UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ANÁLISIS FUNCIONAL DE PROTEÍNAS PARÁLOGAS EN GENOMAS PROCARIONTES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

BIÓLOGO

P R E S E N T A :

ALEJANDRO ALBERTO ÁLVAREZ LUGO

DIRECTOR DE TESIS:
DR. ARTURO CARLOS II BECERRA BRACHO

2016

CIUDAD UNIVERSITARIA, CD. MX.

Hoja de Datos del Jurado

1. Datos del alumno

Álvarez
Lugo
Alejandro Alberto
55 90 33 10
Universidad Nacional Autónoma de México
Facultad de Ciencias
Biología
307521493

2. Datos del Tutor

Doctor
Arturo Carlos II
Becerra
Bracho

3. Datos del Sinodal 1

Doctor
Rafael
Camacho
Carranza

4. Datos del Sinodal 2

Doctora
María
Colín
García

5. Datos del Sinodal 3

Maestro en Ciencias
Ricardo
Hernández
Morales

6. Datos del Sinodal 4

Biólogo
Amadeo Luis
Estrada
Nieto

7. Datos del Trabajo Escrito

Análisis funcional de proteínas parálogas en genomas procariontes
78 pp
2016

A mis padres y a mi hermana, por todo su amor y apoyo incondicional.

A mi abuela Judith, y a la memoria de mis abuelos Alfonso, Armando y Lidia.

Resumen

La duplicación de genes es un proceso fundamental de la evolución biológica debido a que constituye uno de los principales mecanismos por los cuales se pueden originar nuevos genes y nuevas funciones. El análisis de los genomas de organismos contemporáneos no nos permite identificar cuáles son los genes que más se han duplicado, sin embargo, sí podemos saber cuáles son aquellos que se han retenido con mayor frecuencia y que han formado familias de genes parálogos. Por lo tanto, resulta conveniente clasificar a los genes (o sus productos) con base en algún criterio biológico que nos permita identificar categorías funcionales de genes con una mayor proporción de parálogos.

En el presente trabajo elegí dos clasificaciones: (1) una basada en la clasificación enzimática universalmente aceptada (NC-IUBMB, 1992), que consta de seis clases generales y que se analiza desde un punto de vista bioquímico y (2) otra con base en las categorías funcionales presentes en la Enciclopedia de Genes y Genomas de Kyoto (KEGG, por sus siglas en inglés). Debido a que en organismos procariontes existe una mayor diversidad de procesos metabólicos y que sus proteomas suelen estar mejor anotados que los de eucariontes, decidí trabajar únicamente con miembros de los dominios Archaea y Bacteria.

Los resultados indican que de las seis clases enzimáticas, sólo la clase de las óxidorreductasas presenta una proporción de parálogos significativamente mayor al resto de las otras clases, situación que puede explicarse por la presencia de grupos funcionales que ayudan a la catálisis en las enzimas de esta clase y que favorecen la aparición de actividades secundarias. Es probable que haya ocurrido este proceso en etapas tempranas de la evolución biológica. Si dichas actividades secundarias representaban una ventaja para el organismo, eventos de duplicación y posterior divergencia pudieron haber llevado a la especialización funcional en las óxidorreductasas ancestrales.

Por otra parte, mediante el esquema de clasificación de KEGG, los resultados indican que las categorías con mayor proporción de parálogos están asociadas con la interacción organismo-ambiente (por ejemplo, aquellas asociadas al transporte membranal, la transducción de señales y el metabolismo de sustancias xenobióticas). En contraparte, aquellas categorías que abarcan procesos esenciales en los organismos (y que suelen ser de las más conservadas) presentan una proporción muy baja de parálogos.

En conclusión, creo que existe una retención preferencial de genes parálogos en ambos dominios celulares, y aquellas categorías que poseen una mayor proporción de éstos se encuentran relacionadas directamente con actividades celulares que probablemente no están sujetas a presiones fuertes de selección. Por último, los resultados obtenidos con base en cada clasificación reflejan la gran diversidad metabólica presente en procariontes y su capacidad de hacer frente a nuevas condiciones ambientales.

Agradecimientos

Mi más profundo agradecimiento al Dr. Antonio Lazcano, por haberme dado la oportunidad de ingresar y ganarme un lugar dentro de este gran laboratorio. De igual manera, le agradezco enormemente a mi asesor, el Dr. Arturo Becerra, por todo su apoyo y paciencia que ha tenido conmigo a lo largo del desarrollo de este trabajo. A mi amigo y maestro Ricardo Hernández, por su valiosa ayuda en el diseño de algunos de los scripts utilizados para esta tesis.

A mi familia y a Dios, por haber formado parte de mi vida desde que tengo uso de razón.

A mis hermanos, Alan, Hugo, Jos y Rafa: porque no concibo mi vida sin ustedes cuatro, ¡locos!

A mis amigos y colegas del Laboratorio de Origen de la Vida: Beto, Ricardo, Pepe, Isra, Pablo, Coral, Alex, Caro, Mario, Rodrigo, Clau, Wolfgang, Luciana, Héctor, Sara y Ervin, por todo el apoyo y grandes momentos que he compartido con ustedes.

A mis padres, Claudia y Rubén, así como a mi queridísima hermana Daniela, por ser lo más valioso de mi vida. A mi abuela Judith y a mis tíos Alfonso y Rosario, por haber despertado en mí el gusto por la Biología.

A mis amigos de toda la carrera: Lalo, Mike y Sandy, por haber compartido conmigo tantas materias y momentos de diversión.

Índice

Resumen……..……..4
Agradecimientos…….…….6
Índice……7
1. Introducción…….…..9
1.1. Existen diferentes clases de genes homólogos……..……9
1.2. Breve reseña histórica ……..……...10
1.3. Destinos posibles de genes duplicados……11
1.3.1. Neofuncionalización…….……..11
1.3.2. Subfuncionalización……..12
1.3.3. Prevalencia para efectos de dosis……13
1.3.4. Formación de pseudogenes…….……..13
1.4. Sesgos funcionales en la retención de genes duplicados…….……15
1.5. El papel de las duplicaciones génicas en la evolución del metabolismo…...18
1.6. Algunos modelos de duplicación génica…….……..…19
1.6.1. DDC (Duplication-Degeneration-Complementation)……..…19
1.6.2. EAC (Escape from the Adaptive Conflict)……..…20
1.6.3. IAD (Innovation, Amplification and Divergence)……20
1.7. Clasificación funcional de proteínas ……21
Planteamiento del problema y objetivos……..……..24
2. Metodología…….……..26
3. Resultados……....31
3.1 Proporción de enzimas con base en la clasificación IUPAC.……...……31
3.2 Proporción de enzimas con base en las categorías funcionales de KEGG PATHWAY…….……...36
3.3 El número de parálogos presenta una correlación positiva con respecto al tamaño del genoma……..…41
4. Discusión…….…….…44
4.1 Las categorías funcionales con mayor proporción de parálogos están asociadas con la interacción ambiental ……..…44
4.2 El caso del metabolismo y la degradación de xenobióticos ……..……..…46
4.3 La proporción de parálogos para la clase de las óxidorreductasas difiere significativamente del resto de las clases enzimáticas.....49

4.4 La estructura química y función de los metabolitos en los organismos influye en la retención de enzimas parálogas ……..…..……54
Conclusiones ……58
Referencias…..……..60
Anexos……68

1. Introducción

1.1 Clases de genes homólogos.

Los genes homólogos son todos aquellos que se originaron a partir de un gen ancestral común. Sin embargo, existen varias formas a partir de las cuales un gen ancestral puede originar a otro, y por lo tanto, dar lugar a diferentes clases de genes homólogos. Entre los genes homólogos se encuentran los ortólogos, los parálogos (Fitch, 1970; Gogarten y Olendzenski, 1999) y los xenólogos (Fitch, 1970; Olendzenski *et al*., 2005) (Fig. 1). Los genes ortólogos son producto de un evento de especiación, es decir, una especie ancestral diverge y en cada una de las especies resultantes habrá una copia del gen *X*. Esto significa que dicho gen estará presente en el genoma de las dos especies descendientes. Los genes parálogos son aquellos que se duplican en algún momento que no coincide con la especiación, lo cual trae como consecuencia que haya dos copias del mismo gen en el genoma de ese organismo (Fitch, 1970; Gogarten y Olendzenski, 1999).

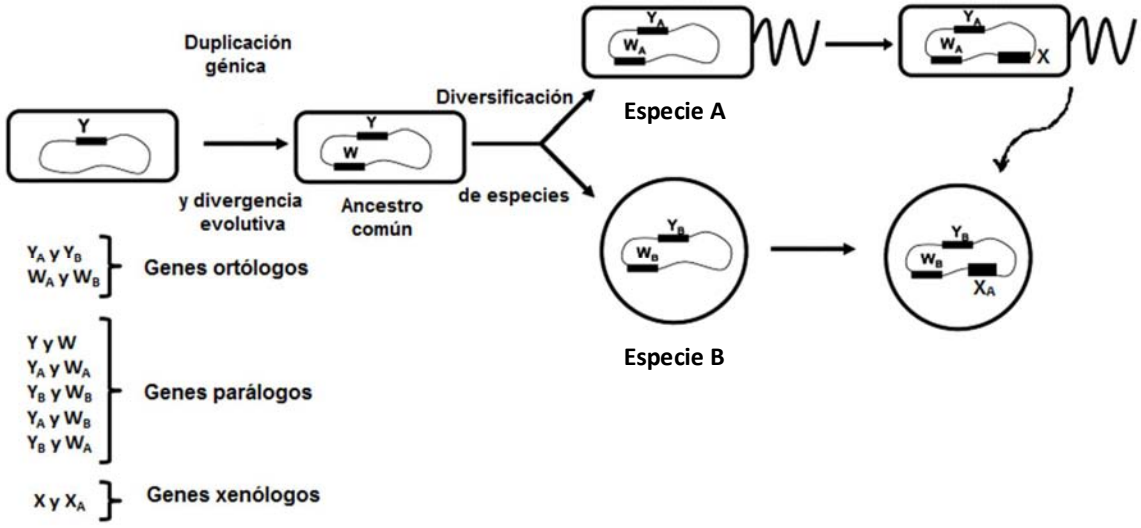


Figura 1. Distinción entre genes ortólogos y parálogos. En una especie inicial, uno de los genes (Y) se duplica y posteriormente diverge (W). Posteriormente, ocurre un evento de especiación a partir de un ancestro común, lo cual genera dos especies distintas (A y B) por medio de un proceso de diversificación. Cada especie poseerá una copia de los genes ancestrales W y Y (denotados como W_A y Y_A para la especie A y W_B y Y_B para la especie B. Adicionalmente, una de las dos especies puede adquirir un nuevo gen (no necesariamente por duplicación) y éste puede ser transferido a otra mediante transporte horizontal, formándose así un par de genes xenólogos (X y X_A). Modificado de Fani y Fondi, 2009.

Existe otro tipo de genes homólogos que no son producto de divergencia de linajes ni de eventos de duplicación génica, sino que son adquiridos por transferencia horizontal, es decir, no son heredados directamente por los progenitores. Se dice que son adquiridos horizontalmente porque el organismo donador y el aceptor coexisten en tiempo y espacio. El mecanismo más común es, quizá, por medio de plásmidos (como ocurre con los genes que confieren resistencia a antibióticos); a los genes adquiridos por esta vía se les denomina xenólogos (Olendzenski *et al*., 2006). Una característica adicional de este tipo de genes es que, en ocasiones, su filogenia no coincide con la de los organismos en que se encuentran (Koonin *et al*., 2001) por lo que representan un problema para la reconstrucción de filogenias de especies, debido a que pueden ser difíciles de identificar y se pueden confundir con ortólogos y/o parálogos (Lawrence y Hendrickson, 2003).

1.2 El estudio de las duplicaciones génicas

Las investigaciones referentes a este tema (que inicialmente concernía a la genética clásica) comenzaron en los albores del siglo XX. En dicha época, la única forma en la que los genetistas podían detectar la duplicación de algún locus (o de loci) era por medio de análisis citogenéticos, los cuales permitían examinar los cromosomas directamente con un microscopio y así detectar segmentos (e incluso cromosomas completos) que aparentemente se hallaban repetidos, aunque aún no se especulaba acerca de la función o importancia que podían tener éstos (Blakeslee *et al*., 1920; Babcock y Collins, 1922). En la década siguiente, gracias a los trabajos de genetistas como Herman Müller y Calvin Bridges, y a la brillantez de teóricos como J. B. S. Haldane, se pudo especular acerca de las implicaciones que tendría la duplicación de uno o más genes (Müller, 1934; Bridges, 1935) e incluso del genoma completo (poliploidía) o de una gran fracción de éste (Haldane, 1933), en la prevalencia de los individuos y de las especies.

