



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

---

# POSGRADO EN CIENCIAS BIOLÓGICAS

Facultad de Ciencias

Estudio de la Duplicación de Genes en Secuencias  
Ancestrales: Análisis de su Relevancia en el Mundo  
de RNA/proteínas

# TESIS

QUE PARA OBTENER EL GRADO ACADÉMICO DE

**MAESTRO EN CIENCIAS BIOLÓGICAS**  
(Biología Experimental)

P R E S E N T A

RICARDO HERNÁNDEZ MORALES

Director de Tesis: Dr. Antonio Lazcano-Araujo Reyes  
Comité tutor: Dr. Diego González Halphen  
Dr. Víctor Manuel Valdés López

MÉXICO, D.F.

Junio, 2011



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

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

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



UNIVERSIDAD NACIONAL  
AUTÓNOMA DE MÉXICO

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

OFICIO FCIE/DEP/295/11

ASUNTO: Oficio de Jurado

**Dr. Isidro Ávila Martínez**  
Director General de Administración Escolar, UNAM  
Presente

Me permito informar a usted que en la reunión ordinaria del Comité Académico del Posgrado en Ciencias Biológicas, celebrada el día **14 de marzo de 2011** se aprobó el siguiente jurado para el examen de grado de **MAESTRO EN CIENCIAS BIOLÓGICAS (BIOLOGÍA EXPERIMENTAL)** del (la) alumno (a) **HERNÁNDEZ MORALES RICARDO** con número de cuenta **95097358** con la tesis titulada "**Estudio de la Duplicación de Genes en Secuencias Ancestrales: Análisis de su Relevancia en el Mundo de RNA/Proteínas**", realizada bajo la dirección del (la) **DR. ANTONIO EUSEBIO LAZCANO-ARAUJO REYES**:

Presidente: DR. ARTURO CARLOS II BECERRA BRACHO  
Vocal: DR. LEÓN PATRICIO MARTÍNEZ CASTILLA  
Secretario: DR. VÍCTOR MANUEL VALDÉS LÓPEZ  
Suplente: DR. ENRIQUE MERINO PÉREZ  
Suplente: DR. DIEGO GONZÁLEZ HALPHEN

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

**Atentamente**  
"POR MI RAZA HABLARA EL ESPÍRITU"  
Cd. Universitaria, D.F., a 24 de mayo de 2011

Dra. María del Coro Arizmendi Arriaga  
Coordinadora del Programa



DIVISIÓN DE ESTUDIOS  
DE POSGRADO

MCAA/MJFM/ASR/ipp

### **Agradecimientos especiales:**

Al Posgrado en Ciencias Biológicas de la UNAM por abrirme sus puertas y permitirme formar parte de la vida académica y científica

Al Consejo Nacional de Ciencia y Tecnología por el apoyo económico otorgado para realizar este trabajo (número de becario 175710)

Al apoyo recibido por los proyectos 50520-Q y 100199 de CONACYT, México

A los miembros del comité tutorial: Dr. Antonio Lazcano-Araujo Reyes, Dr. Diego González Halpeh y Dr. Víctor Manuel Valdés López por sus enseñanzas, consejos y comentarios para la realización de este trabajo y para mi formación como científico

A los miembros del jurado: Dr. Arturo Carlos II Becerra Bracho, Dr. León Patricio Martínez Castilla, Dr. Víctor Manuel Valdés López, Dr. Enrique Merino Pérez y Dr. Diego González Halphen por sus importantes observaciones y aportaciones durante el escrito de la tesis

**A mis padres, hermanos y amigos**

We have formerly seen that parts many times repeated are eminently liable to vary in number and structure; consequently it is quite probable that natural selection, during the long-continued course of modification, should have seized on a certain number of the primordially similar elements, many times repeated, and have adapted them to the most diverse purposes.

**Darwin C., 1859**

## Indice

<b>I Resumen</b>	10
<b>II Abstract</b>	12
<b>III Introducción</b>	13
3.1 Secuencias moleculares como documentos históricos	14
3.2 Establecimiento de una filogenia universal	15
3.3 El último ancestro común (LCA – Last Common Ancestor)	17
3.4 Genómica comparativa y caracterización del último ancestro común	18
3.5 Naturaleza genómica del último ancestro común	21
3.6 Evolución de los genomas	27
3.7 Duplicación interna de genes	29
<b>IV Material y Métodos</b>	31
4.1 Análisis de bases de datos de secuencias altamente conservadas	32
4.2 Búsqueda de proteínas originada por duplicación interna	32
4.3 Distribución funcional de genes originados por un evento de duplicación interna	33
4.4 Secuencias altamente conservadas que interactúan con el metabolismo del RNA y que se originaron por duplicación interna	34
4.5 Visualización de estructura primaria y estructura terciaria de proteínas altamente conservadas con duplicación interna	35
4.6 Antigüedad relativa de secuencias altamente conservadas	36
<b>V Resultados</b>	37
5.1 Proteínas altamente conservadas que se originaron por duplicación interna en las diferentes bases de datos	38
5.2 Secuencias altamente conservadas que están relacionadas con el metabolismo del RNA y que se originaron por duplicación interna	66
5.3 Evidencia de duplicación interna en estructura terciaria de proteínas altamente conservadas	70
5.4 Temporalidad relativa de secuencias altamente conservadas	78
5.5 Pérdida del rastro de la duplicación interna en secuencias conservadas	85
<b>VI Discusión</b>	94
6.1 Secuencias altamente conservadas originadas por duplicación y fusión de genes parálogos en diferentes bases de datos	95
6.2 Diversidad funcional de proteínas conservadas originadas por duplicación interna	97
6.3 Proteínas conservadas que interactúan con el RNA y que se originaron por duplicación interna	98
6.4 Diversidad funcional de proteínas altamente conservadas que están relacionadas con el RNA y que se originaron por duplicación interna	99

6.5 La adquisición de dominios adicionales o la duplicación parcial de un gen podría enmascarar la evidencia tridimensional de las regiones homólogas internas	99
6.6 Antigüedad relativa de secuencias altamente conservadas	100
6.7 Duplicación interna de genes como mecanismo que promueve la complejidad estructural de las proteínas	102
6.8 Conservación de la huella de la duplicación	104
<b>VII Conclusiones</b>	107
<b>VIII Referencias</b>	110



# **I RESUMEN**

Un conjunto de secuencias altamente conservadas, algunas de las cuales están involucradas en el metabolismo del RNA, fueron analizadas para valorar el papel que tuvieron los eventos de duplicación interna de genes durante la evolución celular temprana. Nuestros resultados muestran que algunas secuencias antiguas encontradas en los tres principales linajes de organismos celulares fueron el resultado de eventos de duplicación seguidos por la fusión de los genes duplicados. Entre ellas se incluyen aquellas secuencias relacionadas a procesos biológicos principales como la transcripción, traducción, regulación y biosíntesis y degradación de ribonucleótidos, derivados de ribonucleótidos, y poli-ribonucleótidos, indicando que varios genes contemporáneos fueron originados a partir de un gen ancestral pequeño por el incremento de su tamaño y complejidad estructural vía duplicación interna durante etapas tempranas de la vida. Nuestros resultados proporcionan evidencia directa de que las secuencias altamente conservadas no son igualmente antiguas, sino que surgieron en diferente tiempo geológico, y apoyan la idea de la duplicación interna de genes es un mecanismo muy antiguo el cual podría haber estado actuando en estadios tempranos de la evolución de las proteínas (ej. en un mundo de RNA/proteínas).

# **II ABSTRACT**

A set of highly conserved sequences, some of them involved in RNA metabolism, has been analyzed in order to assess the role of internal gene duplication events that may have taken place during early cell evolution. Our results show that some ancient sequences found in all three major cell lineages are the outcome of duplications followed by fusion events. The sequences which we have found are the outcome of internal duplication events include those related to major biological processes including transcription, translation, regulation and biosynthesis and degradation of ribonucleotides, ribonucleotide-derivatives, and polyribonucleotides. These results indicate that several contemporary genes were originated from a small ancestral gene by increase their size and structural complexity via internal duplication during early Archaean epochs. Our results provide direct evidence that highly conserved sequences are not equally ancient but arose at different geological time, and support the idea that internal gene duplication is an ancient mechanism which could have been acting in early stages of evolution of proteins (e.g. in RNA/protein world).

# **III INTRODUCCIÓN**

### 3.1 Secuencias moleculares como documentos históricos

Los caracteres moleculares son una de las principales herramientas que nos sirven para conocer las relaciones evolutivas entre diferentes seres vivos. El estudio de las relaciones evolutivas entre los organismos, utilizando caracteres moleculares, tuvo sus inicios a principios del siglo XX con el uso de la bioquímica comparada. En el año de 1901 el fisiólogo inglés George H. F. Nuttall comenzó a comparar reacciones inmunológicas entre sueros sanguíneos de distintas especies de animales como un intento para descubrir las relaciones de parentesco entre ellos. Los resultados fueron publicados en su libro “The blood immunity and blood relationship” escrito en 1904 (Nuttall, 1904). Las conclusiones principales de su trabajo establecen que las especies más cercanamente relacionadas presentan una reacción inmune cruzada más fuerte entre suero y antisuero.

La esencia de los trabajos de Nuttall establece el principio más importante en evolución molecular, el cual indica que el grado de similitud entre genes o sus productos refleja el grado de las relaciones evolutivas entre ellos. Sin embargo, la premisa de que la información almacenada en las secuencias podía ser utilizada como registro histórico a partir del cual se pueden establecer inferencias en torno al pasado biológico de los organismos se dedujo años más tarde.

El campo de la evolución molecular comenzó a desarrollarse hacia finales de los años 60s y principios de los 70s con la disponibilidad de secuencias de genes y proteínas. En 1965, Emile Zuckerkandl y Linus Pauling propusieron que las secuencias, ya sea de nucleótidos o de aminoácidos, podían ser utilizadas para estudiar las relaciones evolutivas entre los organismos, ya que en ellas está almacenada la información que guarda el registro de su pasado histórico (Zuckerkandl y Pauling, 1965).

La propuesta de Zuckerkandl y Pauling estuvo basada en los primeros trabajos evolutivos sobre la especificidad de las reacciones serológicas que llevó a cabo el patólogo y biólogo austriaco Karl Landsteiner en 1936. En el segundo capítulo del libro titulado “The specificity of serological reactions”, Landsteiner reconoció que el uso de las diferencias químicas puede utilizarse para medir diferencias entre especies, idea que originalmente había sido propuesta por Reichert y Brown. A principios del siglo XX, Reichert y Brown encontraron que las formas y ángulos de los cristales de las

hemoglobinas eran característicos de cada especie, y que las diferencias entre estas moléculas se mantenía de acuerdo a la distancia de las relaciones evolutivas entre las especies. Con estos antecedentes, Zuckerkandl y Pauling realizaron los primeros trabajos sobre las diferencias entre las secuencias de la hemoglobina perteneciente a diversos grupos de organismos (Morgan, 1998), y reconocieron que mientras más cercanamente relacionados estuvieran los grupos, el número de diferencias entre las secuencias sería cada vez menor. De esta manera, pudieron establecer en términos cuantitativos las relaciones evolutivas de las hemoglobinas, sugiriendo que estas proteínas provienen de un ancestro común, el cual a partir de un evento de duplicación dio surgimiento a las cadenas que la componen (Zuckerkandl *et al.*, 1960).

Con las ideas de Zuckerkandl y Pauling se pudieron conjuntar los campos de la paleontología, la biología evolutiva y la biología molecular, y se inició una nueva área de estudio conocida como “evolución molecular”, la cual nos permite conocer tanto las relaciones evolutivas de organismos completamente diferentes, como su pasado (Jorde *et al.*, 2001), historia y estilo de vida (Fraser *et al.*, 1997), utilizando exclusivamente información o datos moleculares.

### **3.2 Establecimiento de una filogenia universal**

Debido a que las secuencias de genes y proteínas forman parte del registro histórico de un organismo a partir del cual se pueden hacer extrapolaciones a su historia de vida y a su pasado remoto, la disponibilidad de un número cada vez mayor de genomas celulares completamente secuenciados ha ofrecido una gran oportunidad para estudiar diferentes problemas con un enfoque evolutivo en biología molecular (Tekaia *et al.*, 1999), y ha permitido desarrollar bases de datos que han tenido un gran impacto en la filogenia molecular (Pagel, 2000; Kanehisa y Bork, 2003; Wolfe y Li, 2003).

El análisis comparativo de secuencias moleculares ha llegado a ser una aproximación poderosa para determinar las relaciones evolutivas de distintas especies (Fitch, 1987), y para generar ideas importantes de los estadios evolutivos que pudieron haber existido antes de la separación de los tres principales linajes celulares.

En 1977 Carl R. Woese y George E. Fox realizaron lo que probablemente sea uno de los descubrimientos más interesantes que se hayan hecho en los últimos cincuenta años en torno a la diversidad de la vida en la Tierra. Encontraron que la vida en la Tierra se divide en tres grandes grupos con características perfectamente definidas a nivel molecular. Dichos agrupamientos son conocidos hoy como los dominios Archaea, Bacteria y Eucarya (Woese *et al.*, 1990) (Figura 1). Su conclusión se derivó a partir de la comparación de secuencias de varios RNAs de la subunidad pequeña del ribosoma, pertenecientes a diferentes organismos procariontes y eucariontes. El análisis filogenético basado en secuencias de RNA ribosomal (rRNA) reveló que todos los seres vivos contemporáneos pueden ser agrupados en uno de los tres grandes linajes celulares: uno de ellos, el dominio Bacteria, contiene a todas las bacterias típicas que habían sido caracterizadas experimentalmente; el segundo grupo, el dominio Eucarya, está definido por los rRNAs 18S de los animales, plantas, hongos y protistas; mientras que el tercer grupo, el dominio Archaea, estaba caracterizado por un pequeño grupo de bacterias metanógenas anaeróbicas que se habían encontrado hasta ese momento, y que poseían un metabolismo único basado en la reducción del dióxido de carbono a metano. Asimismo, el resultado de sus análisis comparativos abría la posibilidad de una clasificación natural universal monofilética (basada exclusivamente en secuencias moleculares), idea que había sido concebida por Charles Darwin en 1859, en la que manifestaba la descendencia, a partir de un mismo ancestro, de todos los seres vivos que han existido en la Tierra (Darwin, 1859). (Figura 2).

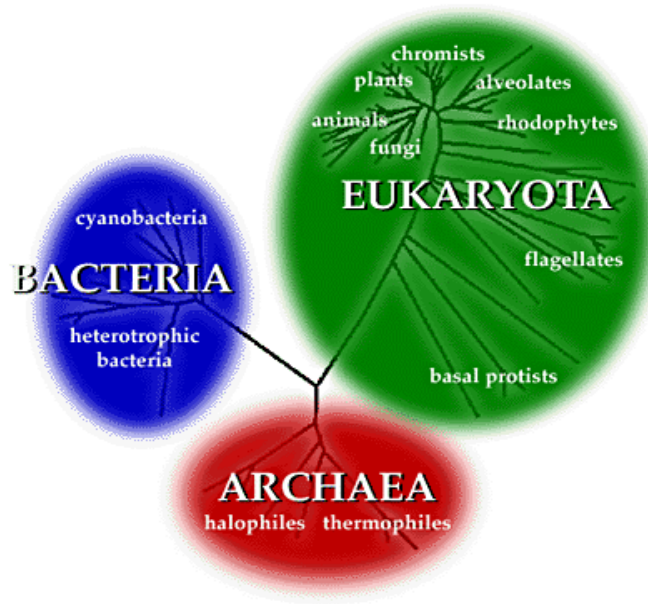


Figura 1. Representación de los tres grandes linajes celulares (Archaea, Bacteria y Eucarya) obtenida a partir de los análisis comparativos de las secuencias de la subunidad pequeña del rRNA. Imagen tomada de [www.ucmp.berkeley.edu](http://www.ucmp.berkeley.edu).



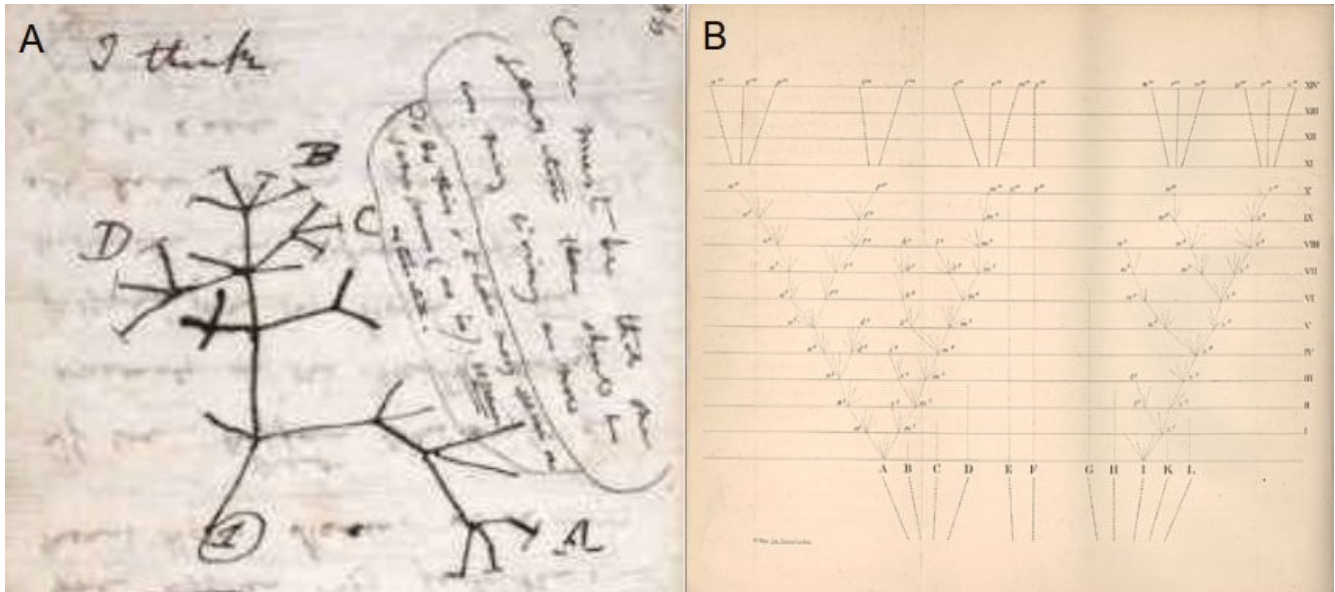


Figura 2. (a) Esquema tomado de los diarios de Charles Darwin en donde se muestra su primer esbozo de un árbol evolutivo. (b) Imagen que representa al árbol de la vida que apareció en el libro *On the Origin of Species by Natural Selection* (Darwin, 1859).

### 3.3 El último ancestro común (LCA – Last Common Ancestor)

Como se mostró por la reconstrucción de un árbol trifurcado no enraizado, basado en la comparación de secuencias de RNA ribosomal 16S/18S, todos los organismos se derivan de una forma ancestral. Al especular sobre la naturaleza del ancestro común a los tres linajes celulares o tres dominios, Carl R. Woese y George E. Fox sugirieron que en el punto de la trifurcación de los tres grandes linajes celulares había existido una entidad biológica hipotética que difería por mucho de la biología de los procariontes contemporáneos. A esta entidad rudimentaria la denominaron “progenote”. El progenote sería entonces, una entidad mucho más simple que no poseía muchas de las características básicas de los procariontes actuales y cuya tasa de evolución difería también de estos mismos organismos unicelulares, además de que la separación evolutiva entre su genotipo y fenotipo aún no se había completado del todo (Woese y Fox, 1977).

Desde una perspectiva evolutiva darwiniana, no era fácil aceptar esta idea. Es evidente que los organismos contemporáneos debieron haber sido precedidos por organismos mucho más simples pero la posibilidad de que el ancestro común de los tres dominios fuera un progenote resultaba difícil de conciliar con la complejidad de los procesos moleculares básicos de cada uno de los linajes. Teniendo

esto en mente, Fitch y Upper acuñaron el término de “cenancestro” para diferenciar las características compartidas del último ancestro común (LCA – por sus siglas en inglés) a todos los seres vivos de aquellas propuestas para el progenote (Fitch y Upper, 1987).

Todos los organismos comparten el mismo código genético básico, las mismas características esenciales de la replicación del genoma y la expresión de genes, reacciones anabólicas básicas, y producción de energía mediada por ATPasas asociadas a membrana. La distribución universal de estas características permitió sugerir que el ancestro común de todas las formas de vida no fue un descendiente inmediato del mundo del RNA, una protocélula, o cualquier otro sistema ancestral, sino que era una entidad que en complejidad biológica no difería significativamente a la de los organismos procariontes actuales (Lazcano *et al.*, 1992; Lazcano, 1995). Por lo tanto, el último ancestro común hace referencia al ancestro compartido más reciente a partir del cual todos los seres vivos han evolucionado, y cuya naturaleza es muy parecida a la de una bacteria contemporánea. Es muy probable que el último ancestro común haya formado parte de una población de entidades similares a él que existieron a través del mismo periodo, es decir, hace más de 3,500 millones de años. Sin embargo, es improbable que todas las características que distinguen al ancestro común estuvieran ya presentes en los primeros seres vivos, por lo que desde una perspectiva evolutiva, es razonable suponer que las primeras formas de vida fueron entidades mucho más simples.

### **3.4 Genómica comparativa y caracterización del último ancestro común**

Desde un punto de vista cladista, es posible reconstruir los rasgos biológicos que podría haber tenido el último ancestro común mediante la comparación de las características homólogas comunes a todos los seres vivos y que han sido producto de una herencia de tipo vertical, sumándole todos aquellos rasgos que estuvieron presentes en el ancestro pero que se perdieron en uno o mas grupos de organismos, menos todos aquellos rasgos que han sido transferidos de manera horizontal entre diferentes especies. Por consiguiente, el último ancestro común a todos los seres vivos podría considerarse como un inventario inferido de las características compartidas entre los organismos existentes, los cuales están localizados en la punta de las ramas de las filogenias moleculares. (Figura 3).

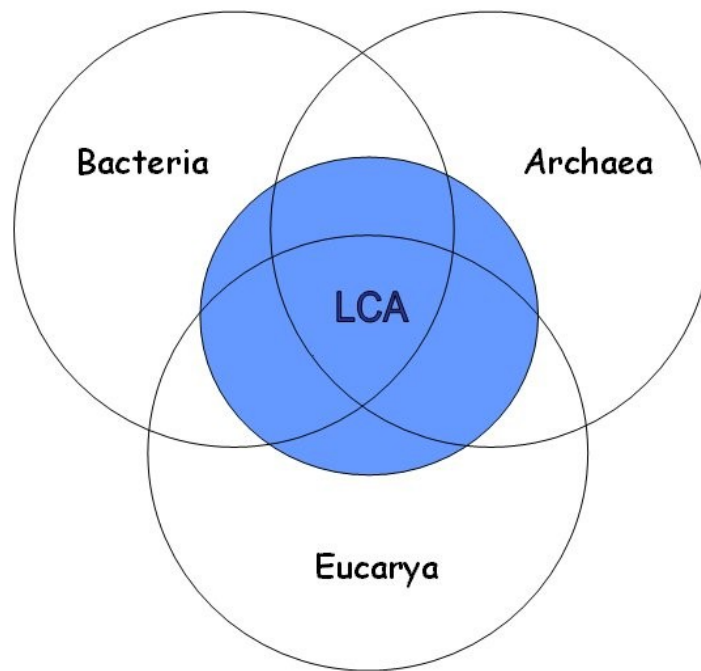


Figura 3. Esquema que representa la reconstrucción del último ancestro común a partir de las características compartidas entre los organismos existentes. La intersección de los conjuntos denotarían a las características homólogas comunes que pudieron haber estado presentes en el LCA.

Como se muestra en la figura 3, los genes altamente conservados que componen al último ancestro común estarían definidos por el grupo de secuencias presentes en la intersección de los conjuntos que representan a los genomas de los dominios Archaea, Bacteria y Eucarya. Sin embargo, estudios basados en la disponibilidad de la información genómica han revelado discrepancias importantes con la topología de los árboles de rRNA, cuestionando la viabilidad de la reconstrucción y apropiada comprensión de la historia biológica temprana (Doolittle, 2000).

En la práctica, la reconstrucción del último ancestro común se ha visto limitada por diversos fenómenos entre los que se encuentran: (a) un muestreo de la biodiversidad sesgado, (b) pérdida polifilética de genes, (c) tasas desiguales de evolución molecular, (d) pérdida secundaria de organelos, (e) convergencia y polifilia, (f) sustitución intragenómica por otros parálogos, (g) innovaciones evolutivas, (h) pérdida de rutas metabólicas, y (i) una transferencia horizontal de genes de una intensidad aún no conocida (Delaye *et al.*, 2004). Ante este inventario de dificultades, y principalmente haciendo énfasis en este último fenómeno, diversos investigadores han sugerido que es casi imposible poder asomarnos al pasado remoto (Doolittle, 1999; Gogarten *et al.*, 2002), por lo que han postulado que más que un árbol trifurcado universal, la historia de la vida debería

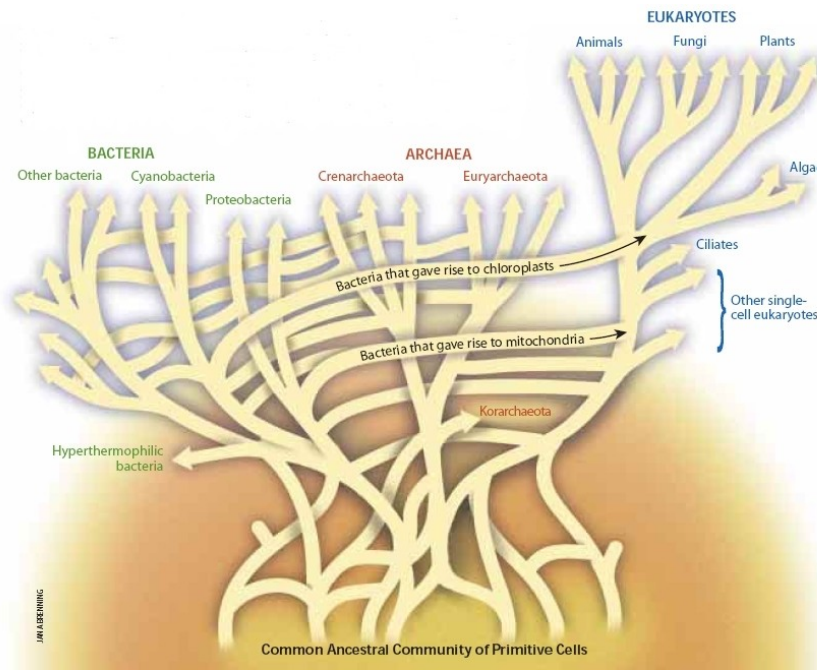


Figura 4. Figura en la que se intenta representar a la historia de la vida como una red de intercambios genéticos debido a la transferencia horizontal de genes masiva que se dio entre los primeros organismos. Imagen tomada del artículo de Scientific American de Ford Doolittle.

representarse como una red de intercambios genéticos (Figura 4). Aunque las dificultades para la reconstrucción del árbol universal son un hecho real, existen evidencias alternas que sustentan la topología global de los tres dominios. En 1999, tres grupos de trabajo (Snel *et al.*, 1999; Tekaia *et al.*, 1999; Fitz-Gibbon y House, 1999) demostraron de manera independiente que el contenido total de secuencias de diversos genomas celulares, representado en un fenograma o árbol genómico, era muy parecido a aquel producido por los análisis comparativos de secuencias de rRNA. Otro ejemplo que refuerza la idea de conservación de una fuerte señal evolutiva presente en el genoma, es aquél que desarrollaron Briones *et al.* en base a la sensibilidad del aparato ribosomal ante distintos antibióticos, encontrando que la topología de sus fenogramas obtenidos era consistente con la topología global reconstruida a partir de la secuencia del RNA ribosomal de la subunidad pequeña (Briones *et al.*, 2005). Posteriormente, Yang *et al.*, con base en la clasificación de la presencia o ausencia de dominios considerados a nivel de superfamilias, encontraron no sólo una clara distinción de los tres principales linajes celulares, sino que además podían distinguir varias de las subdivisiones al interior de cada una de las ramificaciones de los tres dominios (Yang *et al.*, 2005). Por lo tanto, si el pasado biológico no ha sido completamente borrado, entonces la cladística molecular y la genómica comparada pueden proveer pistas para la organización genética y complejidad bioquímica de las entidades de las cuales

el LCA evolucionó.

### 3.5 Naturaleza genómica del último ancestro común

Definir la naturaleza genómica del último ancestro común ha sido una de las principales metas en el estudio de la evolución temprana de la vida en la Tierra. Los estudios de filogenias profundas (Brown *et al.*, 2001; Daubin *et al.*, 2001; Doolittle, 2000; Moreira y López-García, 2006) y la genómica comparativa nos son útiles para generar ideas importantes del complemento génico del ancestro común a todos los seres vivos (Tabla 1).

