol dehydrogenase*	*Amino Acid Metabolism*
K00806	*undecaprenyl pyrophosphate synthetase*	*Biosynthesis of Secondary Metabolites*
K00764	*amidophosphoribosyltransferase*	*Nucleotide Metabolism*
K01698	*porphobilinogen synthase*	*Metabolism of Cofactors and Vitamins*
K00800	*3-phosphoshikimate 1-carboxyvinyltransferase*	*Amino Acid Metabolism*
K00215	*dihydrodipicolinate reductase*	*Amino Acid Metabolism*
K01956	*carbamoyl-phosphate synthase small chain*	*Nucleotide Metabolism*
K02492	*glutamyl-tRNA reductase*	*Metabolism of Cofactors and Vitamins*
K01586	*diaminopimelate decarboxylase*	*Amino Acid Metabolism*
K02500	*cyclase HisF*	*Amino Acid Metabolism*
K00609	*aspartate carbamoyltransferase catalytic chain*	*Nucleotide Metabolism*
K01652	*acetolactate synthase large subunit*	*Amino Acid Metabolism, Carbohydrate Metabolism, Metabolism of Cofactors and Vitamins*
K00765	*ATP phosphoribosyltransferase*	*Amino Acid Metabolism*
K00817	*histidinol-phosphate aminotransferase*	*Amino Acid Metabolism, Biosynthesis of Secondary Metabolites*
K00382	*dihydrolipoamide dehydrogenase*	*Carbohydrate Metabolism, Amino Acid Metabolism*
K08681	*glutamine amidotransferase*	*Metabolism of Cofactors and Vitamins*
K01749	*hydroxymethylbilane synthase*	*Metabolism of Cofactors and Vitamins*
K02946	*small subunit ribosomal protein S10*	*Translation*
K01756	*adenylosuccinate lyase*	*Nucleotide Metabolism*
K09013	*Fe-S cluster assembly ATP-binding protein*	*Membrane Transport*
K02952	*small subunit ribosomal protein S13*	*Translation*
K03637	*molybdenum cofactor biosynthesis protein C*	*Metabolism of cofactors and vitamins*
K02433	*aspartyl-tRNA(Asn)/glutamyl-tRNA (Gln)*	*Translation*

	amidotransferase subunit A	
K00761	*uracil phosphoribosyltransferase*	*Nucleotide Metabolism*
K01711	*GDPmannose 4,6-dehydratase*	*Carbohydrate Metabolism*
K01696	*tryptophan synthase beta chain*	*Amino Acid Metabolism*
K07566	*putative translation factor*	*Translation*
K01845	*glutamate-1-semialdehyde 2,1-aminomutase*	*Metabolism of Cofactors and Vitamins*
K01778	*diaminopimelate epimerase*	*Amino Acid Metabolism*
K01937	*CTP synthase*	*Nucleotide Metabolism*
K00820	*glucosamine--fructose-6-phosphate aminotransferase (isomerizing)*	*Carbohydrate Metabolism*
K00286	*pyrroline-5-carboxylate reductase*	*Amino Acid Metabolism*
K01915	*glutamine synthetase*	*Amino Acid Metabolism, Energy Metabolism*
K01681	*aconitate hydratase 1*	*Carbohydrate Metabolism, Energy Metabolism*
K00939	*adenylate kinase*	*Nucleotide Metabolism*
K02470	*DNA gyrase subunit B*	*Replication and Repair*
K06215	*pyridoxine biosynthesis protein*	*Metabolism of Cofactors and Vitamins*
K00940	*nucleoside-diphosphate kinase*	*Nucleotide Metabolism*
K03686	*molecular chaperone DnaJ*	*Folding, Sorting and Degradation*
K01265	*methionyl aminopeptidase*	*Metabolism; Enzyme Families; Peptidases*
K01881	*prolyl-tRNA synthetase*	*Translation*
K00927	*phosphoglycerate kinase*	*Carbohydrate Metabolism*
K02528	*dimethyladenosine transferase*	*Translation*
K01939	*adenylosuccinate synthase*	*Nucleotide Metabolism*
K01953	*asparagine synthase (glutamine-hydrolysing)*	*Energy Metabolism*
K00265	*glutamate synthase (NADPH) large chain*	*Energy Metabolism*
K01890	*phenylalanyl-tRNA synthetase beta chain*	*Translation*
K01480	*agmatinase*	*Amino Acid Metabolism*
K00058	*D-3-phosphoglycerate dehydrogenase*	*Amino Acid Metabolism*
K00758	*thymidine phosphorylase*	*Nucleotide Metabolism*
K00858	*NAD+ kinase*	*Metabolism of Cofactors and Vitamins*
K00611	*ornithine carbamoyltransferase*	*Amino Acid Metabolism*
K04487	*cysteine desulfurase*	*Metabolism of Cofactors and Vitamins*
K01807	*ribose 5-phosphate isomerase A*	*Carbohydrate Metabolism*
K00602	*phosphoribosylaminoimidazolecarboxamide formyltransferase*	*Nucleotide Metabolism, Metabolism of Cofactors and Vitamins*
K01689	*enolase*	*Carbohydrate Metabolism*
K00873	*pyruvate kinase*	*Carbohydrate Metabolism, Nucleotide Metabolism*