Años más tarde, a principios de la década de los 70, Susumu Ohno publicó una pequeña obra intitulada *Evolution by Gene Duplication*, la cual conjuntó un gran número de trabajos que sobre genes y proteínas duplicadas. Esta obra representó un parteaguas para la biología debido a que colocó a la duplicación génica como uno de los procesos más importantes en la evolución biológica. Lo más importante es quizá su propuesta de que eventos distintos de duplicación de genoma completo (WGD, por sus siglas en inglés) habían sido determinantes en la diversificación de grandes linajes como el de los vertebrados (Ohno, 1970; Taylor y Raes, 2004).

Durante las tres décadas posteriores, gracias al aumento en la secuenciación de genes y genomas completos, se fue desarrollando una serie de herramientas bioinformáticas que permitió abordar el problema de las duplicaciones génicas por medio de métodos *in silico*, lo cual facilitó que se pudiera detectar específicamente qué genes han sido producto de duplicación dentro de un mismo genoma, así como la comparación entre los genomas de organismos diferentes. En el primer caso, cuando en un mismo organismo existe un conjunto de genes que son homólogos, se trata de una familia de genes parálogos, los cuales son producto de diferentes eventos de duplicación a lo largo de la historia evolutiva del organismo (Fig. 2).

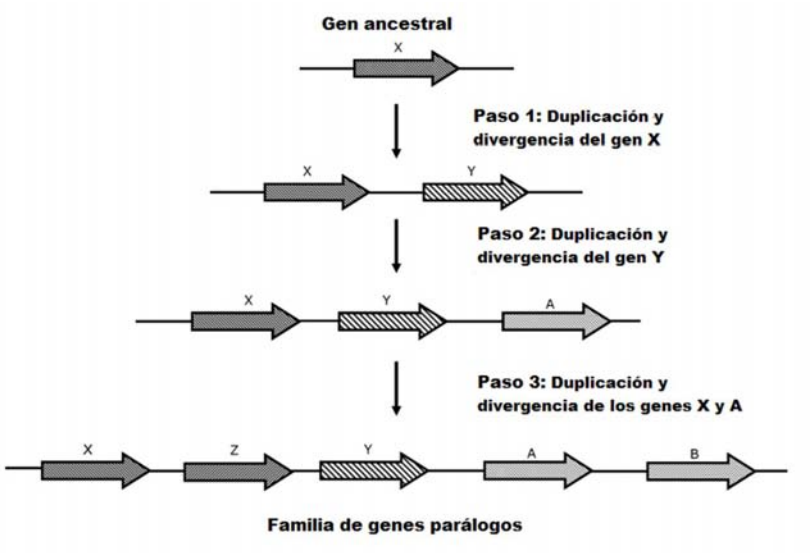


Figura 2. Esquema en el que se muestra, de manera simplificada, el proceso de formación de una familia de genes parálogos (modificado de Fani y Fondi, 2009).

1.3 Destinos de genes duplicados

1.3.1 Neofuncionalización

En su obra *Evolution by Gene Duplication*, Susumu Ohno estableció por primera vez que el papel principal de la duplicación génica en la evolución biológica era la creación de un nuevo gen a partir de una copia reduntante originada por duplicación (Ohno, 1970) (Fig. 3d). Este proceso está basado en lo que él llama mutaciones “prohibidas” que involucran una tasa de sustitución no sinónima asimétrica en una de las copias y se asume que puede adquirir mutaciones libremente debido a que las presiones de selección sobre ella

están más relajadas (Ohno, 1970; Maere y van de Peer, 2010). Por el contrario, la otra copia permanece casi sin cambios y su tasa de sustitución no sinónima es menor que la de su parálogo (Maere y van de Peer, 2010).

De acuerdo con el modelo original, una vez que un gen se duplica, una de las copias puede perder su funcionalidad debido a mutaciones y probablemente las seguirá acumulando hasta que haya adquirido una función totalmente diferente a la de la otra copia y gracias a esto pueda ser fijado en el genoma (Ohno, 1970; Conant y Wolfe, 2008). Sin embargo, conforme fueron avanzando los estudios dentro de este campo, se observó que la preservación de genes parálogos con base en el modelo tradicional de Ohno ocurre muy rara vez (Maere y van de Peer, 2010) y en la mayoría de los casos el gen con una tasa de sustitución asimétrica adquiere una función químicamente relacionada a la original y no una función nueva y diferente, como el caso de los genes de opsina de humanos (sensibles al rojo y al verde), los cuales provienen de un evento de duplicación previo a la diversificación de homínidos y monos del viejo mundo, y gracias a los cuales podemos percibir un rango muy amplio de colores (Zhang, 2003).

1.3.2 Subfuncionalización

Como su nombre lo indica, el término subfuncionalización hace referencia a la división de funciones a partir de un gen ancestral, lo cual implica que éste llevaba a cabo por lo menos dos funciones distintas antes de la duplicación. Una vez que ocurre este proceso, cada una de las copias parálogas se encargará de una sola de las funciones originales, lo cual permitirá la preservación de ambas copias debido a que conservarán las funciones que llevaba a cabo el gen original (Maere y van de Peer, 2010) (Fig. 3c). Un ejemplo clásico es el par de proteínas parálogas *RNASE1* y *RNASE1B*. Ambas presentan actividad ribonucleolítica y se encuentran en los monos douc langur, los cuales pertenecen a la subfamilia Colobinae. Estas proteínas tienen una función en común: degradar RNA, pero mientras que *RNASE1* posee la capacidad de degradar RNA de doble cadena (dsRNA), *RNASE1B* únicamente degrada RNA de una sola cadena (ssRNA), y además, su actividad óptima ocurre en un pH ácido, a diferencia de su copia *RNASE1* (Zhang *et al*., 2002). Al comparar ambos genes con el ortólogo en otras especies de monos, los autores identificaron que la copia *RNASE1B* era la que se había especializado hacia una función digestiva, a costa de perder la capacidad de degradar dsRNA (Zhang *et al*., 2002; Zhang, 2003). Por lo tanto, se puede asumir que el gen

ancestral poseía actividad dsRNAsa y RNAsa incluso en un pH ácido, pero después del evento de duplicación una de las copias optimizó la primera función, mientras que la otra se especializó en la segunda función (Zhang *et al*., 2002).

Aunque casos como el anterior sean la evidencia más fehaciente del proceso de subfuncionalización, éste puede darse también a nivel de expresión génica ya sea espacial o temporalmente, es decir, que cada una de las copias parálogas se exprese en un sitio distinto de un mismo organismo o que se exprese en el mismo sitio pero en diferentes tiempos. Un caso muy conocido es el de los genes *engrailed-1* y *engrailed-1b* del pez cebra (Force *et al*., 1999; Zhang, 2003; Taylor y Raes, 2004). Ambos son factores de transcripción, resultado de una duplicación cromosómica segmental y son homólogos al gen *engrailed-1* del ratón pero, mientras que en el ratón dicho gen se expresa tanto en el apéndice pectoral como en la espina dorsal, en el pez cebra *engrailed-1* se expresa en el apéndice pectoral, y *engrailed-1b* en la espina dorsal (Force *et al*., 1999; Zhang, 2003).

1.3.3 Conservación de la función original

En ciertos casos, la retención de un gen duplicado que tenga la misma función que la copia original puede ser benéfica para el organismo (Ohno, 1970; Zhang, 2003; Hahn, 2009). ¿En qué casos puede ser así? Por ejemplo, cuando se requiere una mayor cantidad de proteínas, como el caso de las histonas, proteínas ribosomales, o transcritos de RNA, en particular, el RNA ribosomal (rRNA) (Zhang, 2003; Maere y van de Peer, 2010) (Fig. 3b). Existen dos mecanismos por los cuales es posible que dos genes parálogos permanezcan idénticos en cuanto a función y con muy pocos cambios en cuanto a secuencia. El primero de ellos se conoce comúnmente como evolución concertada (Liao, 2008), e implica eventos frecuentes de conversión génica. El segundo establece que la selección natural purificadora, que actúa en contra de aquellas mutaciones que alteren la función nativa de la proteína, es la que evita que un par de genes parálogos diverjan mucho (Zhang, 2003). En pocas palabras, el requerimiento de dosis adicionales de determinado producto o transcrito es la presión selectiva que permite la coexistencia de más de una copia que lleve a cabo la misma función.

1.3.4 Formación de pseudogenes

El destino más común para los genes que se duplican es que no encuentren un nicho

disponible y por lo tanto, no tengan ningún valor adaptativo para el organismo en cuestión. La acumulación de mutaciones en su mayoría deletéreas llevará a una de las copias a que pierda totalmente su función (Liberles *et al*, 2010) (Fig. 3a). Se piensa que, en ausencia de cualquier tipo de selección, este proceso suele ocurrir en los primeros millones de años posteriores a la duplicación (Lynch y Connery, 2000; Zhang, 2003). Quizá el ejemplo más conocido sea el de la familia de los receptores olfatorios, que consta de aproximadamente mil secuencias diferentes en varios grupos de mamíferos, con la diferencia de que en nosotros aproximadamente el 60% constituye pseudogenes, seguido por un 30% en chimpancés y en ratones solo un 20% (Zhang, 2003; Taylor y Raes, 2004). La mejor explicación para este hecho es que el uso de nuestro sentido del olfato se ha reducido mucho gracias a que puede ser compensado por otros sentidos (Zhang, 2003).

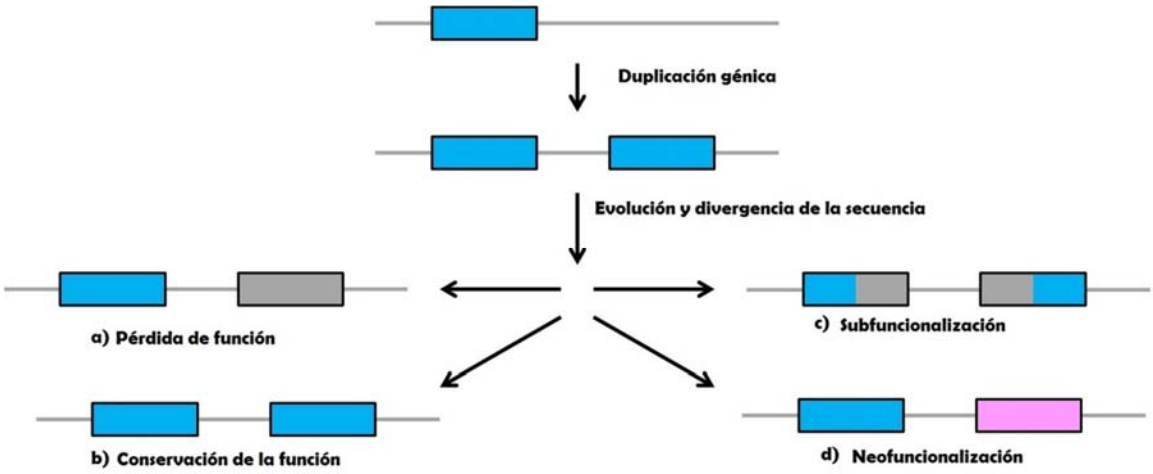


Figura 3. Destinos posibles de los genes duplicados. A partir de un gen ancestral, pueden surgir nuevos genes mediante procesos de duplicación y posterior divergencia. En a) se representa el destino más común para los genes duplicados, que consiste en la pérdida de función para una de las copias, la cual puede ser eliminada por completo del genoma o permanecer como pseudogén. Otro de los destinos posibles, el cual se muestra en b), no implica modificación ni pérdida de la función original sino la conservación de ésta. Este hecho se asocia muchas veces con el requerimiento de una dosis mayor de determinado producto génico, misma que representa la presión selectiva que mantiene a la copia, y aunque ocurren mutaciones, éstas no tienen un efecto negativo sobre la función del gen. También puede darse el caso, como se ilustra en c), que cada una de las copias pierda una de las funciones ancentrales que llevaba a cabo el gen previamente al evento de duplicación, a lo cual se le conoce como subfuncionalización. Finalmente, d) representa el destino menos común, llamado neofuncionalización, que implica la pérdida total (o en

ocasiones parcial) de la función original debido a mutaciones, mismas que a su vez serán responsables de que dicho gen adquiera una nueva función.