Tabla 1. Estimaciones del complemento génico del último ancestro común basados en análisis genómicos cuantitativos. Modificado de Becerra *et al.* (2007)

Referencia	Características del ancestro	Metodología	Número de secuencias y categorías funcionales
	<b>LCA</b>		<b>80 COGs universalmente distribuidos</b>
Harris <i>et al.</i> (2003)	Eficiente transcripción y estructura del ribosoma; funciones ligadas a membrana; capaz de sintetizar cadenas largas de DNA	Identificación de COGs universalmente conservados 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) Manejo de proteínas (2/80) Otros (2/80)
	<b>LUCA</b>		<b>~600 genes asignados a LUCA (COGs)</b>
Mirkin <i>et al.</i> (2003)	Genes suficientes para mantener funcionalmente a un organismo	Escenarios parsimoniosos para conjuntos individuales de COGs basados en árboles de especies	Transcripción y traducción (112/600) Replicación y reparación de DNA (30/600) Proteínas asociadas a membrana y metabolismo (287/600) Manejo de proteínas (25/600) Otras (94/600)
	<b>LUCA</b>		<b>~63 genes universales (proteínas)</b>
Koonin (2003)	Simple con pocos genes; carente de un genoma de DNA y de un sistema de replicación	Comparación de las secuencias de ~100 genomas	Transcripción y traducción (56/63) Replicación y reparación de DNA (3/63) Proteínas asociadas a membrana (3/63) Manejo de proteínas (1/63)
	<b>LCA</b>		<b>49 plegamientos de proteínas universalmente distribuidos (superfamilias SCOP)</b>
Yang <i>et al.</i> (2005)	Con una maquinaria genética sofisticada de equipo estructural	Distribución 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)
	<b>LCA</b>		<b>~115 dominios de proteínas (Pfam)</b>
Delaye <i>et al.</i> (2005)	Similares en complejidad genética a las células actuales	Comparación de secuencias de 20 genomas con BLAST e identificación de ortólogos	Transcripción y traducción (56/115) Replicación y reparación de DNA (6/115)

		utilizando la base de datos Pfam	Proteínas asociadas a membrana (7/115) Metabolismo de nucleótidos y azúcares (33/115) Metabolismo de aminoácidos (12/115) Manejo de proteínas (1/115)
	<b>LUCA</b>		<b>20 motivos descritos (octapéptidos en proteínas con estructura 3D)</b>
Sobolevski y Trifonov (2006)	La cantidad total de los octámeros deben de estar en el orden de magnitud de miles	Identificación de octámeros prácticamente omnipresentes en motivos de proteínas.	Transcripción y traducción; replicación y reparación de DNA; manejo de proteínas; proteínas asociadas a membrana
	<b>LUCA</b>		<b>~1000 genes con un mínimo de 561 a 669 secuencias/categorías funcionales (proteínas)</b>
Ouzounis <i>et al.</i> (2006)	Organismos complejos en sus genomas, similares a los procariontes de vida libre actuales	Identificación 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)
	<b>LUCA</b>		<b>140 dominios de proteínas ancestrales (superfamilias CATH)</b>
Ranea <i>et al.</i> (2006)	Entidad genéticamente compleja, con prácticamente todas las características presentes en organismos actuales	Distribución de las superfamilias CATH en 114 genomas 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)

Se han hecho varios intentos para tratar de reconstruir las posibles características que podrían haber estado presentes en este organismo ancestral (Mushegian y Koonin, 1996; Koonin, 2003; Harris *et al.*, 2003; Delaye y Lazcano, 2000; Anantharaman *et al.*, 2002; Delaye *et al.*, 2005). Basados en la biología comparativa, distintos investigadores han caracterizado parte de la naturaleza del genoma del último ancestro común. Como se muestra en la tabla 1, los resultados de los diferentes intentos para caracterizar al ancestro de todos los seres vivos incluyen secuencias de genes de procesos biológicos básicos representados de manera incompleta; entre ellos encontramos a la transcripción, traducción, metabolismo energético, biosíntesis de nucleótidos y aminoácidos, y plegamientos de proteínas, también como algunas secuencias relacionadas con la replicación, reparación, y transporte celular. Sin embargo, si el término distribución universal se restringe a su sentido más obvio, es decir, a los rasgos encontrados en todos los genomas completamente secuenciados, entonces no es sorprendente que el repertorio resultante esté formado por relativamente pocas características y por procesos bioquímicos representados de manera incompleta (Tabla 1). Además, debemos tomar en cuenta que las reconstrucciones del complemento de genes de ancestros distantes son meras aproximaciones

estadísticas del pasado biológico, ya que su exactitud depende de múltiples factores entre los que se incluyen: (1) sesgos en la construcción de las bases de datos de genomas; (2) transferencia horizontal de genes; (3) variaciones significativas en las tasas de sustitución de diferentes proteínas; (4) el grado de pérdida secundaria de genes; y también (5) las metodologías utilizadas durante el análisis (Becerra *et al.*, 1997; Mirkin *et al.*, 2003).

A pesar de las diferentes aproximaciones metodológicas, diversos investigadores han encontrado que una gran parte de las secuencias altamente conservadas que compondrían a este organismo ancestral están relacionadas de una forma u otra con el metabolismo del RNA (Tabla 2), el cual ha sido ampliamente definido como el compendio de todos los procesos celulares que involucran al RNA, incluyendo la transcripción, procesamiento y modificación de transcritos, traducción, síntesis, degradación y regulación de esta molécula y de los ribonucleótidos (Delaye y Lazcano, 2000; Anantharaman *et al.*, 2002; Delaye *et al.*, 2005; Becerra *et al.*, 2007) (Figura 5).

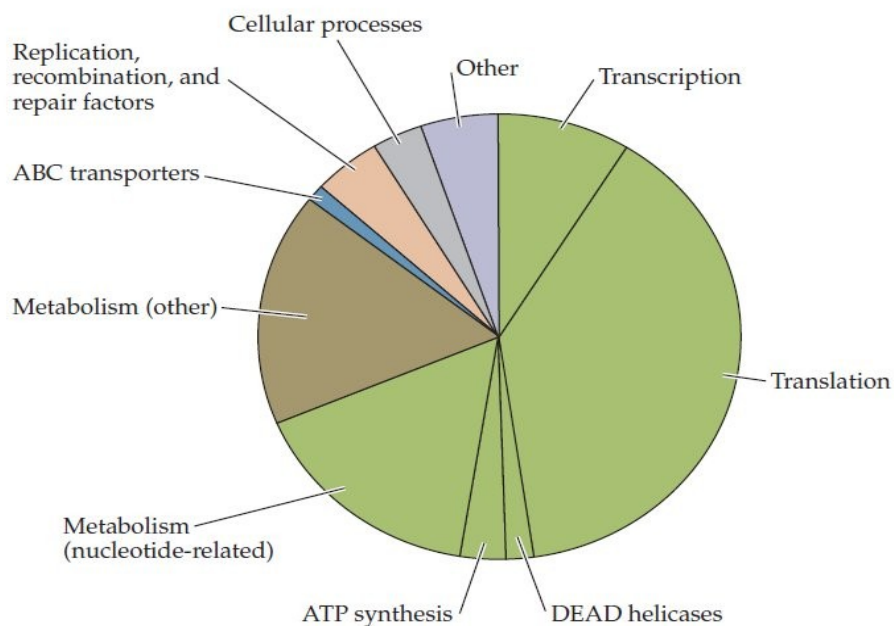


Figura 5. Figura en la que se muestra que la mayoría de los dominios de secuencias altamente conservadas corresponden a proteínas que interactúan con el RNA (color verde). Tomada de Delaye *et al.* (2005)



Tabla 2. Lista de genes altamente conservados y dominios conservados de los genomas de *E. coli*, *M. jannaschii* y *S. cerevisiae*. Las regiones en color gris denotan a las secuencias que están relacionadas en más de una forma con el metabolismo del RNA (Delaye *et al.*, 2005)

Description	Pfam domain	<i>E. coli</i>	<i>M. jannaschii</i>	<i>S. cerevisiae</i>
<b>Transcription</b>				
RNA polymerase β	RNA_pol_Rpb2_1; RNA_pol_Rpb2_2; RNA_pol_Rpb2_3; RNA_pol_Rpb2_6; RNA_pol_Rpb2_7	b3987	MJ1040, MJ1041	YOR151C, YOR207C, YPR010C
RNA polymerase β'	RNA_pol_Rpb1_1; RNA_pol_Rpb1_2; RNA_pol_Rpb1_3; RNA_pol_Rpb1_4; RNA_pol_Rpb1_5	b3988	MJ1042, MJ1043	YDL140C, YDR118C, YQB341W
<b>Translation</b>				
aminoacyl-tRNA synthetase class I	tRNA-synt_1c; tRNA-synt_1c_C; Arg_tRNA_synt_N; tRNA-synt_1d; tRNA-synt_1d_C; tRNA-synt_1	b2400, b0144, b1876, b3384, b2114, b0642, b0026, b4258, b0526	MJ1377, MJ0237, MJ1415, MJ1263, MJ0647, MJ1007, MJ0633	YOL033W, YGL245W, YDR168W, YDR341C, YHR091C, YDR288W, YOL097C, YGR264C, YGR171C, YPL040C, YBL078C, YGR094W, YLR382C, YPL160W, YNL247W
aminoacyl-tRNA synthetase class II	tRNA-synt_2d; tRNA_antl; tRNA-synt_2; tRNA-synt_2b; HGTP_antocodon; tRNA-synt_2c; DHHA1	b1714, b1713, b0930, b1886, b4129, b0893, b0194, b2697, b2890, b4155, b1719, b2514	MJ0487, MJ1108, MJ1556, MJ1077, MJ0564, MJ1238, MJ0226, MJ1197, MJ1000	YFL022C, YPR047W, YLR060W, YHR019C, YLL018C, YCR024C, YPL104W, YDR023W, YHR011W, YKL104C, YOR335C, YNL040W, YDR037W, YNL073W, YER087W, YIL078W, YHR020W, YPR033C
tRNA pseudouridine 55 synthase	TruB_N	b3186	MJ0148	YNL292W, YLR175W
dimethyladenosine transferase	RnaAD	b0051	MJ1029	YPL288W
elongation factors (EF-G, EF-Tu, and other GTP binding proteins)	GTP_EFTU; GTP_EFTU_D2;	b3340, b3339, b3980, b4375, b2599, b3971, b3188, b3590, b2751	MJ1048, MJ0324, MJ0495, MJ0262, MJ1261, MJ0325	YDR385W, YOR133W, YBR118W, YPR080W, YLR069C, YJL102W, YOR187W, YNL183C, YKL173W, YDR172W, YKR084C, YLR289W, YAL035W, YOL023W, YER025W, YLR244C, YBL091C, YER078C, YFR006W, YHR013C
methionine aminopeptidase	Peptidase_M24	b0188, b2385, b3847, b2908	MJ1329, MJ0806	YLR244C, YBL091C, YER078C, YFR006W, YHR013C
ribosomal-protein-alanine acetyltransferase	Acetyltransf_1	b4373, b2434**, b1448, b4012	MJ1530, MJ1207	YHR013C
Sun5, RNA-binding several different RNA-binding proteins containing the S1 domain	Sun5_yclO_yrdC S1	b3282, b3164, b0911	MJ0062, MJ0117	YGL189W, YJR007W, YMR229C
ribosomal proteins (small subunit) *	Ribosomal_S5; Ribosomal_S6_C; Ribosomal_S6; KH_2; Ribosomal_S3_C; S4; Ribosomal_S7; Ribosomal_S8; Ribosomal_S9; Ribosomal_S10; Ribosomal_S11; Ribosomal_S12; Ribosomal_S13; Ribosomal_S19	b3303, b0169, b3314, b3296, b3341, b3306, b3230, b3321, b3297, b3342, b3298, b3316	MJ0475, MJ0982, MJ0461, MJ0180, MJ1047, MJ0470, MJ0195, MJ0322, MJ0191, MJ1046, MJ0189, MJ0180	YGL123W, YBR251W, YLR048W, YGR214W, YHL004W, YNL178W, YPL081W, YBR189W, YNL137C, YHR148W, YJR123W, YJL190C, YLR067W, YDL083C, YMR143W, YBR146W, YHL015W, YJL191W, YCR031C, YNR036C, YGR118W, YPR132W, YDR450W, YML026C, YNL081C, YOL040C, YNR037C
ribosomal proteins (large subunit) *	Ribosomal_L1; Ribosomal_L2; Ribosomal_L2_C; Ribosomal_L6; Ribosomal_L11_N; Ribosomal_L11; Ribosomal_L5; Ribosomal_L5_C; Ribosomal_L14	b3984, b3317, b3305, b3983, b3308, b3310	MJ0510, MJ0179, MJ0176, MJ0471, MJ0469, MJ0466	YGL135W, YPL220W, YEL050C, YFR031C-A, YIL018W, YGR220C, YNL087W, YGL147C, YDR237W, YGR085C, YPR102C, YKL170W, YBL087C, YER117W



Description	Pfam domain	<i>E. coli</i>	<i>M. jannaschii</i>	<i>S. cerevisiae</i>
Metabolism				
thymidylate kinase [EC:2.7.4.9]	Thymidylase_kin	b1098	MJ0293	YJF057W
dihydroorotate cooxidase [EC:1.3.3.1]	DHO_oh	b0245, b2147**	MJ0664	YK0216W
oxotale	Primosytran	b3942	MJ1109, MJ1656	YML109W, YMR271C
phosphoribosyltransferases [EC:2.4.2.10]				
ispartate [EC:2.1.3.2] and ornithine [EC:2.1.3.3]	OTCase_N, OTCase	b4245, b4254, b0273, b2670**	MJ1581, MJ0881	YJL130C, YJL088W
carbamoyl- transferase catalytic chain				
carbamoyl-phosphate synthase small chain [EC:6.3.5.5]	CPCase_tm_chain; GATase	b0032, b3360, b2507, b1263	MJ1019, MJ0236, MJ1575, MJ1131	YJL130C, YCR303W, YKL211C, YMR217W
ribose-phosphate pyrophosphokinase [EC:2.7.6.1]	Primosytran	b1207	MJ1368	YER099C, YBL068W, YHL011C, YK0181W, YOL061W
IMP dehydrogenase [EC:1.1.1.205] and hypoxanthine proteins	MPDH (1st. half); CBS, CBS; MPDH (2nd. half)	b2508	MJ1616, MJ0188, MJ1292, MJ0653, MJ0100, MJ1225, MJ0622, MJ0392, MJ0868, MJ1404, MJ0556	YAR073W, YHR216W, YML058C
adenylo succinate lyase [EC:4.3.2.2]	Lyase_1	b1121, b3963, b1611, b4139	MJ0629, MJ0791	YLR058W, YPL262W
amido-phosphoribosyltransferase [EC:2.4.2.14]	GATase_2	b2312	MJ0204	YMR300C
pyruvate kinase [EC:2.7.1.40]	Primosytran PK, PK_C	b1676, b1854	MJ0108	YAL038W, YCR347C
thioredoxin reductase and other reductases [EC:1.8.1.9]	Pyr_redox	b0988, b0905, b0116, b3500, b0304, b3962, b0305, b2711, b2703, b2542	MJ1536, MJ0649, MJ0651	YHR106W, YDR333W, YPL018C, YPL091W, YPL017C, YJR137C
glucosamine- fructose-6- phosphate aminotransferase (isomerizing) [EC:2.6.1.18]	GATase_2; SfS; SfS	b3729, b3731**	MJ1420, MJ1116	YK109C, YMR065W, YMR064W
metal dependent hydrolase superfamily [EC:3.5.-.-]	Amidohydro_1	b0873**	MJ1490	YFR07C
UDP-galactose 4-epimerase [EC:5.1.3.2] and others	Epimerase	b0759, b3619, b3041, b3788	MJ0211, MJ1055	YBR019C
nucleosyltransferase activity [EC: 2.7.7.-]	NTP_transferase; Hexapep	b0009, b3789, b1236, b2042, b3730, b3430	MJ1101, MJ1334	YDL055C, YDR211W
hypoxanthine nucleoside- triphosphatase [EC:3.6.1.15]	Hamtp_Ho	b2954**	MJ0226	YJR069C**
enolase [EC:4.2.1.11]	Enolase_N; Enolase_C	b2729	MJ0232, MJ0199**	YGR264W, YHR174W, YMR323W, YCR393W, YPL281C YCR012W
phosphoglycerate kinase [EC:2.7.2.3]	PGK	b2926	MJ0641	YCR012W
phosphomannomutase [EC:5.4.2.8]	PGM_PMM_I; PGM_PMM_II; PGM_PMM_III; PGM_PMM_IV	b2048, b0688, b3176	MJ1100, MJ0292	YMR278W, YMR105C, YKL127W
sugar transferases	Glycos_transf_1	b2044, b3631	MJ1607, MJ1178, MJ1069, MJ1089	YPL175W
sugar transferases	Glycos_transf_2	b2254, b2351, b0363, b1022, b3615 b1103**	MJ1222, MJ0544	YPL227C, YPR183W
nucleotide-binding proteins	HIT	b2913, b1300, b3553, b2300, b1033	MJ0686** MJ1010	YDL125C, YDR005C YER081W, YL074C, YCR308C, YNL274C, YGL185C, YPL119C
phosphoglycerate dehydrogenase [EC:1.1.1.95]	2-Hackd_oh; 2-Hackd_oh_C; ACT			
NAD synthetase [EC:6.3.1.5, 6.3.5.1]	NAD_synthase	b1740	MJ1352	YHR074W
flavoprotein enzymes	Flavoprotein	b0639	MJ0913	YK1088W, YKR072C, YCR054C
UPP synthetase [EC:2.5.1.-]	Phenyltransf	b0174	MJ1372	YMR101C, YBR002C
tryptophan synthase (β-chain) [EC:4.2.1.20]	PALP	b1261, b3117, b2421, b3772, b2871, b2414	MJ1032, MJ1465	YGL026C, YCL064C, YKL216C, YGR155W, YER058W, YGR012W
tryptophan synthase (α-chain) [EC:4.2.1.20]	Trp_syntA	b1260	MJ1038	YGL026C
histidinol-phosphate aminotransferase [EC:2.6.1.9]	Aminotran_1_2	b0021, b2379, b0600, b2290, b1439, b1622, b4340	MJ0255, MJ0001, MJ1391, MJ0684, MJ1479	YJL116W, YJL060W, YDR111C, YLR089C
Polyprenyl synthase [EC: 2.5.1.-]	polyprenyl_synt	b0421, b3187	MJ0650	YJL167W, YPL069C, YBR003W
probable glyoxylase II [EC:3.1.2.8]	Lactamase_B	b0927**, b0212	MJ0686**	YDR272W
probable peroxiredoxin [EC:1.6.4.-]	AhpC-TSA	b0605, b2480	MJ0736	YIL010W, YBL064C, YML028W, YDR453C

Description	Pfam domain	<i>E. coli</i>	<i>M. jannaschii</i>	<i>S. cerevisiae</i>
<b>DEAD helicases</b>	DEAD; Helicase_C	b0797, b3182, b1343, b3780, b2576, b3822, b1653	MJ0669, MJ1401, MJ1574, MJ0294, MJ0383, MJ1124	YJL139C, YKR059W, YDR021W, YPL119C, YOR204W, YGL078C, YNL112W, YDL160C, YLL008W, YHR065C, YDL064W, YHR166W, YJL033W, YDR243C, YOR046C, YBR237W, YMR290C, YGL171W, YFL002C, YDL031W, YDR194C, YLR276C, YBR142W, YKR024C, YGL064C, YNR038W, YDR291W, YGL251C, YER172C, YMR180C, YGR271W
<b>ATP synthesis (atpA, atpB)</b>	ATP-synt_ab_N; ATP-synt_ab; ATP-synt_ab_C	b3734, b3732, b1941	MJ0217, MJ0216	YBL099W, YJR121W, YDL185W, YBR127C
<b>Replication, recombination and repair factors</b>				
<b>ATPase family proteins (clamp-loading, <math>\gamma</math> <math>\tau</math> subunits)</b>	AAA	b0470, b3178, b0892	MJ1422, MJ0884, MJ1156, MJ1176, MJ1494	YJR066W, YNL290W, YOL094C, YMR089C, YER017C, YDL126C, YBR080C, YPR024W, YGR270W, YPR173C, YKL145W, YGL048C, YOR259C, YDL007W, YOR117W, YDR394W, YLR397C, YNL329C, YLL034C, YKL197C, YGR028W, YPL074W, YER047C, YDR375C, YBR188W
<b>Ribonuclease III endonuclease III</b>	RNase III HhH-GPD (1rst. half); HHH; HhH-GPD (2nd. half);	b0163 b1633, b2961	MJ0135 MJ1434, MJ0613	YNL072W YAL015C, YOL043C
<b>DNA topoisomerase I and III</b>	Toprim; Topoisom_bac	b1274, b1763	MJ1652, MJ1512	YLR234W
<b>ABC transporters</b>	ABC_tran	b0448, b1290, b1291, b1496, b1682, b1709, b1756, b2201, b3479, b4058, b4096, b0066, b0127, b0151, b0199, b0262, b0366, b0449, b0490, b0495, b0588, b0652, b0760, b0794, b0809, b0820, b0829, b0855, b0864, b0879, b0886, b0887, b0914, b0933, b0949, b1117, b1126, b1246, b1247, b1318, b1441, b1483, b1484, b1513, b1858, b1900, b1917, b2129, b2149, b2180, b2306, b2422, b2547, b2677, b3201, b3271, b3352, b3450, b3454, b3455, b3463, b3460, b3486, b3540, b3541, b3567, b3725, b3749, b4035, b4067, b4067, b4106, b4228, b4267, b4391	MJ1023, MJ1088, MJ1242, MJ1287, MJ1367, MJ1508, MJ1572, MJ1662	YKR104W, YLL015W, YNR070W, YOR011W, YOR328W, YPL058C, YPL147W, YCR011C, YDR091C, YFR009W, YGR281W, YHL035C, YKL209C, YLL048C, YLR189W, YLR249W, YMR301C, YNL014W, YOL075C, YOR153W, YPL226W, YPL270W
<b>Protein management signal recognition particle protein</b>	SRP54_N; SRP54; SRP_SPB	b2610, b3464	MJ0101, MJ0291	YPR088C, YDR292C
<b>chaperonin Cpn60</b>	Cpn60_TCP1	b4143	MJ0999	YLR259C, YJR064W, YJL111W, YDL143W, YIL142W, YDR188W, YJL014W, YDR212W, YJL008C

Delage *et al.* (2005) encontraron genes altamente conservados que forman parte de la maquinaria del degradosoma del RNA (ej. gen de la enolasa y gen de la RNA helicasa tipo DEAD), idea que ha sido interpretada como la evolución temprana de un mecanismo de control para la expresión génica a nivel de RNA debido a su papel en la hidrólisis del mRNA y el reciclaje de ribonucleótidos. Asimismo, los resultados de su trabajo aportan evidencia adicional (1) a la hipótesis

de que durante la evolución celular temprana las moléculas de RNA jugaron un papel muy importante, (2) la existencia de un mundo antiguo de RNA/proteína, (3) la complejidad biológica del último ancestro común, y (4) evidencia de la naturaleza bioquímica del genoma cenancestral.

No obstante, si se analiza el conjunto de secuencias altamente conservadas y aquellas que están involucradas con el metabolismo del RNA, se puede observar que entre ellas se incluyen moléculas grandes y complejas estructuralmente, por lo que es razonable suponer desde una perspectiva evolutiva, que éstas debieron haber sido precedidas por secuencias más sencillas.

¿Cuáles podrían ser los mecanismos que promueven la complejidad tanto estructural como funcional de las proteínas? ¿Existe evidencia de que un conjunto de secuencias altamente conservadas hayan sido precedidas por moléculas más simples? Estas serían algunas de las interrogantes planteadas que nos proporcionarían pistas de la organización genética y complejidad bioquímica de las entidades primitivas de las cuales el LCA evolucionó.

### **3.6 Evolución de los genomas**

Existen varios mecanismos por los cuales las secuencias ancestrales pudieron haber evolucionado para dar lugar a los genomas actuales. Uno de los principales eventos que ha contribuido a la formación de la diversidad en proteínas son las duplicaciones, evento en el cual se generan copias idénticas de material genético.

Se ha postulado que la duplicación de genes y genomas es un factor importante en la evolución de los organismos, por lo que ha sido reconocido como una de las fuerzas principales en la expansión del material genético (Ohno, 1970). Asimismo, se ha demostrado en los últimos años que el fenómeno de duplicación es uno de los mecanismos biológicos altamente relevantes y predominantes por el cual la ampliación de una secuencia se da, y quizá una de las fuerzas principales para la acreción y conformación de los genomas.

La comparación de secuencias de proteínas ha confirmado el papel que han tenido las duplicaciones antiguas de genes en la evolución de los genomas (Becerra y Lazcano, 1998). Las familias de genes que han sido producto de eventos de duplicación y divergencia, conocidas como

familias de genes parálogos, nos puede brindar pistas de la organización genética y complejidad bioquímica de las entidades primitivas de las cuales el cenanestro evolucionó (Islas *et al.*, 2007).

El número de secuencias que han experimentado alguna duplicación previa a la divergencia de los tres linajes celulares incluyen genes que codifican para una variedad de enzimas que participan en diferentes procesos bioquímicos, tales como la traducción, la replicación del DNA, fijación del CO<sub>2</sub>, metabolismo de nitrógeno, y rutas biosintéticas. Los análisis de los genomas completos han revelado que las secuencias que han sido resultado de la expansión de genes parálogos en una época pre-ancestral se pueden clasificar en tres grandes grupos (Islas *et al.*, 2007):

- (a) Familias de genes formadas por múltiples duplicaciones de una secuencia, así como los transportadores ABC, ATPasas tipo-P, y permeasas acopladas a iones (Clayton *et al.*, 1997);
- (b) Familias formadas por un número pequeño de secuencias parálogas. Entre la lista se incluye a la pareja de genes homólogos que codifica para los factores de elongación EF-Tu y EF-G (Iwabe *et al.*, 1989), también como la secuencias duplicadas que codifican para las subunidades alfa y beta de la ATPasa tipo-F (Gogarten *et al.*, 1989). La extraordinaria conservación de estos duplicados sugiere que el último ancestro común estuvo precedido por una célula más simple, con un genoma más pequeño en el cual existía únicamente una copia de cada uno de estos genes, es decir, por células en las cuales la síntesis proteínica involucró únicamente un factor de elongación y ATPasas con habilidades reguladoras limitadas, con un solo tipo ancestral de subunidades hidrolíticas; y
- (c) Secuencias formadas por módulos homólogos arreglados en tandem los cuales experimentaron eventos de fusión, tales como las proteínas disulfuro oxidoreductasas (PDO) (Ren *et al.*, 2000), la subunidad grande de la carbamoil fosfato sintasa (Alcántara *et al.*, 2000), y HisA, una isomerasa de la biosíntesis de la histidina (Alifano *et al.*, 1996). Esto indica que el tamaño y la estructura de un número de proteínas son el resultado evolutivo de duplicaciones parálogas seguidas por eventos de fusión, que han tenido lugar previamente a la divergencia de los tres reinos primarios.

En este trabajo intentamos reconocer cuál es el papel que ha jugado la duplicación de genes seguidos de eventos de fusión en la evolución temprana de las proteínas (caso c), ya que es uno de los tipos de duplicación relativamente poco estudiado y cuya importancia ha sido reconocida únicamente de manera individual.

### 3.7 Duplicación interna de genes

La duplicación interna de genes es un mecanismo por el cual se genera una copia idéntica de un segmento de DNA formando módulos homólogos arreglados en tandem, los cuales posteriormente sufren eventos de fusión. Este fragmento duplicado traerá como consecuencia la elongación de la secuencia génica la cual codificará para una proteína cuya longitud se ve así duplicada. La longitud del fragmento intragénico duplicado se ha manejado indistintamente en la literatura, esto es, que puede abarcar desde unos pocos aminoácidos repetidos a lo largo de la proteína, uno o más dominios duplicados dentro de la misma, o hasta la duplicación del gen casi completo. Los primeros dos eventos han sido ampliamente estudiados por diversos investigadores y han sido reconocidos como factores importantes en la evolución del genoma (Barker *et al.*, 1978; Heringa y Argos, 1993; Heringa y Taylor, 1997; Heringa, 1998; Marcotte *et al.*, 1998; Kurtz y Schleiermacher, 1999; Pelligrini *et al.*, 1999; Andrade *et al.*, 2001; Kurtz *et al.*, 2001). Sin embargo, poco se ha dicho sobre la duplicación casi completa del gen, siendo un evento muy poco estudiado y sólo ha llegado a ser reconocido de manera individual en cierto tipo de proteínas. En este trabajo, las secuencias repetidas encontradas dentro de un gen que han sido producto de duplicaciones y la duplicación casi completa o total de un gen serán llamados como repeticiones internas y duplicación interna de genes respectivamente, siguiendo una clasificación equivalente a la que se ha manejado en el estudio de 163 proteínas realizado por Barker *et al.* (1978).