K00560	*Thymidylate synthase*	*Nucleotide Metabolism, Metabolism of Cofactors and Vitamins*
K00052	*3-isopropylmalate dehydrogenase*	*Amino Acid Metabolism*
K01495	*GTP cyclohydrolase I*	*Metabolism of Cofactors and Vitamins*
K02933	*large subunit ribosomal protein L6*	*Translation*
K02988	*small subunit ribosomal protein S5*	*Translation*
K01955	*carbamoyl-phosphate synthase large chain*	*Nucleotide Metabolism*
K03665	*GTP-binding protein HflX*	*Unclassified*
K01791	*UDP-N-acetylglucosamine 2-epimerase*	*Carbohydrate Metabolism*
K01945	*phosphoribosylamine--glycine ligase*	*Nucleotide Metabolism*
K01867	*tryptophanyl-tRNA synthetase*	*Amino Acid Metabolism, Translation*
K01735	*3-dehydroquinate synthase*	*Amino Acid Metabolism*
K00784	*ribonuclease Z*	*Translation*
K00059	*3-oxoacyl-[acyl-carrier protein] reductase*	*Lipid Metabolism*
K02867	*large subunit ribosomal protein L11*	*Translation*
K00941	*phosphomethylpyrimidine kinase*	*Metabolism of Cofactors and Vitamins*
K02906	*large subunit ribosomal protein L3*	*Translation*
K00852	*ribokinase*	*Carbohydrate Metabolism*
K04567	*lysyl-tRNA synthetase, class II*	*Translation*
K02931	*large subunit ribosomal protein L5*	*Translation*
K02986	*small subunit ribosomal protein S4*	*Translation*
K02992	*small subunit ribosomal protein S7*	*Translation*
K03750	*molybdopterin biosynthesis protein MoeA*	*Metabolism of cofactors and vitamins*
K01840	*phosphomannomutase*	*Carbohydrate Metabolism*
K01783	*ribulose-phosphate 3-epimerase*	*Carbohydrate Metabolism*
K01784	*UDP-glucose 4-epimerase*	*Carbohydrate Metabolism*
K01823	*isopentenyl-diphosphate delta-isomerase*	*Biosynthesis of Secondary Metabolites*
K01893	*asparaginyl-tRNA synthetase*	*Translation*
K00809	*deoxyhypusine synthase*	*Translation*
K00943	*dTMP kinase*	*Nucleotide Metabolism*
K09129	*hypothetical protein*	*Unclassified*
K03593	*ATP-binding protein involved in chromosome partitioning*	*Replication and Repair*
K01710	*dTDP-glucose 4,6-dehydratase*	*Biosynthesis of Secondary Metabolites, Biosynthesis of Polyketides*
K00283	*glycine dehydrogenase subunit 2*	*Amino Acid Metabolism*
K01598	*phosphopantothenoylcysteine decarboxylase*	*Metabolism of Cofactors and Vitamins*
K00797	*spermidine synthase*	*Amino Acid Metabolism, Metabolism of Other Amino Acids*
K00605	*aminomethyltransferase*	*Amino Acid Metabolism, Metabolism of*