1.4 Sesgos funcionales en la retención de genes duplicados

Al buscar genes parálogos dentro del genoma de algún organismo, observamos aquellos que han sido retenidos después de eventos de duplicación, sin embargo, esto no significa que estos genes sean los que más se duplican, debido a que en principio, éste es un proceso aleatorio y todos los genes poseen la misma probabilidad de duplicarse. La diferencia es que no todos poseen la misma probabilidad de amplificar su número de copias ni de ser retenidos después de un evento de duplicación.

Actualmente, existe un gran número de trabajos en los que se ha tratado de identificar aquellas clases de genes que presentan una mayor retención después de un evento de duplicación. Para ello, los genes se han clasificado con base en criterios biológicamente importantes. Sin embargo, las clasificaciones suelen ser muy diferentes (Conant y Wagner, 2002; Papp *et al*., 2003; Gevers *et al.*, 2004; Maere *et al*., 2005; He y Zhang, 2005, 2006; Gout *et al*., 2009) y el número de genes reportados como duplicados puede variar dependiendo de los criterios utilizados previamente al análisis y del método bioinformático que se utilice (Zhang, 2003). Por ejemplo, Balázs Papp y su grupo lograron identificar que los genes involucrados en la formación de complejos proteínicos (como en el caso de las chaperonas) presentan un número similar de parálogos. La explicación más recurrente tiene que ver con efectos de dosaje, es decir, si una de las proteínas formadoras del complejo presentara una alta duplicabilidad, cualquiera de estas copias podría ocupar su lugar en el complejo, lo cual representaría un desbalance para el organismo, al modificar la concentración de ciertas proteínas. Por esta razón, además de tener una proporción similar de parálogos, ésta es también relativamente baja para este grupo de genes (Papp *et al.,* 2003). Al año siguiente, Marland *et al.* (2004) identificaron en *Saccharomyces cerevisiae* y *Escherichia coli* que los genes que codifican para enzimas del metabolismo central presentan mayor duplicabilidad con respecto a aquellos del metabolismo no central y a los no metabólicos. Sin embargo, puede haber un sesgo debido a que en estos organismos no hay muchos genes reguladores (tales como factores de transcripción, genes involucrados en transducción de señales, etc.). Asimismo, He y Zhang clasificaron a un grupo de genes parálogos de varios eucariontes con base en su complejidad (aquellos con las secuencias más largas, mayor número de dominios

funcionales y mayor número de motivos reguladores en *cis*), lo cual les permitió postular que aquellos genes con mayor complejidad representarían a la mayoría de los duplicados (He y Zhang, 2005). Un año más tarde, estos autores realizaron un estudio similar (esta vez sólo en *S. cerevisiae* y en otras especies del mismo género) en el que clasificaron a los genes en tres grupos: (1) unicopia en cada especie, (2) unicopia en *S. cerevisiae* pero con por lo menos dos homólogos en otras especies y (3) duplicados específicos para cada linaje en por lo menos uno de los genomas que no sea *S. cerevisiae*. Gracias a esto, pudieron identificar que los genes no esenciales son los que tienden a retenerse en mayor proporción dentro de un evento de duplicación gracias a que poseen un mayor número de reguladores en *cis* (He y Zhang, 2006).

Una limitante de algunos de estos trabajos es que utilizaron únicamente el conjunto de genes de un organismo, u organismos filogenéticamente cercanos, por lo que no se pueden extrapolar, de una manera más general, los patrones de retención de genes duplicados hacia grupos filogenéticamente más distantes. En general, aquellas categorías funcionales con un mayor número de duplicados retenidos suelen variar entre organismos de diferentes dominios celulares e incluso entre organismos pertenecientes al mismo dominio, mientras que hay otras cuya proporción de retención es muy parecida entre organismos filogenéticamente distantes, como los que participan en el transporte a través de membrana (Conant y Wagner, 2002). Además, existen genes metabólicos duplicados (los cuales codifican para enzimas) que son específicos a nivel de especie, mismos que generalmente confieren características metabólicas únicas al organismo en el que se encuentran (Serres *et al*., 2009).

Otro sesgo importante es que la mayoría de estos estudios se ha hecho utilizando únicamente eucariontes, o eucariontes y bacterias (Conant y Wagner, 2002; Marland *et al*., 2004). Una diferencia fundamental entre eucariontes y procariontes es que en este último grupo no se observan huellas de eventos de duplicación a gran escala (específicamente WGD), por lo cual las categorías con más parálogos en eucariontes no son las mismas para los procariontes. Uno de los primeros estudios para el cual se utilizaron únicamente organismos procariontes fue hecho por el grupo de Yves van de Peer en el 2004, en el cual primeramente se halló una correlación entre el tamaño del genoma y el número de genes parálogos en dicho genoma y posteriormente se identificaron dos grupos de duplicados que son retenidos: aquellos que presentan una retención preferencial y aquellos que no. En el primer grupo se ubican los genes que participan en el metabolismo de aminoácidos, regulación de la transcripción y transporte

de iones inorgánicos; mientras que en el segundo están los involucrados en metabolismo de carbohidratos, mecanismos de defensa, y en la producción y conversión de energía (Gevers *et al*., 2004). Varios años más tarde Bratlie *et al.* (2010) encontraron que la mayor cantidad de parálogos pertenecen a las categorías de producción de energía, movimiento celular, transporte de iones, transducción de señales y los involucrados en mecanismos de defensa. Estos grupos de genes podrían proporcionar alguna ventaja competitiva (Bratlie *et al*., 2010), las cuales son fundamentales para la adaptación a nuevos ambientes. En ambos estudios se concluyó que las categorías de genes con mayor número de parálogos tienen que ver con la adaptación a diferentes ambientes, los cuales suelen ser más cambiantes en procariontes (Gevers *et al.,* 2004; Bratlie *et al.,* 2010).

Un aspecto interesante sería el conocer si en eucariontes la retención de genes duplicados se ve también favorecida por la adaptación a un ambiente diferente. Sin embargo, existe una limitante, ya que no se han hecho estudios a gran escala (es decir, que involucren decenas de genomas) debido principalmente al gran tamaño de sus genomas y a la mala caracterización de muchas de sus proteínas. Aun así, existen evidencias que apuntan fuertemente a que la estrategia de duplicación en respuesta al cambio de ambiente es un proceso que ocurre en los tres dominios celulares. Tal es el caso de los genes parálogos *adh1* y *adh2* de la levadura *S. cerevisiae*. El primero de ellos cataliza la formación de etanol a partir de acetaldehído, mientras que el segundo lleva a cabo la reacción inversa. Gracias a una serie de métodos paleogenéticos/paleobioquímicos, el grupo de Steven Benner logró recrear al gen ancestral *adhA*, el cual catalizaba la misma reacción que actualmente lleva a cabo ADH1. Por medio del método “silent nucleotide dating” determinaron que el evento de duplicación ocurrió en el periodo Cretácico, periodo durante el cual se piensa que *S. cerevisiae* no consumía ni acumulaba etanol (debido a que sólo convertía el acetaldehído en etanol). Además, este hecho coincide con la diversificación de las plantas con frutos carnosos (Thomson *et al*., 2005), lo cual permitió llegar a la conclusión de que la acumulación de etanol fue una estrategia para defender un recurso tan valioso para las levaduras como lo son los frutos carnosos y posteriormente utilizarlo como fuente de carbono, una vez que ya no hubiera competencia por el recurso (Piskur *et al*., 2006).

Finalmente, también se ha tratado de correlacionar el nivel de expresión génica con la tasa de duplicación. Para esto, Davis y Petrov decidieron considerar las tasas de expresión (TE) como una propiedad previa a la duplicación. De esta forma, pudieron hallar que los genes que presentan una mayor tasa de expresión, son a su vez los que dan

lugar a genes parálogos que se retienen en mayor proporción (Davis y Petrov, 2004).

1.5 El papel de las duplicaciones génicas en la evolución del metabolismo

Al parecer, el proceso de retención de genes duplicados no es homogéneo para todas las clases de genes, sino que depende en gran medida de la red metabólica, transcripcional o de señalización a la cual estén asociados (Roth *et al*., 2007), así como de las condiciones ambientales a las cuales se encuentren sometidos diferentes organismos (Conant y Wagner, 2002).

Para el caso concreto del metabolismo, existen varias hipótesis que tratan de explicar los orígenes de éste (Fani y Fondi, 2009). Dos de éstas propuestas las que consideran la duplicación génica como el proceso central en la evolución del metabolismo. La primera de ellas, llamada hipótesis de evolución retrógrada (Horowitz, 1945), plantea el surgimiento de rutas metabólicas enteras a partir de la necesidad de sintetizar algún producto A, el cual previamente habría sido agotado del medio. Para esto, sería necesaria la acción de una enzima que pudiera sintetizar A a partir de los sustratos B y C. Con el tiempo, alguno de estos dos iría disminuyendo su concentración en el medio y sería necesario sintetizarlos a partir de precursores D y E. Esto sería posible gracias a la duplicación de la enzima que catalizaba la reacción B + C = A, misma que sufriría mutaciones necesarias para llevar a cabo la reacción D + E = B (o C). De esta forma, el proceso podría continuar hasta que se ensamblara toda una ruta metabólica, y las enzimas más recientes serían aquellas que catalizaran la primera reacción de dicha ruta (Horowitz, 1945), aunque estudios recientes han demostrado que prácticamente ninguna ruta metabólica está compuesta en su totalidad por enzimas homólogas (Caetano-Anollés, G. *et al*., 2009) y son muy pocos los casos en los que reacciones sucesivas en una misma ruta son catalizadas por enzimas parálogas (Lazcano y Miller, 1999). Una alternativa a dicha hipótesis fue vislumbrada casi dos décadas después (Waley, 1969; Ycas, 1974) y conjuntada un par de años más tarde (Jensen, 1976). A esta hipótesis se le conoce como hipótesis del mosaico (Lazcano y Miller, 1999; Fani y Fondi, 2009), cuya premisa se basa en un conjunto pequeño de enzimas poco específicas, las cuales podían llevar a cabo diversas reacciones cada una. A través de duplicaciones génicas y mutaciones divergentes sería posible que cada una de las copias fuera aumentando su especificidad y la eficiencia de la reacción catalizada, lo cual a su vez sería el motivo para la conservación de dichos genes, lo cual provocó el incremento en el tamaño del genoma.

Además, el ensamble de nuevas rutas metabólicas no ocurre de manera secuencial (como en la hipótesis retrógrada), y las copias provenientes de una enzima ancestral no necesariamente van a participar en la misma ruta metabólica (Jensen, 1976) debido a que la divergencia funcional puede derivar en que cada una de las copias duplicadas lleve a cabo funciones muy diferentes (Hughes, 1994).

1.6 Modelos de duplicación génica

1.6.1 DDC (Duplication-Degeneration-Complementation)

Después de la publicación del libro *Evolution by Gene Duplication*, quedó establecido el modelo clásico de neofuncionalización, el cual a grandes rasgos establece que después de que ocurre la duplicación de un gen, ambas copias se conservan siempre y cuando una de ellas adquiera un número suficiente de mutaciones benéficas que le confieran una nueva función (Ohno, 1970; Force *et al*., 1999). No fue sino hasta casi treinta años más tarde cuando se propuso un modelo diferente, en el que se consideraba a las mutaciones degenerativas como un factor que podía facilitar la preservación de un par de genes parálogos, los cuales se caracterizaban por tener más de una función (Force *et al*., 1999). Una característica importante de dicho modelo es que las mutaciones ocurren de manera neutral y que el par de duplicados retiene las funciones originales (Maere y van de Peer, 2010) y aunque originalmente se propuso que las subfunciones presentes en el gen que se duplica eran de tipo regulatorio (asociadas a expresión génica) (Force *et al*., 1999) posteriormente se asoció cada vez más con funciones de tipo catalítico (Zhang, 2003).