Son varios los mecanismos que pueden generar duplicaciones internas. Uno de los mecanismos hipotéticos que ha sido reconocido en eucariontes, y que se cree que podría funcionar de manera semejante en procariontes, es aquél en el que está involucrado un evento de entre-cruzamiento desigual, seguido por una mutación que genera la sustitución de uno de los nucleótidos (Figura 6). Se ha encontrado de manera individual que un gran número de proteínas pertenecientes a los organismos actuales, principalmente eucariontes, contienen duplicaciones internas.

Estas observaciones sugieren que la duplicación interna de genes ha jugado un papel importante en el incremento de la complejidad funcional del gen y en su evolución, ya que este tipo de duplicación podría permitir la adquisición de una función adicional por la modificación de un segmento redundante o por el incremento de sitios activos. Así, muchos genes complejos encontrados

en los genomas actuales podrían haber evolucionado a partir de genes pequeños primordiales vía duplicación interna y su modificación subsecuente (Li, 1983).

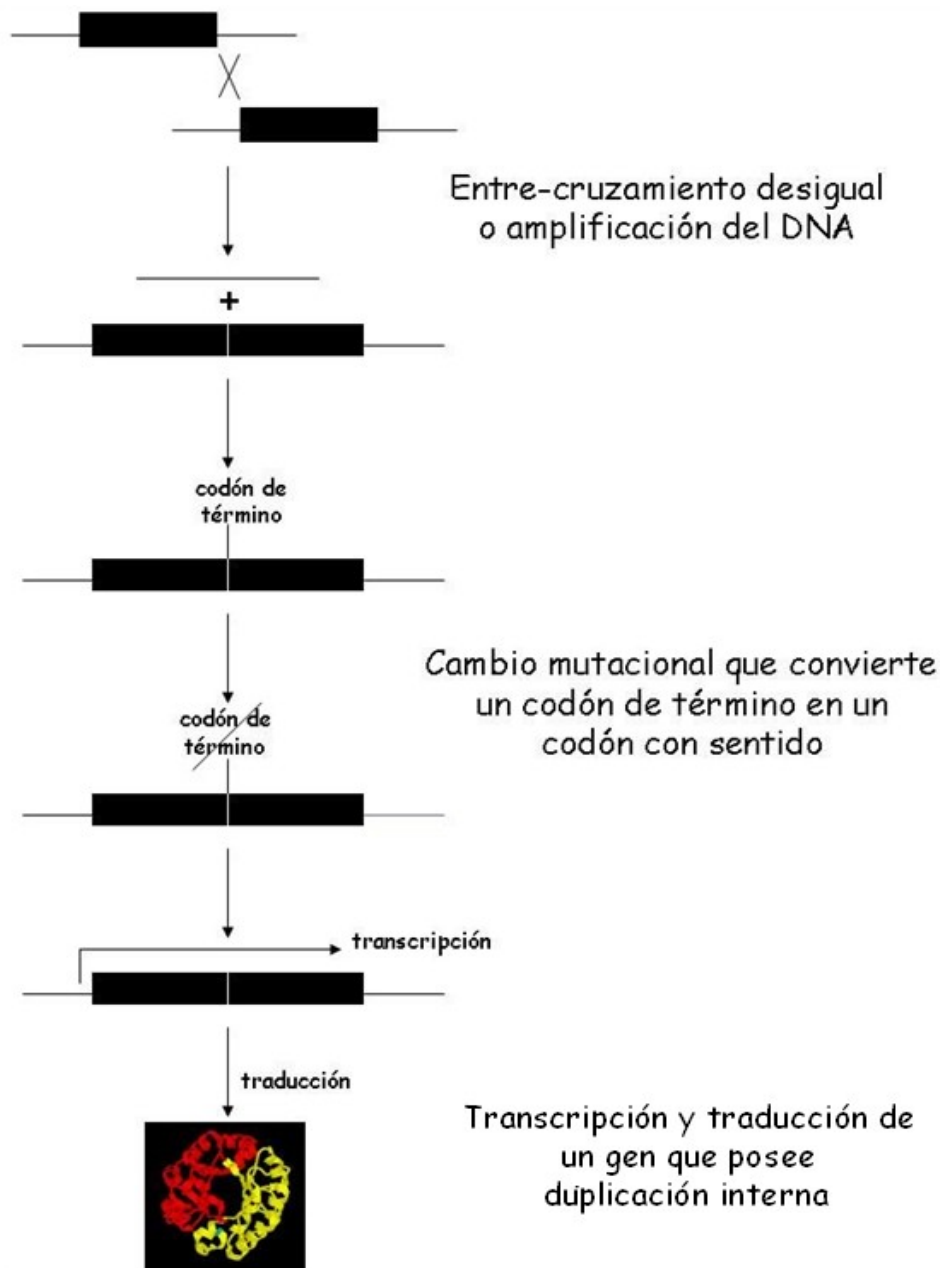


Figura 6. Modelo hipotético que muestra el mecanismo de duplicación interna. El primer paso muestra el mecanismo por el que un gen puede duplicarse y en el que la posición de los genes duplicados se mantiene de manera adyacente. Dicho mecanismo puede involucrar el fenómeno de entre-cruzamiento desigual o la amplificación del DNA durante el periodo de replicación. En el segundo paso se muestra otro de los cambios que pueden ocurrir durante la replicación: la mutación. Esta clase de mutación no sinónima alteraría al codón de término, transformándolo en un codón que codificaría para un aminoácido. El último paso indica la transcripción y traducción del gen fusionado, el cual codificará para una proteína que posee dos fragmentos que fueron producto de la duplicación.

# **IV MATERIAL Y MÉTODOS**

## 4.1 Análisis de bases de datos de secuencias altamente conservadas

Debido a que la reconstrucción del complemento de genes de ancestros distantes se basan en aproximaciones estadísticas del pasado biológico, ya que su exactitud depende de múltiples factores que influyen directamente en los estimados teóricos del genoma del último ancestro común, se han encontrado diferencias cualitativas y cuantitativas en los resultados arrojados por las diferentes aproximaciones metodológicas utilizadas.

En nuestro trabajo se analizaron diferentes conjuntos de secuencias altamente conservadas de los estimados del complemento de genes del LCA de diversos investigadores. Las bases de datos que se analizaron son (1) el conjunto de secuencias que caracterizan la naturaleza del último ancestro común de Delaye *et al.*, (2005) (2) el conjunto de secuencias que forman parte del núcleo génico del ancestro universal de Harris *et al.*, (2003) y (3) el conjunto de secuencias presentes en el último ancestro común universal obtenidas bajo una serie de eventos parsimoniosos de Mirkin *et al.* (2003).

## 4.2 Búsqueda de proteínas originadas por duplicación interna

La comparación de la secuencia de una proteína altamente conservada contra si misma, nos permite detectar la existencia de un cierto grado de similitud interno en la secuencia, es decir, si existen niveles de parecido entre la mitad amino terminal y la mitad carboxilo terminal. De ser así, ello indicaría que probablemente ambos segmentos tuvieron un origen común, esto es, que pudieron haberse originado mediante un evento en el cual uno de los segmentos se duplicó y posteriormente la fusión de ambos segmentos duplicados llevó a conformar a una sola unidad de expresión genética.

La búsqueda de segmentos duplicados dentro de una secuencia de aminoácidos se llevó a cabo mediante la comparación de la secuencia contra ella misma para cada una de las proteínas altamente conservadas. Esto se logró por medio de la herramienta de alineación local LFASTA versión v2.Ou66 (Pearson y Lipman, 1988), la cual forma parte de la paquetería FASTA2.

LFASTA compara dos secuencias de manera local, es decir, hace búsquedas de similitud con base en fragmentos, lo que permite encontrar regiones idénticas y similares a lo largo de las



secuencias. Los siguientes parámetros fueron utilizados: una matriz de sustitución tipo PAM250, penalización de “-12” al encontrar un gap y “-2” al extenderlo, tamaño de la palabra de búsqueda igual a 1 ( $ktup = 1$ ), y un registro de búsqueda con regiones similares mayor a 25.

Debido a que nuestro interés se centra particularmente en la duplicación intragénica de fragmentos grandes, pero no en una sucesión de repeticiones internas pequeñas, decidimos analizar únicamente aquellas proteínas cuyos segmentos duplicados abarcaran la mayor parte de la longitud total del gen. El interés de encontrar la duplicación casi completa del gen se centra en la posibilidad de tener sitios funcionales duplicados. De esta forma, mediante el programa computacional DUPINT.pl, desarrollado en el laboratorio y el cual fue escrito en la plataforma de lenguaje Perl, se extrajeron todos aquellos segmentos cuyo fragmento duplicado abarcara por lo menos dos terceras partes de la longitud total de la secuencia y cuyas regiones de aminoácidos que se encontraran sobrelapadas no fueran mayores a un 20%, además de que el porcentaje de identidad entre los segmentos duplicados fuera mayor o igual a 25%.

### **4.3 Distribución funcional de genes originados por un evento de duplicación interna**

Se han desarrollado múltiples sistemas de clasificación para agrupar a un conjunto de secuencias. Algunas de ellas se han organizado en términos de su función celular (Riley *et al.*, 1997). Otras han sido clasificadas en funciones con una distribución universal, basada en la comparación de secuencias y anotación funcional (Kyrpides *et al.*, 1999), otras más se han agrupado utilizando las categorías funcionales de la base de datos COGs (Harris *et al.*, 2003; Mirkin *et al.*, 2003). Sin embargo, muy pocos sistemas de clasificación han desarrollado agrupamientos basados en una perspectiva evolutiva, la cual es muy importante para comprender la historia evolutiva de las secuencias; por ejemplo, el sistema de clasificación utilizado por Delaye *et al.* (2005) enfatiza aquellas secuencias que están relacionadas con el metabolismo del RNA, indicando que algunas de ellas pudieron haberse originado en un mundo de RNA/proteína.

La clasificación funcional de aquellos genes altamente conservados y que se originaron por un evento de duplicación interna en las diversas bases de datos, está basada en la distribución de funciones metabólicas descritas para el conjunto de proteínas ortólogas (COGs), identificando aquellas secuencias

que se relacionan con el metabolismo del RNA. Las múltiples y diversas categorías funcionales que caracterizan a la base de datos COGs permitió clasificar a todas las secuencias de las diversas bases de datos en agrupamientos funcionales equivalentes, lo cual hubiera sido problemático ante un número reducido de categorías funcionales o ante categorías funcionales muy simples y elementales.

Entre las categorías comprendidas en la base de datos de COGs se encuentran: (J) traducción, estructura ribosomal y biogénesis; (A) procesamiento y modificación del RNA; (K) transcripción; (L) replicación, recombinación y reparación; (B) estructura y dinámica de la cromatina; (D) control del ciclo celular, división celular, partición de cromosomas; (Y) estructura nuclear; (V) mecanismos de defensa; (T) mecanismos de transducción de señales; (M) pared celular, membrana, envoltura; (N) motilidad celular; (Z) citoesqueleto; (W) estructura extracelular; (U) tráfico intracelular, secreción, y transporte vesicular; (O) modificación postraduccional, chaperonas; (C) producción y conversión de energía; (G) transporte y metabolismo de carbohidratos; (E) transporte y metabolismo de aminoácidos; (F) transporte y metabolismo de nucleótidos; (H) transporte y metabolismo de coenzimas; (I) transporte y metabolismo de lípidos; (P) transporte y metabolismo de iones inorgánicos; (Q) biosíntesis de metabolitos secundarios, transporte y catabolismo; (R) función general; (S) función desconocida.

Asimismo, se dividieron las secuencias conservadas de las diferentes bases de datos en dos grupos: (1) aquellas que se encuentran relacionadas en más de una forma con el metabolismo del RNA y (2) aquellas que no se relacionan directamente con el metabolismo del RNA (restantes). Esta división nos permitió hacer análisis bajo una perspectiva evolutiva, ya que dentro del conjunto de secuencias que se relacionan con el RNA podrían encontrarse las proteínas más antiguas que podemos reconocer.

#### **4.4 Secuencias altamente conservadas que interactúan con el metabolismo del RNA y que se originaron por duplicación interna**

Debido a que el objetivo fundamental de nuestro trabajo consiste en dilucidar el papel que ha desempeñado la duplicación interna de genes en la evolución temprana de las proteínas, analizamos todas aquellas secuencias que estuvieran relacionadas con el metabolismo del RNA a partir del conjunto de secuencias altamente conservadas, y cuyo probable origen se sitúa antes de la divergencia de los tres grandes linajes celulares, ej. en un mundo de RNA/proteínas.

Se seleccionaron todas aquellas secuencias de proteínas que estuvieran relacionadas en mas de una forma con el metabolismo del RNA de las bases de datos desarrolladas por Mirkin *et al.* (2003), Harris *et al.* (2003) y Delaye *et al.* (2005), y se buscaron aquellas que han sido producto de la duplicación y posterior fusión de genes.

#### **4.5 Visualización de estructura primaria y estructura terciaria de proteínas altamente conservadas con duplicación interna**

Existen tres métodos bioinformáticos por los cuales se puede inferir la homología interna de una proteína. Dos de ellos están basados exclusivamente en la estructura primaria, por lo que la detección del fenómeno de duplicación interna se lleva a cabo mediante algoritmos de alineación local. La alineación local puede ser de dos tipos: (a) uno en donde se muestre la alineación aminoácido por aminoácido indicándonos las similitudes e identidades, criterio utilizado como primera búsqueda en este trabajo; y (b) otro en donde el archivo resultado es una matriz (dot-plot) en la que se muestran líneas diagonales que representan a las identidades. Este último método es el más fácil, el más simple y el más completo para comparar dos secuencias e identificar regiones de similitud, repeticiones internas y eventos de re-arreglos (Gibbs y McIntyre, 1970). El tercer método de detección es (c) aquél en el que se hace uso de la estructura terciaria de una proteína, es decir, observando la regiones homólogas encontradas en la estructura tridimensional de la molécula, lo cual depende de su disponibilidad en las bases de datos experimentales.

Como métodos complementarios a la alineación de la estructura primaria, y para asegurar la validez de nuestros resultados, se utilizaron tanto el método visual PLALIGN versión 2.Ou66 (Huang y Miller, 1991), con una matriz de sustitución tipo BLOSUM 50, penalización de “-12” al encontrar un gap y “-2” al extenderlo como parámetros; y la búsqueda del cristal de la molécula bajo estudio en la base de datos Protein Data Bank (PDB) (Berman *et al.*, 2000). Las regiones homólogas internas en el cristal fueron visualizadas con el programa “RasMol Molecular Graphics v.2.7.5” (Sayle y Milner-White, 1995).

## **4.6 Antigüedad relativa de secuencias altamente conservadas**

No todas las secuencias altamente conservadas son coetáneas, es decir, no todas ellas se originaron en la misma época. Como un primer intento para establecer la edad relativa de este conjunto de secuencias, identificamos a todas aquellas secuencias que participan en reacciones que requieren de oxígeno molecular de la base de datos de compuestos del KEGG (Kanehisa y Goto, 2000; Kanehisa, y Bork, 2003; Kanehisa *et al.*, 2006; Kanehisa *et al.*, 2010). Esta aproximación nos indicaría que estas proteínas debieron haberse originado cuando este gas se acumuló en la atmósfera de la Tierra.

# **V RESULTADOS**

## 5.1 Proteínas altamente conservadas que se originaron por duplicación interna en las diferentes bases de datos

Con el objetivo de detectar duplicaciones internas en las secuencias altamente conservadas, hemos analizado diferentes estimaciones del complemento génico del último ancestro común. Entre estas se encuentran (1) el conjunto de secuencias obtenidas a partir de la comparación de 20 genomas celulares de vida libre que caracterizan al último ancestro común (Delaye *et al.*, 2005); (2) el conjunto de secuencias que forman parte del núcleo génico del ancestro universal obtenidas a partir de la identificación de grupos de genes ortólogos universalmente conservados que muestran filogenias de los tres dominios celulares (Harris *et al.*, 2003); y (3) el conjunto de secuencias presentes en el último ancestro común universal obtenidas a partir de la construcción de escenarios parsimoniosos para un conjunto de COGs basados en árboles de especies (Mirkin *et al.*, 2003).

Los resultados muestran que aproximadamente 38% de las secuencias altamente conservadas fueron resultado de un evento de duplicación interna en la base de datos de Mirkin *et al.*, 21% en la de Delaye *et al.*, y 38% en la de Harris *et al.* (Figura 7).

Únicamente un núcleo de 48 secuencias conservadas son comunes a las tres bases de datos analizadas, 23 de las cuales presentan duplicación interna (Tabla 3).

El número de secuencias conservadas que se comparten tan sólo en dos bases de datos está representado por un total de 96, de las cuales 64 son comunes a las bases de Mirkin *et al.* y Delaye *et al.* Las 32 secuencias restantes se localizan en las bases de Mirkin *et al.* y Harris *et al.* No se encontraron secuencias compartidas únicamente entre las bases de Delaye *et al.* y Harris *et al.* (Tabla 4).

De las 64 secuencias presentes en Mirkin *et al.* y Delaye *et al.*, 30 han sido originadas por duplicación interna, mientras que 16 de las 32 secuencias localizadas en Mirkin *et al.* y Harris *et al.* se originaron por el mismo mecanismo.

Existen 479 secuencias conservadas que se encuentran exclusivamente en una base de datos,

siendo la base de Mirkin *et al.* la que contiene la mayor parte de ellas (427), y Delaye *et al.* la que contiene al resto (52). Debido a que la totalidad de la base de datos de Harris *et al.* está contenida en las secuencias analizadas por Mirkin *et al.*, no se encontraron secuencias exclusivas en esta base (Tabla 5). Se encontró que 167 secuencias únicas de la base de Mirkin *et al.* han sido el resultado de duplicación y fusión de genes, mientras que 6 secuencias exclusivas de Delaye *et al.* se han originaron por el mismo mecanismo.

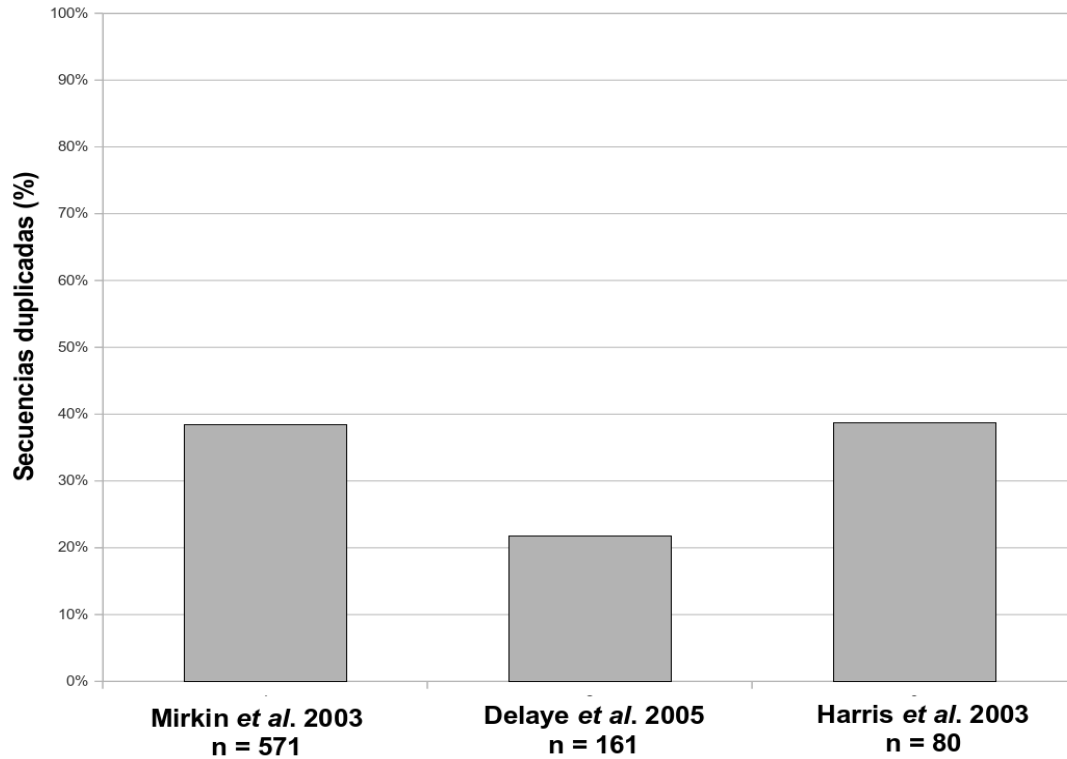


Figura 7. Porcentaje de secuencias que se han originado por duplicación interna en cada una de las bases de datos analizadas. “n” indica el número total de secuencias conservadas en cada una de las bases.

Tabla 3. Conjunto de COGs (secuencias) altamente conservados que son comunes a las tres bases de datos. “M, D y H” representan a la base de datos de Mirkin *et al.* (2003), Delaye *et al.* (2005) and Harris *et al.* (2003) respectivamente. “+” denota la presencia del COG en la base de datos. El símbolo “++” indica que por lo menos una secuencia del COG presenta evidencia de duplicación interna. Las filas sombreadas muestran a las secuencias conservadas que están involucradas con el metabolismo del RNA.

COG ID	Activity or Function	M	D	H
COG0006	Xaa-Pro aminopeptidase	+	+	+
COG0541	Signal recognition particle GTPase	+	+	+
COG0552	Signal recognition particle GTPase	+	+	++
COG0459	Chaperonin GroEL (HSP60 family)	+	+	+
COG0492	Thioredoxin reductase	+	++	+
COG0073	EMAP domain	++	+	++
COG0125	Thymidylate kinase	++	+	++
COG0550	Topoisomerase IA	+	++	+
COG1109	Phosphomannomutase	+	++	+
COG0085	DNA-directed RNA polymerase beta subunit/140 kD subunit (split gene in Mjan, Mthe, Aful)	+	+	+
COG0086	DNA-directed RNA polymerase beta' subunit/160 kD subunit (split gene in archaea and Syn)	+	+	+
COG0013	Alanyl-tRNA synthetase	+	+	+
COG0018	Arginyl-tRNA synthetase	+	+	+
COG0030	Dimethyladenosine transferase (rRNA methylation)	++	++	++
COG0008	Glutamyl- and glutaminyl-tRNA synthetases	+	+	+
COG0124	Histidyl-tRNA synthetase	+	+	+
COG0060	Isoleucyl-tRNA synthetase	+	+	+
COG0495	Leucyl-tRNA synthetase	+	+	+
COG0024	Methionine aminopeptidase	++	++	++
COG0143	Methionyl-tRNA synthetase	+	+	+
COG0016	Phenylalanyl-tRNA synthetase alpha subunit	+	+	+
COG0072	Phenylalanyl-tRNA synthetase beta subunit	+	+	+
COG0442	Prolyl-tRNA synthetase	+	+	+



COG ID	Activity or Function	M	D	H
COG0081	Ribosomal protein L1	+	++	+
COG0080	Ribosomal protein L11	++	+	++
COG0093	Ribosomal protein L14	++	+	++
COG0090	Ribosomal protein L2	+	+	+
COG0094	Ribosomal protein L5	++	+	++
COG0097	Ribosomal protein L6	++	++	++
COG0051	Ribosomal protein S10	+	++	+
COG0100	Ribosomal protein S11	++	++	++
COG0048	Ribosomal protein S12	++	+	++
COG0099	Ribosomal protein S13	++	++	++
COG0185	Ribosomal protein S19	+	+	+
COG0052	Ribosomal protein S2	+	++	+
COG0092	Ribosomal protein S3	++	++	++
COG0522	Ribosomal protein S4 and related proteins	+	+	+
COG0098	Ribosomal protein S5	+	++	+
COG0049	Ribosomal protein S7	++	+	++
COG0096	Ribosomal protein S8	++	++	++
COG0103	Ribosomal protein S9	+	+	+
COG0172	Seryl-tRNA synthetase	+	+	+
COG0441	Threonyl-tRNA synthetase	+	+	+
COG0480	Translation elongation and release factors (GTPases)	+	+	+
COG0532	Translation initiation factor 2 (GTPase)	+	++	+
COG0180	Tryptophanyl-tRNA synthetase	+	+	+
COG0525	Valyl-tRNA synthetase	+	+	+
COG0050	GTPases - translation elongation factors	+	+	+

Tabla 4. Conjunto de COGs (secuencias) altamente conservados que son comunes a dos bases de datos. “M, D y H” representan a la base de datos de Mirkin *et al.* (2003), Delaye *et al.* (2005) and Harris *et al.* (2003) respectivamente. “+” denota la presencia del COG en la base de datos. El símbolo “++” indica que por lo menos una secuencia del COG presenta evidencia de duplicación interna. Las celdas vacías representan la ausencia del COG en la base de datos. Las filas sombreadas muestran a las secuencias conservadas que están involucradas con el metabolismo del RNA.

COG ID	Activity or Function	M	D	H
COG0547	Anthranilate phosphoribosyltransferase	+	+	
COG0165	Argininosuccinate lyase	+	+	
COG0031	Cysteine synthase	+	+	
COG0112	Glycine hydroxymethyltransferase	+		+
COG0079	Histidinol-phosphate aminotransferase/Tyrosine aminotransferase	+	+	
COG0078	Ornithine carbamoyltransferase	+	+	
COG0436	PLP-dependent aminotransferases	++	+	
COG0111	Phosphoglycerate dehydrogenase and related dehydrogenases	+	+	
COG1171	Threonine dehydratase	+	+	
COG0159	Tryptophan synthase alpha chain	+	+	
COG0133	Tryptophan synthase beta chain	++	++	
COG0512	Anthranilate/para-aminobenzoate synthases component II	++	+	
COG1124	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component	+	+	
COG0444	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component	+	+	
COG0505	Carbamoylphosphate synthase small subunit	+	+	
COG0037	Predicted ATPase of the PP-loop superfamily implicated in cell cycle control	++		++
COG0481	Membrane GTPase LepA	+	+	
COG0201	Preprotein translocase subunit SecY	+		++
COG0449	Glucosamine 6-phosphate synthetase, contains amidotransferase and phosphosugar isomerase domains	+	+	
COG0463	Glycosyltransferases involved in cell wall biogenesis	++	+	
COG0438	Predicted glycosyltransferases	++	+	
COG0451	Nucleoside-diphosphate-sugar epimerases	++	+	
COG0465	ATP-dependent Zn proteases	+	+	

COG ID	Activity or Function	M	D	H
COG0533	Metal-dependent proteases with possible chaperone activity	+		+
COG0450	Peroxiredoxin	+	+	
COG1225	Peroxiredoxin	+	+	
COG0526	Thiol-disulfide isomerase and thioredoxins	++		++
COG0142	Geranylgeranyl pyrophosphate synthase	+	+	
COG0171	NAD synthase	+	+	
COG0452	Phosphopantothenoylcysteine synthetase/decarboxylase	++	+	
COG1249	Dihydrolipoamide dehydrogenase/glutathione oxidoreductase and related enzymes	+	+	
COG0056	F0F1-type ATP synthase alpha subunit	+	+	
COG0055	F0F1-type ATP synthase beta subunit	+	+	
COG0636	F0F1-type ATP synthase c subunit/Archaeal/vacuolar-type H <sup>+</sup> -ATPase subunit K	++		++
COG0114	Fumarase	+	+	
COG1052	Lactate dehydrogenase and related dehydrogenases	+	+	
COG1136	ABC-type transport systems, involved in lipoprotein release, ATPase components	++	+	
COG0456	Acetyltransferases	+	+	
COG0517	CBS domains	++	++	
COG0396	Iron-regulated ABC transporter ATPase subunit SufC	++	+	
COG0012	Predicted GTPase	++		++
COG0446	Uncharacterized NAD(FAD)-dependent dehydrogenases	++	+	
COG0491	Zn-dependent hydrolases, including glyoxylases	++	+	
COG1121	ABC-type Mn/Zn transport systems, ATPase component	++	+	
COG1117	ABC-type phosphate transport system, ATPase component	+	++	
COG0575	CDP-diglyceride synthetase	++		++
COG0020	Undecaprenyl pyrophosphate synthase	++	+	
COG0015	Adenylosuccinate lyase	++	+	
COG0540	Aspartate carbamoyltransferase, catalytic chain	+	+	
COG0167	Dihydroorotate dehydrogenase	++	++	

COG ID	Activity or Function	M	D	H
COG0518	GMP synthase - Glutamine amidotransferase domain	++	+	
COG0519	GMP synthase - PP-ATPase domain	+	+	
COG0034	Glutamine phosphoribosylpyrophosphate amidotransferase	+	+	
COG0516	IMP dehydrogenase/GMP reductase	+	++	
COG0461	Orotate phosphoribosyltransferase	++	++	
COG0005	Purine nucleoside phosphorylase	+	+	
COG0127	Xanthosine triphosphate pyrophosphatase	+	+	
COG0462	Phosphoribosylpyrophosphate synthetase	+	++	
COG0258	5'-3' exonuclease (including N-terminal domain of PolI)	+		+
COG0470	ATPase involved in DNA replication	++		++
COG0592	DNA polymerase sliding clamp subunit (PCNA homolog)	+		+
COG0177	Predicted EndoIII-related endonuclease	++	++	
COG0468	RecA/RadA recombinase	+		+
COG0164	Ribonuclease HII	+	+	
COG0513	Superfamily II DNA and RNA helicases	++	++	
COG1131	ABC-type multidrug transport system, ATPase component	++	++	
COG1132	ABC-type multidrug/protein/lipid transport system, ATPase component	++	+	
COG0126	3-phosphoglycerate kinase	+	++	
COG1130	ABC-type sugar/spermidine/putrescine/iron/thiamine transport systems, ATPase component	+	+	
COG0148	Enolase	+	++	
COG0469	Pyruvate kinase	+	++	
COG0202	DNA-directed RNA polymerase alpha subunit/40 kD subunit	+		+
COG0250	Transcription antiterminator	++		++
COG0173	Aspartyl-tRNA synthetase	+	+	
COG0215	CysteinyI-tRNA synthetase	+	+	
COG1190	Lysyl-tRNA synthetase class II	+	+	
COG0130	Pseudouridine synthase	+	++	

COG ID	Activity or Function	M	D	H
COG0101	Pseudouridylate synthase (tRNA psi55)	+		+
COG0244	Ribosomal protein L10	+		+
COG0102	Ribosomal protein L13	+		+
COG0200	Ribosomal protein L15	+		+
COG0197	Ribosomal protein L16/L10E	+		+
COG0256	Ribosomal protein L18	++		++
COG0091	Ribosomal protein L22	++		++
COG0089	Ribosomal protein L23	++		++
COG0198	Ribosomal protein L24	++		++
COG0255	Ribosomal protein L29	++		++
COG0087	Ribosomal protein L3	+		+
COG0088	Ribosomal protein L4	+		+
COG0539	Ribosomal protein S1	++	++	
COG0199	Ribosomal protein S14	++		++
COG0184	Ribosomal protein S15P/S13E	++		++
COG0186	Ribosomal protein S17	+		+
COG0231	Translation elongation factor P/translation initiation factor eIF-5A	++		++
COG0361	Translation initiation factor IF-1	+		+
COG0162	Tyrosyl-tRNA synthetase	+		+

Tabla 5. Conjunto de COGs (secuencias) altamente conservados que son exclusivos para una sola bases de datos. “M, D y H” representan a la base de datos de Mirkin *et al.* (2003), Delaye *et al.* (2005) and Harris *et al.* (2003) respectivamente. “+” denota la presencia del COG en la base de datos. El símbolo “++” indica que por lo menos una secuencia del COG presenta evidencia de duplicación interna. Las celdas vacías representan la ausencia del COG en la base de datos. Las filas sombreadas muestran a las secuencias conservadas que están involucradas con el metabolismo del RNA.