		Cofactors and Vitamins, Energy Metabolism
K00721	*dolichol-phosphate mannosyltransferase*	*Glycan Biosynthesis and Metabolism*
K01810	*glucose-6-phosphate isomerase*	*Carbohydrate Metabolism*
K00012	*UDPglucose 6-dehydrogenase*	*Carbohydrate Metabolism*
K01530	*phospholipid-translocating ATPase*	*Energy Metabolism*
K00226	*dihydroorotate oxidase*	*Nucleotide Metabolism*
K00796	*dihydropteroate synthase*	*Metabolism of Cofactors and Vitamins*
K02858	*3,4-dihydroxy 2-butanone 4-phosphate synthase*	*Metabolism of Cofactors and Vitamins*
K01079	*phosphoserine phosphatase*	*Amino Acid Metabolism*
K03386	*peroxiredoxin (alkyl hydroperoxide reductase subunit C)*	*Folding, Sorting and Degradation*
K01581	*ornithine decarboxylase*	*Amino Acid Metabolism, Metabolism of Other Amino Acids*
K01657	*anthranilate synthase component I*	*Amino Acid Metabolism*
K02967	*small subunit ribosomal protein S2*	*Translation*
K01679	*fumarate hydratase*	*Carbohydrate Metabolism, Energy Metabolism*
K01658	*anthranilate synthase component II*	*Amino Acid Metabolism*
K00928	*aspartate kinase*	*Amino Acid Metabolism*
K03751	*molybdopterin biosynthesis protein MoeB*	*Metabolism of cofactors and vitamins*
K00133	*aspartate-semialdehyde dehydrogenase*	*Amino Acid Metabolism*
K02501	*amidotransferase HisH*	*Amino Acid Metabolism*
K07304	*peptide-methionine (S)-S-oxide reductase*	*Folding, Sorting and Degradation*
K03687	*molecular chaperone GrpE*	*Folding, Sorting and Degradation*
K01469	*5-oxoprolinase (ATP-hydrolysing)*	*Metabolism of Other Amino Acids*
K00641	*homoserine O-acetyltransferase*	*Energy Metabolism*
K01738	*cysteine synthase*	*Metabolism of Other Amino Acids, Energy Metabolism*
K00788	*thiamine-phosphate pyrophosphorylase*	*Metabolism of Cofactors and Vitamins*
K02303	*precorrin-2 dehydrogenase*	*Metabolism of Cofactors and Vitamins*
K00145	*N-acetyl-gamma-glutamyl-phosphate reductase*	*Amino Acid Metabolism*
K00053	*ketol-acid reductoisomerase*	*Amino Acid Metabolism, Metabolism of Cofactors and Vitamins*
K01653	*acetolactate synthase small subunit*	*Amino Acid Metabolism, Carbohydrate Metabolism, Metabolism of Cofactors and Vitamins*
K03564	*bacterioferritin comigratory protein*	*Folding, Sorting and Degradation*
K00147	*glutamate-5-semialdehyde dehydrogenase*	*Amino Acid Metabolism*
K00014	*shikimate 5-dehydrogenase*	*Amino Acid Metabolism*
K01714	*dihydrodipicolinate synthase*	*Amino Acid Metabolism*
K03147	*thiamine biosynthesis protein ThiC*	*Metabolism of Cofactors and Vitamins*