De acuerdo con este modelo, existen dos fases desde que el gen se duplica hasta su fijación. En la primera, una de las copias adquiere una mutación en la secuencia codificante, de manera que se vuelve un alelo nulo y el producto pierde su funcionalidad. Otra posibilidad es que ocurra una mutación en una copia la cual le conferirá una nueva función. Finalmente, una opción adicional hace referencia a una serie de mutaciones degenerativas que provocan que ambas copias pierdan alguna de las subfunciones, pero que la acción sinérgica de ambas pueda restaurar la totalidad de las subfunciones originales (Lynch y Force, 2000). La segunda fase implica que los parálogos que hayan sido preservados, debido a que aún son funcionales, sigan acumulando mutaciones degenerativas hasta que no queden subfunciones redundantes en ambas copias (Force *et al*., 1999).

1.6.2 EAC (Escape from the Adaptive Conflict)

Este modelo, generalmente considerado como una forma de neofuncionalización, genera cierta controversia debido a que algunas de sus características hacen que se confunda con un modo de subfuncionalización. Su postulado central establece que, previamente a la duplicación, aparece una nueva función en un gen ancestral (debida a mutaciones), convirtiéndolo así en un gen bifuncional (Des Marais y Rausher, 2008; Conant y Wolfe, 2008). Sin embargo, la optimización de la nueva actividad se verá restringida debido a que posiblemente las mutaciones requeridas para esto afectan a la actividad original, generando un conflicto adaptativo entre ambas actividades (Hittinger y Carroll, 2007). La solución a este conflicto será entonces la duplicación del gen ancestral, seguida por la acción de la selección natural positiva sobre ambas copias, gracias a la cual, en una se optimizará la función ancestral, mientras que la otra mantendrá la función adquirida previamente a la duplicación (Hittinger y Carroll, 2007; Des Marais y Rausher, 2008).

De lo anterior, resulta la controversia respecto a si es un caso de neo o subfuncionalización, debido a que posteriormente a la duplicación, cada copia optimizará solo una función. La diferencia radica en el tipo de mutaciones: en el modelo clásico de subfuncionalización (DDC), las mutaciones involucradas son neutras o ligeramente deletéreas, mientras que en EAC involucra fundamentalmente sustituciones adaptativas (Des Marais y Rausher, 2008). Y, si bien se trata de un caso de neofuncionalización, existen ciertas diferencias con respecto al modelo original propuesto por Ohno (1970). De acuerdo con este último, el cambio adaptativo es posterior a la duplicación y la nueva función aparecerá gracias a la acción de la selección direccional, mientras que la selección purificadora actuará sobre la copia que mantenga la función ancestral. En el modelo EAC, la diferencia fundamental es que la nueva función es previa a la duplicación. Además, el cambio adaptativo ocurre sobre ambas copias; una optimizará la función ancestral (lo cual no ocurre bajo el modelo clásico) mientras que la otra optimizará la nueva función (Des Marais y Rausher, 2008).

1.6.3 IAD (Innovation, Amplification and Divergence)

Este modelo, propuesto por Pilar Francino hace una década, considera a las “explosiones puntuadas de amplificación génica” como el mecanismo principal para la evolución de

nuevos genes con nuevas funciones (Francino, 2005). El postulado principal de este modelo puede resumirse de la siguiente forma: ante la aparición de una nueva presión de selección (la cual generalmente está asociada a un cambio en el ambiente de determinado organismo) ocurre un proceso de duplicación génica masiva, generando un gran número de copias parálogas (amplificación génica). Evidentemente, no todas las copias prevalecerán, y la mayoría de ellas pasará por un periodo de no-funcionalización, a partir del cual se convertirán en pseudogenes o simplemente se perderán. Sin embargo, una o más copias habrán adquirido una nueva función con valor adaptativo, lo cual permitirá su fijación, además de que se requiere de un mínimo de especialización hacia una determinada función, lo cual no necesariamente implica duplicación génica previa debido a que la función requerida puede ser una actividad secundaria en el gen en cuestión (por ejemplo, que tenga cierto grado de promiscuidad catalítica) (Copley, 2003; Francino, 2005; Näsvall *et al*., 2012). Este modelo representa una alternativa para superar los obstáculos impuestos por el modelo clásico de neofuncionalización (Ohno, 1970) y se basa en tres postulados centrales: (1) explosiones puntuadas de amplificación génica, lo cual llevará a la evolución de nuevas funciones y a la subsecuente fijación de aquellos parálogos con valor adaptativo; (2) un periodo inicial de selección natural, positiva o purificadora, en todas las copias (incluidas aquellas que se volverán pseudogenes) y (3) formación y pérdida de muchos pseudogenes (Francino, 2005).

1.7 Clasificación funcional de proteínas

Una de las características de los estudios comparativos de proteínas es la necesidad de agrupar a éstas en un sistema de clasificación que permita identificar rasgos comunes en distintas proteínas. Una de las primeras clasificaciones fue la establecida por el Comité de Nomenclatura de la Unión Internacional de Bioquímica y Biología Molecular, la cual únicamente considera a las proteínas con actividad catalítica (enzimas) (Boyce y Tipton, 2001). En total se distinguen seis clases principales (Tabla 1), organizadas de acuerdo con un sistema numérico de cuatro dígitos en el que el primero de éstos indica el tipo de reacción catalizada, el segundo y el tercero representan a las subclases y sub- subclases respectivamente y el cuarto es el dígito exclusivo para cada enzima. Por ejemplo, en la enzima cuyo número de identificación es el 1.1.1.1, el primer dígito indica que pertenece a la clase de las óxidorreductasas. El segundo dígito indica que se encuentra dentro de la subclase de enzimas que actúan sobre grupos de donadores CH-OH, mientras que el

tercer dígito o dígito de la sub-subclase indica que utilizan NAD^+ o $NADP^+$ como aceptor. El último digito es el número de serie para cada enzima, y en este caso señala que se trata de la alcohol deshidrogenasa dependiente de NAD+. Este sistema es muy útil porque de esta forma se garantiza un número exclusivo para cada enzima diferente (Michal y Schomburg, 2012).

Tabla 1. Clasificación enzimática de acuerdo con el sistema establecido por el Comité de Nomenclatura de la Unión Internacional de Bioquímica y Biología Molecular (NC-IUBMB, 1992).

Número de la clase	Tipo de reacción catalizada	Nombre común de la clase	Significado del segundo dígito	Significado del tercer dígito
1.X.Y.Z	Oxidación / reducción	Óxidorreductasas	Donador de H^+ o e^- que se oxida	Aceptor
2.X.Y.Z	Transferencia de grupos	Transferasas	Grupo transferido	Información adicional del grupo transferido
3.X.Y.Z	Hidrólisis	Hidrolasas	Naturaleza del enlace hidrolizado	Naturaleza del sustrato
4.X.Y.Z	Rompimiento de moléculas de manera no hidrolítica	Liasas	Naturaleza del enlace roto	Información adicional del grupo eliminado
5.X.Y.Z	Isomerización	Isomerasas	Tipo de isomerización	Tipo de sustrato
6.X.Y.Z	Síntesis	Ligasas	Tipo de enlace formado	*Aplica solo en ligasas C - N

Sin embargo, este sistema de clasificación no considera al resto de proteínas que no llevan a cabo una reacción catalítica, como en el caso de las proteínas ribosomales, lo cual hace necesario generar clasificaciones diferentes que engloben a toda la diversidad proteínica. Una clasificación alternativa es aquella que agrupa a las proteínas de acuerdo al proceso celular en el que participan, lo cual puede derivar en clasificaciones tanto generales como específicas (a nivel de rutas metabólicas individuales). Por ello, lo ideal sería encontrar una clasificación intermedia de modo que pueda ser lo suficientemente informativa, como la clasificación que desarrolló Mónica Riley (1993) en la que clasificaba a todos los genes de *Escherichia coli* que habían sido anotados correctamente hasta ese momento en seis categorías generales que a su vez están divididas en subcategorías diferentes, pero sin llegar al nivel de rutas individuales (Riley, 1993). Por lo tanto, años más tarde, el grupo de Eugene Koonin desarrolló un sistema para identificar genes conservados en los tres dominios celulares y agruparlos en grupos de genes ortólogos (COGs, por sus siglas en inglés) (Tatusov *et al.*, 1997). Este sistema de clasificación

funcional consta de 15 categorías y ha sido ampliamente utilizado. Alrededor de la misma época, fue creada la base de datos Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa y Goto, 2000) la cual es un compendio de varias bases de datos entre las que se incluyen aquellas que agrupan a las rutas metabólicas en una serie de categorías más generales (KEGG PATHWAY) (Ogata *et al.*, 1999), similares en número y contenido a las de la base de datos COGs.

Existen otras bases de datos que poseen una clasificación funcional para las proteínas de un organismo. Entre éstas destaca la del proyecto Ontología Génica (Gene Ontology), misma que trata de unificar la descripción de los genes así como de sus productos en cualquier organismo con base en tres categorías: función molecular, proceso biológico en el que participa y componente celular en el que se localiza (Ashburner *et al.*, 2000). Otras clasificaciones como la de MetaCyc (Karp *et al.*, 2002) tienen la ventaja de basarse principalmente en datos experimentales, pero debido a esto, únicamente cuentan con información referente a unos cuantos organismos. Incluso la base de datos BRENDA, misma que inicialmente surgió como un recurso para proporcionar información meramente bioquímica sobre cada enzima descrita, cuenta actualmente con una clasificación funcional para las enzimas similar a aquella de la base de datos KEGG (Chang *et al.*, 2014).

2. Planteamiento del problema y objetivos

En este trabajo me propuse identificar si existen sesgos en cuanto a la proporción de proteínas parálogas dentro de las diferentes categorías funcionales. Aunque se han publicado varios trabajos cuyo objetivo principal ha sido similar (Conant y Wagner, 2002; Marland *et al.*, 2004; Gevers *et al.*, 2004; Roth *et al.*, 2007; Bratlie *et al.*, 2010), prácticamente cada autor ha utilizado un sistema de clasificación diferente para los productos génicos en cuestión. En particular, los trabajos de Gevers *et al* (2004) y de Bratlie *et al.* (2010) son los más parecidos a esta tesis en el sentido de que se utilizaron únicamente organismos procariontes, aunque el sistema de clasificación que ambos autores utilizan se basa en la base de datos COGs (Tatusov *et al*., 1997; 2003), que a pesar de ser un recurso muy utilizado para identificar la naturaleza de secuencias proteínicas, presenta ciertos errores e inconsistencias debido a que la anotación de los productos génicos se hace automáticamente (Cruz-González, 2015). Debido lo anterior, considero que el presente análisis ofrece ciertas ventajas tales como: 1) el uso de una base de datos curada manualmente (KEGG PATHWAY) y que además contiene información de las rutas metabólicas específicas para cada organismo (Ogata *et al*., 1999; Kanehisa y Goto, 2000) para identificar la categoría a la cual pertenecen las proteínas identificadas como parálogas: 2) la selección de genomas se hizo con base en un criterio biológico, debido a que Gevers *et al.* (2004) y Bratlie *et al.* (2010) emplean todo tipo de organismos procariontes de los algunos pueden sesgar los resultados, particularmente los parásitos y organismos con genomas reducidos; 3) la selección de una muestra más representativa de organismos, la cual a pesar de ser más pequeña que en los otros dos trabajos, incluye un número similar de organismos de los dominios Archaea y Bacteria.

Además, a la fecha no se ha publicado trabajo alguno en el que se aborde el problema de la retención de genes parálogos desde una perspectiva meramente enzimática, lo cual automáticamente excluye a todas aquellas proteínas sin actividad catalítica. Esta propuesta constituye una idea novedosa y original, y debe analizarse desde una perspectiva bioquímica.

El presente trabajo tiene dos objetivos generales los cuales son:

1) Identificar los genes parálogos de 20 genomas procariontes y comparar la proporción de éstos, tomando como base las categorías funcionales de KEGG

PATHWAY (Ogata *et al*., 1999; Kanehisa y Goto, 2000).

2) Analizar la proporción de genes parálogos con base en las clases enzimáticas (de acuerdo con la clasificación establecida por el NC-IUBMB, 1992).

Finalmente, este trabajo permitirá conocer si existen categorías funcionales con una tendencia a retener una gran cantidad de parálogos, e igualmente, si existen otras en las que no ocurra este fenómeno. Además, proporciona la primera explicación respecto a si alguna(s) clase(s) enzimática(s) posee(n) una proporción mayor de parálogos con respecto a las demás.