COG ID	Activity or Function	M	D	H
COG0337	3-dehydroquinate synthetase	+		
COG0065	3-isopropylmalate dehydratase large subunit	+		
COG0066	3-isopropylmalate dehydratase small subunit	+		
COG0685	5,10-methylenetetrahydrofolate reductase	+		
COG0128	5-enolpyruvylshikimate-3-phosphate synthase	++		
COG1176	ABC-type spermidine/putrescine transport system, permease component I	++		
COG1177	ABC-type spermidine/putrescine transport system, permease component II	+		
COG0040	ATP phosphoribosyltransferase (histidine biosynthesis)	+		
COG0440	Acetolactate synthase, small subunit	+		
COG0548	Acetylglutamate kinase	+		
COG0002	Acetylglutamate semialdehyde dehydrogenase	++		
COG0624	Acetylmithine deacetylase/Succinyl-diaminopimelate desuccinylase and related deacylases	+		
COG0531	Amino acid transporters	++		
COG0137	Argininosuccinate synthase	+		
COG0136	Aspartate-semialdehyde dehydrogenase	+		
COG0527	Aspartokinases	+		
COG1605	Chorismate mutase	++		
COG0082	Chorismate synthase	+		
COG0626	Cystathionine beta-lyases/cystathionine gamma-synthases	++		
COG1104	Cysteine sulfinate desulfinate/cysteine desulfurase and related enzymes	++		
COG0019	Diaminopimelate decarboxylase	+		
COG0014	Gamma-glutamyl phosphate reductase	+		
COG0263	Glutamate 5-kinase	+		

COG ID	Activity or Function	M	D	H
COG0334	Glutamate dehydrogenase/leucine dehydrogenase	+		
COG0067	Glutamate synthase domain 1	+		
COG0069	Glutamate synthase domain 2	+		
COG0070	Glutamate synthase domain 3	++		
COG0118	Glutamine amidotransferase	++		
COG0174	Glutamine synthase	+		
COG0509	Glycine cleavage system H protein (lipoate-binding)	+		
COG0404	Glycine cleavage system T protein (aminomethyltransferase)	++		
COG1003	Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain	+		
COG0403	Glycine cleavage system protein P (pyridoxal-binding), N-terminal domain	+		
COG0665	Glycine/D-amino acid oxidases (deaminating)	+		
COG0141	Histidinol dehydrogenase	+		
COG0460	Homoserine dehydrogenase	++		
COG0083	Homoserine kinase	+		
COG0131	Imidazoleglycerol-phosphate dehydratase	++		
COG0107	Imidazoleglycerol-phosphate synthase	++		
COG0134	Indole-3-glycerol phosphate synthase	++		
COG0473	Isocitrate/isopropylmalate dehydrogenase	+		
COG0119	Isopropylmalate/homocitrate/citramalate synthases	+		
COG0620	Methionine synthase II (cobalamin-independent)	+		
COG2873	O-acetylhomoserine sulfhydrylase	+		
COG0160	PLP-dependent aminotransferases	+		
COG0139	Phosphoribosyl-AMP cyclohydrolase	+		
COG0140	Phosphoribosyl-ATP pyrophosphohydrolase	+		
COG0135	Phosphoribosylanthranilate isomerase	++		
COG0106	Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase	++		
COG0077	Prephenate dehydratase	+		

COG ID	Activity or Function	M	D	H
COG0287	Prephenate dehydrogenase	+		
COG0345	Pyrroline-5-carboxylate reductase	+		
COG0520	Selenocysteine lyase	+		
COG0075	Serine-pyruvate aminotransferase/archaeal aspartate aminotransferase	+		
COG0169	Shikimate 5-dehydrogenase	++		
COG0703	Shikimate kinase	+		
COG0421	Spermidine synthase	+		
COG0498	Threonine synthase	+		
COG0145	N-methylhydantoinase A		+	
COG0410	ABC-type branched-chain amino acid transport systems, ATPase component		+	
COG0411	ABC-type branched-chain amino acid transport systems, ATPase component		+	
COG1027	Aspartate ammonia-lyase		+	
COG1125	ABC-type proline/glycine betaine transport systems, ATPase components		+	
COG1126	ABC-type polar amino acid transport system, ATPase component		+	
COG1168	PLP-dependent aminotransferase		+	
COG3842	ABC-type spermidine/putrescine transport systems, ATPase components		+	
COG4161	arginine transporter ATP-binding subunit		+	
COG4598	histidine/lysine/arginine/ornithine transporter subunit		+	
COG4608	oligopeptide ABC transporter ATP-binding protein		+	
COG0329	Dihydrodipicolinate synthase/N-acetylneuraminate lyase	+		
COG0147	Anthranilate/para-aminobenzoate synthases component I	+		
COG0115	Branched-chain amino acid aminotransferase/4-amino-4-deoxychorismate lyase	++		
COG0059	Ketol-acid reductoisomerase	+		
COG0591	Na <sup>+</sup> /proline, Na <sup>+</sup> /panthothenate symporters and related permeases	++		
COG1387	Histidinol phosphatase and related hydrolases of the PHP family	++		
COG0493	NADPH-dependent glutamate synthase beta chain and related oxidoreductases	+		
COG1063	Threonine dehydrogenase and related Zn-dependent dehydrogenases	+		



COG ID	Activity or Function	M	D	H
COG0747	ABC-type dipeptide/oligopeptide/nickel transport systems, periplasmic components	+		
COG0601	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	++		
COG1173	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	++		
COG1101	Various ABC transport systems, ATPase components		+	
COG0458	Carbamoylphosphate synthase large subunit (split gene in MJ)	++		
COG0129	Dihydroxyacid dehydratase/phosphogluconate dehydratase	+		
COG1192	ATPases involved in chromosome partitioning	++		
COG0489	ATPases involved in chromosome partitioning	+		
COG0206	Cell division GTPase	++		
COG1196	Chromosome segregation ATPases	++		
COG1077	HSP70 class molecular chaperones involved in cell morphogenesis	++		
COG0239	Integral membrane protein possibly involved in chromosome condensation	++		
COG0445	NAD/FAD-utilizing enzyme apparently involved in cell division	+		
COG0424	Nucleotide-binding protein implicated in inhibition of septum formation	+		
COG2884	Predicted ATPase involved in cell division		+	
COG0706	Preprotein translocase subunit YidC	+		
COG0681	Signal peptidase I	+		
COG1989	Signal peptidase, cleaves prepilin-like proteins	+		
COG1253	Hemolysins and related proteins containing CBS domains	+		
COG1157	Flagellar biosynthesis/type III secretory pathway ATPase		+	
COG3839	ABC-type sugar transport systems, ATPase components		+	
COG0616	Periplasmic serine proteases (ClpP class)	+		
COG0750	Predicted membrane-associated Zn-dependent proteases 1	+		
COG0729	Predicted outer membrane protein	+		
COG0794	Predicted sugar phosphate isomerase involved in capsule formation	+		
COG0668	Small-conductance mechanosensitive channel	++		
COG0472	UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N- acetylglucosamine-1-phosphate transferase	++		

COG ID	Activity or Function	M	D	H
COG1087	UDP-glucose 4-epimerase		+	
COG1088	dTDP-D-glucose 4,6-dehydratase		+	
COG1207	N-acetylglucosamine-1-phosphate uridyltransferase (contains nucleotidyltransferase and I-patch acetyltransferase domains)		+	
COG1209	dTDP-glucose pyrophosphorylase		+	
COG1210	UDP-glucose pyrophosphorylase		+	
COG2222	Predicted phosphosugar isomerases		+	
COG1208	Nucleoside-diphosphate-sugar pyrophosphorylases involved in lipopolysaccharide biosynthesis/translation initiation factor eIF2B subunits	+		
COG0466	ATP-dependent Lon protease, bacterial type	+		
COG1219	ATP-dependent protease Clp, ATPase subunit	+		
COG0542	ATPases with chaperone activity, ATP-binding subunit	+		
COG0234	Co-chaperonin GroES (HSP10)	+		
COG0229	Conserved domain frequently associated with peptide methionine sulfoxide reductase	+		
COG0695	Glutaredoxin and related proteins	+		
COG0330	Membrane protease subunits, stomatin/prohibitin homologs	++		
COG0443	Molecular chaperone	++		
COG0071	Molecular chaperone (small heat shock protein)	++		
COG0576	Molecular chaperone GrpE (heat shock protein)	+		
COG0484	Molecular chaperones (contain C-terminal Zn finger domain)	+		
COG0760	Parvulin-like peptidyl-prolyl isomerase	++		
COG0225	Peptide methionine sulfoxide reductase	+		
COG0652	Peptidyl-prolyl cis-trans isomerase (rotamase) - cyclophilin family	++		
COG0638	Proteasome protease subunit	++		
COG1404	Subtilisin-like serine proteases	+		
COG0265	Trypsin-like serine proteases, typically periplasmic, contain C-terminal PDZ domain	+		
COG0501	Zn-dependent protease with chaperone function	++		
COG0108	3,4-dihydroxy-2-butanone 4-phosphate synthase	+		
COG0163	3-polyprenyl-4-hydroxybenzoate decarboxylase	++		

COG ID	Activity or Function	M	D	H
COG0043	3-polyprenyl-4-hydroxybenzoate decarboxylase and related decarboxylases	+		
COG0382	4-hydroxybenzoate polyprenyltransferase and related prenyltransferases	++		
COG0190	5,10-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase	+		
COG0212	5-formyltetrahydrofolate cyclo-ligase	++		
COG0801	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase	+		
COG0156	7-keto-8-aminopelargonate synthetase and related enzymes	+		
COG0161	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	+		
COG0502	Biotin synthase and related enzymes	+		
COG0340	Biotin-(acetyl-CoA carboxylase) ligase	++		
COG0113	Delta-aminolevulinic acid dehydratase	+		
COG0237	Dephospho-CoA kinase	++		
COG0262	Dihydrofolate reductase	++		
COG0294	Dihydropteroate synthase and related enzymes	++		
COG0476	Dinucleotide-utilizing enzymes involved in molybdopterin and thiamine biosynthesis family 2	+		
COG0196	FAD synthase	+		
COG0285	Folypolyglutamate synthase	+		
COG0807	GTP cyclohydrolase II	++		
COG0001	Glutamate-l-semialdehyde aminotransferase	++		
COG0373	Glutamyl-tRNA reductase	+		
COG0351	Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase	++		
COG0413	Ketopantoate hydroxymethyltransferase	++		
COG0320	Lipoate synthase	+		
COG0095	Lipoate-protein ligase A	+		
COG2226	Methylase involved in ubiquinone/menaquinone biosynthesis	+		
COG0846	NAD-dependent protein deacetylases, SIR2 family	+		
COG0157	Nicotinate-nucleotide pyrophosphorylase	+		
COG1057	Nicotinic acid mononucleotide adenylyltransferase	++		

COG ID	Activity or Function	M	D	H
COG1488	Nicotinic acid phosphoribosyltransferase	+		
COG0414	Panthenate synthetase	++		
COG0181	Porphobilinogen deaminase	+		
COG0276	Protoheme ferro-lyase (ferrochelatase)	+		
COG0117	Pyrimidine deaminase	++		
COG1985	Pyrimidine reductase, riboflavin biosynthesis	+		
COG0307	Riboflavin synthase alpha chain	++		
COG0054	Riboflavin synthase beta-chain	++		
COG0499	S-adenosylhomocysteine hydrolase	+		
COG0192	S-adenosylmethionine synthetase	+		
COG0422	Thiamine biosynthesis protein ThiC	+		
COG0407	Uroporphyrinogen-III decarboxylase	++		
COG1587	Uroporphyrinogen-III synthase	+		
COG3840	ABC-type thiamine transport systems, ATPase components		+	
COG4138	vitamin B12-transporter ATPase		++	
COG0543	2-polyprenylphenol hydroxylase and related flavodoxin oxidoreductases	++		
COG1060	Thiamine biosynthesis enzyme ThiH and related uncharacterized enzymes	++		
COG1048	Aconitase A	+		
COG1454	Alcohol dehydrogenase IV	+		
COG0372	Citrate synthase	+		
COG1290	Cytochrome b subunit of the bc complex	++		
COG1271	Cytochrome bd-type quinol oxidase, subunit 1	+		
COG0644	Dehydrogenases (flavoproteins)	+		
COG0508	Dihydrolipoamide acyltransferases	++		
COG0356	F0F1-type ATP synthase a subunit	+		
COG0712	F0F1-type ATP synthase delta subunit (mitochondrial oligomycin sensitivity protein)	++		
COG0355	F0F1-type ATP synthase epsilon subunit (mitochondrial delta subunit)	++		

COG ID	Activity or Function	M	D	H
COG0224	F <sub>0</sub> F <sub>1</sub> -type ATP synthase gamma subunit	+		
COG0277	FAD/FMN-containing dehydrogenases	+		
COG0633	Ferredoxin	++		
COG1143	Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)	++		
COG0240	Glycerol 3-phosphate dehydrogenase	+		
COG0554	Glycerol kinase	+		
COG0584	Glycerophosphoryl diester phosphodiesterase	+		
COG0221	Inorganic pyrophosphatase	+		
COG0538	Isocitrate dehydrogenases	+		
COG1304	L-lactate dehydrogenase (FMN-dependent) and related alpha-hydroxy acid dehydrogenases	+		
COG0039	Malate/lactate dehydrogenases	+		
COG0281	Malic enzyme	+		
COG1012	NAD-dependent aldehyde dehydrogenases	++		
COG1034	NADH dehydrogenase/NADH:ubiquinone oxidoreductase 75 kD subunit (chain G)	+		
COG0377	NADH:ubiquinone oxidoreductase 20 kD subunit and related Fe-S oxidoreductases	+		
COG0852	NADH:ubiquinone oxidoreductase 27 kD subunit	+		
COG0649	NADH:ubiquinone oxidoreductase 49 kD subunit 7	+		
COG1301	Na <sup>+</sup> /H <sup>+</sup> -dicarboxylate symporters	+		
COG0822	NifU homologs involved in Fe-S cluster formation	+		
COG0778	Nitroreductase	+		
COG0667	Predicted oxidoreductases (related to aryl-alcohol dehydrogenases)	++		
COG0674	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, alpha subunit	+		
COG1013	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, beta subunit	+		
COG1014	Pyruvate:ferredoxin oxidoreductase and related 2-oxoacid:ferredoxin oxidoreductases, gamma subunit	++		
COG0723	Rieske Fe-S protein	+		
COG0479	Succinate dehydrogenase/fumarate reductase Fe-S protein	+		
COG2009	Succinate dehydrogenase/fumarate reductase cytochrome b subunit	++		

COG ID	Activity or Function	M	D	H
COG1053	Succinate dehydrogenase/fumarate reductase, flavoprotein subunits	+		
COG0074	Succinyl-CoA synthetase alpha subunit	+		
COG0045	Succinyl-CoA synthetase beta subunit	+		
COG1251	NAD(P)H-nitrite reductase		+	
COG4987	cysteine/glutathione ABC transporter membrane/ATP-binding component		+	
COG4988	cysteine/glutathione ABC transporter membrane/ATP-binding component		+	
COG0062	Uncharacterized ACR	++		
COG1354	Uncharacterized ACR	++		
COG1739	Uncharacterized ACR	++		
COG0327	Uncharacterized ACR	+		
COG0432	Uncharacterized ACR	+		
COG2078	Uncharacterized ACR	+		
COG0011	Uncharacterized ACR	+		
COG0391	Uncharacterized ACR	++		
COG1432	Uncharacterized ACR	+		
COG0217	Uncharacterized ACR	+		
COG1624	Uncharacterized ACR	++		
COG2110	Uncharacterized ACR related to the C-terminal domain of histone macroH2A1	+		
COG0842	ABC-type multidrug transport system, permease component	++		
COG0110	Acetyltransferases (the isoleucine patch superfamily)	++		
COG1853	Conserved protein/domain typically associated with flavoprotein oxygenases, DIM6/NTAB family	+		
COG2070	Dioxygenases related to 2-nitropropane dioxygenase	+		
COG0618	Exopolyphosphatase-related proteins	+		
COG2262	GTPases	+		
COG1073	Hydrolases of the alpha/beta superfamily	+		
COG1234	Metal-dependent hydrolases of the beta-lactamase superfamily III	++		
COG0714	MoxR-like ATPases	++		

COG ID	Activity or Function	M	D	H
COG1881	Phospholipid-binding protein	+		
COG1832	Predicted CoA-binding protein	+		
COG0486	Predicted GTPase	+		
COG0536	Predicted GTPase	+		
COG0218	Predicted GTPases	+		
COG1418	Predicted HD superfamily hydrolase	+		
COG1611	Predicted Rossmann fold nucleotide-binding protein	++		
COG0220	Predicted S-adenosylmethionine-dependent methyltransferase	+		
COG0042	Predicted TIM-barrel enzymes, possibly dehydrogenases, nifR3 family	+		
COG2220	Predicted Zn-dependent hydrolases of the beta-lactamase fold	++		
COG0612	Predicted Zn-dependent peptidases	++		
COG0388	Predicted amidohydrolase	+		
COG0579	Predicted dehydrogenase	+		
COG0673	Predicted dehydrogenases and related proteins	+		
COG1355	Predicted dioxygenase	+		
COG0325	Predicted enzyme with a TIM-barrel fold	+		
COG1752	Predicted esterase of the alpha-beta hydrolase superfamily	+		
COG1011	Predicted hydrolases of the HAD superfamily	++		
COG0561	Predicted hydrolases of the HAD superfamily	++		
COG0596	Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily)	++		
COG0061	Predicted kinase	+		
COG0719	Predicted membrane components of an uncharacterized iron-regulated ABC-type transporter SufB	+		
COG1266	Predicted metal-dependent membrane protease	++		
COG0628	Predicted permease	++		
COG0730	Predicted permeases	++		
COG0637	Predicted phosphatase/phosphohexomutase	+		
COG0546	Predicted phosphatases	+		

COG ID	Activity or Function	M	D	H
COG0824	Predicted thioesterase	++		
COG0693	Putative intracellular protease/amidase	+		
COG0496	Survival protein, predicted acid phosphatase	+		
COG0457	TPR-repeat-containing proteins	++		
COG1268	Uncharacterized ACR	++		
COG0705	Uncharacterized membrane protein (homolog of Drosophila rhomboid)	++		
COG1546	Uncharacterized protein (competence- and mitomycin-induced)	++		
COG1994	Zn-dependent proteases	++		
COG0488	ATPase components of ABC transporters with duplicated ATPase domains		++	
COG1135	Uncharacterized ABC-type transport system ATPase component		+	
COG1201	Lhr-like helicases		+	
COG1341	Predicted GTPase or GTP-binding protein		+	
COG4172	putative ATP-binding component of a transport system		++	
COG4178	putative ATP-binding component of a transport system		+	
COG4619	putative ABC transporter ATP-binding protein YbbL		+	
COG0803	ABC-type Mn/Zn transport system, periplasmic Mn/Zn-binding (lipo)protein (surface adhesin A)	+		
COG1108	ABC-type Mn <sup>2+</sup> /Zn <sup>2+</sup> transport systems, permease components	++		
COG0226	ABC-type phosphate transport system, periplasmic component	+		
COG0581	ABC-type phosphate transport system, permease component	++		
COG0573	ABC-type phosphate transport system, permease component	++		
COG0529	Adenylylsulfate kinase and related kinases	++		
COG0004	Ammonia permeases	+		
COG0530	Ca <sup>2+</sup> /Na <sup>+</sup> antiporter	++		
COG2217	Cation transport ATPases	+		
COG1230	Co/Zn/Cd efflux system component	++		
COG2608	Copper chaperone	++		
COG0471	Di- and tricarboxylate transporters	++		



COG ID	Activity or Function	M	D	H
COG0569	K <sup>+</sup> transport systems, NAD-binding component	++		
COG0475	Kef-type K <sup>+</sup> transport systems, membrane components	++		
COG1226	Kef-type K <sup>+</sup> transport systems, predicted NAD-binding component	++		
COG0598	Mg <sup>2+</sup> and Co <sup>2+</sup> transporters	+		
COG0704	Phosphate uptake regulator	++		
COG0306	Phosphate/sulphate permeases	++		
COG0053	Predicted Co/Zn/Cd cation transporters	++		
COG0428	Predicted divalent heavy-metal cations transporter	++		
COG0607	Rhodanese-related sulfurtransferases	++		
COG0605	Superoxide dismutase	+		
COG0168	Trk-type K <sup>+</sup> transport systems, membrane components	+		
COG0155	Sulfite reductase hemoprotein beta-component		+	
COG1116	ABC-type nitrate/sulfonate/taurine/bicarbonate transport systems, ATPase components		+	
COG1118	ABC-type sulfate/molybdate transport systems, ATPase component		+	
COG1119	ABC-type molybdenum transport system, ATPase component/photorepair protein PhrA		++	
COG3638	ABC-type phosphate/phosphonate transport system, ATPase component		+	
COG3841	ABC-type iron transport systems, ATPase components		+	
COG1120	ABC-type cobalamin/Fe <sup>3+</sup> -siderophores transport systems, ATPase components		++	
COG0331	(acyl-carrier-protein) S-malonyltransferase	++		
COG0204	1-acyl-sn-glycerol-3-phosphate acyltransferase	+		
COG2084	3-hydroxyisobutyrate dehydrogenase and related proteins	+		
COG0183	Acetyl-CoA acetyltransferases	+		
COG0825	Acetyl-CoA carboxylase alpha subunit	+		
COG0365	Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases	+		
COG0511	Biotin carboxyl carrier protein	++		
COG0439	Biotin carboxylase	+		
COG1024	Enoyl-CoA hydratase/carnithine racemase	+		

COG ID	Activity or Function	M	D	H
COG0671	Membrane-associated phospholipid phosphatase	++		
COG1260	Myo-inositol-1-phosphate synthase	+		
COG0558	Phosphatidylglycerophosphate synthase	++		
COG0688	Phosphatidylserine decarboxylase	+		
COG1183	Phosphatidylserine synthase	++		
COG1502	Phosphatidylserine/phosphatidylglycerophosphate/cardiolipin synthases and related enzymes	++		
COG0736	Phosphopantetheinyl transferase (holo-ACP synthase)	++		
COG0304	3-oxoacyl-(acyl-carrier-protein) synthase	+		
COG0236	Acyl carrier protein	++		
COG0318	Acyl-CoA synthetases (AMP-forming)/AMP-acid ligases II	++		
COG0737	5'-nucleotidase/2',3'-cyclic phosphodiesterase and related esterases	+		
COG0138	AICAR transformylase/IMP cyclohydrolase PurH (only IMP cyclohydrolase domain in Aful)	+		
COG0503	Adenine/guanine phosphoribosyltransferases and related PRPP-binding proteins	++		
COG0563	Adenylate kinase and related kinases	++		
COG0104	Adenylosuccinate synthase	+		
COG0504	CTP synthase (UTP-ammonia lyase)	+		
COG0295	Cytidine deaminase	+		
COG0274	Deoxyribose-phosphate aldolase	+		
COG0044	Dihydroorotase and related cyclic amidohydrolases	+		
COG0299	Folate-dependent phosphoribosylglycinamide formyltransferase PurN	+		
COG0194	Guanylate kinase	+		
COG0105	Nucleoside diphosphate kinase	++		
COG0284	Orotidine-5'-phosphate decarboxylase	++		
COG0151	Phosphoribosylamine-glycine ligase	+		
COG0150	Phosphoribosylaminoimidazol (AIR) synthetase	+		
COG0026	Phosphoribosylaminoimidazole carboxylase (NCAIR synthetase)	+		
COG0152	Phosphoribosylaminoimidazolesuccinocarboxamide (SAICAR) synthase	+		

COG ID	Activity or Function	M	D	H
COG0041	Phosphoribosylcarboxyaminoimidazole (NCAIR) mutase	++		
COG1828	Phosphoribosylformylglycinamidine (FGAM) synthase, PurS component	+		
COG0047	Phosphoribosylformylglycinamidine (FGAM) synthase, glutamine amidotransferase domain	+		
COG0046	Phosphoribosylformylglycinamidine (FGAM) synthase, synthetase domain	++		
COG1351	Predicted alternative thymidylate synthase	+		
COG0209	Ribonucleotide reductase alpha subunit	+		
COG0208	Ribonucleotide reductase beta subunit	+		
COG0035	Uracil phosphoribosyltransferase	+		
COG0572	Uridine kinase	++		
COG0528	Uridylate kinase	+		
COG0756	dUTPase	+		
COG0402	Cytosine deaminase and related metal-dependent hydrolases	+		
COG0248	Exopolyphosphatase	+		
COG0537	Diadenosine tetraphosphate (Ap4A) hydrolase and other HIT family hydrolases	++		
COG0590	Cytosine/adenosine deaminases	+		
COG0419	ATPase involved in DNA repair	++		
COG0188	DNA gyrase (topoisomerase II) A subunit	+		
COG0187	DNA gyrase (topoisomerase II) B subunit	+		
COG0323	DNA mismatch repair enzyme (predicted ATPase)	+		
COG0749	DNA polymerase I - 3'-5' exonuclease and polymerase domains	+		
COG0847	DNA polymerase III epsilon subunit and related 3'-5' exonucleases	+		
COG0358	DNA primase (bacterial type)	+		
COG0420	DNA repair exonuclease	++		
COG0648	Endonuclease IV	+		
COG0582	Integrase	++		
COG0350	Methylated DNA-protein cysteine methyltransferase	++		
COG0084	Mg-dependent DNase	+		

COG ID	Activity or Function	M	D	H
COG1525	Micrococcal nuclease (thermonuclease) homologs	++		
COG0249	MutS-like ATPases involved in mismatch repair, family 2	+		
COG0675	Predicted transposases	++		
COG0328	Ribonuclease HI	+		
COG0629	Single-stranded DNA-binding protein	+		
COG0608	Single-stranded DNA-specific exonuclease	++		
COG0210	Superfamily I DNA and RNA helicases	+		
COG1112	Superfamily I DNA and RNA helicases and helicase subunits	+		
COG1573	Uracil-DNA glycosylase	+		
COG0551	Zn-finger domain associated with topoisomerase type I	++		
COG0178	Excinuclease ATPase subunit		+	
COG0514	Superfamily II DNA helicase		+	
COG1194	A/G-specific DNA glycosylase		+	
COG2812	DNA polymerase III, gamma/tau subunits		+	
COG0494	NTP pyrophosphohydrolases including oxidative damage repair enzymes	++		
COG1335	Amidases related to nicotinamidase	+		
COG0534	Na <sup>+</sup> -driven multidrug efflux pump	++		
COG4167	putative ATP-binding protein of peptide transport system		+	
COG4170	putative ATP-binding protein of peptide transport system		+	
COG4181	putative ABC transporter ATP-binding protein YbbA		+	
COG1028	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	++		
COG0500	SAM-dependent methyltransferases	++		
COG0784	CheY-like receiver domain	++		
COG0639	Diadenosine tetraphosphatase and related serine/threonine protein phosphatases	+		
COG2203	GAF domain	++		
COG0631	Protein serine/threonine phosphatases	+		
COG0394	Protein-tyrosine-phosphatase	+		