K00795	*geranyltranstransferase*	*Biosynthesis of Secondary Metabolites*
K01834	*phosphoglycerate mutase*	*Carbohydrate Metabolism*
K00606	*3-methyl-2-oxobutanoate hydroxymethyltransferase*	*Metabolism of Cofactors and Vitamins*
K01012	*biotin synthetase*	*Metabolism of Cofactors and Vitamins*
K01649	*2-isopropylmalate synthase*	*Amino Acid Metabolism, Carbohydrate Metabolism*
K00793	*riboflavin synthase alpha chain*	*Metabolism of Cofactors and Vitamins*
K01496	*phosphoribosyl-AMP cyclohydrolase*	*Amino Acid Metabolism*
K04518	*prephenate dehydratase*	*Amino Acid Metabolism*
K01938	*formate--tetrahydrofolate ligase*	*Carbohydrate Metabolism, Metabolism of Cofactors and Vitamins*
K01533	*Cu2+-exporting ATPase*	*Energy metabolism*
K00282	*glycine dehydrogenase subunit 1*	*Amino Acid Metabolism*
K00573	*protein-L-isoaspartate(D-aspartate) O-methyltransferase*	*Folding, Sorting and Degradation*
K06287	*septum formation protein*	*Unclassified*
K06969	*putative SAM-dependent methyltransferase*	*Unclassified*
K01959	*pyruvate carboxylase subunit A*	*Carbohydrate Metabolism*
K00833	*adenosylmethionine-8-amino-7-oxononanoate aminotransferase*	*Metabolism of Cofactors and Vitamins*
K01489	*cytidine deaminase*	*Nucleotide Metabolism*
K00958	*sulfate adenylyltransferase*	*Nucleotide Metabolism, Metabolism of Other Amino Acids, Energy Metabolism*
K00261	*glutamate dehydrogenase (NAD(P)+)*	*Amino Acid Metabolism, Metabolism of Other Amino Acids, Energy Metabolism*
K01754	*threonine dehydratase*	*Amino Acid Metabolism*
K07588	*LAO/AO transport system kinase*	*Amino acid metabolism*
K02437	*glycine cleavage system H protein*	*Amino acid metabolism*
K00333	*NADH dehydrogenase I chain D*	*Energy Metabolism*
K07305	*peptide-methionine (R)-S-oxide reductase*	*Folding, Sorting and Degradation*
K00074	*3-hydroxybutyryl-CoA dehydrogenase*	*Xenobiotics Biodegradation and Metabolism, Carbohydrate Metabolism*
K03787	*5'-nucleotidase*	*Nucleotide Metabolism, Metabolism of Cofactors and Vitamins*
K07560	*D-tyrosyl-tRNA(Tyr) deacylase*	*Translation*
K00762	*orotate phosphoribosyltransferase*	*Nucleotide Metabolism*
K02377	*GDP-L-fucose synthase*	*Carbohydrate Metabolism*
K00288	*methylenetetrahydrofolate dehydrogenase (NADP+)*	*Carbohydrate Metabolism, Metabolism of Cofactors and Vitamins*
K01551	*arsenite-transporting ATPase*	*Unclassified*
K03644	*lipoic acid synthetase*	*Metabolism of Cofactors and Vitamins*
K03521	*electron transfer flavoprotein beta subunit*	*Energy metabolism*

K00020	*3-hydroxyisobutyrate dehydrogenase*	*Amino Acid Metabolism*
K01638	*malate synthase*	*Carbohydrate Metabolism*
K01647	*citrate synthase*	*Carbohydrate Metabolism*
K00162	*pyruvate dehydrogenase E1 component, beta subunit*	*Carbohydrate Metabolism, Amino Acid Metabolism*
K03522	*electron transfer flavoprotein alpha subunit*	*Energy metabolism*
K00390	*phosphoadenosine phosphosulfate reductase*	*Energy Metabolism*
K00003	*homoserine dehydrogenase*	*Amino Acid Metabolism*
K03781	*catalase*	*Amino Acid Metabolism, Energy Metabolism*
K00266	*glutamate synthase (NADPH) small chain*	*Energy Metabolism*
K00931	*glutamate 5-kinase*	*Amino Acid Metabolism*
K00763	*nicotinate phosphoribosyltransferase*	*Metabolism of Cofactors and Vitamins*
K01092	*myo-inositol-1(or 4)-monophosphatase*	*Biosynthesis of Secondary Metabolites, Carbohydrate Metabolism, Signal Transduction*
K01465	*dihydroorotase*	*Nucleotide Metabolism*
K02427	*cell division protein methyltransferase FtsJ*	*Translation*
K01599	*uroporphyrinogen decarboxylase*	*Metabolism of Cofactors and Vitamins*
K03695	*ATP-dependent Clp protease ATP-binding subunit ClpB*	*Folding, Sorting and Degradation*
K00022	*3-hydroxyacyl-CoA dehydrogenase*	*Lipid Metabolism, Amino Acid Metabolism, Carbohydrate Metabolism, Xenobiotics Biodegradation and Metabolism*
K00681	*gamma-glutamyltranspeptidase*	*Metabolism of Other Amino Acids, Lipid Metabolism*
K03841	*fructose-1,6-bisphosphatase I*	*Carbohydrate Metabolism*
K01966	*propionyl-CoA carboxylase beta chain*	*Amino Acid Metabolism, Carbohydrate Metabolism*
K00966	*mannose-1-phosphate guanylyltransferase*	*Carbohydrate Metabolism*
K01256	*membrane alanyl aminopeptidase*	*Metabolism of Other Amino Acids*
K00838	*aromatic amino acid aminotransferase I*	*Amino Acid Metabolism*
K00759	*adenine phosphoribosyltransferase*	*Nucleotide Metabolism*
K03458	*nucleobase:cation symporter-2, NCS2 family*	*Unclassified*
K01338	*ATP-dependent Lon protease*	*Metabolism; Enzyme Families; Peptidases*
K00026	*malate dehydrogenase*	*Carbohydrate Metabolism, Energy Metabolism*
K03685	*ribonuclease III*	*Translation*
K01007	*pyruvate,water dikinase*	*Carbohydrate Metabolism, Energy Metabolism*
K01790	*dTDP-4-dehydrorhamnose 3,5-epimerase*	*Biosynthesis of Secondary Metabolites, Biosynthesis of Polyketides*
K04488	*nitrogen fixation protein NifU and related proteins*	*Energy Metabolism*