2. Metodología

Los proteomas de los organismos utilizados en el presente análisis fueron obtenidos a partir del sitio de archivos de texto plano del Centro Nacional de Información Biotecnológica (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/genomes/). Se seleccionaron los organismos de tal manera que se abarcara la mayor diversidad de phyla para cada dominio. En el caso de Bacteria, únicamente se seleccionó un individuo por phylum, mientras que para Archaea se seleccionó más de un organismo por phylum para que la muestra tuviera un número similar de organismos. Para ambas muestras, los organismos seleccionados debían cumplir con las siguientes características: a) que su genoma (y proteoma) esté completamente secuenciado; b) ser organismos con estilo de vida libre (no parásitos ni endosimbiontes); c) cuando menos un representante por cada phylum y d) que cada uno de los phyla incluidos tuviera por lo menos dos especies diferentes con genoma completamente secuenciado. Tomando en cuenta todas estas características, la lista final de los organismos seleccionados se presenta en la Tabla 2.

Tabla 2. Lista de organismos perteneciententes a los dominios Archaea y Bacteria, cuyos proteomas fueron utilizados para el análisis posterior. Junto a la columna con los nombres de cada organismo se indica el código de tres letras para cada uno, de acuerdo con la base de datos KEGG.

Bacteria		Archaea	
Aquifex aeolicus	**aae**	*Acidilobus saccharovorans* 345-15	**asc**
Bacillus subtilis subsp. *subtilis* 168	**bsu**	*Ferroglobus placidus* DSM 10642	**fpl**
Chlorobium chlorochromatii	**cch**	*Haloarcula marismortui* ATCC 43049	**hma**
Deferribacter desulfuricans SSM1	**ddf**	*Methanococcus maripaludis* C5	**mmp**
Escherichia coli str. K-12 substr. MG1655	**eco**	*Methanosarcina acetivorans* C2A	**mac**
Spirochaeta smaragdinae DSM 11293	**sma**	*Nitrosopumilus maritimus* SCM1	**nmr**
Streptomyces coelicolor A3 (2)	**sco**	*Pyrobaculum calidifontis* JCM 11548	**pcl**
Synechocystis sp. PCC 6803 substr. GT-I	**syn**	*Pyrococcus furiosus* DSM 3638	**pfu**
Thermodesulfovibrio yellowstonii DSM 11347	**tye**	*Sulfolobus solfataricus* 98/2	**sol**
Thermus thermophilus HB27	**tth**		
Salinibacter ruber DSM 13855	**sru**		

Análisis bioinformático. El propósito de este análisis fue obtener una lista de genes duplicados (parálogos) por cada organismo. Para ello se utilizó el algoritmo BLAST (Altschul *et al.,* 1990). Los parámetros utilizados fueron: (a) Un valor de corte (E-value) de $1e^{-07}$, con el objetivo de evitar un número grande de falsos positivos. Este valor de corte fue el mismo para todos los genomas analizados debido a que el tamaño de éstos no es muy diferente entre sí, sino que todos caen dentro del mismo orden de magnitud (Conant y Wagner, 2002). (b) Un porcentaje de identidad igual o mayor a 20. (c) Un valor de query

coverage (cobertura de la secuencia query) igual o superior a 75, el cual representa el porcentaje de la secuencia query que se encuentra alineado con la secuencia blanco.

Asignación de número enzimático y/o categoría funcional. Se utilizó la herramienta en línea *db2db*, (http://biodbnet.abcc.ncifcrf.gov/) (Mudunuri *et al.,* 2009). Dicha herramienta permite asignar automáticamente la información correspondiente al número enzimático y la categoría funcional, entre otras opciones, a partir de una lista inicial de identificadores (en este caso los correspondientes al GI Number). Es importante mencionar que únicamente asigna la información a aquellos "hits" que presenten la información correspondiente al número enzimático y a la(s) ruta(s) metabólica(s) en la(s) que participa(n), lo cual excluye a aquellas proteínas catalogadas como hipotéticas, con función desconocida, o cuya definición haya sido asignada a partir de los grupos de ortólogos (COGs). Para poder obtener esta información, obtuvimos los códigos equivalentes de KEGG Gene ID (identificadores utilizados en la base de datos KEGG) para cada lista de valores GI Number (una por organismo), tanto para las de genes duplicados como para aquellas con los identificadores del proteoma completo. Una vez hecho esto, se introdujeron los valores de KEGG Gene ID en el campo de búsqueda de *db2db* para que hiciera la asignación del número enzimático, la ruta metabólica en la cual participan y, en muchos casos, de ambos valores para cada gen (Fig. S1). En el caso del número enzimático, se utilizaron los valores establecidos por el Comité de Nomenclatura de la Unión Internacional de Bioquímica y Biología Molecular (Michal y Schomburg, 2012), mientras que para la categoría funcional, se utilizó inicialmente la información de KEGG PATHWAY (Ogata *et al*., 1999; Kanehisa y Goto, 2000), la cual corresponde al nombre de cada una de las rutas metabólicas identificadas para cada organismo.

Clasificación operativa. El propósito de este trabajo fue tratar de identificar si hay alguna tendencia de retención de genes parálogos con base en el número enzimático de las enzimas que codifican, así como corroborar lo propuesto por otros autores respecto a que existe una tendencia a retener genes parálogos que intervienen en la interacción de un organismo con su ambiente. Por lo tanto, para la clasificación con base en el número enzimático se consideraron las seis grandes clases de reacciones enzimáticas (oxidorreductasas, transferasas, hidrolasas, liasas, isomerasas y ligasas). La razón de esto es que cada clase de reacción engloba reacciones individuales con una química muy

parecida, y debido a que a la fecha no existe un estudio similar a éste, el hecho de considerar las seis categorías de manera general representa una buena aproximación para identificar si hay algún patrón de retención hacia cierta clase o clases. Para el caso de las categorías funcionales, los resultados obtenidos con la herramienta *db2db* fueron reagrupados en categorías más generales (aquellas presentadas como subcategorías en la base de datos KEGG PATHWAY (Tabla 3), con el fin de que hubiera el mismo número de categorías funcionales en cada organismo. Entre las categorías utilizadas están: 1) metabolismo, 2) procesamiento de información genética, 3) procesamiento de información ambiental y 4) procesos celulares, dentro de cada cual se incluye al conjunto de subcategorías utilizadas en este trabajo. Cabe destacar que existen más categorías generales en esta base de datos además de las cuatro anteriores, pero ninguna de éstas está identificada en organismos procariontes, y debido a esto, no fueron contempladas para el análisis.

Tabla 3. Subcategorías de KEGG PATHWAY con base en las cuales se clasificaron las enzimas consideradas en el análisis. Para el caso de las últimas dos categorías generales existe un número mayor de subcategorías, pero debido a que ninguna de éstas se encuentra identificada en los organismos de la muestra, no se consideraron para el análisis.

Categoría general	**Clave de subcategoría**	**Subcategoría**
Metabolismo		
	1.1	Metabolismo de carbohidratos
	1.2	Metabolismo energético
	1.3	Metabolismo de lípidos
	1.4	Metabolismo de nucleótidos
	1.5	Metabolismo de aminoácidos
	1.6	Metabolismo de otros aminoácidos
	1.7	Metabolismo y biosíntesis de glicanos
	1.8	Metabolismo de cofactores y vitaminas
	1.9	Metabolismo de terpenos y policétidos
	1.10	Biosíntesis de otros metabolitos secundarios
	1.11	Metabolismo y degradación de xenobióticos
Procesamiento de información genética		
	2.1	Transcripción
	2.2	Traducción
	2.3	Plegamiento, clasificación y degradación
	2.4	Replicación y reparación
Procesamiento de información ambiental		
	3.1	Transporte membranal
	3.2	Transducción de señales
Procesos celulares		
	4.1	Motilidad celular

Conteo por número enzimático y categoría funcional. En ambos casos, el conteo se hizo de manera manual bajo el siguiente criterio. Para el caso de los números enzimáticos, si la proteína en cuestión presentaba más de uno, pero todos coincidían en el primer dígito, se contó como una sola vez, asumiendo que la función general es en esencia la misma y que podría tratarse de una enzima poco específica. En el caso de que presentara más de un número enzimático pero que el primer dígito fuera diferente, se contó como dos o más hits. Posteriormente, se hizo un conteo con base en la categoría funcional. En este caso, debido a que hay enzimas identificadas en más de una ruta metabólica, el criterio de selección se hizo a nivel subcategoría, es decir, si participaban en más de una ruta pero ambas pertenecen a la misma subcategoría (por ejemplo, si participaba en las rutas de glucólisis/gluconeogénesis y metabolismo de piruvato, las cuales entran en la subcategoría de metabolismo de carbohidratos), se contó como una. En caso de que participara en rutas de subcategorías diferentes, se contó como un resultado para cada subcategoría.

Proporción de parálogos. Una vez que se realizó el conteo para ambas clasificaciones, se obtuvo la proporción de cada categoría funcional y de cada número enzimático con base en la sumatoria de las seis categorías enzimáticas y de las categorías funcionales consideradas (18 en total). Se hizo un análisis similar para los resultados pertenecientes a la parte del proteoma completo, y una vez obtenido esto, se calculó el valor proporcional con base en la siguiente fórmula:

$$proporción\ de\ duplicados = \frac{\#\ de\ duplicados\ en\ cada\ categoría}{\#\ total\ de\ genes\ para\ dicha\ categoría}$$

Fórmula 1. Ecuación general para calcular la proporción de genes duplicados (parálogos) dentro del proteoma de cada organismo

La razón de trabajar con un valor en porcentaje es que no podemos emplear el número de genes totales y duplicados totales para cierta categoría en cada organismo porque en cada uno el tamaño del genoma es diferente, así como el número de genes por cada categoría, además de que existe una correlación positiva entre el tamaño del genoma y el número de parálogos, por lo que organismos con genomas más grandes tendrán a su vez un número mayor de parálogos (Gevers *et al*., 2004; Bratlie *et al*., 2010). Por lo tanto, el valor proporcional obtenido mediante la fórmula 1 es una manera de normalizar los datos

(al convertirlos a valores porcentuales) y soluciona el problema mencionado anteriormente debido a que nos permite tener una idea general de cuáles son las categorías que presentan un mayor (cercano a 1) o menor (cercano a 0) número de duplicados, con lo cual se pueden hacer comparaciones intra e interdominio. Un valor cercano a 1 indica que la categoría en cuestión presenta un gran número de parálogos, mientras que un valor cercano a 0 indica lo contrario.

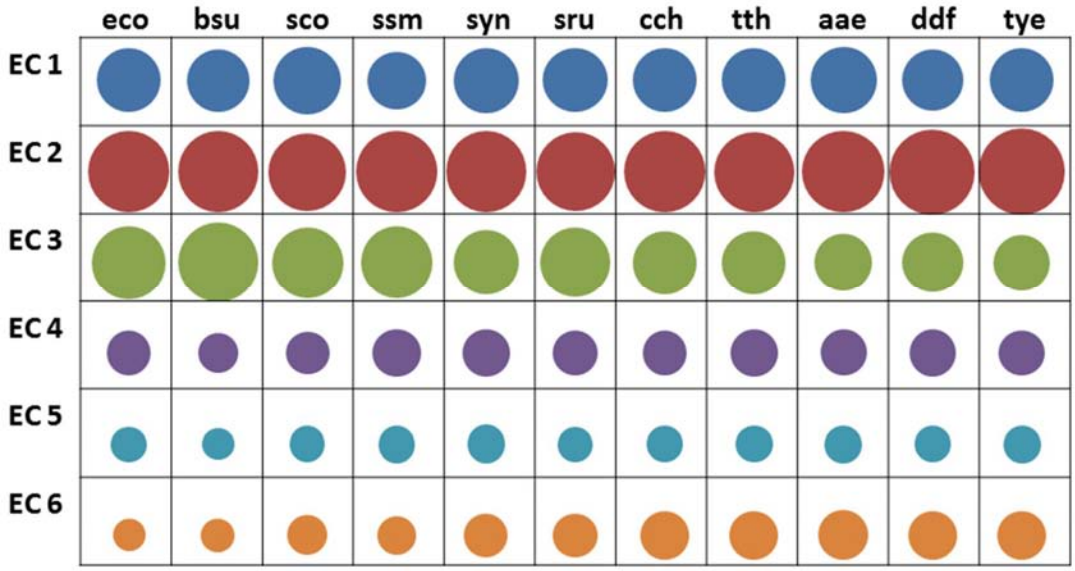
Análisis estadístico. Para poder identificar si existen diferencias significativas en cuanto a la proporción de parálogos de cada clase enzimática y de cada categoría funcional, se realizó un ANOVA de una vía para cada conjunto de datos (4 en total: 2 para la clasificación por número enzimático y 2 para la clasificación de KEGG). Posteriormente se realizó una prueba de F para identificar si las varianzas eran iguales y finalmente una prueba de t pareada para identificar si había diferencias significativas entre las medias. El valor de la significancia estadística para todas las pruebas fue de $P \leq 0.05$ y fueron hechas con Analysis ToolPak, incluido dentro de la paquetería Microsoft Excel 2010.