COG ID	Activity or Function	M	D	H
COG0467	RecA-superfamily ATPases implicated in signal transduction	++		
COG0642	Signal transduction histidine kinase	++		
COG0515	Serine/threonine protein kinases	+		
COG0589	Universal stress protein UspA and related nucleotide-binding proteins	++		
COG0664	cAMP-binding domains - Catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	++		
COG0205	6-phosphofructokinase	+		
COG0363	6-phosphogluconolactonase/Glucosamine-6-phosphate isomerase/deaminase	+		
COG0483	Archaeal fructose-1,6-bisphosphatase and related enzymes of inositol monophosphatase family	+		
COG1472	Beta-glucosidase-related glycosidases	+		
COG0191	Fructose/tagatose bisphosphate aldolase	++		
COG0364	Glucose-6-phosphate 1-dehydrogenase	+		
COG0166	Glucose-6-phosphate isomerase	+		
COG0057	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase	+		
COG0036	Pentose-5-phosphate-3-epimerase	+		
COG0574	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase	+		
COG0063	Predicted sugar kinase	++		
COG0647	Predicted sugar phosphatases of the HAD superfamily	+		
COG0120	Ribose 5-phosphate isomerase	+		
COG0235	Ribulose-5-phosphate 4-epimerase and related epimerases and aldolases	++		
COG0176	Transaldolase	+		
COG0021	Transketolase	+		
COG0149	Triosephosphate isomerase	+		
COG0033	Phosphoglucomutase		+	
COG0448	ADP-glucose pyrophosphorylase		+	
COG1129	ABC-type sugar (aldose) transport system, ATPase component		++	
COG0697	Permeases of the drug/metabolite transporter (DMT) superfamily	++		
COG0477	Permeases of the major facilitator superfamily	+		

COG ID	Activity or Function	M	D	H
COG1758	DNA-directed RNA polymerase subunit K/omega	++		
COG0557	Exoribonucleases	+		
COG1321	Mn-dependent transcriptional regulator	+		
COG1293	Predicted RNA-binding protein homologous to eukaryotic snRNP	+		
COG0640	Predicted transcriptional regulators	++		
COG1475	Predicted transcriptional regulators	++		
COG0195	Transcription elongation factor	+		
COG1309	Transcriptional regulator	++		
COG1940	Transcriptional regulators	++		
COG0571	dsRNA-specific ribonuclease	+		
COG1167	Transcriptional regulators containing a DNA-binding HTH domain and an aminotransferase domain (MocR family) and their eukaryotic orthologs		+	
COG0454	Histone acetyltransferase HPA2 and related acetyltransferases	++		
COG0553	Superfamily II DNA/RNA helicases, SNF2 family	+		
COG0621	2-methylthioadenine synthetase	+		
COG1670	Acetyltransferases, including N-acetylases of ribosomal proteins	++		
COG0154	Asp-tRNAAsn/Glu-tRNAGln amidotransferase A subunit and related amidases	+		
COG0064	Asp-tRNAAsn/Glu-tRNAGln amidotransferase B subunit (PET112 homolog)	+		
COG0721	Asp-tRNAAsn/Glu-tRNAGln amidotransferase C subunit	+		
COG0223	Methionyl-tRNA formyltransferase	+		
COG0193	Peptidyl-tRNA hydrolase	++		
COG2519	Predicted SAM-dependent methyltransferase involved in tRNA-Met maturation	+		
COG2890	Predicted rRNA or tRNA methylase	+		
COG0482	Predicted tRNA(5-methylaminomethyl-2-thiouridylate) methyltransferase, contains the PP-loop ATPase domain	++		
COG0216	Protein chain release factor A	+		
COG1186	Protein chain release factor B	+		
COG0564	Pseudouridylate synthases, 23S RNA-specific	+		

COG ID	Activity or Function	M	D	H
COG0009	Putative translation factor (SUA5)	++		
COG0251	Putative translation initiation inhibitor	++		
COG0343	Queuine/archaeosine tRNA-ribosyltransferase	+		
COG0203	Ribosomal protein L17	++		
COG0335	Ribosomal protein L19	++		
COG0211	Ribosomal protein L27	++		
COG0227	Ribosomal protein L28	++		
COG1841	Ribosomal protein L30/L7E	++		
COG0333	Ribosomal protein L32	+		
COG0267	Ribosomal protein L33	++		
COG0230	Ribosomal protein L34	++		
COG0291	Ribosomal protein L35	++		
COG0257	Ribosomal protein L36	+		
COG0222	Ribosomal protein L7/L12	++		
COG0228	Ribosomal protein S16	+		
COG0238	Ribosomal protein S18	+		
COG0360	Ribosomal protein S6	++		
COG0233	Ribosome recycling factor	+		
COG2265	SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase	+		
COG0182	Translation initiation factor eIF-2B alpha subunit	+		
COG0566	rRNA methylases	+		
COG0144	tRNA and rRNA cytosine-C5-methylases	+		
COG0324	tRNA delta(2)-isopentenylpyrophosphate transferase	+		
COG0617	tRNA nucleotidyltransferase/poly(A) polymerase	++		
COG0017	Aspartyl/asparaginyl-tRNA synthetases		+	

La clasificación funcional de aquellas secuencias que se originaron por un evento de duplicación interna y que se encuentran altamente conservadas en las diversas bases de datos esta basada en la distribución de funciones metabólicas descritas para el conjunto de grupos de genes ortólogos (COGs) utilizados por Mirkin *et al.* (2003). Las múltiples y diversas categorías funcionales que caracterizan a la base de datos COGs permitió clasificar a las secuencias de todas las bases de datos en agrupamientos funcionales equivalentes, lo cual hubiera sido problemático ante un número reducido de categorías funcionales o ante categorías funcionales muy simples y elementales.

El agrupamiento funcional de las secuencias altamente conservadas originadas por eventos de duplicación y fusión muestra un tendencia hacia la presencia de un mayor número de genes involucrados en procesos metabólicos universales conservados. De manera no sorprendente, el mayor número de secuencias duplicadas es aquel en el que están involucrados procesos universales como el de la traducción. La presencia de un gran número de genes duplicados involucrados en dicho proceso se debe a que es la categoría con un mayor número de representantes, es decir, es la categoría con el mayor número de genes altamente conservados (Figuras 8, 9 y 10).

Debido a la enorme diferencia en el número de secuencias disponibles por proceso metabólico, en este análisis no encontramos algún sesgo evidente entre el número de genes duplicados y la función metabólica en la cual participan.



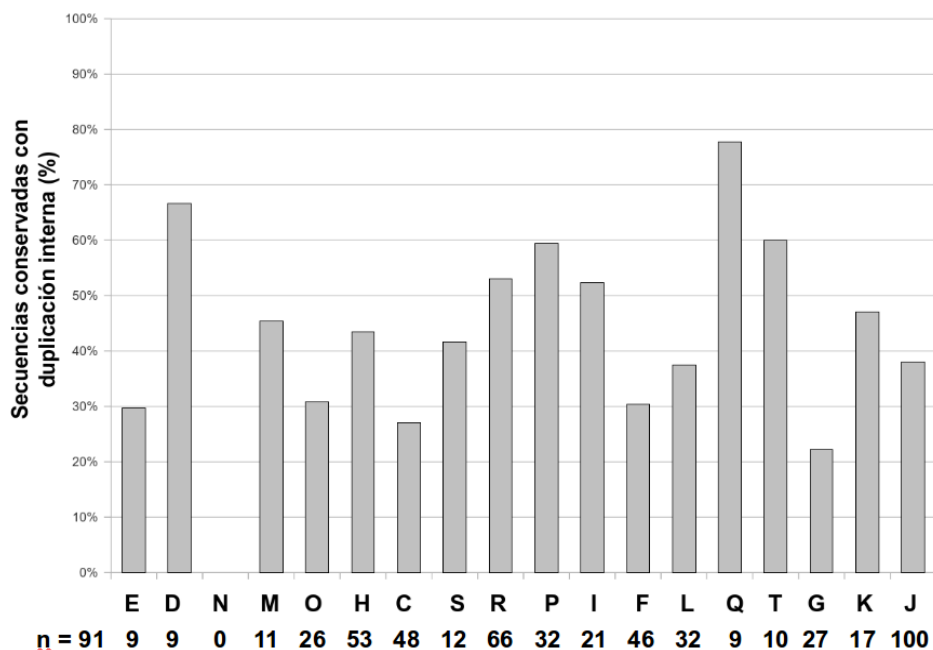


Figura 8. Distribución funcional por proceso metabólico de secuencias duplicadas en la base de datos de Mirkin *et al.* “n” indica el número total de secuencias por proceso metabólico. Los siguiente procesos fueron considerados: E – metabolismo de aminoácidos, D – división celular, N – motilidad celular, M – biogénesis de la pared celular, O – chaperonas, H – metabolismo de coenzimas, C – conversión de energía, S – función desconocida, R – función general, P – transporte de iones, I – metabolismo de lípidos, F – metabolismo de nucleótidos, L – replicación y reparación, Q – metabolismo secundario, T – transducción de señales, G – metabolismo de azúcares, K – transcripción, J – traducción.

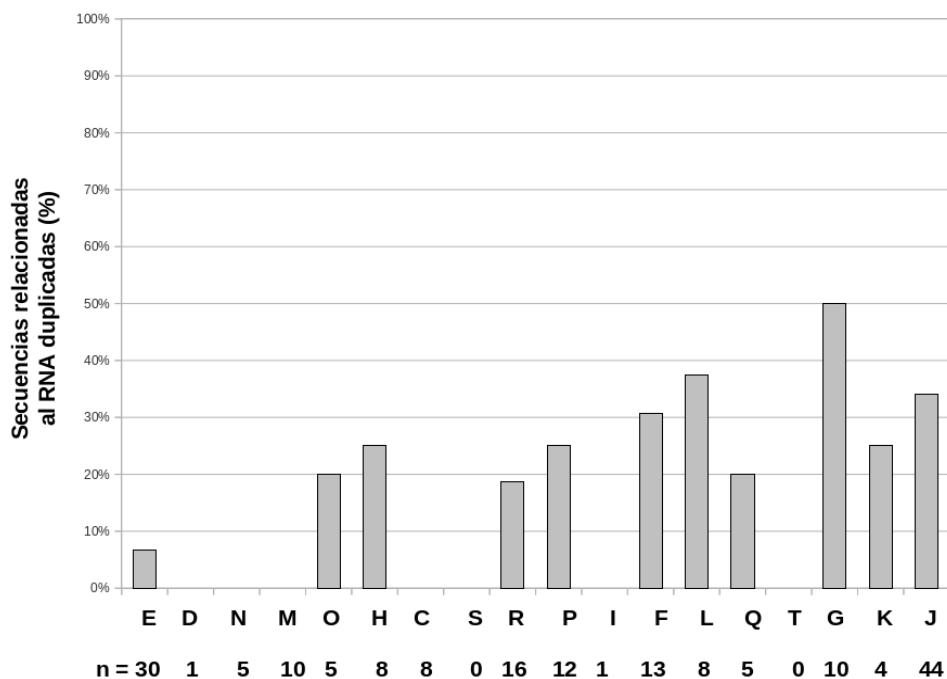


Figura 9. Distribución funcional por proceso metabólico de secuencias duplicadas en la base de datos de Delaye *et al.* “n” indica el número total de secuencias por proceso metabólico. Los siguiente procesos fueron considerados: E – metabolismo de aminoácidos, D – división celular, N – motilidad celular, M – biogénesis de la pared celular, O – chaperonas, H – metabolismo de coenzimas, C – conversión de energía, S – función desconocida, R – función general, P – transporte de iones, I – metabolismo de lípidos, F – metabolismo de nucleótidos, L – replicación y reparación, Q – metabolismo secundario, T – transducción de señales, G – metabolismo de azúcares, K – transcripción, J – traducción.

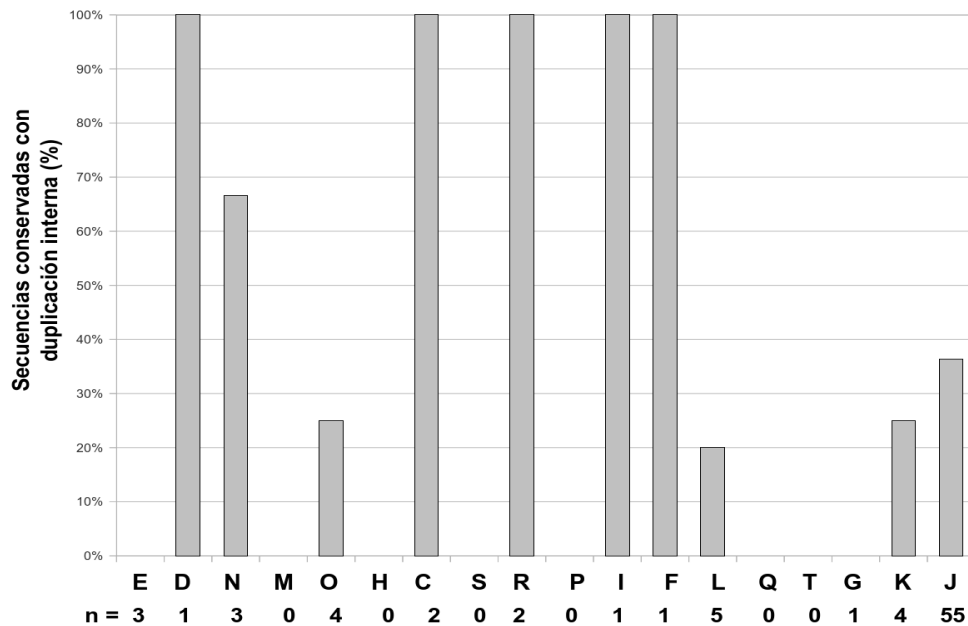


Figura 10. Distribución funcional por proceso metabólico de secuencias duplicadas en la base de datos de Harris *et al.* “n” indica el número total de secuencias por proceso metabólico. Los siguientes procesos fueron considerados: E – metabolismo de aminoácidos, D – división celular, N – motilidad celular, M – biogénesis de la pared celular, O – chaperonas, H – metabolismo de coenzimas, C – conversión de energía, S – función desconocida, R – función general, P – transporte de iones, I – metabolismo de lípidos, F – metabolismo de nucleótidos, L – replicación y reparación, Q – metabolismo secundario, T – transducción de señales, G – metabolismo de azúcares, K – transcripción, J – traducción.

## 5.2 Secuencias altamente conservadas que están relacionadas con el metabolismo del RNA y que se originaron por duplicación interna.

En este trabajo se ha intentado dilucidar el papel que ha jugado la duplicación interna de genes en la evolución temprana de las proteínas. Para ello se analizó al conjunto de secuencias que se encuentran altamente conservadas en los tres grandes linajes celulares (Archaea, Bacteria y Eucaria), con énfasis en aquellas secuencias que tienen una estrecha relación funcional con el metabolismo del RNA.

La clasificación de secuencias altamente conservadas en aquellas que se relacionan con el metabolismo del RNA muestra que aproximadamente el 34% se originó por un evento de duplicación intragénica en la base de Mirkin *et al.* (2003), alrededor de 24% en Delaye *et al.* (2005), y un 36% en Harris *et al.* (2003) (Figura 11).

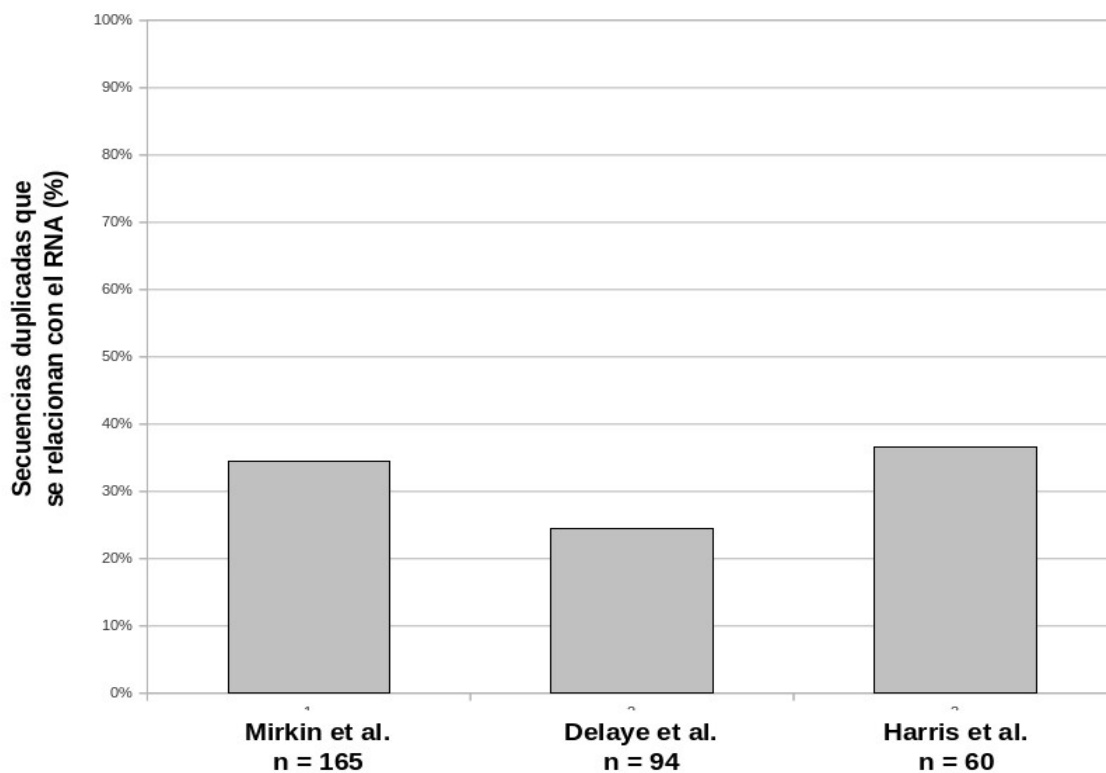


Figura 11. Porcentaje de secuencias que están relacionados con el metabolismo del RNA y que se han originado por duplicación interna en cada una de las diferentes bases de datos analizadas. “n” indica el número total de secuencias conservadas relacionadas al RNA en cada una de las bases.

Como se muestra en la figura 12, el número de secuencias que ha experimentado duplicación interna previo a la divergencia de los tres linajes celulares incluye proteínas que participan en: (1) procesos informacionales como la síntesis de proteínas, (2) procesos de regulación de la expresión del material genético, y (3) en funciones celulares relacionadas con la biosíntesis de diversas moléculas.

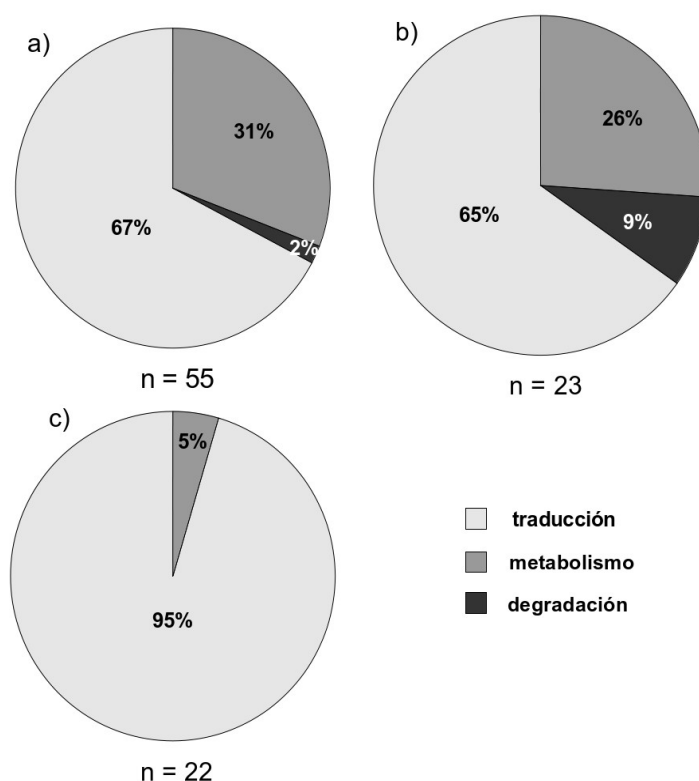


Figura 12. Distribución funcional de proteínas conservadas que se relacionan con el RNA y que originaron por duplicación interna. en moléculas que participan en procesos informacionales como la síntesis de proteínas, en funciones celulares relacionadas con la biosíntesis de diversas moléculas o en procesos de regulación de la expresión del material genético. a), b) y c) representan a la base de datos de Mirkin *et al.* (2003), Delaye *et al.* (2005) y Harris *et al.* (2003) respectivamente.

La lista de proteínas altamente conservadas, que interactúan con el RNA y que surgen mediante un evento de duplicación interna está conformada por un total de 57 secuencias en la base de datos de Mirkin *et al.* (2003), 23 secuencias polipeptídicas en la base de Delaye *et al.* (2005) y 22 secuencias en Harris *et al.* (2003) (Tablas 6, 7 y 8).

Tabla 6. Distribución funcional de proteínas altamente conservadas con duplicación interna en la base de datos de Mirkin *et al.* (2003).

Nombre de la proteína	Proceso
Acetiltransferasas, incluyen N-acetilinas de proteínas ribosomales	Traducción
Dimetiladenosina transferasa (metilación de rRNA)	Traducción
Metionina aminopeptidasa	Traducción
Peptidil-tRNA hidrolasa	Traducción
tRNA metiltransferase, contiene el dominio PP-loop ATPase	Traducción
Factor de traducción putativo (SUA5)	Traducción
Inhibidor de inicio de la traducción putativo	Traducción
Factor de elongación P/factor de inicio de la traducción eIF-5A	Traducción
tRNA nucleotidiltransferasa/poli(A) polimerasa	Traducción
Proteína ribosomal L11	Traducción
Proteína ribosomal L14	Traducción
Proteína ribosomal L17	Traducción
Proteína ribosomal L18	Traducción
Proteína ribosomal L19	Traducción

Proteína ribosomal L22	Traducción
Proteína ribosomal L23	Traducción
Proteína ribosomal L24	Traducción
Proteína ribosomal L27	Traducción
Proteína ribosomal L28	Traducción
Proteína ribosomal L29	Traducción
Proteína ribosomal L30/L7E	Traducción
Proteína ribosomal L33	Traducción
Proteína ribosomal L34	Traducción
Proteína ribosomal L35	Traducción
Proteína ribosomal L5	Traducción
Proteína ribosomal L6	Traducción
Proteína ribosomal L7/L12	Traducción
Proteína ribosomal S1	Traducción
Proteína ribosomal S11	Traducción
Proteína ribosomal S12	Traducción
Proteína ribosomal S13	Traducción
Proteína ribosomal S14	Traducción
Proteína ribosomal S15P/S13E	Traducción
Proteína ribosomal S3	Traducción
Proteína ribosomal S6	Traducción
Proteína ribosomal S7	Traducción
Proteína ribosomal S8	Traducción
Superfamilia II de DNA y RNA helicasas	Degradación
ATP sintasa subunidad c tipo F0F1/Archaeal/ATPasa subunidad K tipo vacuolar	Metabolismo
ATP sintasa subunidad delta tipo F0F1	Metabolismo
ATP sintasa subunidad epsilon tipo F0F1 (subunidad delta mitocondrial)	Metabolismo
Carbamoylfosfato sintasa subunit grande	Metabolismo
Adenina/guanina fosforibosiltransferasas y proteínas de unión a PRPP relacionadas	Metabolismo
Adenilato cinasa y cinasas relacionadas	Metabolismo
Adenilosuccinato liase	Metabolismo
Diadenosina tetrafosfato (Ap4A) hidrolasa y otras hidrolasas de la familia HIT	Metabolismo
Dihidroorotato deshidrogenasa	Metabolismo
GMP sintasa – dominio Glutamina amidotransferasa	Metabolismo
Nucleosido difosphate cinasa	Metabolismo
Orotato fosforibosiltransferasa	Metabolismo
Orotidina-5'-fosfato descarboxilasa	Metabolismo
Fosforibosilcarboxiaminoimidazol (NCAIR) mutasa	Metabolismo
Fosforibosilformilglicinamida (FGAM) sintasa, dominio sintetasa	Metabolismo
Uridina cinasa	Metabolismo
Dominios CBS	Metabolismo

Tabla 7. Distribución funcional de proteínas altamente conservadas con duplicación interna en la base de datos de Delaye *et al.* (2005).

Nombre de la Proteína	Proceso
Metionina aminopeptidasa	Traducción
Dimethyladenosina transferasa	Traducción
tRNA pseudouridina 55 sintasa	Traducción
Factor de inicio de la traducción IF-2	Traducción
Proteína ribosomal S11	Traducción
Proteína ribosomal S10	Traducción
Proteína ribosomal S5	Traducción
Proteína ribosomal S8	Traducción
Proteína ribosomal S2	Traducción
Proteína ribosomal S3	Traducción
Proteína ribosomal S13	Traducción
Proteína ribosomal L1	Traducción
Proteína ribosomal L8	Traducción
Proteína ribosomal L6	Traducción
Proteína ribosomal S1	Traducción
Enolasa	Metabolismo y Degradación
RNA helicase DeaD ATP dependiente	Transcripción, Traducción y Degradación
Orotato fosforibosiltransferasa	Metabolismo

Tioredoxina reductasa	Metabolismo
IMP deshidrogenasa	Metabolismo
Dihidroorotato oxidasa	Metabolismo
Ribosa-fosfato pirofosfocinasa	Metabolismo
Piruvato cinasa	Metabolismo

Tabla 6. Distribución funcional de proteínas altamente conservadas con duplicación interna en la base de datos de Harris *et al.* (2003).