K01703	*3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit*	*Amino Acid Metabolism*
K00616	*transaldolase*	*Carbohydrate Metabolism*
K01803	*triosephosphate isomerase (TIM)*	*Carbohydrate Metabolism*
K00872	*homoserine kinase*	*Amino Acid Metabolism*
K07010	*putative glutamine amidotransferase*	*Metabolism; Enzyme Families; Peptidases*
K00789	*S-adenosylmethionine synthetase*	*Metabolism of Other Amino Acids*
K00033	*6-phosphogluconate dehydrogenase*	*Carbohydrate Metabolism*
K01262	*X-Pro aminopeptidase*	*Metabolism; Enzyme Families; Peptidases*
K00140	*methylmalonate-semialdehyde dehydrogenase*	*Amino Acid Metabolism, Carbohydrate Metabolism*
K00864	*glycerol kinase*	*Lipid Metabolism*
K00626	*acetyl-CoA C-acetyltransferase*	*Lipid Metabolism, Amino Acid Metabolism, Carbohydrate Metabolism, Xenobiotics Biodegradation and Metabolism, Biosynthesis of Secondary Metabolites*
K00161	*pyruvate dehydrogenase E1 component, alpha subunit*	*Carbohydrate Metabolism*
K01745	*histidine ammonia-lyase*	*Amino Acid Metabolism, Energy Metabolism*
K01814	*phosphoribosylformimino-5-aminoimidazole carboxamide ribotide*	*Amino Acid Metabolism*
K01443	*N-acetylglucosamine-6-phosphate deacetylase*	*Carbohydrate Metabolism*
K00700	*1,4-alpha-glucan branching enzyme*	*Carbohydrate Metabolism*
K01712	*urocanate hydratase*	*Amino Acid Metabolism*
K01468	*imidazolonepropionase*	*Amino Acid Metabolism*
K01609	*indole-3-glycerol phosphate synthase*	*Amino Acid Metabolism*
K00794	*riboflavin synthase beta chain*	*Metabolism of Cofactors and Vitamins*
K01903	*succinyl-CoA synthetase beta chain*	*Carbohydrate Metabolism, Energy Metabolism*
K00652	*8-amino-7-oxononanoate synthase*	*Metabolism of Cofactors and Vitamins*
K00826	*branched-chain amino acid aminotransferase*	*Amino Acid Metabolism, Metabolism of Cofactors and Vitamins*
K06180	*ribosomal large subunit pseudouridine synthase D*	*Translation*
K03575	*A/G-specific adenine glycosylase*	*Replication and Repair*
K00772	*5'-methylthioadenosine phosphorylase*	*Amino Acid Metabolism*
K04043	*molecular chaperone DnaK*	*Folding, Sorting and Degradation*
K01142	*exodeoxyribonuclease III*	*Replication and Repair*
K03660	*N-glycosylase/DNA lyase*	*Replication and Repair*
K01740	*O-acetylhomoserine (thiol)-lyase*	*Amino Acid Metabolism*
K03703	*excinuclease ABC subunit C*	*Replication and Repair*
K02291	*phytoene synthase*	*Biosynthesis of Polyketides*