3. Resultados

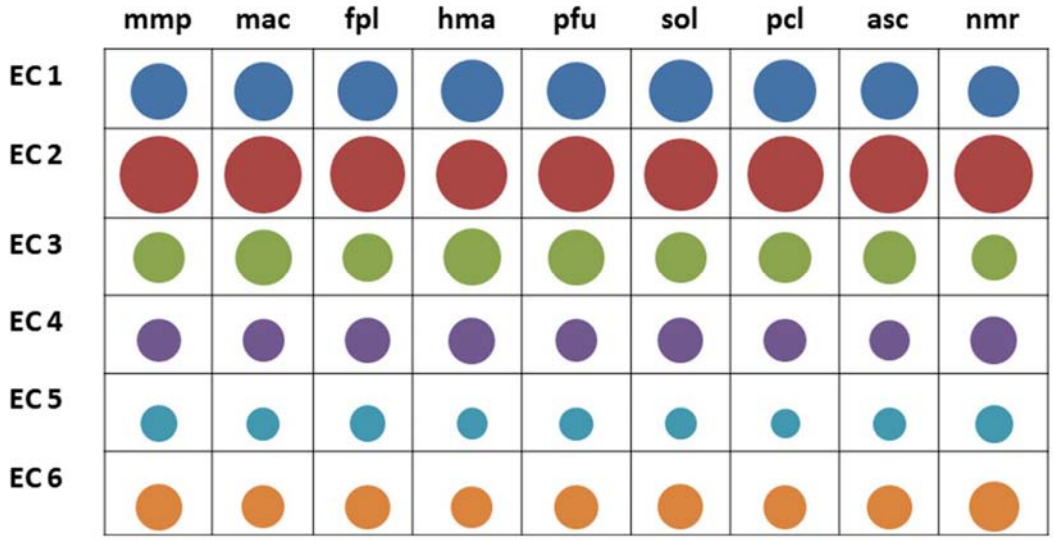
3.1 Proporción de cada clase de enzimas con base en su abundancia

En la Figura 4 se presentan los gráficos de burbuja obtenidos a partir de los datos de las tablas S1 y S2, mismos que representan la proporción de cada clase enzimática con respecto al total de enzimas en los proteomas de bacterias (4A) y arqueas (4B), mientras que en la figura 5 se muestra la proporción de parálogos por cada clase con respecto al total de enzimas parálogas (5A y 5B, respectivamente), obtenida a partir de los datos de las tablas S3 y S4. Claramente, se pueden distinguir dos grupos dentro de de ambas figuras: el primero que está conformado por las clases 1, 2 y 3, mismas que en conjunto representan en promedio el 70% del total de enzimas, y el grupo 2, conformado por las clases 4, 5 y 6 y que en conjunto no representan más del 30% del total de enzimas. Esto concuerda con la distribución general de enzimas (Tabla 11), que está sesgada hacia las tres primeras clases. Teniendo esto en cuenta, no resulta extraño que la clase EC 2 sea la que se encuentra en mayor proporción en todos los proteomas (Figura 4), debido a que es la clase con un mayor número de enzimas reportadas, aunque esta tendencia no se observa para la proporción de cada clase enzimática para cada conjunto de parálogos (Figuras 5A y 5B). Resulta evidente que la proporción de enzimas de la clase 1 aumenta considerablemente y presenta valores similares (y en ocasiones mayores) a los de la clase 2, mientras que el resto de las clases presenta valores similares a su contraparte de la Figura 4.

En la Figura 6 se muestran las gráficas de burbujas construidas a partir de los datos de las tablas S5 y S6, la cual representa los valores de la duplicabilidad (o proporción de retención) de cada clase enzimática, tanto para bacterias como para arqueas (Fig. 6A y 6B respectivamente). En todos los organismos el valor de proporción de parálogos para la clase 1 es siempre mayor al del resto de las clases, y es en la única clase en la que ocurre esto, lo cual nos dice que a pesar de que esta clase no se encuentra en una gran proporción en cada proteoma, sí es la que presenta una proporción mayor de enzimas duplicadas.

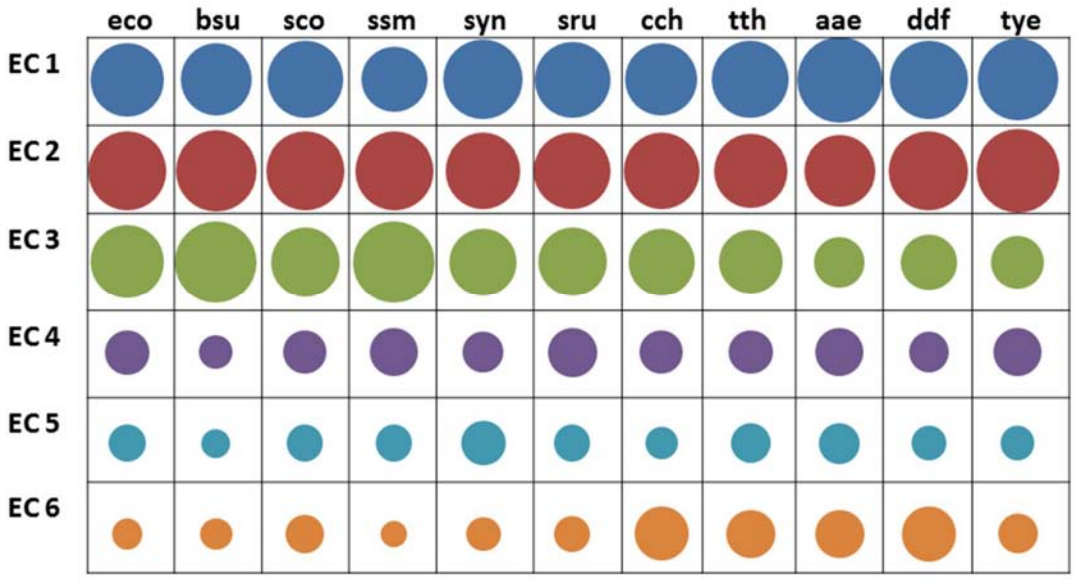


A)

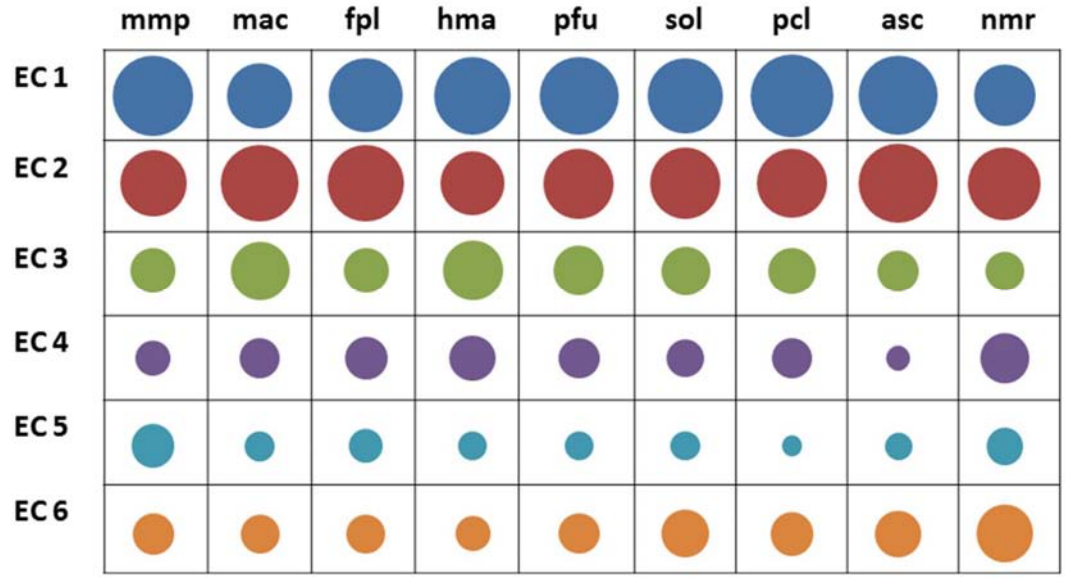


B)

Figura 4. Gráficos de burbujas en los que se representa la proporción porcentual de cada clase enzimática respecto al total descrito en el proteoma para cada organismo. Éstas fueron construidas con base en las tablas S1 y S2, y cada columna representa el 100% de enzimas en dicho organismo. La figura 4A corresponde al dominio Bacteria, mientras que la figura 4B al dominio Arquea.

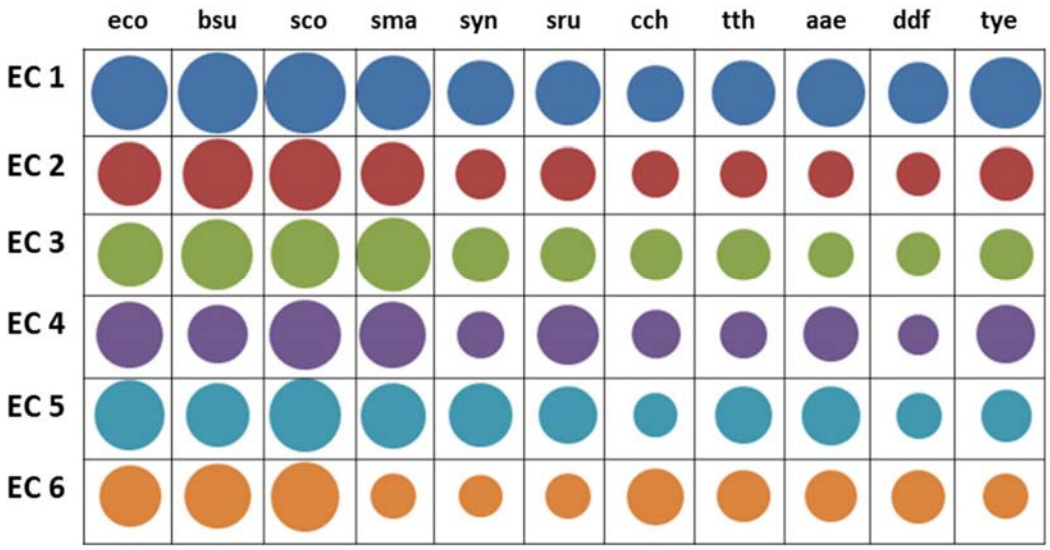


A)

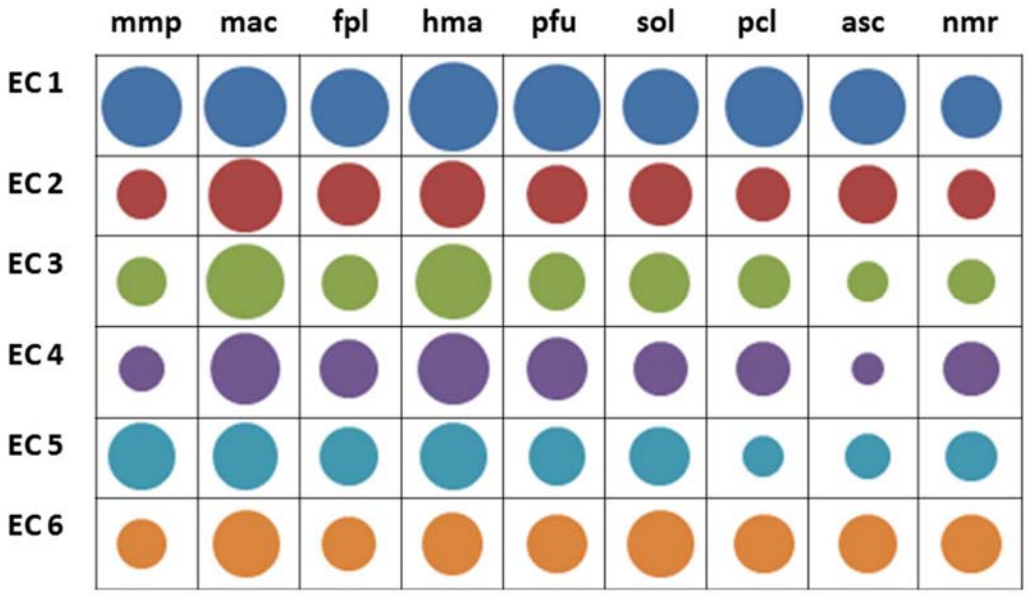


B)

Figura 5. Gráficos de burbujas en los que se representa la proporción porcentual de cada clase enzimática respecto al total descrito en el conjunto de parálogos para cada organismo. Éstas fueron construidas con base en las tablas S3 y S4, y cada columna representa el 100% de enzimas en dicho organismo. La figura 5A corresponde al dominio Bacteria, mientras que la figura 5B al dominio Arquea.