Nombre de la Proteína	Proceso
Proteína ribosomal S12	Traducción
Proteína ribosomal S7	Traducción
Proteína ribosomal S8	Traducción
Proteína ribosomal S13	Traducción
Proteína ribosomal S11	Traducción
Proteína ribosomal S15	Traducción
Proteína ribosomal S14	Traducción
Proteína ribosomal S3	Traducción
Proteína ribosomal L29	Traducción
Proteína ribosomal L11	Traducción
Proteína ribosomal L23	Traducción
Proteína ribosomal L22	Traducción
Proteína ribosomal L14	Traducción
Proteína ribosomal L5	Traducción
Proteína ribosomal L6	Traducción
Proteína ribosomal L24	Traducción
Proteína ribosomal L18	Traducción
Proteína parecida al factor de elongación P	Traducción
Metionina aminopeptidasa	Traducción
Dimetiladenosina transferasa	Traducción
Fenilalanina tRNA sintetasa, subunidad beta	Traducción
ATP sintasa, subunit C (unión al sector F0)	Metabolismo

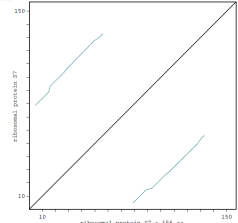
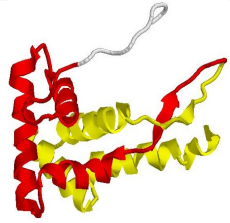
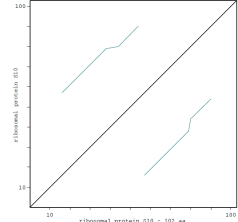
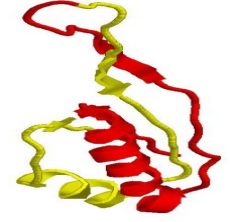
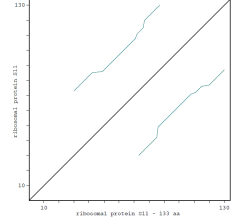
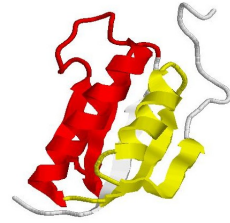
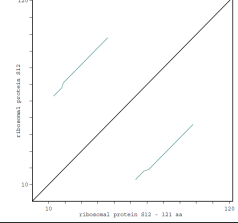
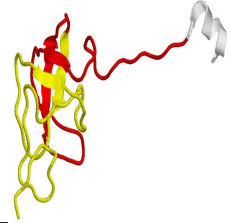
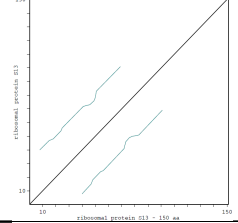
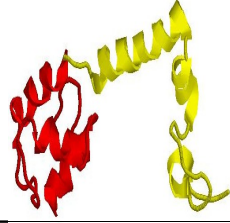
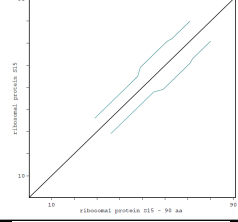
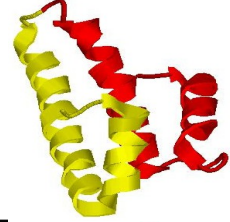
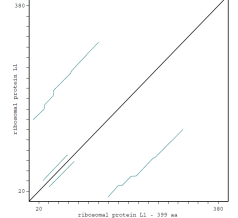
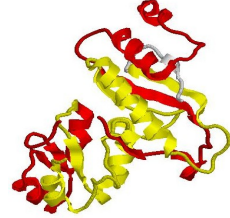
### 5.3 Evidencia de la duplicación interna en la estructura terciaria de proteínas altamente conservadas

Como es sabido, los dominios son las unidades estructurales, funcionales y evolutivas de las cuales están conformadas las proteínas (Murzin *et al.*, 1995; Orengo *et al.*, 1997; Riley y Labedan, 1997). Al parecer, en la naturaleza existe un repertorio limitado de familias de dominios (Chothia, 1992; Wolf *et al.*, 2000) los cuales han sido duplicados, recombinados, fusionados y fisionados para formar a la diversidad abundante de proteínas contemporáneas (Apic *et al.*, 2001; Chothia y Gough,

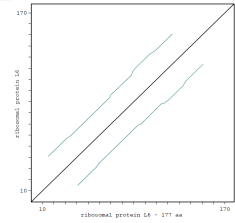

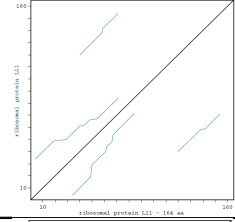
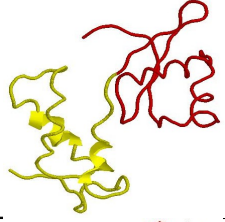
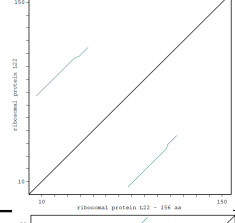
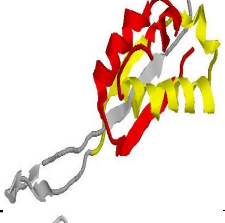
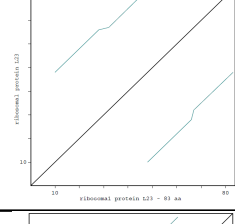
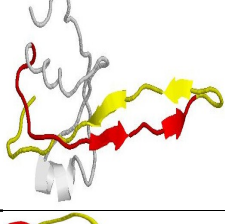
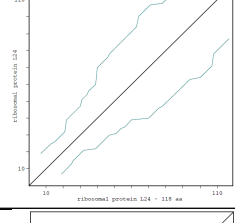
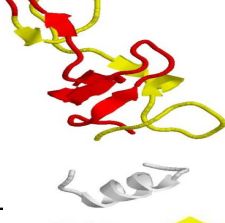
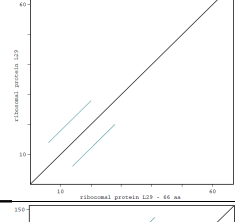
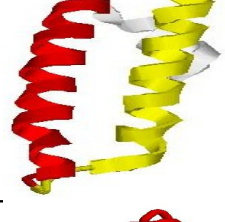
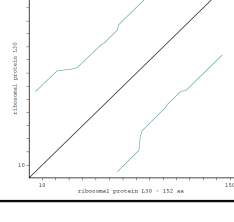
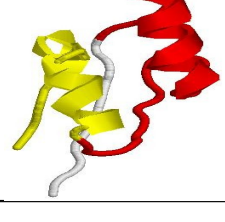
2009).

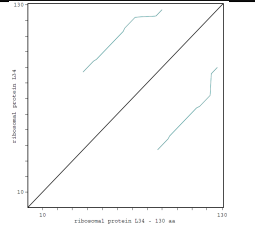

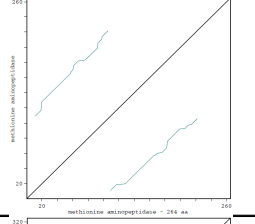
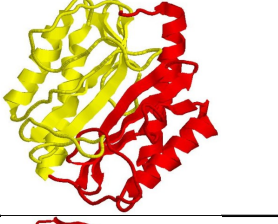
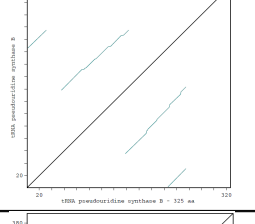
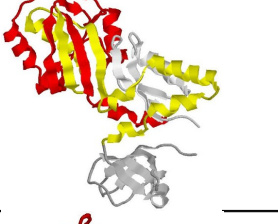
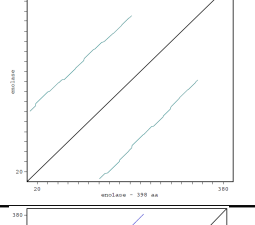
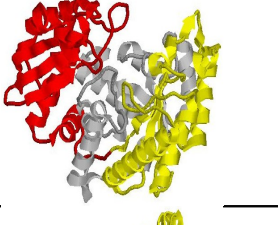
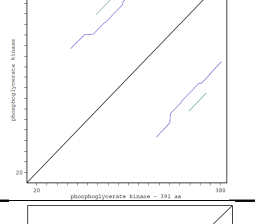
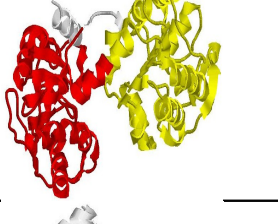
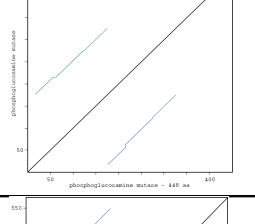
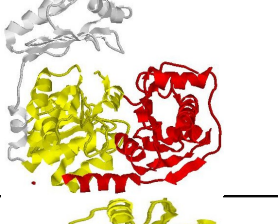
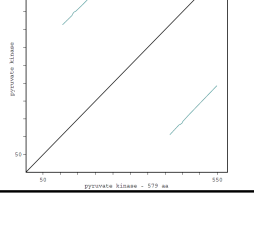

La figura 13 muestra tanto la estructura terciaria como la matriz que representa a la alineación en estructura primaria de las secuencias que han sido producto de duplicación interna. En algunos casos es claramente notable la duplicación del dominio o dominios que forman parte de la estructura terciaria de la proteína, ejemplos notables de ello son las proteínas ribosomales S5, S6, S7, S10, S11, S12, S13, S15, L1, L6, L11, L24, L29, L30 y L34 que conforman al ribosoma, la metionina aminopeptidasa, la fosfoglicerato cinasa, la piruvato cinasa, ribosa-fosfato pirofosfocinasa, timidilato cinasa, la subunidad grande de la carbamoil-fosfato sintasa, la dihidropteroato sintasa, la ferredoxina, la N-(5'-fosfo-L-ribosil-formimino)-5-amino-1-(5'-fosforibosil)-4-imidazolcarboxamida isomerasa, la imidazol glicerol fosfato sintasa, la indol-3-glicerol fosfato sintasa, la cetopantoato hidroximetiltransferasa, la fosfatidilserina cardiopina sintasa, la fosforibosilantranilato isomerasa, la fosforibosilformilglicinamidina sintasa II, la tioesterasa, la shikimato deshidrogenasa, la proteína de estrés universal UspA, la uroporfirinógeno descarboxilasa, la endonucleasa III, la proteína de unión a ATP del sistema de transporte de vitamina B12, la proteína acarreadora de acilos, y la subunidad C de la ATP sintasa (sector F0). Sin embargo, existen otros casos en los que los dominios duplicados no se evidencian a primera vista en la estructura terciaria. Posibles ejemplos de ello son las proteínas ribosomales L22, L23, la tRNA pseudouridina 55 sintasa, la enolasa, la fosfomanomutasa, la tioredoxina reductasa, la dihidroorotato deshidrogenasa, y la DNA topoisomerasa I. En estos casos, La evidencia de la duplicación interna sería más fácil de reconocer a nivel de estructura primaria, por lo que sería detectada fácilmente mediante la comparación interna de la secuencia.

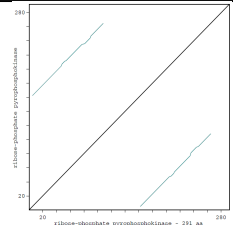
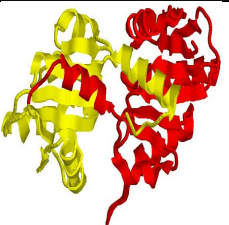
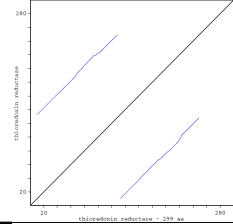
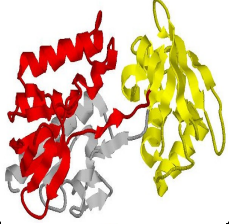
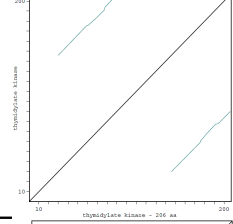
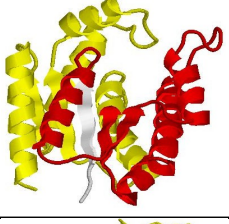
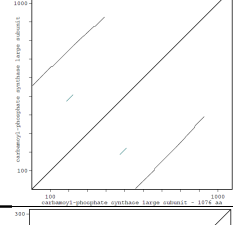
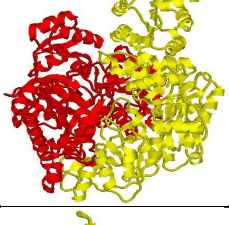
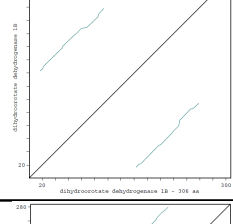
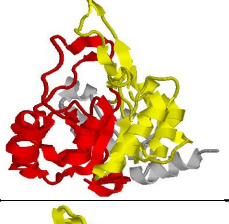
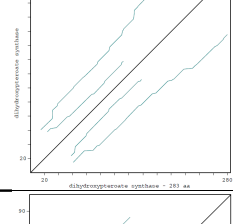
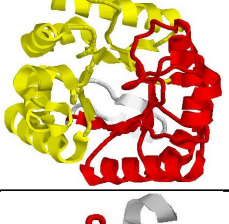
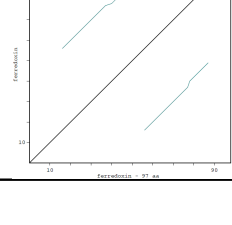
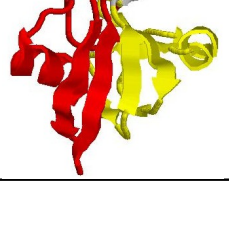
Proteína	Función	Alineación	Estructura terciaria
Proteína ribosomal S5	Ribosoma		
Proteína ribosomal S6	Ribosoma		

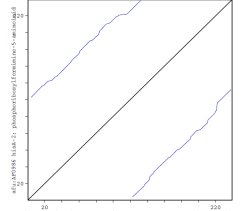
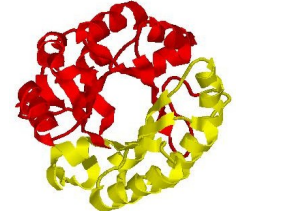
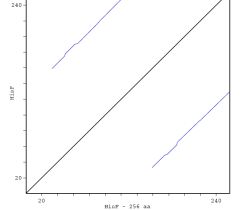
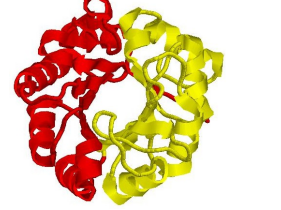
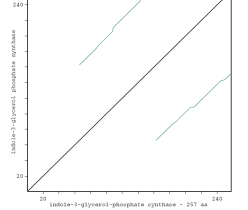
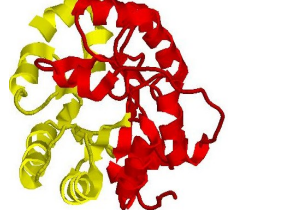
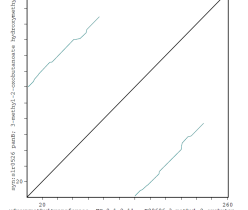

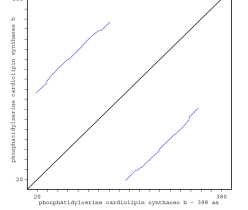
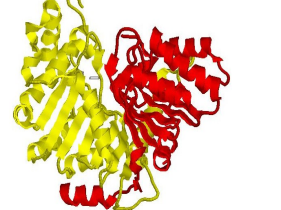
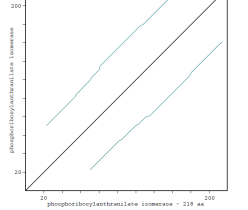
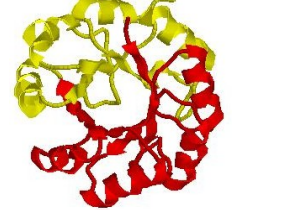
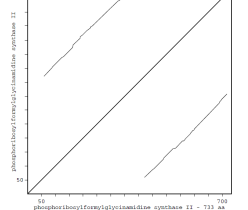
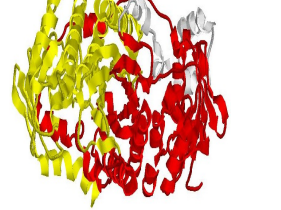
Proteína ribosomal S7	Ribosoma		
Proteína ribosomal S10	Ribosoma		
Proteína ribosomal S11	Ribosoma		
Proteína ribosomal S12	Ribosoma		
Proteína ribosomal S13	Ribosoma		
Proteína ribosomal S15	Ribosoma		
Proteína ribosomal L1	Ribosoma		



Proteína ribosomal L6	Ribosoma		
Proteína ribosomal L11	Ribosoma		
Proteína ribosomal L22	Ribosoma		
Proteína ribosomal L23	Ribosoma		
Proteína ribosomal L24	Ribosoma		
Proteína ribosomal L29	Ribosoma		
Proteína ribosomal L30	Ribosoma		

Proteína ribosomal L34	Ribosoma		
Metionina aminopeptidasa	Libera amino ácidos de la región amino-terminal, preferencialmente metionina, de péptidos nacientes		
tRNA pseudouridina sintasa	tRNA uridina $\Leftrightarrow$ tRNA pseudouridina		
Enolasa	2-fosfo-D-glicerato $\Leftrightarrow$ fosfoenolpiruvato + H2O		
Fosfoglicerato cinasa	ATP + 3-fosfo-D-glicerato $\Leftrightarrow$ ADP + 3-fosfo-D-glicerol fosfato		
Fosfomanomutasa	alpha-D-manosa 1-fosfato $\Leftrightarrow$ D-manosa 6-fosfato		
Piruvato cinasa b	ATP + piruvato $\Leftrightarrow$ ADP + fosfoenolpiruvato		

Ribosa-fosfato pirofosfocinasa	ATP + D-ribosa 5-fosfato $\rightleftharpoons$ AMP + 5-fosfo-alpha-D-ribosa 1-difosfato		
Tiorredoxina reductasa	tiorredoxina + NADP+ $\rightleftharpoons$ tiorredoxina disulfuro + NADPH + H+		
Timidilato cinasa	ATP + dTMP $\rightleftharpoons$ ADP + dTDP		
Carbamoil-fosfato sintasa subunidad grande	2 ATP + L-glutamina + HCO3- + H2O = 2 ADP + fosfato + L-glutamato + carbamoil fosfato / 2 ATP + HCO3- $\rightleftharpoons$ 2 ADP + fosfato + carbamoil fosfato		
Dihidroorotato deshidrogenasa 1B	(S)-dihidroorotato + O2 = orotato + H2O2		
Dihidropteroato sintasa	(2-amino-4-hidroxi-7,8-dihidropteridin-6-il)metil difosfato + 4-aminobenzoato $\rightleftharpoons$ difosfato + 7,8-dihidropteroato		
Ferrodoxina	Transfiere electrones en una amplia variedad de reacciones metabólicas		

<p>N-(5'-fosfo-L-ribosil-formimino)-5-amino-1-(5'-fosforibosil)-4-imidazolcarboxamida isomerasa</p>	<p>1-(5-fosforibosil)-5-[(5-fosforibosilamino)metilideneamino]imidazol-4-carboxamida <math>\Leftrightarrow</math> 5-[(5-fosfo-1-desoxiribulos-1-ilamino)metilideneamino]-1-(5-fosforibosil)imidazol-4-carboxamida</p>		
<p>Imidazol glicerol fosfato sintasa</p>	<p>N-(5'-fosfo-D-1'-ribulosilformimino)-5-amino-1-(5''-fosfo-D-ribosil)-4-imidazolcarboxamida + L-Glutamina <math>\Leftrightarrow</math> D-eritro-1-(Imidazol-4-il)glicerol 3-fosfato + 1-(5'-fosforibosil)-5-amino-4-imidazolcarboxamida + L-Glutamato</p>		
<p>Indole-3-glycerol phosphate synthase</p>	<p>1-(2-carboxifenilamino)-1-desoxi-D-ribulosa 5-fosfato <math>\Leftrightarrow</math> 1-C-(indol-3-il)glicerol 3-fosfato + CO2 + H2O</p>		
<p>Cetopantoato hidroximetiltransferasa</p>	<p>5,10-metilenotetrahidrofolato + 3-metil-2-oxobutanoato + H2O <math>\Leftrightarrow</math> tetrahidrofolato + 2-dehidropantoato</p>		
<p>Fosfatidilserina cardiolipina sintasa b</p>	<p>Fosfatidilglicerol + CDP-diacilglicerol <math>\Leftrightarrow</math> Cardiolipina + CMP</p>		
<p>Fosforibosilantranilato isomerasa</p>	<p>N-(5-fosfo-beta-D-ribosil)antranilato <math>\Leftrightarrow</math> 1-(2-carboxifenilamino)-1-desoxy-D-ribulosa 5-fosfato</p>		
<p>Fosforibosilformilglicinamida sintasa II</p>	<p>ATP + N2-formil-N1-(5-fosfo-D-ribosil)glicinamida + L-glutamina + H2O <math>\Leftrightarrow</math> ADP + fosfato + 2-(formamido)-N1-(5-fosfo-D-ribosil)acetamidina + L-glutamato</p>		

Tioesterasa	$\text{acil-CoA} + \text{H}_2\text{O} \rightleftharpoons \text{CoA} + \text{a carboxilato}$		
Shikimato deshidrogenasa	$\text{shikimato} + \text{NADP}^+ \rightleftharpoons \text{3-deshidrosikimato} + \text{NADPH} + \text{H}^+$		
Proteína de estrés universal UspA	Requerida para resistir a agentes que dañan al DNA		
Uroporfirinogeno descarboxilasa	$\text{uroporfirinogeno III} \rightleftharpoons \text{coproporfirinogeno III} + 4 \text{CO}_2$		
DNA topoisomerasa I	Rompimiento de DNA de cadena sencilla independiente de ATP		
Endonucleasa III	El enlace 3' C-O-P en el DNA es roto por una reacción de eliminación beta, dejando un azúcar 3'-terminal insaturado y un producto con 5'-fosfato terminal		
Proteína de unión a ATP del sistema de transporte de Vitamina B12	$\text{ATP} + \text{H}_2\text{O} + \text{vitamina B12}_{\text{fuera}} \rightleftharpoons \text{ADP} + \text{fosfato} + \text{vitamina B12}_{\text{dentro}}$		

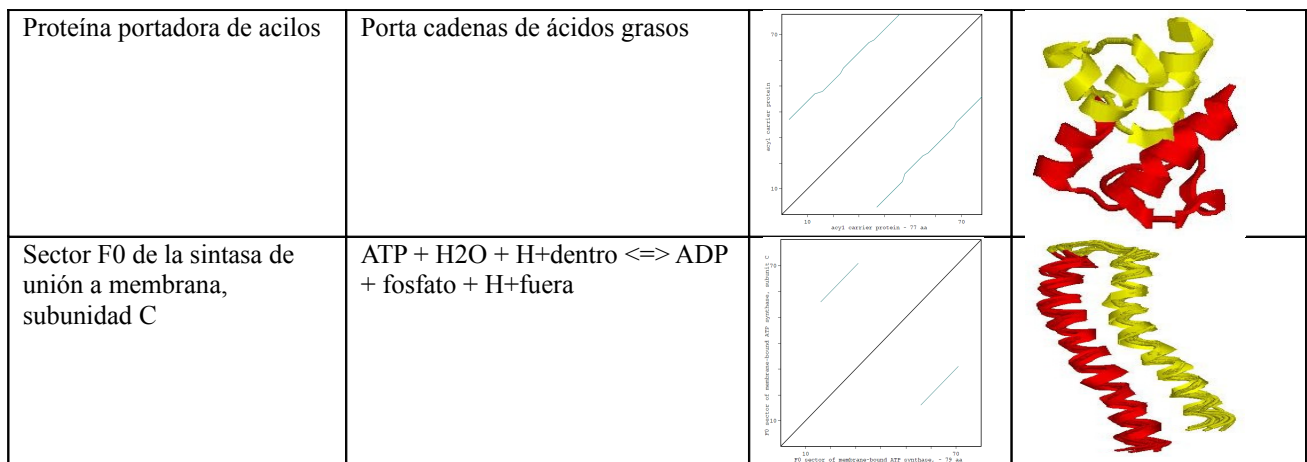


Figura 13. Descripción, alineación en dot-plot, y cristal de las proteínas que se encuentran conservadas y que se originaron por eventos de duplicación interna. Los cristales de las estructuras terciarias de las proteínas se obtuvieron de la base de datos PDB y visualizados con RasMol.

Para el resto de las proteínas analizadas en donde la estructura terciaria aún no ha sido cristalizada, o en donde se ha cristalizado únicamente un fragmento de la molécula, la alineación es la única herramienta disponible que tenemos para proponer homología al interior de una secuencia.

## 5.4 Temporalidad relativa de secuencias altamente conservadas

Del conjunto de secuencias conservadas (COGs conservados), 12 están involucradas en reacciones en las cuales el oxígeno molecular está presente (Tabla 9). De los 12 COG, 4 están relacionados con el metabolismo del RNA y 3 de ellos presentan evidencia de la duplicación interna. Los 8 COGs restantes no están relacionados con el RNA, y únicamente uno de ellos fue originado por duplicación interna. Como se muestra en la tabla 9, los COG que llevan a cabo reacciones dependientes de oxígeno pueden ser agrupados en tres categorías:

- (a) COGs que llevan a cabo únicamente reacciones dependientes de oxígeno, es decir, aparentemente no hay otras reacciones que catalicen.
- (b) COGs que únicamente llevan a cabo reacciones dependientes de oxígeno pero que tienen una contraparte enzimática que cataliza la misma reacción de manera independiente del oxígeno.
- (c) COGs que llevan a cabo tanto reacciones dependientes de oxígeno como otras reacciones en las

cuales el oxígeno no es necesario.

La existencia de múltiples reacciones asociadas con COG puede ser explicada debido a que cada COG está compuesto de grupos de genes homólogos, por lo que cada COG podría tener más de una reacción enzimática asociada a ellos.



Tabla 9. COGs altamente conservados los cuales llevan a cabo reacciones en donde el oxígeno molecular está presente. Las columnas indican (de izquierda a derecha): el COG conservado; si el COG está relacionado con el metabolismo del RNA; si el COG se originó por duplicación interna; la reacción dependiente de oxígeno que lleva a cabo; y si llevan a cabo otra reacción química independiente del oxígeno molecular. El símbolo \* indica si la enzima tiene una contraparte que lleva a cabo la misma reacción en condiciones anóxicas.

COG conservado	Relacionado con el metabolismo del RNA	Evidencia de duplicación interna	Reacción dependiente de O <sub>2</sub>	Reacción independiente de O <sub>2</sub>
COG0155	si	no	EC:1.14.13.83 (metabolismo de porfirina y clorofila)  $precorrin-3A + NADH + H^+ + O_2 \Leftrightarrow precorrin-3B + NAD^+ + H_2O$	EC:1.7.7.1 (metabolismo de nitrógeno)  $NH_3 + 2 H_2O + 6 \text{ ferredoxina oxidada} \Leftrightarrow \text{nitrito} + 6 \text{ ferredoxina reducida} + 7 H^+$  EC:1.8.1.2 (metabolismo de azufre)  $Sulfuro de hidrógeno + 3 NADP^+ + 3 H_2O \Leftrightarrow sulfito + 3 NADPH + 3 H^+$
COG0167	si	si	* EC:1.3.3.1 (metabolismo de pirimidinas) $(S)\text{-dihydroorotato} + O_2 \Leftrightarrow \text{orotato} + H_2O_2$	no
COG0446	si	si	* EC:1.5.3.1 (metabolismo de glicina, serina y treonina) $sarcosina + H_2O + O_2 \Leftrightarrow \text{glicina} + \text{formaldehido} + H_2O_2$	EC:1.3.1.34 $trans\text{-}2,3\text{-didehidroacil-CoA} + NADP^+ = trans,trans\text{-}2,3,4,5\text{-tetradehidroacil-CoA} + NADPH + H^+$ EC:1.18.1.3 (metabolismo de ácidos grasos) $ferredoxina reducida + NAD^+ \Leftrightarrow \text{ferredoxina oxidada} + NADH + H^+$ EC:1.4.99.5 (metabolismo de cianoaminos) $glicina + 2 \text{ Aceptores} \Leftrightarrow \text{cianuro de hidrógeno} + CO_2 + H_2\text{-}2 \text{ Aceptores}$
COG0451	si	si	EC:1.1.1.170 (biosíntesis de esteroides) $3\beta\text{-hidroxi-}4\beta\text{-metil-}5\alpha\text{-colest-}7\text{-ene-}4\alpha\text{-carboxilato} + NAD(P)^+ \Leftrightarrow 4\alpha\text{-metil-}5\alpha\text{-colest-}7\text{-en-}3\text{-}1 + CO_2 + NAD(P)H$	EC:1.1.1.219 (biosíntesis de flavonoides) $cis\text{-}3,4\text{-leucopelargonidina} + NADP^+ \Leftrightarrow (+)\text{-dihidrokaempferol} + NADPH + H^+$ EC:4.2.1.45 (metabolismo de amino-azúcares y azúcar-nucleotido) $CDP\text{-glucosa} \Leftrightarrow CDP\text{-}4\text{-dehidro-}6\text{-desoxy-D-glucosa} + H_2O$ EC:5.1.3.12 (metabolismo de amino-azúcares y azúcar-nucleotido) $UDP\text{-glucuronato} \Leftrightarrow UDP\text{-L-iduronato}$ EC:1.1.1.271 (metabolismo de manosa y fructosa y



				metabolismo de amino-azúcares y azúcar-nucleotido) <i>GDP-L-fucosa + NADP<sup>+</sup> ⇌ GDP-4-dehidro-6-desoxy-D-manosa + NADPH + H<sup>+</sup></i> EC:5.1.3.20 (biosíntesis de lipopolisacáridos) <i>ADP-D-glicero-D-mnno-heptosa ⇌ ADP-L-glicero-D-mano-heptosa</i> EC:3.13.1.1 (metabolismo de amino-azúcares y azúcar-nucleotido y metabolismode glicerolípidos) <i>UDP-glucosa + sulfito ⇌ UDP-6-sulfoquinovosa + H<sub>2</sub>O</i>
COG0277	no	no	* EC:1.1.3.15 (meabolismo de glioxilata y dicarboxilato) <i>(S)-2-hidroxi ácido + O<sub>2</sub> ⇌ a 2-oxo ácido + H<sub>2</sub>O<sub>2</sub></i>	EC:1.1.1.28 (metabolismo de piruvato) <i>(R)-lactato + NAD<sup>+</sup> ⇌ piruvato + NADH + H<sup>+</sup></i>
COG0605	no	no	EC:1.15.1.1 <i>2 O<sub>2</sub>- + 2 H<sub>+</sub> ⇌ O<sub>2</sub> + H<sub>2</sub>O<sub>2</sub></i>	no
COG0665	no	no	* EC:1.5.3.1 (metabolismo de glicina, serina y treonina) <i>sarcosina + H<sub>2</sub>O + O<sub>2</sub> ⇌ glicina + formaldehido + H<sub>2</sub>O<sub>2</sub></i> EC:1.5.3.- <i>tRNA que contiene 5-carboximetilaminometil-2-tiouridina + O<sub>2</sub> + H<sub>2</sub>O ⇌ tRNA que contiene 5-aminometil-2-tiouridina + Glioxilato + H<sub>2</sub>O<sub>2</sub></i> EC:1.4.3.19 (metabolismo de glicina, serina y treonina) <i>glicina + H<sub>2</sub>O + O<sub>2</sub> ⇌ glioxilato + NH<sub>3</sub> + H<sub>2</sub>O<sub>2</sub></i> <i>D-alanina + H<sub>2</sub>O + O<sub>2</sub> ⇌ piruvato + NH<sub>3</sub> + H<sub>2</sub>O<sub>2</sub></i> <i>sarcosina + H<sub>2</sub>O + O<sub>2</sub> ⇌ glioxilato + metilamina + H<sub>2</sub>O<sub>2</sub></i> <i>N-etilglicina + H<sub>2</sub>O + O<sub>2</sub> ⇌ glioxilato + etilamina + H<sub>2</sub>O<sub>2</sub></i> EC:1.4.3.- (metabolismo de arginina y prolina) <i>gamma-L-Glutamilputrescina + H<sub>2</sub>O + O<sub>2</sub> ⇌ gamma-Glutamil-gamma-aminobutiraldehido + NH<sub>3</sub> + H<sub>2</sub>O<sub>2</sub></i>	EC:1.4.99.1 (metabolismo de fenilalanina y metabolismo de nitrógeno) <i>a D-amino ácido + H<sub>2</sub>O + acceptor ⇌ a 2-oxo ácido + NH<sub>3</sub> + acceptor reducido</i> EC:1.5.99.1 (metabolismo de glicina,serina y treonina) <i>sarcosina + acceptor + H<sub>2</sub>O ⇌ glicina + formaldehido + acceptor reducido</i> EC:1.5.99.2 (metabolismo de glicina, serina y treonina) <i>N,N-dimetilglicina + acceptor + H<sub>2</sub>O ⇌ sarcosina + formaldehido + acceptor reducido</i> EC:1.4.99.5 (metabolismo de cianoamino) <i>glicina + 2 acceptor ⇌ cianuro de hidrógeno + CO<sub>2</sub> + H<sub>2</sub>-2 acceptor</i>
COG1271	no	no	EC:1.10.3.- (biosíntesis de betalaina) <i>Dopaxantina + O<sub>2</sub> ⇌ 2 Dopaxantina quinona + 2 H<sub>2</sub>O</i> <i>2 Dopamina + O<sub>2</sub> ⇌ 2 Dopamina quinona + 2 H<sub>2</sub>O</i>	no
COG1304	no	no	EC:1.13.12.- (metabolismo de nitrógeno y	EC:1.1.2.3 (metabolismo de piruvato)