K02551	*2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase*	*Metabolism of Cofactors and Vitamins*
K03186	*3-octaprenyl-4-hydroxybenzoate carboxy-lyase UbiX*	*Metabolism of Cofactors and Vitamins*
K03657	*DNA helicase II / ATP-dependent DNA helicase PcrA*	*Replication and Repair*
K04077	*chaperonin GroEL*	*Folding, Sorting and Degradation*
K03702	*excinuclease ABC subunit B*	*Replication and Repair*

La tabla contiene todas las proteínas encontradas en el estudio. Al estar presentes en alguno de los análisis correspondientes se puede concluir que es probable que estuvieran en el LCA. La primera columna contiene los números K, ortólogos del KEGG, la segunda el nombre de la proteína tomando como referencia a *E. coli* y la tercera el proceso metabólico donde participan.

Apéndice C. Porcentaje de proteínas hipotéticas conservadas y totales para cada genoma utilizado.

Nombre del organismo	**Porcentaje de hipotéticas conservadas**	**Porcentaje de hipotéticas por genoma**
Thermotoga petrophila	0.33	16.86
Nanoarcheaum equitans	1.84	50.75
Encephalitozoon cuniculi	0.22	46.04
Thermus thermophilus HB27	0.16	28.73
Thermofilum pendens	0.32	33.26
Trichomonas vaginalis	0.01	80.82
Deinococcus radiudurans	0.24	39.2
Aeropyrum pernix	0.45	38.82
Dictyostelium discoideum	0.3	64.42
Roseiflexus castenholzii	0.08	28.34
Hyperthermus butylicus	0.74	42.13
Rattus norvergicus	0	9.95
Rubrobacter xylanophilus	0.18	17.8
Sulfolobus solfatacarius	0.75	35.81
Propionobacter acnes	0	25.82
Thermoplasma volcanicum	0	20.35
Theileria parva	0.17	72.2
Nocardioides spjs614	0.29	28.23
Picrophilus torridus	0.62	21.17
Nematostella vectensis	1.02	18.68
Corynebacterium diphteriae	0.87	35.43

Thermococcus kodakaraensis	0.24	35.99
Sacharomices cerevisae	0	0
Mycobacterium vanbaalenii	0.06	28.62
Methanococcus vannielii	0.24	24.61
Ostreococcus tauri	0	0.01
Bifidobacterium longum ncc2705	1.75	33.14
Archaeoglobus fulgidus	0.23	35.99
Psychomitrella patens	0.19	90.91
Artrobacter aurescens	0.09	24
Methanobacteria thermoautotrophicus	0.44	36.41
Oryza sativa	0	88.85
Streptomyces griseus	0.12	36.52
Methanosarcina mazei	0.3	39.23
Plasmodium yoelii	0.19	76.73
Thermoanaerobacter tengcongensis	0.27	28.98
Methanocuellus marisnigri	0.52	30.9
Mus musculus	0	6.57
Clostridium acetobutylicum	0.19	27.65
Methanocorpusculum labreanum	3.58	48.25
Appis mellifera	0.33	6.76
Mycoplasma pneumoniae	0	0.44
Natronomonas pharaonis	0	35.05
Crypstosporidium parvum	0	43.21
Staphylococcus haemolyticus	0	45.48

Pycobaculum calidifontis	0	35.55
Kluveromyces waltii	0	0
Lactobacillus sakei	0	19.58
Ignococcus hospitalis	0	35.56
Ostreococcus lucimarinus	0	0.05
Enterococcus faecalis	0	35.25
Sulfolobus acidocaldarius	0.53	42.33
Arabidopsis thailiana	0	26.42
Streptococcus mutans	0.36	28.06
Thermoplasma acidophilum	0.82	32.86
Entamoeba histolytica	0.04	55.07
Prochlorococcus marinus 9515	0.45	35.05
Thermococcus onnurineus	2.88	26.37
Equus caballus	0.81	9.11
Fusobacterium nucleatum	0	25.69
Methanococcus aeolicus	0.29	23.09
Drosophila pseudoscura	0	0
Leptospira biflexa (ames)	0	32.25
Methanobrevibacter smithii	0	25.32
Crypstosporidium hominis	0.13	59.12
Rodhopirellula baltica	0.08	52.57
Methanosarcina acetivorans	0.28	46.52
Candida glabrata	1.08	57.21
Akkermansia muciniphila	0	29.19