A)



B)

Figura 6. Gráficos de burbujas en los que se representa la proporción de parálogos para cada una de las clases enzimáticas. Éstas fueron construidas a partir de los datos de las tablas S5 y S6, y a diferencia de las figuras anteriores, cada columna no representa el 100% de enzimas totales por cada organismo, sino que cada celda representa la proporción (0-1) de duplicados para dicha clase enzimática (Fórmula 1). La figura 6A corresponde al dominio Bacteria, mientras que la figura 6B al dominio Arquea.

Para identificar si existen diferencias significativas entre la proporción de parálogos de cada clase enzimática, se realizó un análisis de varianza de una vía (con un intervalo de confianza de 95%) para cada conjunto de valores de proporción de parálogos, es decir, uno para cada dominio (Tablas S5 y S6). Para ambos casos se encontró que existen diferencias significativas entre las medias de la proporción de cada clase enzimática, aunque los valores de *P* son diferentes. En el caso de Archaea se obtuvo un valor de $P = 1.6 \times 10^{-6}$, mientras que para Bacteria el valor obtenido fue de $P = 2.25 \times 10^{-3}$. Posteriormente se realizó una prueba de *F* para cada una con el fin de identificar si las varianzas son iguales o desiguales. Los resultados de las pruebas de t pareada indican que sí existe una diferencia significativa entre la proporción de parálogos de la clase de las oxidorreductasas (EC 1) con respecto al resto de las clases, y esto ocurre tanto para Archaea como para Bacteria (Tablas 4 y 5). Con respecto a las demás clases enzimáticas, en ninguno de los casos se encontró que hubiera diferencias significativas entre ellas. Por lo tanto, con estos resultados es posible decir que existe una retención preferencial de parálogos a nivel de la estructura primaria únicamente para la clase de oxidorreductasas.

Tabla 4. Resultados obtenidos mediante la prueba de t pareada, con el fin de identificar diferencias significativas entre la proporción de parálogos de cada clase enzimática. Los resultados corresponden al dominio Archaea y sólo aquellos significativamente diferentes se muestran en color gris claro. Se utilizó un valor de *P* = 0.05 y un valor crítico de t = 2.12

Archaea	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
EC 1						
EC 2	5.39431051					
EC 3	4.51571368	0.16869613				
EC 4	5.22924914	0.52075999	0.27504377			
EC 5	5.8652265	0.52124933	0.2369429	-0.07317992		
EC 6	5.98058871	-0.10969092	-0.26374792	-0.66915742	-0.71531524	

Tabla 5. Resultados obtenidos mediante la prueba de t pareada, con el fin de identificar diferencias significativas entre la proporción de parálogos de cada clase enzimática. Los resultados corresponden al dominio Bacteria y sólo aquellos significativamente diferentes se muestran en color gris claro. Se utilizó un valor de *P* = 0.05 y un valor crítico de t = 2.09

Bacteria	EC 1	EC 2	EC 3	EC 4	EC 5	EC 6
EC 1						
EC 2	3.60054942					
EC 3	2.80860989	-0.57347935				
EC 4	3.56149805	-0.16543265	0.43558331			
EC 5	2.92945829	-0.69979466	-0.08041473	-0.55723914		
EC 6	4.34952364	0.51513969	1.08656226	0.71143684	1.26422857	

3.2 Proporción de enzimas con base en las categorías funcionales de KEGG

Una de las razones por las cuales se escogió el sistema de clasificación de rutas metabólicas de la base de datos KEGG fue para poder contrastar los resultados obtenidos con los trabajos de Gevers *et al*. (2004) y Bratlie *et al*. (2010), en los cuales los autores clasifican al conjunto de proteínas de acuerdo con la base de datos COG (Tatusov *et al*., 1997, 2003) la cual a pesar de ser un recurso muy utilizado, suele presentar ciertos errores en cuanto a la correcta anotación de las proteínas y de las rutas metabólicas en las que participan (Cruz-González, 2015). A pesar de que los nombres de las categorías COGs y KEGG son similares en muchos casos, existen otros en los que no coinciden, tal como la de metabolismo y degradación de xenobióticos (KEGG), la cual no tiene una contraparte con el mismo nombre en COGs aunque algunas de las proteínas de la categoría en KEGG se encuentran en la categoría de mecanismos de defensa (COGs). En la Tabla 6 se presenta una comparación resumida entre las principales características de los tres trabajos.

Tabla 6. Comparación de algunas características de nuestro trabajo con los de Gevers *et al*., (2004) y Bratlie *et al*., (2010). Una de las principales diferencias es el uso del parámetro "query coverage", el cual no es considerado en ninguno de los otros trabajos. Además, aunque nuestra muestra es más reducida, ésta contiene un número similar de organismos de los dominios Arquea y Bacteria, los cuales se analizan por separado.

	Gevers *et al*., 2004	**Bratlie *et al*., 2010**	**Álvarez, 2016**
Método de identificación de parálogos	blastp *all against all* intraproteoma	blastp *all against all* intraproteoma	blastp *all against all* intraproteoma
Parámetros	30% **ident.**; 150 ó + residuos alineados	75**% ident.** entre secuencias	**E-value:** 1e-07; > 20**% ident.**; **qcovs:** 75 ó +
¿Organismos parásitos?	Sí	Sí	No
Clasificación funcional con base en:	COGs	COGs + GO terms	KEGG PATHWAY
Diversidad procarionte	Bacteria	Bacteria + Archaea	Bacteria + Archaea
¿Consideraron plásmidos?	No	No	No

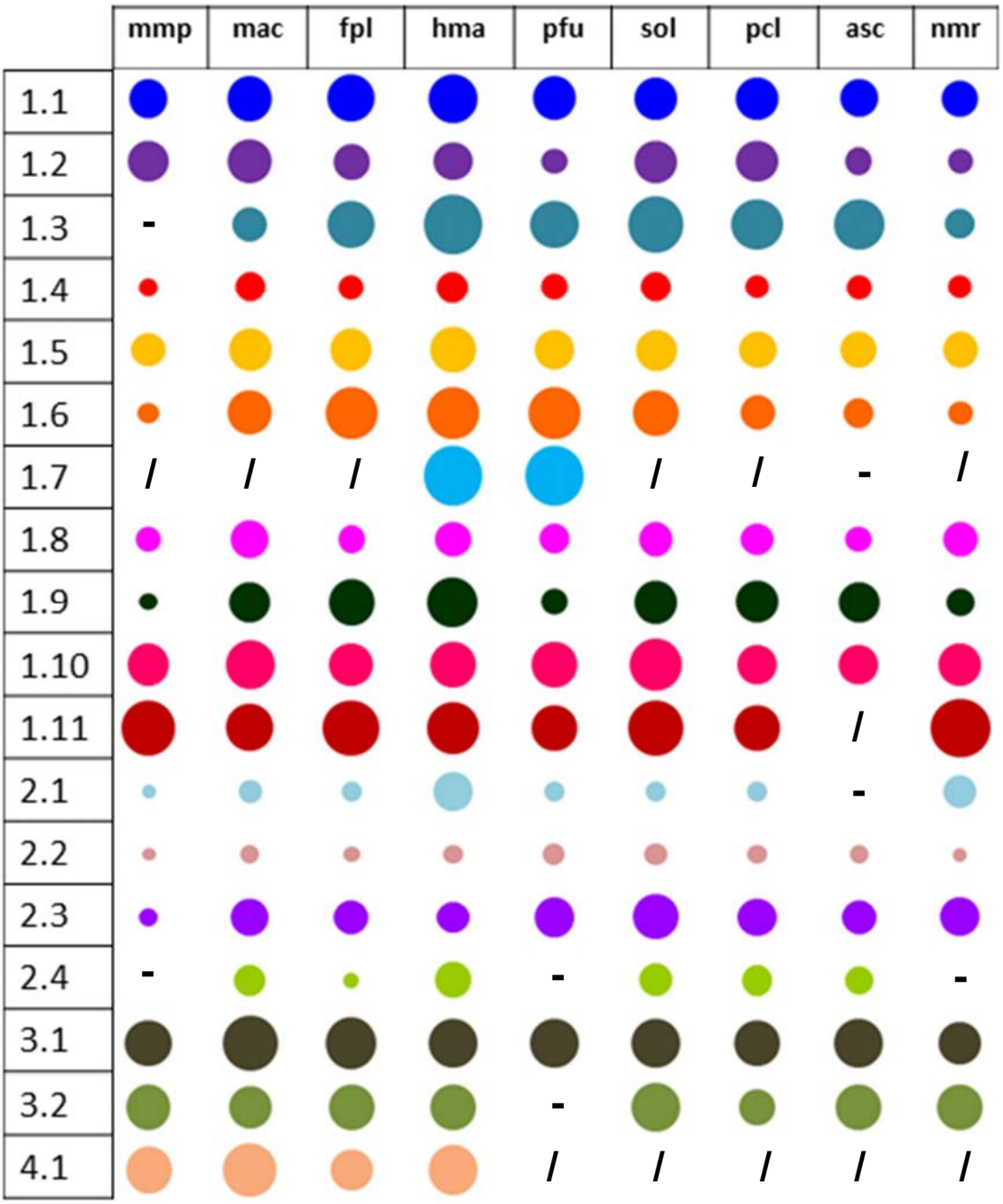


Figura 7. Gráfico de burbujas en el que se representa la proporción de parálogos para cada una de las categorías funcionales dentro del dominio Arquea, la cual se construyó a partir de la tabla S7. Un guión (**-**) indica que no se detectaron parálogos dentro de dicha categoría funcional, mientras que una diagonal (*/*) indica la ausencia de dicha categoría para el organismo en cuestión, de acuerdo con la base de datos KEGG PATHWAY. Los números al principio de cada fila indican la categoría funcional, de acuerdo con la Tabla 3.