			metabolismo de tirosina) $NH_3 + O_2 + Ubiquinol \rightleftharpoons Hidroxilamina + H_2O + Ubiquinona$ $Homogentisato + O_2 \rightleftharpoons Gentisato aldehido + CO_2 + H_2O$	$(S)\text{-lactato} + 2\text{ ferricitocromo } c \rightleftharpoons \text{piruvato} + 2\text{ ferrocitocromo } c + 2\text{ H}^+$ EC:5.3.3.2 (biosíntesis del esqueleto terpenoide) $isopentenil\ difosfato \rightleftharpoons dimetilalil\ difosfato$
COG1853	no	no	EC:1.14.13.3 (metabolismo de tirosina) $4\text{-hidroxifenilacetato} + NADH + H^+ + O_2 \rightleftharpoons 3,4\text{-dihidroxifenilacetato} + NAD^+ + H_2O$	EC:1.5.1.- (metabolismo de triptófano y metabolismo de metano) $2\text{-Oxoadipato} + NH_3 + NAD^+ \rightleftharpoons 2\text{-Aminomuconato} + NADH + H^+ + H_2O$ $2\text{-Oxoadipato} + NH_3 + NADP^+ \rightleftharpoons 2\text{-Aminomuconato} + NADPH + H^+ + H_2O$ $5,10\text{-Metilenetetrahidrometanopterina} + NADP^+ \rightleftharpoons 5,10\text{-Meteniltetrahidrometanopterina} + NADPH$ $\gamma\text{-Coniceína} + NADPH + H^+ \rightleftharpoons \text{Coniina} + NADP^+$ EC:1.16.8.1 (metabolismo de porfirina y clorofila) $2\text{ cob(II)irinic ácido } a,c\text{-diamida} + FMN + 2\text{ H}^+ \rightleftharpoons 2\text{ cob(II)irinic ácido } a,c\text{-diamido} + FMNH_2$
COG2070	no	no	EC:1.13.12.16 (metabolismo de nitrógeno) $etilnitronato + O_2 + FMNH_2 = \text{acetaldehido} + \text{nitrito} + FMN + H_2O$	no
COG1028	no	si	EC:1.1.1.270 (biosíntesis de esteroides) $4\alpha\text{-metil-5}\alpha\text{-colest-7-en-3}\beta\text{-ol} + NADP^+ \rightleftharpoons 4\alpha\text{-metil-5}\alpha\text{-colest-7-en-3-1} + NADPH + H^+$	EC:1.1.1.30 (interconversiones de pentosa y glucuronato y metabolismo de ascorbato y aldarato) $3\text{-dehidro-L-gulonato} + NAD(P)^+ \rightleftharpoons (4R,5S)\text{-4,5,6-trihidroxi-2,3-dioxohexanoato} + NAD(P)H + H^+$ EC:1.1.1.36 (metabolismo de butanoato) $(R)\text{-3-hidroxiacil-CoA} + NADP^+ \rightleftharpoons 3\text{-oxoacil-CoA} + NADPH + H^+$ EC:1.1.1.47 (ruta de pentosa fosfato) $\beta\text{-D-glucosa} + NAD(P)^+ \rightleftharpoons \text{D-glucono-1,5-lactona} + NAD(P)H + H^+$ EC:1.1.1.69 $\text{D-gluconato} + NAD(P)^+ \rightleftharpoons 5\text{-dehidro-D-gluconato} + NAD(P)H + H^+$ EC:1.1.1.125 (interconversiones de pentosa y glucuronato) $2\text{-desoxi-D-gluconato} + NAD^+ \rightleftharpoons 3\text{-dehidro-2-desoxi-D-gluconato} + NADH + H^+$ EC:1.1.1.140 (metabolismo de fructosa y manosa) $\text{D-sorbitol } 6\text{-fosfato} + NAD^+ \rightleftharpoons \text{D-fructosa } 6\text{-fosfato} + NADH + H^+$ EC:1.1.1.153 (biosíntesis de folato)

				<p>(1) <math>7,8\text{-dihidrobiopterina} + \text{NADP}^+ \Leftrightarrow \text{sepiapterina} + \text{NADPH} + \text{H}^+</math>  <math>\text{tetrahidrobiopterina} + 2 \text{NADP}^+ \Leftrightarrow 6\text{-piruvoil-5,6,7,8-tetrahidropterina} + 2 \text{NADPH} + 2 \text{H}^+</math>  EC:1.1.1.159  <math>3\alpha,7\alpha,12\alpha\text{-trihidroxi-5}\beta\text{-colanato} + \text{NAD}^+ \Leftrightarrow 3\alpha,12\alpha\text{-dihidroxi-7-oxo-5}\beta\text{-colanato} + \text{NADH} + \text{H}^+</math>  EC:1.3.1.28 (biosíntesis de grupo sideroforo)  <math>(2S,3S)\text{-2,3-dihidro-2,3-dihidroxibenzoato} + \text{NAD}^+ \Leftrightarrow 2,3\text{-dihidroxibenzoato} + \text{NADH} + \text{H}^+</math>  EC:1.3.1.33 (metabolismo de porfirina y clorofila)  <math>\text{clorofilina a} + \text{NADP}^+ \Leftrightarrow \text{protoclorofilina} + \text{NADPH} + \text{H}^+</math>  EC:1.1.1.4 (metabolismo de butanoato)  <math>(R,R)\text{-butano-2,3-diol} + \text{NAD}^+ \Leftrightarrow (R)\text{-acetoina} + \text{NADH} + \text{H}^+</math>  EC:1.1.1.102 (metabolismo de esfingolípidos)  <math>\text{Esfingina} + \text{NADP}^+ \Leftrightarrow 3\text{-dehidrosfingina} + \text{NADPH} + \text{H}^+</math>  EC:1.3.1.25 (degradación de benzoato via hidroxilación y degradación de fluorobenzoato y benzoato via ligación CoA)  <math>(1R,6S)\text{-1,6-dihidroxiciclohexa-2,4-dieno-1-carboxilato} + \text{NAD}^+ \Leftrightarrow \text{catecol} + \text{CO}_2 + \text{NADH} + \text{H}^+</math>  EC:1.1.1.101 (metabolismo de glicerofosfolípidos y metabolismo de lípidos-eter)  <math>1\text{-palmitoilglicerol 3-fosfato} + \text{NADP}^+ \Leftrightarrow \text{palmitoilglicerona fosfato} + \text{NADPH} + \text{H}^+</math>  EC:1.1.1.206 (biosíntesis de tropano, piperidina y piridina alkaloid)  <math>\text{tropina} + \text{NADP}^+ \Leftrightarrow \text{tropinona} + \text{NADPH} + \text{H}^+</math>  EC:1.1.1.100 (biosíntesis de ácidos grasos y biosíntesis de ácidos grasos insaturados)  <math>a(3R)\text{-3-hidroxiacil-[proteína acarreadora de acilos]} + \text{NADP}^+ \Leftrightarrow a\text{-3-oxoacil-[proteína acarreadora de acilos]} + \text{NADPH} + \text{H}^+</math>  EC:1.5.1.- (metabolismo de triptófano y metabolismo de metano)  <math>2\text{-Oxoadipato} + \text{NH}_3 + \text{NAD}^+ \Leftrightarrow 2\text{-Aminomuconato} +</math></p>
--	--	--	--	--

				<p> <math>NADH + H^+ + H_2O</math>  <math>2\text{-Oxoadipato} + NH_3 + NADP^+ \rightleftharpoons 2\text{-Aminomuconato} + NADPH + H^+ + H_2O</math>  <math>5,10\text{-Metilendetraidrometanopterina} + NADP^+ \rightleftharpoons 5,10\text{-Meteniltetraidrometanopterina} + NADPH</math>  <math>\gamma\text{-Coniceina} + NADPH + H^+ \rightleftharpoons \text{Coniina} + NADP^+</math>            EC:1.5.1.3 (biosíntesis de folato)  <math>5,6,7,8\text{-tetraidrofolato} + NADP^+ = 7,8\text{-dihidrofolato} + NADPH + H^+</math> </p>
--	--	--	--	--

## **5.5 Pérdida del rastro de la duplicación interna en secuencias conservadas**

El análisis de los resultados provenientes de la comparación de secuencias a nivel de estructura primaria de las diferentes bases de datos, muestra que sólo en algunas de ellas se conserva el rastro de la duplicación interna, mientras que en el resto de las secuencias ortólogas la huella de la duplicación se ha perdido a pesar de que la secuencia está ampliamente distribuida en los tres grandes linajes celulares, posee un alto grado de similitud, tiene aproximadamente la misma longitud y está descrita como parte del mismo conjunto de genes ortólogos. (Tabla 10).

Las tablas 3 y 4 muestran claramente que aunque la secuencia puede estar presente en mas de una base de datos, sólo en algunas de ellas se mantiene la evidencia de la duplicación. Tanto la metodología empleada como el análisis de los resultados arrojados por esa metodología no nos permiten asegurar cuál podría ser la explicación que subyace al fenómeno de la pérdida de la huella de la duplicación.

Tabla 10. Conservación de la huella de la duplicación interna. “+” indica que la secuencia está presente en el genoma. “++” indica que la secuencia conserva la evidencia de la duplicación interna. Las celdas vacías denotan la ausencia de la secuencia en el genoma. *Escherichia coli* K-12 MG1655 (eco); *Escherichia coli* O157:H7 EDL933 (ece); *Buchnera aphidicola* APS (buc); *Haemophilus influenzae* Rd KW20 (hin); *Pasteurella multocida* (pnu); *Xylella fastidiosa* 9a5c (xfa); *Vibrio cholerae* O1 (vch); *Pseudomonas aeruginosa* PAO1 (pae); *Neisseria meningitidis* MC58 (nme); *Neisseria meningitidis* Z2491 (nma); *Helicobacter pylori* 26695 (hpy); *Helicobacter pylori* J99 (hpi); *Campylobacter jejuni* NCTC11168 (cje); *Rickettsia prowazekii* (rpr); *Mesorhizobium loti* (mlo); *Caulobacter crescentus* CB15 (ccr); *Bacillus subtilis* (bsu); *Bacillus halodurans* (bha); *Lactococcus lactis* subsp. *lactis* IL1403 (lla); *Streptococcus pyogenes* SF370 (spy); *Mycoplasma genitalium* (mge); *Mycoplasma pneumoniae* (mpn); *Ureaplasma parvum* serovar 3 ATCC 700970 (uur); *Mycobacterium tuberculosis* H37Rv (mtu); *Mycobacterium leprae* TN (mle); *Chlamydia trachomatis* D/UW-3/CX (ctr); *Chlamydomphila pneumoniae* CWL029 (cpn); *Borrelia burgdorferi* B31 (bbu); *Treponema pallidum* subsp. *pallidum* Nichols (tpa); *Synechocystis* sp. PCC6803 (syn); *Deinococcus radiodurans* (dra); *Aquifex aeolicus* (aae); *Thermotoga maritima* (tma); *Methanocaldococcus jannaschii* (mja); *Methanothermobacter thermautotrophicus* (mth); *Archaeoglobus fulgidus* (afu); *Halobacterium* sp. NRC-1 (hal); *Thermoplasma acidophilum* (tac); *Thermoplasma volcanium* (tvo); *Pyrococcus horikoshii* (pho); *Pyrococcus abyssi* (pab); *Aeropyrum pernix* (ape); *Saccharomyces cerevisiae* (sce)

COG	eco	ece	buc	hin	pnu	xfa	vch	pae	nme	nma	hpy	hpi	cje	rpr	mlo	ccr	bsu	bha	lla	spy	mge	mpn	uur	mtu	mle	ctr	cpn	bbu	tpa	syn	dra	aae	tma	mja	mth	afu	hal	tac	tvo	pho	pab	ape	sce					
COG0001	+	+			+	+	+	+	+	+	+	+	+		+	+	++	+					+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				
COG0002	+	+	+		+	+	+	+	+	+			+		+	+	+	+	+	+				+	+						+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++		
COG0009	+	+	+	+	+	+	+	+	+	+				+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
COG0012	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
COG0015	+	+	+	+	+	+	+	+	+	+	+	+	+		+	++	+	+	+	+				+	+					+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG0020	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG0024	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+		
COG0030	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+		
COG0037	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
COG0041	+	+		+	+	+	+	+	+	+			+		+	+	+	+	+	+	+			+	+						+	+	+	+	+	++	+	+	+	+	+	+	+	+	+			
COG0046	+	+		+	+	+	+	+	+	+			+		+	++	+	+	+	+					++	+					+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+		
COG0048	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
COG0049	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
COG0053	+	+			+		+						+	+	+	+	++	+	+	+				+					+	++	+	+	+	+	++	++	+					+	+	+	+	+		
COG0054	+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+				+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+		
COG0062	+	+			+	+	+				+	+	+		++	+								+	+					+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
COG0063	+	+			++	+	++	+	+	+	+	+	++				+		+					+	+				+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
COG0070	+	+			+	+	+						+		+	+	+	+	+	+				+	+						+	+	+	++	+	+	+									+	+	
COG0071	+	+	+		+	+	+						+	+	+	++	+							+	+					+	+	+	+	+	+	++	+	+	+	+	+	++	+	+	+	+		
COG0073	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	
COG0080	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+
COG0089	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+	

















COG1121	+	+	+	+	+		+	+	+	+			+	+	+	+	+	+					+	+	+	+	+	++		+	+	+			+	+	+							
COG1131	++	++		+	++	++	+	++	+	+	+	+		+	+	+	+	+	+		+	+	+	+	+	+		+	+	+	++	+	+	+	++	+								
COG1132	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+						
COG1136	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+						
COG1143	+	+	+		+		+	++	++	+	+	+	+	++	+									+	+	+	+	+	++	++	+	+	+	+	+	+								
COG1173	+	+		+	+		+	+		+	+	++	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++		+	+	+	+	+	+	+	+							
COG1176	+	+		+	+		+	+	+			++	+		+	+	+	+	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+	+							
COG1183					+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+						
COG1192	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+						
COG1196						+	+	+	+			+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	+	++		+	+	+	+	++	+					
COG1226	++	++		+	+	+	+	+	+	+	+	+	++	++	+	+	+								+	+					++	+	+	+	+	++	+	++	+	+	+	+	+	+
COG1230	+	+		+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+									+			+			
COG1234	+	+		+									+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+				
COG1266		+				+	++						+	++	+	+	+	+						++		+	+	++	++	+	++	++	+	+	+	+	+	+	+	+				
COG1268													+	+		+	+	++	+							+	+	+	+	++		+	+	+			+	+						
COG1290						+	+	+	+	+	+	+	+	+	+	++									+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG1309	+	+		+	+	+	+	++	+	+		+		++	++	+	++	++	+					+	+	+					+	++	++	+		+	+	+		+				
COG1354						+	+	+	+				+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	++	+	+		+	+	+	+	+	+	+	+	+	+			
COG1387	+	+				+																		++			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG1475	+	+				+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG1502	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	++	+	+					+			++	+	+		++		+	+	+	+	+	+	+	+	+				
COG1525				+	+			+	+	+		+	+	+	+										+	+	+	+	+	+	+	+	+	+	+	++		+	+	+	+			
COG1546	+	+				++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG1587	+	+			+	+	+	+	+	+	+	+	+	+	++	+									+	+	++	+	+	++	++	+	++	+	+	+	+	+	+	+				
COG1605	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+			
COG1611	+	+			+	+	+	+	+	+		+	+	+	++	+	+							+	+	+	+	+	+	+	+	+	+	+	+	+	+	++	+	+	+			
COG1624													+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				
COG1670	+	+			+		+	+	+	+	+	+	+	+	++	+									+	+	+	+	+	+	+	+	+	+	++	+	+	+	+	+				
COG1739	+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	++				+	+	+	+	+	+	+	+	+	+			
COG1758	+	+		+	+	+	+	+	+	+	+	+	+	+	+	++	+							+	+			+	+	+	+	+	+	+	+	+	+	+	+	+	+			
COG1841	+	+	+	+	+	+	+	+	+			+	+	+	+	+	+	+						+	+		+	+	++	+	+	++	+	+	++	+	+	+	+	+	++			
COG1940	+	+		+	+								+	++	+	+	+	+						++	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+				
COG1985	+	+	+	+	+	+	+	++	+	+	+	+	+	+	+	+								+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+			



# VI DISCUSIÓN

Durante los últimos años ha aparecido una gran cantidad de información acerca de la importancia que han tenido los distintos fenómenos genéticos en la evolución del genoma (ej. duplicación de genes, transferencia horizontal de genes, etc.) tanto de manera funcional como en el incremento de su tamaño y en la complejidad del mismo (Hughes, 1994; Lawrence y Roth, 1996; de la Cruz y Davies, 2000; Eisen, 2000; Ochman *et al.*, 2000; Ragan, 2001; Brosius, 2003; Kunin y Ouzounis, 2003; Tautz *et al.*, 1986; Hanckock, 1995; Becerra, 2000). Sin embargo, muy poco se ha comentado sobre la repercusión que ha tenido el mecanismo de la duplicación interna de genes en la evolución temprana de las proteínas.

Debido al interés que tenemos en conocer los mecanismos moleculares que actuaron en los estadios tempranos de la vida y que dieron paso a la complejidad de las distintas proteínas que componen a los genomas actuales, hemos intentado reconocer cuál es el papel que ha jugado el fenómeno de la duplicación interna en la evolución de las proteínas. Por esta razón, decidimos analizar exclusivamente a las secuencias que se encuentran altamente conservadas que se han obtenido en los procesos de reconstrucción del genoma del último ancestro común.

## **6.1 Secuencias altamente conservadas originadas por duplicación y fusión de genes parálogos en diferentes bases de datos**

Los resultados de los análisis de la estructura primaria indican que aproximadamente 21% de todas las secuencias conservadas de la base de datos de Delaye *et al.* (2005) fueron originadas por el mecanismo de duplicación interna. La base de datos de Harris *et al.* (2003) muestra que aproximadamente 38% de la secuencias que la componen fueron originadas por el mismo mecanismo, mientras que la base de datos de Mirkin *et al.* (2003) contiene 38% de secuencias altamente conservadas con homología interna. A pesar de que existe un porcentaje diferente de secuencias antiguas originadas por duplicación interna en estas bases de datos, existen un conjunto de secuencias conservadas entre ellas. De igual forma, los resultados muestran que las secuencias altamente conservadas que se relacionan con el metabolismo del RNA y que se originaron por duplicación interna varían en cada base de datos. Por ejemplo, alrededor del 24% de las secuencias conservadas originadas por duplicación interna y que están involucradas con el metabolismo del RNA en la base de Delaye *et al.* (2005), aproximadamente un 36% en la base de Harris *et al.* (2003), y un 34% en la base de Mirkin

*et al.* (2003).

Las diferencias en el número de secuencias duplicadas en las diferentes bases de datos pueden estar relacionadas con múltiples factores entre los que se destacan:

(1) Diferencias metodológicas utilizadas durante la reconstrucción del ancestro común a todos los seres vivos. Por ejemplo, Mirkin *et al.* (2003) encuentran un gran número de genes altamente conservados (572 COGs para  $g = 1$ ; donde  $g$  es igual a la penalidad de la ganancia \*revisar artículo para mayor información) mediante algoritmos que evalúan la ganancia y pérdida de genes, siendo ésta última favorecida y promoviendo un incremento en la constitución genética del último ancestro común universal. Delaye *et al.* (2005) encontraron un menor número de genes conservados (161 COGs), debido a que cada gen debe estar presente en cada uno de los genomas celulares de vida libre que se analizaron. Si el término de distribución universal es restringido a su sentido más obvio, es decir, que las secuencias deben encontrarse en todos los genomas analizados, entonces, es completamente comprensible que el repertorio resultante esté formado por un menor número de secuencias. Por lo tanto, genes originados por duplicación interna encontrados en la base de datos de Mirkin *et al.* están ausentes en sus resultados. De igual forma, Harris *et al.* (2003) encuentran muy pocos genes conservados (80 COGs), ya que cada gen debe estar universalmente conservado en todos los organismos que componen a la base de datos COG, base cuyo contenido de genes ortólogos se encuentran presentes en unos pocos linajes celulares.

(2) Los niveles de astringencia utilizados. Aunque los valores de los parámetros en la búsqueda de duplicaciones internas fueron los mismos para todas las bases de datos, los distintos valores de corte para la búsqueda de genes conservados que se utilizaron fueron diferentes, reflejando el número variable de secuencias conservadas que caracterizan al ancestro común de los tres linajes celulares.

(3) Pérdida de la huella de la duplicación interna. De manera sorprendente, encontramos que solo ciertas secuencias, de un conjunto de secuencias homólogas, conservan el rastro de la duplicación interna. Aparentemente, no existe alguna relación o sesgo entre la preservación de la huella de la duplicación y el organismo que la porta, es decir, no parece haber una relación entre la pérdida del rastro de la secuencia duplicada y posición filogenética del organismo. Debido a que cada metodología utilizó una base de datos distinta, y esta está compuesta por secuencias provenientes de diferentes genomas



celulares, observamos que la huella de la duplicación se conserva sólo en ciertos genomas, los cuales pudieron haber estado presentes en una base de datos específica. Por ejemplo, la secuencia de la enzima tiorredoxina reductasa, de la base de datos de Delaye *et al.* (2005), preserva el rastro de la duplicación interna, mientras que las secuencias ortólogas a ella han perdido la huella de la duplicación. Analizando con detenimiento encontramos que la secuencia que conserva la huella de la duplicación pertenece al genoma de *Schizosaccharomyces pombe*, genoma presente en la base de datos de Delaye *et al.* (2005), pero ausente en la base de datos de COGs utilizada por Mirkin *et al.* (2003) y Harris *et al.* (2003).

Es bien sabido que muchas otras secuencias han sido originadas mediante un evento de duplicación interna, sin embargo, en este trabajo no fueron contempladas debido a que aparentemente no se encuentran altamente conservadas en los genomas que representan a los tres dominios celulares.

Asimismo, es importante distinguir varias causas por las que la exactitud de nuestros resultados pueden verse afectados. La transferencia horizontal de genes, sesgos en la construcción de las bases de datos de genomas, variaciones significativas en las tasas de sustitución de diferentes proteínas, y la metodología utilizada durante los análisis serían las principales causas de falsos negativos en nuestros resultados.

## **6.2 Diversidad funcional de proteínas conservadas originadas por duplicación interna**

La lista de secuencias altamente conservadas que se originaron por duplicación interna de las tres bases de datos incluye genes que codifican para una variedad de enzimas que participan en procesos metabólicos ampliamente diversos. La diversidad funcional de las enzimas que poseen duplicación interna nos indica que este mecanismo no está restringido a un sólo tipo de proteínas, sino que parece haber actuado en distintos intervalos del tiempo evolutivo de diversos polipéptidos ancestrales para formar a varias de las enzimas actuales, contribuyendo de esta manera a la conformación de una parte de la diversidad de proteínas.

Los resultados muestran que el tamaño y la estructura de un conjunto grande de proteínas conservadas son el resultado evolutivo de duplicaciones de genes parálogos seguidos por eventos de

fusión que tuvieron lugar previamente a la divergencia de los tres reinos primarios. Asimismo, la identificación de secuencias formadas por módulos homólogos fusionados en tandem proveen evidencia de la existencia de genes funcionales más pequeños durante el Precámbrico temprano, indicando que el último ancestro común fue precedido por células con genes y genomas más sencillos.

De manera no sorprendente, el número más grande de secuencias duplicadas están involucradas en el proceso universal conservado de la síntesis de proteínas, sesgo que se ve reflejado por el número de secuencias involucradas en dicho proceso.

### **6.3 Proteínas conservadas que interactúan con el RNA y que se originaron por duplicación interna**

La lista de proteínas altamente conservadas que sintetizan, degradan o interactúan con el RNA y que parecen haber surgido mediante un evento de duplicación interna está conformada por un total de 23 secuencias polipeptídicas, representando cerca de un 24% del conjunto total de secuencias analizadas de la base de datos de Delaye *et al.* (2005), 57 secuencias que representan aproximadamente un 34% del conjunto total de la base de Mirkin *et al.* (2003), y 22 secuencias de la base de Harris que representan un 36%.

El conjunto de secuencias que han experimentado duplicación intragénica previamente a la divergencia de los tres dominios celulares incluye genes que codifican para una variedad de enzimas que participan en procesos celulares universales. Como se muestra en la figura 12 y en las tablas 3, 4 y 5, entre los procesos se encuentran: (1) secuencias relacionadas con procesos informacionales como la traducción, (2) secuencias relacionadas con la regulación de la expresión del material genético, y (3) varias más involucradas en funciones celulares relacionadas con la biosíntesis de diversas moléculas.

De manera no sorprendente, entre el número mayor de secuencias duplicadas se encuentran aquellas involucradas en el proceso de traducción, reflejando quizá un sesgo por el número de secuencias conservadas involucradas en este proceso. El resto de las secuencias participa en procesos de degradación y en procesos metabólicos.

## **6.4 Diversidad funcional de proteínas altamente conservadas que están relacionadas con el RNA y que se originaron por duplicación interna**

Las secuencias altamente conservadas y que probablemente fueron originadas a partir de un evento de duplicación interna ostentan una gran variedad de propiedades catalíticas. Estas propiedades abarcan desde enzimas que participan en la transcripción, traducción, metabolismo de azúcares, biosíntesis de nucleótidos, regulación de la expresión génica, hasta la biosíntesis de algunos aminoácidos y otras moléculas.

La variedad de enzimas originadas a partir de la duplicación y fusión de un gen en secuencias relacionadas con el RNA, muestra la importancia del papel que ha desempeñado este mecanismo en la evolución temprana de las proteínas.

## **6.5 La adquisición de dominios adicionales o la duplicación parcial de un gen podría enmascarar la evidencia tridimensional de las regiones homólogas internas**

Todas las proteínas consisten de uno o mas dominios (Murzin *et al.*, 1995; Chothia y Gough, 2009), los cuales forman parte de su estructura tridimensional y funcional. Por las propiedades que los dominios les confieren a las proteínas, han llegado a ser vistos como la unidad evolutiva de éstas. Parece que existe un repertorio limitado de estos bloques estructurales, los cuales han sido duplicados, combinados, fusionados y fisionados para formar la amplia gama de proteínas en un genoma (Apic, *et al.*, 2001). Se ha postulado que un gran porcentaje de los dominios que componen a las proteínas descienden de un ancestro común mediante eventos de duplicación (Chothia y Gough, 2009).