Holoarcula marismortui	0.16	44.03
Cyanidioschyzon merolae	0.35	22.66
Chloroherpeton thalassium	0	22.99
Pyrobaculum arsenaticum	0.24	36.71
Trypanosoma cruzi	0.01	56.19
Cytophaga hutchinsonii	0.06	42.32
Staphylothermos marinus	0.2	32.36
Sus scrofa	0	4.09
Porphyromonas gingivalis W83	0.31	34.15
Sulfolobus tokodaii	0.57	55.65
Drosophila melanogaster	0	0
Flavobacterium psychrophilum	0.34	36.77
Pyrococcus horikoshii	0.45	45.93
Paramecium tetraurelia	0.07	95.08
Geobacter sulfurreducens	0.32	27.55
Methanococcus jannaschii	0.31	36.06
Candida albicans	0	23.62
Myxococcus xanthus	0.04	38.64
Methanospharae stadmanae	1.83	32.14
Chlamydomonas reindhardtii	0.32	67.04
Helicobacter hepaticus	0.81	46.08
Methanosarcina barkeri	0.19	43.29
Trypanosoma brucei	0.12	64.89
Rhodospirillum rubrum	0.12	21.74

Methanopyrus kandleri	0	31
Ornitorhynchus anatius	0.62	21.55
Orientia tsutsugamushi Ikeda	0	18.02
Halobacterium NCR1	0.17	45.12
Anopheles gambiae	0	0
Paracoccus denitrificans	0	20.29
Pyrobaculum islandicum	0.29	35.14
Tetrahymenia thermophila	0.02	69.73
Rhizobium etli CFN 42	0	0.34
Methalosphaera sedula	0.57	31.03
Monosiga brevicollis	1.49	87.81
Brucella suis 1330	0.1	30.9
Mehtanococcoides burtonii	0.44	29.96
Yarrowia lipolitica	0	0
Methylobacterium radiotolerans	0.05	30.99
Methanosaeta thermophila	0.24	24.82
Leishmania mayor	0.13	63.19
Xanthomonas oryzae KACC10331	0.28	26.14
Candidatus Methanosphaerela palustris	0	25.8
Monodeplhis domestica	0.88	20.83
Acinetobacter baumannii ATCC 17978	0.45	19.63
Holoquadratum walbsyi	0	33.86
Aedes aegypti	0.05	49.76
Chromobacterium violaceum	0.07	33.79

Pyrobaculum aerophilum	0.33	45.87
Theileria anmulata	0.24	53.99
Burkholderia thailandensis	0.08	22.12
Methanococcus maripaludis s2	0.45	26.07
Nuerospora crassa	0.24	71.07
Pseudomonas mendocina	0.32	20.72
Methanospirillum hungatei	0.29	33.2
Gallus galllus	0.1	10.71
Alteromonas macleodii	0	24.02
Methanococcus maripaludis c5	0	27.33
Caenorhabditis elegans	0.19	55.25
Shewanella denitrificans	0.11	24.83
Candidatus Methanoregula boonei	0	31.71
Plasmodium falsiparum 3d7	0.12	64.02
Photobacterium profundum	0.15	35.63
Methanococcus maripaludis c6	0.36	30.67
Aspergillus fumigatus	0	27.23
Serratia proteamaculans	0.12	16.29
Candidatus methanogenic archaeon RC-1	0.08	42.82
Xenopus laevis	2.48	23.09

La tabla contiene la información correspondiente a las proteínas hipotéticas del presente estudio. La primera columna contiene el nombre del organismo con que se trabajo. La segunda el porcentaje de proteínas hipotéticas conservadas y que no contienen ninguna anotación formal, con respecto al total de las proteínas hipotéticas en cada genoma. La

tercera lista contiene el porcentaje de hipotéticas en cada genoma, tomando como total todas las proteínas del genoma. Para realizar los cálculos de esta tabla se tomaron en cuenta aquellas proteínas que correspondieran a a la anotación *hypothetical protein* y además que no estuvieran asociadas a ningún número de ortólogo KO o correspondieran a alguna proteína putativa, que a su vez esto le aseguraría tener un número de KO asignado.