La figura 13 muestra tanto la estructura terciaria como la alineación de la estructura primaria, en una matriz “dot-plot”, de las secuencias conservadas que han sido producto de duplicación interna en las diferentes bases de datos. En algunos casos se nota claramente la duplicación del dominio o dominios que forman parte de la estructura terciaria de la proteína, confirmando que efectivamente la alineación de tipo local es una herramienta extremadamente poderosa para detectar regiones homólogas dentro de una secuencia. Ejemplos notables los encontramos en las proteínas ribosomales S5, S6, S7, S10, S11, S12, S13, S15, L1, L6, L11, L24, L29, L30 y L34 que conforman al ribosoma, la metionina

aminopeptidasa, la fosfoglicerato cinasa, la piruvato cinasa, ribosa-fosfato pirofosfocinasa, timidilato cinasa, la subunidad grande de la carbamoil-fosfato sintasa, la dihidropteroato sintasa, la ferredoxina, la N-(5'-fosfo-L-ribosil-formimino)-5-amino-1-(5'-fosforibosil)-4-imidazolcarboxamida isomerasa, la imidazol glicerol fosfato sintasa, la indol-3-glicerol fosfato sintasa, la cetopantoato hidroximetiltransferasa, la fosfatidilserina cardiolipina sintasa, la fosforibosilantranilato isomerasa, la fosforibosilformilglicinamidina sintasa II, la tioesterasa, la shikimato deshidrogenasa, la proteína de estrés universal UspA, la uroporfirinógeno descarboxilasa, la endonucleasa III, la proteína de unión a ATP del sistema de transporte de vitamina B12, la proteína acarreadora de acilos, y la subunidad C de la ATP sintasa (sector F0). Sin embargo, existen otros casos en donde la detección de la homología interna a nivel de estructura terciaria no es tan evidente. Las causas podrían ser (1) la adquisición de uno o mas dominios adicionales que enmascararían la evidencia tridimensional de las regiones homólogas intragénicas, lo cual dificultaría su visualización, o (2) por la duplicación parcial del gen, es decir, que el segmento duplicado no corresponda a la longitud total del gen. La evidencia de la duplicación interna sería más fácil de reconocer a nivel de estructura primaria, por lo que sería detectada fácilmente mediante la comparación interna de la secuencia. Posibles ejemplos los encontramos en las proteínas ribosomales L22, L23, la tRNA pseudouridina 55 sintasa, la enolasa, la fosfomanomutasa, la tioredoxina reductasa, la dihidroorotato deshidrogenasa, y la DNA topoisomerasa I

En aquellos casos en los que la estructura terciaria aún no está disponible o en donde solamente un fragmento ha sido cristalizado, la alineación es el único medio por el cual podemos inferir la homología interna.

A pesar de que algunos de los dominios que componen a varias de las secuencias analizadas han sido clasificados como distintos (de origen independiente) por la base de datos Pfam, los valores de la alineación local resultan ser razonablemente buenos para proponer su origen común.

## **6.6 Antigüedad relativa de secuencias altamente conservadas**

Nuestros resultados muestran que existen un conjunto de secuencias conservadas (COGs conservados) que están involucradas en reacciones en las cuales el oxígeno molecular está presente.

Estas secuencias pueden agruparse en tres categorías:

(a) COGs que llevan a cabo únicamente reacciones dependientes de oxígeno, es decir, aparentemente no hay otras reacciones que catalicen, indicando que algunos genes conservados que participan en reacciones en las cuales se requiere oxígeno debieron haberse originado cuando éste gas se estaba acumulando en el ambiente terrestre.

(b) COGs que únicamente llevan a cabo reacciones dependientes de oxígeno pero que tienen una contraparte enzimática que cataliza la misma reacción de manera independiente del oxígeno, indicando que enzimas anaeróbicas han sido remplazadas en muchas reacciones por versiones aeróbicas más eficientes e irreversibles que utilizan oxígeno como lo han demostrado Raymond y Blankenship (2004).

(c) COGs que llevan a cabo tanto reacciones dependientes de oxígeno como otras reacciones en las cuales el oxígeno no es necesario y en donde la ruta metabólica es aparentemente más antigua, indicando que estos COGs realizaban reacciones bajo condiciones anoxigénicas y que posteriormente fueron reclutados para catalizar reacciones dependientes de oxígeno. Probablemente, éste conjunto de COGs se originaron antes de que ocurriera la oxidación de la atmósfera de la Tierra.

Los resultados obtenidos en este trabajo proveen evidencia directa de que las secuencias altamente conservadas no son igualmente antiguas sino que cada una de ellas surgió en un tiempo geológico diferente. Entonces, el conjunto de secuencias conservadas, podría ser dividido en tres grupos de acuerdo a su antigüedad: (1) aquellas que se relacionan con el RNA, siendo quizá las más antiguas, provenientes de un mundo de RNA/proteínas, ej. algunas proteínas ribosomales, (2) aquellas que surgieron cuando el DNA pasó a ser la molécula informacional, ej. DNA topoisomerasa I, y (3) aquellas que surgieron cuando el oxígeno se estaba acumulando en la atmósfera de la Tierra, ej. dihidroorotato deshidrogenasa (Figura 14).

La existencia de múltiples reacciones asociadas a los COGs puede ser explicada debido a que cada COG está compuesto de grupos de genes homólogos, por lo que cada COG podría tener más de una reacción enzimática asociada a ellos.

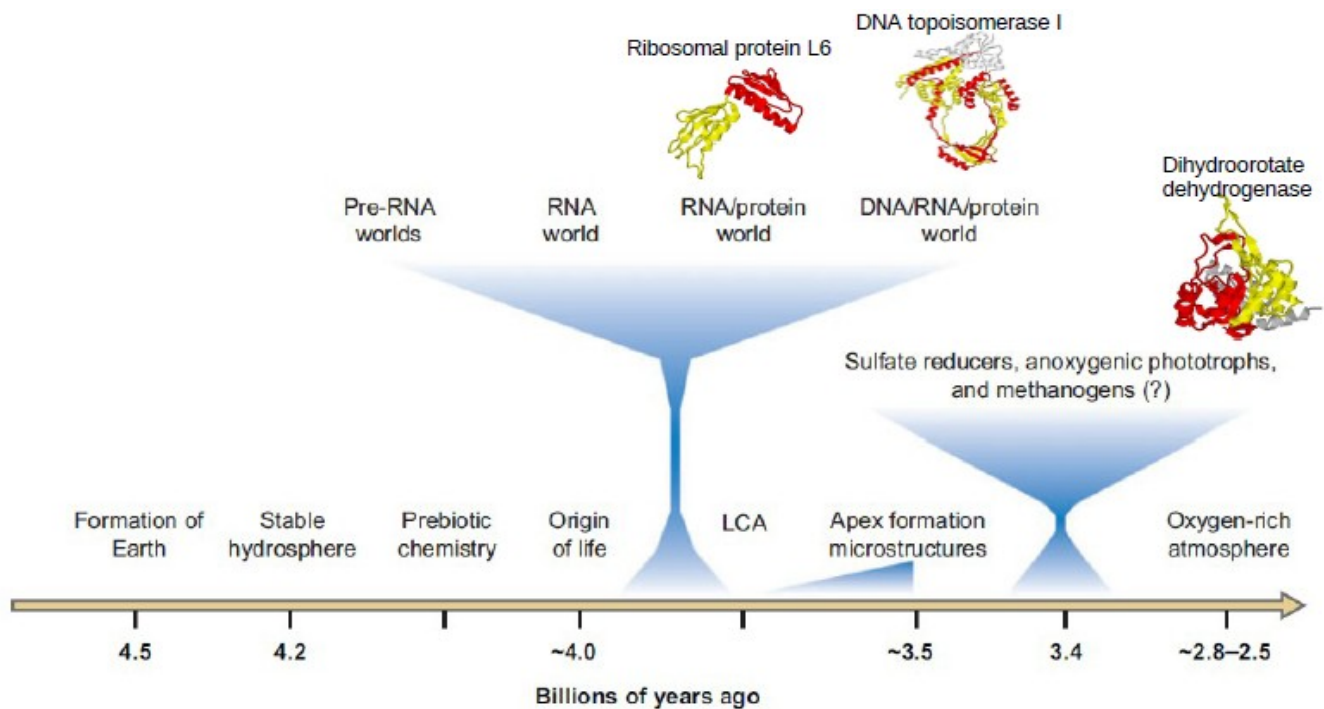


Figura 14. Esquema que representa la temporalidad relativa de secuencias altamente conservadas. Figura modificada de Becerra *et al.* 2007.

## 6.7 Duplicación interna de genes como mecanismo que promueve la complejidad estructural de las proteínas

Un genoma es el producto del surgimiento, pérdida y/o adquisición de genes a través de su historia evolutiva, es decir, los genomas pueden ser vistos como el producto evolutivo de un conjunto de elementos genéticos cuyo origen pertenece a distintas épocas. Si los genomas celulares almacenan una riqueza de información histórica, entonces los análisis genómicos pueden proveer pistas para la organización genética y la complejidad bioquímica de entidades hipotéticas de las cuales los organismos existentes evolucionaron.

El conjunto de proteínas altamente conservadas que están relacionadas con el metabolismo del RNA y que componen al genoma del último ancestro común son de una naturaleza estructural compleja. Desde una perspectiva evolutiva, es razonable suponer que la mayoría de ellas debieron haber sido precedidas por moléculas cuya estructura fuera más simple. Si las secuencias conservadas en los tres dominios celulares fueron precedidas por secuencias más sencillas, ¿existe alguna evidencia de estas secuencias?

Cuando analizamos la estructura primaria y terciaria de las proteínas que se encuentran altamente conservadas y de aquellas que están relacionadas en más de una forma con el metabolismo del RNA, encontramos que una porción significativa fueron producto del fenómeno de duplicación interna.

Con base en este resultado, es razonable suponer que estas moléculas podrían haber evolucionado a partir de otras cuya estructura fue más simple, quizá, con una longitud de la mitad de tamaño de las enzimas actuales. Por lo tanto, se puede asegurar que la duplicación interna de genes ha jugado un papel importante en el incremento de la complejidad del gen y en su evolución, ya que este mecanismo podría permitir la adquisición de una nueva función o incrementar la velocidad de la función mediante la modificación de un segmento redundante o por el incremento de sitios activos. Ello indica que una proporción importante de los genes complejos encontrados en los genomas actuales podrían haber evolucionado a partir de genes pequeños primordiales vía duplicación interna y subsecuente modificación durante los estadios tempranos de la vida.

La conservación de secuencias relacionadas con RNA apoyan la hipótesis de que el último ancestro común fue el resultado evolutivo de un mundo de RNA/proteína. Si algunas de las proteínas altamente conservadas que interactúan con el RNA y que se originaron vía duplicación interna estuvieron presentes en el mundo de RNA/proteínas, entonces es comprensible suponer que este mecanismo podría haber estado actuando en estadios en los que los polímeros de RNA se encontraban interactuando con las primeras proteínas (Figura 15).

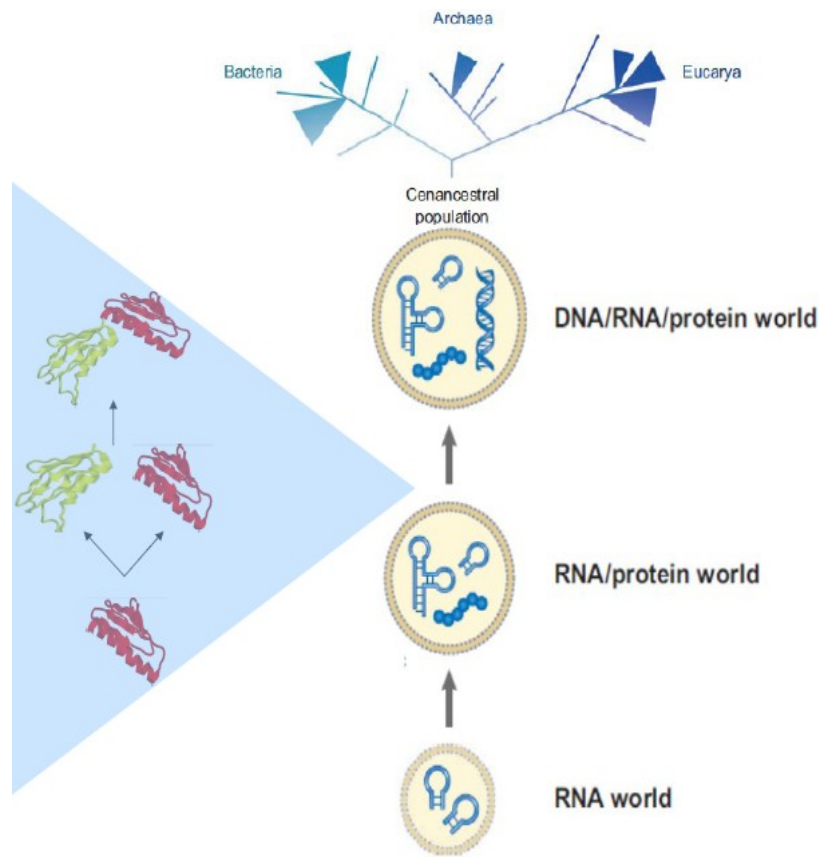


Figura 15. Antigüedad del mecanismo que genera duplicaciones internas. Figura modificada de Becerra *et al.* 2007.

## 6.8 Conservación de la huella de la duplicación

Cuando se analizaron los resultados arrojados por el programa LFASTA para cada una de las secuencias conservadas, encontramos que sólo en algunos representantes de un determinado COG el rastro de la duplicación interna fue evidente, mientras que en el resto de las secuencias, dentro del mismo COG, la huella de la duplicación se ha perdido a pesar de que la secuencia está ampliamente distribuida en los tres grandes linajes celulares, posee un alto grado de similitud, tiene aproximadamente la misma longitud y está descrita como parte del mismo conjunto de genes ortólogos.

Aparentemente, la pérdida del rastro de la duplicación no está relacionada con la distribución filogenética. De igual forma, la conservación de la huella de la duplicación interna de ciertas enzimas no parece tener relación aparente con la importancia funcional o actividad catalítica que desempeña. El análisis de la estructura terciaria de algunas secuencias que aparentemente han perdido la señal y de



aquellas que no la han perdido, pero que codifican para el mismo polipéptido, muestran que la topología es la misma, indicando que es más importante la conservación de la estructura y plegamiento tridimensional que la conservación de la estructura primaria.

Tanto la metodología empleada como los análisis de los resultados arrojados por esa metodología no nos permiten asegurar cuál podría ser la explicación que yace al fenómeno de la pérdida de la huella de la duplicación, por lo que en nuestras perspectivas a futuro intentaremos comprender el por qué ciertos genes han perdido el rastro de la duplicación.

Diversos trabajos acerca de la evolución en la duplicación de genes han sido enfocados principalmente a reconocer cómo es que un gen duplicado diverge tanto a nivel de secuencia como a nivel de función. Se ha postulado que después de que la duplicación se lleva a cabo, una serie de mutaciones ocurren en una de las copias del gen, mientras que la otra copia amortigua los efectos deletereos (Nei y Roychoudhury, 1973; Li, 1980). De esta manera, los genes recién duplicados pueden experimentar una evolución asimétrica en su secuencia, es decir, ambos genes poseerán tasas desiguales de divergencia. Así, uno de los duplicados evolucionará mucho más rápido que el otro (Dermitzakis y Clarck, 2001; Van de Peer *et al.*, 2001; Li y Tosoi, 2002; Kondrashov *et al.*, 2002; Wagner, 2002). Sin embargo, no sabemos si algo similar está ocurriendo a nivel interno de las secuencias, en donde cada una de las mitades podría estaría evolucionando a una tasa de mutación distinta de la que se estaría llevando a cabo en la otra mitad.

# VII CONCLUSIONES

El análisis comparativo de secuencias moleculares ha llegado a ser una aproximación poderosa para determinar las relaciones evolutivas de distintas especies y para generar ideas importantes de los estadios evolutivos que pudieron haber existido antes de la separación de los tres principales linajes celulares.

La reconstrucción de los rasgos biológicos que podría haber tenido el último ancestro común han mostrado que una gran parte de las secuencias altamente conservadas están relacionadas en mas de una forma con el metabolismo del RNA. Sin embargo, si se analiza este conjunto de secuencias se puede observar que entre ellas se incluyen moléculas grandes y complejas estructuralmente, por lo que es razonable suponer desde una perspectiva evolutiva, que éstas debieron haber sido precedidas por secuencias más sencillas.

El análisis de los genomas completos nos han revelado que múltiples secuencias pueden ser resultado de la expansión de genes parálogos en una época previa al LCA (Becerra *et al.* 2007). En particular, la búsqueda de secuencias formadas por módulos homólogos fusionados, arreglados en tandem, nos pueden dar pistas de una organización genética más sencilla y una complejidad bioquímica más simple de las entidades primitivas de las cuales el LCA evolucionó.

El hallazgo de un gran número de genes originados por duplicación interna y que forman parte del genoma de los organismos actuales, nos indica que este fenómeno se ha suscitado frecuentemente durante la evolución de las proteínas, siendo uno de los mecanismos que ha enriquecido el repertorio genético de los organismos y uno de los principales medios de amplificación génica para la formación de nuevos genes y nuevos procesos bioquímicos durante los estadios de la evolución temprana de la vida. Esto indica que el tamaño y la estructura de varias proteínas muy antiguas son el resultado evolutivo de duplicaciones parálogas seguidas por eventos de fusión que han tomado lugar previo a la divergencia de los tres reinos primarios, es decir, varios de los genes estructuralmente complejos encontrados en los genomas contemporáneos podrían haber evolucionado a partir de genes más sencillos vía duplicación interna y subsecuente modificación.

El proceso de duplicación intragénica es un mecanismo extremadamente antiguo, el cual podría haber actuado antes de la divergencia de los tres principales linajes celulares. Como hemos reportado en este trabajo, es posible que el fenómeno de duplicación interna haya actuado en estadios aún más

tempranos de la evolución celular, quizá en un mundo en el cual los genomas de DNA aún no habían aparecido y en donde los genomas celulares eran de RNA.

Para nuestras futuras perspectivas pretendemos:

- (1) conocer la causa de la pérdida de la huella de la duplicación en secuencias ortólogas de ciertos genomas
- (2) conocer los módulos más antiguos que se pueden reconocer a partir de la comparación de secuencias
- (3) analizar todas las estructuras terciarias de las bases de datos y reconocer en cuáles de ellas se puede distinguir su origen por duplicación interna
- (4) analizar secuencias de genomas completos con el objetivo de identificar qué otras proteínas se han originado por el mecanismo de duplicación interna
- (5) reconocer otros mecanismos de evolución en secuencias altamente conservadas, ej. reclutamiento de dominios

# VIII REFERENCIAS

- Alcántara, C., Cervera, J., Rubio, V. 2000. Carbamate kinase can replace in vivo carbamoyl phosphate synthase. Implications for the evolution of carbamoyl phosphate biosynthesis. *FEBS Lett* 484, 261-264
- Alifano, P., Fani, R., Liò, P., Lazcano, A., Bazzicalupo, M. 1996. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol. Rev.* 60, 44-69
- Anantharaman, V., Koonin, E.V. and Aravind, L. 2002. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acid Res.* 30, 1427-1464
- Andrade, M.A., Perez-Iratxeta, C., Ponting, C.P. 2001. Protein repeats: structures, functions, and evolution. *J Struct Biol* 134, 117-131
- Apic, G., Gough, J. and Teichmann, S.A. 2001. An insight into domain combinations. *Bioinformatics* 17, S83-S89
- Barker, W.C., Ketcham, L.K. and Dayhoff, M.O. 1978. A comprehensive examination of protein sequences for evidence of internal gene duplication. *J. Mol. Evol.* 10, 265-281
- Becerra, A., Islas, S., Leguina, J.I., Silva, E., Lazcano, A. 1997. Phylogenetic gene losses can bias backtrack characterizations of the cenancestor. *J. Mol. Evol.* 45, 115-118
- Becerra y Lazcano. 1998. The role of gene duplication in the evolution of purine nucleotide salvage pathways. *Orig Life Biosph.* 28, 539-553
- Becerra, A. 2000. El papel de las secuencias simples en la evolución temprana de la vida. Tesis doctoral. Facultad de Ciencias, UNAM.
- Becerra, A., Delaye, L., Islas, S. And Lazcano, A. 2007. The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. *Annu. Rev. Ecol. Evol. Syst.* 38, 361-379
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and

Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Research* 28, 235-242

Briones, C., Manrubia, S.C., Lazaro, E., Lazcano, A., Amils, R. 2005. Reconstructing evolutionary relationships from functional data: a consistent classification of organisms based on translation inhibition response. *Mol Phylogenet Evol.* 4, 371-81

Brosius, J. 2003. Gene duplication and other evolutionary strategies: from the RNA world to the future. *Journal of Structural and Functional Genomics* 3, 1-17

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J. 2001. Universal trees based on large combined protein sequence datasets. *Nat. Genet.*, 28, 281-285

Chothia, C. 1992. One thousand families for the molecular biologist. *Nature* 357, 543-544

Chothia, C. and Gough, J. 2009. Genomic and structural aspects of protein evolution. *Biochem J.* 419, 15-28

Clayton, R.A., White, O., Ketchum, K.A., Venter, C.J. 1997. The genome from the third domain of life. *Nature* 387, 459-462

Darwin C. 1859. *On the origin of species by means of Natural Selection, or the preservation of favoured races in struggle for life.* 1<sup>st</sup> edition. Londres: John Murray

Daubin, V., Gouy, M., Perriere, G. 2001. A phylogenomic approach to bacterial phylogeny: evidence for a core of genes sharing a common history. *Genome Res.* 12, 1080-1090

de la Cruz, F. and Davies, J. 2000. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* 8, 128-133

Delaye *et al.*, 2004. The nature of the last common ancestor. In Lluís Ribas de Pouplana, Ph.D. (ed) *The genetic code and the origin of life.* (Landes Bioscience and Kluwer academic) pp. 34-47

Delaye, L., Becerra, A. and Lazcano, A. 2005. The last common ancestor: what's in a name? *Ori Life Evol Biosph* 35, 537-554

Delaye, L. and Lazcano, A. 2000. RNA-binding peptides as molecular fossils In J. Chela-Flores, G. Lemerchand, and J. Oró (eds), *Origins from the Big-Bang to Biology: Proceedings of the First Ibero-American School of Astrobiology*: Kluwer Academic Publishers, Dordrecht. pp. 285-288

Dermitzakis, E.T. and Clark, A.G. 2001. Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.* 18, 557-562

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* 284: 2124-2128

Doolittle, W.F. 2000. The nature of the universal ancestor and the evolution of the proteome. *Curr. Opinion Struct. Biol.* 10, 355-358

Eisen, J.A. 2000. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Curr. Opin. Genet. Dev.* 10, 606-611

Fitch, W.M. and Upper, K. 1987. The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. *Cold Spring Harbor Symposia on Quantitative Biology* 52, 759-767

Fitz-Gibbon, S.T. and House, C.H. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* 27, 4218-22

Fraser et al., 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature.* 390, 580-6.

Gibbs, A.J. and McIntyre, G.A. 1970. The diagram method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* 16, 1-11

Gogarten, J.P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E.J. 1989. Evolution of the vacuolar H<sup>+</sup>



-ATPase, implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. USA* 86, 6661-6665

Gogarten, J.P., Doolittle, W.F. and Lawrence, J.G. 2002. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* 19, 2226-2238

Hancock, J.M. 1995. The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* 41, 1038-1047

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

Heringa, J. 1998. Detection of internal repeats: how common are they? *Cur. Op. Struct. Biol.* 8, 338-345

Heringa, J. and Argos, P. 1993. A method to recognize distant repeats in protein sequences. *Proteins Struct. Funct. Genet.* 17, 391-411

Heringa, J. and Taylor, W.R. .1997. Three-dimensional domain duplication, swapping and stealing. *Cur. Op. Struct. Biol.* 7, 416-421

Huang, X. and Miller, W. 1991. A time-efficient, linear-space local similarity algorithm. *Advances in Applied Mathematics* 357

Hughes, A.L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc R Soc Lond B* 256, 119-124

Islas, S., Hernández-Morales, R., Lazcano, A. 2007. Question 7: comparative genomics and early cell evolution: a cautionary methodological note. *Ori Life Evol Biosph* 37, 415-418

Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S., Miyata, T. 1989. Evolutionary relationship of archeobacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* 86, 9355-9359

- Jorde, L.B., Watkins, W.S., Bamshad, M.J. 2001. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet.* 10, 2199-207
- Kanehisa, M. and Goto, S. 2000. Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27-30
- Kanehisa, M., and Bork, P. 2003. Bioinformatics in the post-sequence era. *Nat Genet.* 33, 305-10
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. 2006. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354-357
- Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355-D360
- Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I. and Koonin, E.V. 2002. Selection in the evolution of gene duplications. *Genome Biology* 3, 1-9
- Koonin, E.V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews* 1, 127-136
- Kunin, V. and Ouzounis, C.A. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 13, 1589 - 1594
- Kurtz, S. and Schleiermacher, C. 1999. REPuter--fast computation of maximal repeats in complete genomes. *Bioinformatics* 15, 426-427
- Kurtz, S., Choudhuri, J., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633-4642
- Kyrpides, N., Overbeek, R. and Ouzounis, C. 1999. Universal protein families and the functional

content of the last universal common ancestor. *J. Mol. Evol.* 49, 413-423

Lawrence, J.G. and Roth, J.R. 1996. Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* 143, 1843-1860

Lazcano, A. 1995. Cellular evolution during the early Archean: what happened between progenote and the cenancestor? *Microbiologia SEM* 11: 185-198

Lazcano, A., Fox, G.E. and Oró, J. 1992, Life before DNA: the origin and early evolution of early archean cells. In R. P. Mortlock : (ed), *The Evolution of Metabolic Function* : CRC Press, Boca Raton, FL,, pp. 237-295

Li, W.-H. 1980. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. *Genetics* 95, 237-258

Li, W.-H. 1983. Evolution of duplicate genes and pseudogenes. *Evolution of genes and proteins*. Sinauer, Sunderland, MA. 14-37.

Li, Y.J. and Tsoi, S.C.-M. 2002. Phylogenetic analysis of vertebrate lactate dehydrogenase (LDH) multigene families. *J. Mol. Evol.* 54, 614-624

Marcotte, E. M., Pellegrini, M., Yeates, T. O. and Eisenberg, D. A. 1998. Census of protein repeats. *J. Molec. Biol.* 293, 151-160

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

Moreira, D. and López-García, P. 2006. The last common ancestor. *Earth Moon Planets* 98, 187-193

Morgan, G.J. 1998. Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959-1965. *Journal of the history of biology* 31, 155-178

- Murzin, A., Brenner, S.E., Hubbard, T., Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536- 540
- Mushegian, A.R. and Koonin, E.V. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* 93, 10268-10273
- Nei, M. and Roychoudhury, A.K. 1973. Probability of fixation of nonfunctional genes at duplicated loci. *Am. Nat.* 107, 362-372
- Nyunoya, H., Lusty, C.J. 1983. The *carB* gene of *Escherichia coli*: a duplicated gene coding for the large subunit of carbamoyl-phosphate synthetase. *Proc. Natl. Acad. Sci.* 80, 4629-4633
- Nutall G.H.F. *et al.*, 1904. Blood immunity and blood relationship, a demonstration of certain blood-relationships amongst animals by means of the precipitin test for blood. University press
- Ochman H., Lawrence J.G. and Groisman E.A. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299-304
- Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, NY.
- Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. 1997. CATH a hierarchic classification of protein domain structures. *Structure* 5, 1093-1108
- Pagel, M. 2000. Phylogenetic evolutionary approaches to bioinformatics. *Brief Bioinform.* 1, 117-30
- Pearson, W.R., Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci* 85, 2444-2448
- Pellegrini, M., Marcotte, E. M. and Yeates, T. O. 1999. A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins: Struct. Funct. Genet.* 35, 440-446

- Ragan, M.A. 2001. Detection of lateral gene transfer among microbial genomes. *Curr Opin Genet Dev* 11, 620-626
- Raymond, J. and Blankenship, R.E. (2004) Biosynthetic pathways, gene replacement and antiquity of life. *Geobiology* 2, 199-203
- Ren, B., Tibbelin, G., de Pascale, D., Rossi, M., Bartolucci, S., Ladenstein, R., 1998. A protein disulfide oxidoreductase from the archaeo *Pyrococcus furiosus* contains two thioredoxin fold units. *Nat. Struct. Biol.* 7, 602-611
- Riley, M. and Labedan, B. 1997. Protein evolution viewed through *Escherichia coli* protein sequences: Introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.* 268, 857-868
- Snel, B., Bork, P., Huynen, M.A. 1999. Genome phylogeny based on gene content. *Nat Genet.* 21, 108-110
- Tautz, D., Trick, M. and Dover, G.A. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322, 652-656
- Tekaia, F., Lazcano, A., Dujon, B. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* 9, 550-7
- Van de Peer, Y., Taylor, J.S., Braasch, I. and Meyer, A. 2001. The ghost of selection past: Rates of evolution and functional divergence of anciently duplicated genes. *J. Mol. Evol.* 53, 436-446
- Wagner, A. 2002. Asymmetric functional divergence of duplicate genes in yeast. *Molecular Biology and Evolution* 19, 1760-1768
- Woese, C.R. and Fox, G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* 74, 5088-5090
- Woese C.R., Kandler O., Wheelis M.L. 1990. Towards a natural system of organisms, proposal for the

domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* 87:4576–79

Wolf, Y.I., Grishin, N.V. and Koonin, E.V. 2000. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* 299, 897-905

Wolfe, K.H. and Li, W.H. 2003. Molecular evolution meets the genomics revolution. *Nat Genet.* 33, 255-65

Yang, S., Doolittle, R.F., Bourne, P.E. 2005. Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A.* 102, 373-8

Zuckermandl, E., Pauling, L. 1965. Molecules as documents of evolutionary history. *J Theor Biol.* 8, 357-66

Zuckermandl, E., Jones, R.T. and Pauling L. 1960. A Comparison of animal hemoglobins by tryptic peptide pattern analysis. *Proc. Nat. Acad. Sci.* 46, 1349-1360