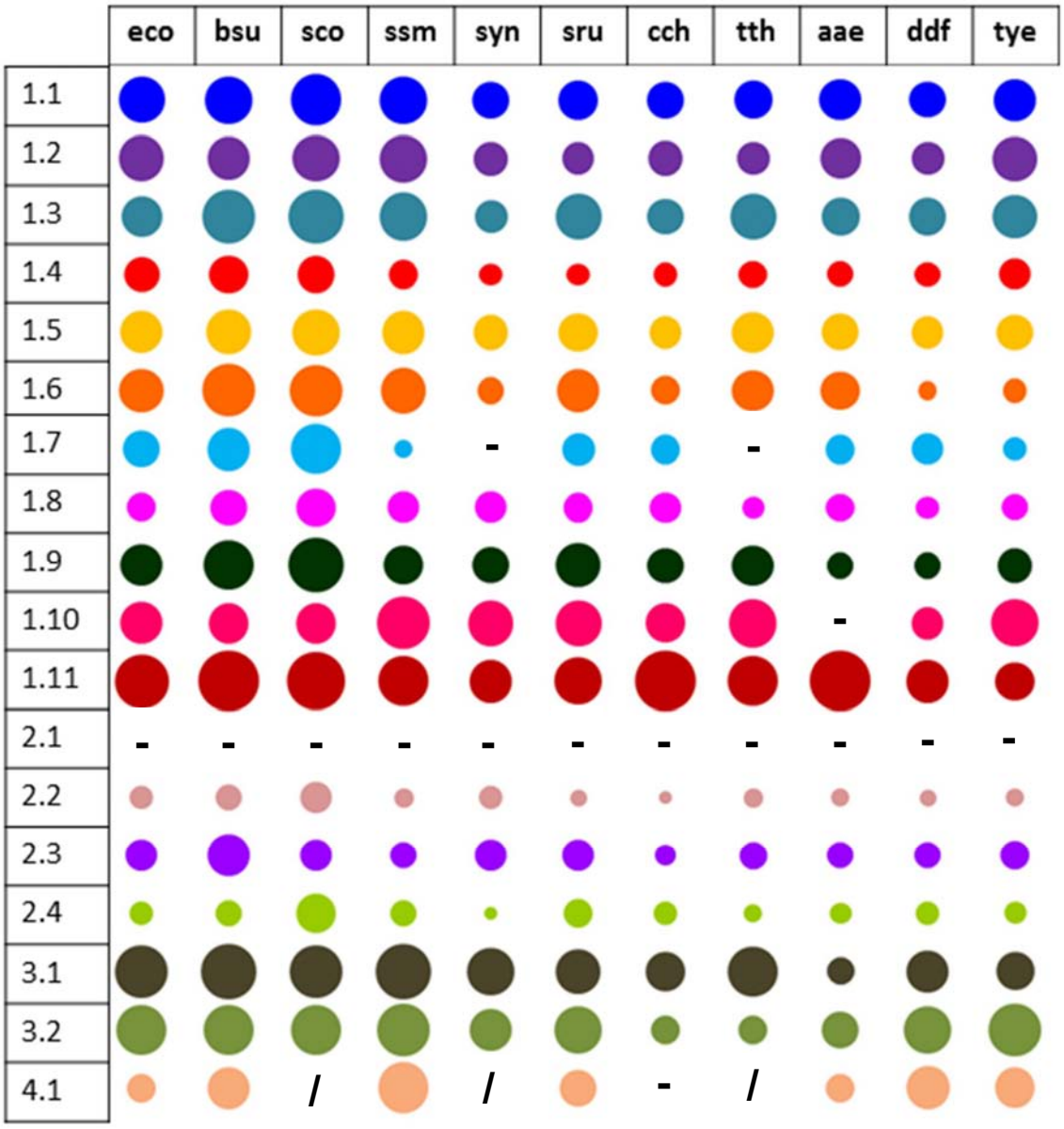


Figura 8. Gráfico de burbujas en el que se representa la proporción de parálogos para cada una de las categorías funcionales dentro del dominio Bacteria, la cual se construyó a partir de la tabla S8. Un guión (**-**) indica que no se detectaron parálogos dentro de dicha categoría funcional, mientras que una diagonal (*/*) indica la ausencia de dicha categoría para el organismo en cuestión, de acuerdo con la base de datos KEGG PATHWAY. Los números al principio de cada fila indican la categoría funcional, de acuerdo con la Tabla 3.

Los resultados de este análisis se presentan en las Figuras 7 y 8, así como en la Tabla 7. Se realizó también una serie de pruebas de t pareada con el fin de identificar diferencias significativas entre cada una de las categorías, mismas que se presentan en el material suplementario (Tablas S9 y S10). A pesar de que cada dominio presenta un patrón globalmente diferente, sí hay varios puntos coincidentes. En primer lugar, en ambos dominios la categoría con mayor proporción de parálogos es la de metabolismo y degradación de xenobióticos, seguida por la de transporte membranal y metabolismo de lípidos. El resto de las categorías metabólicas (verde) se distribuyen de manera diferente, con excepción de metabolismo de cofactores y vitaminas y metabolismo de nucleótidos, mismas que evidentemente presentan una proporción de parálogos muy baja. Asimismo, en ambos dominios las subcategorías de procesamiento de la información genética son las que presentan la menor proporción de parálogos. La subcategoría de transporte membranal posee uno de los valores más altos de proporción de parálogos en ambos dominios. Estos resultados muestran que la subcategoría de transducción de señales está entre aquellas con mayor proporción de parálogos en el dominio Bacteria. Finalmente, la subcategoría de motilidad celular presenta un valor alto sólo en Archaea; en el caso de Bacteria presenta un valor intermedio. Para el dominio Archaea, las categorías de biosíntesis y metabolismo de glucanos, así como la de motilidad celular, a pesar de tener una alta proporción de parálogos, fueron excluidas del análisis debido a que sólo están presentes en unos cuantos organismos y por lo tanto no pueden extrapolarse a todo el dominio.

Resulta interesante el conjunto de categorías conformado por las últimas seis en cada dominio (Tabla 7) debido a que son las mismas para ambos dominios. A partir de éstas, es posible decir que la categoría de procesamiento de la información genética es la que presenta una menor proporción de parálogos, dado que sus cuatro subcategorías se hallan en la parte inferior de la Tabla 7. Además, las categorías de metabolismo de nucleótidos y metabolismo de cofactores y vitaminas son, para ambos dominios, las subcategorías metabólicas con los valores más bajos de proporción de parálogos.

Tabla 7. Proporción porcentual de parálogos para cada una de las categorías funcionales en Arquea y Bacteria. Para cada dominio, los resultados se muestran en orden descendente y solo aquellos cinco con los valores más altos se muestran en negritas. Se resaltan en verde aquellas subcategorías que pertenecen al metabolismo; en violeta las de procesamiento de información genética; en magenta las de procesamiento de la información ambiental y en azul la única que se ubica dentro de procesos celulares. Los valores en rojo representan categorías presentes solo en algunos miembros (menos de la mitad) de Archaea, por lo cual no se consideraron para este apartado debido a que podrían sesgar los resultados. En vez de dichas categorías, se consideraron las dos siguientes con la proporción más alta.

Archaea		Bacteria	
0.77	Xenobiotics biodegradation and metabolism	**0.73**	Xenobiotics biodegradation and metabolism
0.67	Glycan biosynthesis and metabolism	**0.58**	Membrane transport
0.66	Membrane transport	**0.55**	Signal transduction
0.65	Cell motility	**0.52**	Lipid metabolism
0.59	Lipid metabolism	**0.48**	Carbohydrate metabolism
0.55	Biosynthesis of other secondary metabolites	0.47	Biosynthesis of other secondary metabolites
0.51	Carbohydrate metabolism	0.44	Metabolism of terpenoids and polyketides
0.50	Signal transduction	0.43	Energy metabolism
0.47	Metabolism of other amino acids	0.42	Cell motility
0.44	Amino acid metabolism	0.42	Metabolism of other amino acids
0.42	Metabolism of terpenoids and polyketides	0.41	Amino acid metabolism
0.38	Energy metabolism	0.26	Glycan biosynthesis and metabolism
0.38	Folding, sorting and degradation	0.24	Folding, sorting and degradation
0.28	Metabolism of cofactors and vitamins	0.24	Metabolism of cofactors and vitamins
0.20	Nucleotide metabolism	0.24	Nucleotide metabolism
0.17	Replication and repair	0.17	Replication and repair
0.16	Transcription	0.12	Translation
0.10	Translation	0.00	Transcription

En la Tabla 8 se presenta una comparación de los resultados con los de Gevers *et al.* (2004) y Bratlie *et al*. (2010) Para fines prácticos, únicamente se muestran las cinco categorías con mayor número de parálogos y las cinco con el menor número. Dentro del primer grupo, la categoría de transporte membranal es la única que coincide en los tres trabajos. De manera similar, se puede considerar como compartida al conjunto conformado por las categorías de Mecanismos de defensa (COGs) y Metabolismo y degradación de xenobióticos (KEGG) debido a que muchas de las enzimas que entran en esta última están involucrados en procesos de defensa (Copley, 2000; Wackett, 2004; Jansen *et al*., 2005; Russell *et al*., 2011). Además, dentro del segundo grupo existen tres categorías que coinciden en los tres trabajos: metabolismo de nucleótidos, metabolismo de cofactores y vitaminas y traducción. Debido al posible error de anotación dentro de la

categoría de transcripción para Bacteria, la categoría de traducción podría ser aquella con el menor número de parálogos en todos los casos.

Tabla 8. Comparación de los resultados de Gevers *et al*. (2004) y Bratlie *et al*. (2010) con los nuestros. Los resultados se muestran de acuerdo al número o proporción de parálogos por categoría en orden descendente, y en el caso de nuestro trabajo, éstos fueron divididos en Archaea y Bacteria. Aquellas categorías que están presentes en los tres trabajos se marcan con un color diferente. En el caso de aquellas marcadas en amarillo, no se trata de una categoría equivalente en nombre, varias de las enzimas asignadas a metabolismo y degradación de xenobióticos (KEGG) pertenecen también a mecanismos de defensa (COG).

Gevers et al., 2004	Bratlie et al., 2010	Álvarez, 2016	
Sólo Bacteria	**Mayormente Bacteria**	**Archaea**	**Bacteria**
Categorías con mayor número de parálogos			
Metabolismo de aminoácidos	Producción y conversión de energía	Metabolismo y degradación de xenobióticos*	Metabolismo y degradación de xenobióticos*
Transcripción	Movilidad celular	Transporte membranal	Transporte membranal
Metabolismo de iones inorgánicos	Metabolismo y transporte de iones inorgánicos	Metabolismo de lípidos	Transducción de señales
Metabolismo de carbohidratos	Transducción de señales	Biosíntesis de metabolitos secundarios	Metabolismo de lípidos
Mecanismos de defensa*	Mecanismos de defensa*	Metabolismo de carbohidratos	Metabolismo de carbohidratos
Categorías con menor número de parálogos			
Replicación, recombinación y reparación	Metabolismo de coenzimas	Metabolismo de cofactores y vitaminas	Metabolismo de cofactores y vitaminas
Biogénesis de membrana y pared celular	Control de ciclo celular	Metabolismo de nucleótidos	Metabolismo de nucleótidos
Transporte y metabolismo de nucleótidos	Modificaciones post-traduccionales	Replicación y reparación	Replicación y reparación
Transporte y metabolismo de coenzimas	Metabolismo y transporte de nucleótidos	Transcripción	Traducción
Traducción , estructura y biogénesis ribosomal	Traducción	Traducción	Transcripción **

3.3 Correlación positiva entre el número de parálogos y el tamaño del genoma

A partir de los datos del tamaño de proteoma y número de proteínas parálogas (Tabla 9) se construyó la gráfica de la Figura 9, con el propósito de confirmar la propuesta de Gevers *et al*., (2004) respecto a la relación entre el número de parálogos y el tamaño del genoma. Como se observa en la Figura 9, existe una correlación positiva entre el tamaño

del genoma y el número de parálogos, sin importar el dominio al que pertenezcan los organismos considerados (con un valor de R= 0.99).

Tabla 9. Muestra total de organismos utilizados dentro de este análisis. Éstos se encuentran acomodados de acuerdo a la cantidad de proteínas identificadas en el proteoma, en orden ascendente.

Organismo junto con su respectivo código KEGG de tres letras		**Dominio**	**No. de proteínas en el proteoma**	**No. de parálogos**	**Valor prop. de parálogos**
Aquifex aeolicus	**aae**	Bacteria	1497	378	0.253
Acidilobus saccharovorans 345-15	**asc**	Archaea	1499	332	0.221
Nitrosopumilus maritimus SCM1	**nmr**	Archaea	1796	521	0.290
Methanococcus maripaludis C5	**mmp**	Archaea	1813	527	0.291
Thermus thermophilus HB27	**tth**	Bacteria	1982	566	0.286
Chlorobium chlorochromatii	**cch**	Bacteria	1999	542	0.271
Thermodesulfovibrio yellowstonii DSM 11347	**tye**	Bacteria	2028	685	0.338
Deferribacter desulfuricans SSM1	**ddf**	Bacteria	2117	636	0.300
Pyrococcus furiosus DSM 3638	**pfu**	Archaea	2122	727	0.343
Pyrobaculum calidifontis JCM 11548	**pcl**	Archaea	2149	613	0.285
Ferroglobus placidus DSM 10642	**fpl**	Archaea	2480	817	0.329
Sulfolobus solfataricus 98/2	**sol**	Archaea	2679	1042	0.389
Salinibacter ruber DSM 13855	**sru**	Bacteria	2801	851	0.304
Synechocystis sp. PCC 6803 substr. GT-I	**syn**	Bacteria	3169	1096	0.346
Escherichia coli str. K-12 substr. MG1655	**eco**	Bacteria	4140	1766	0.427
Bacillus subtilis subsp. *subtilis* 168	**bsu**	Bacteria	4175	1757	0.421
Spirochaeta smaragdinae DSM 11293	**sma**	Bacteria	4219	2000	0.474
Haloarcula marismortui ATCC 43049	**hma**	Archaea	4243	1890	0.445
Methanosarcina acetivorans C2A	**mac**	Archaea	4540	2302	0.507
Streptomyces coelicolor A3 (2)	**sco**	Bacteria	7767	4023	0.518

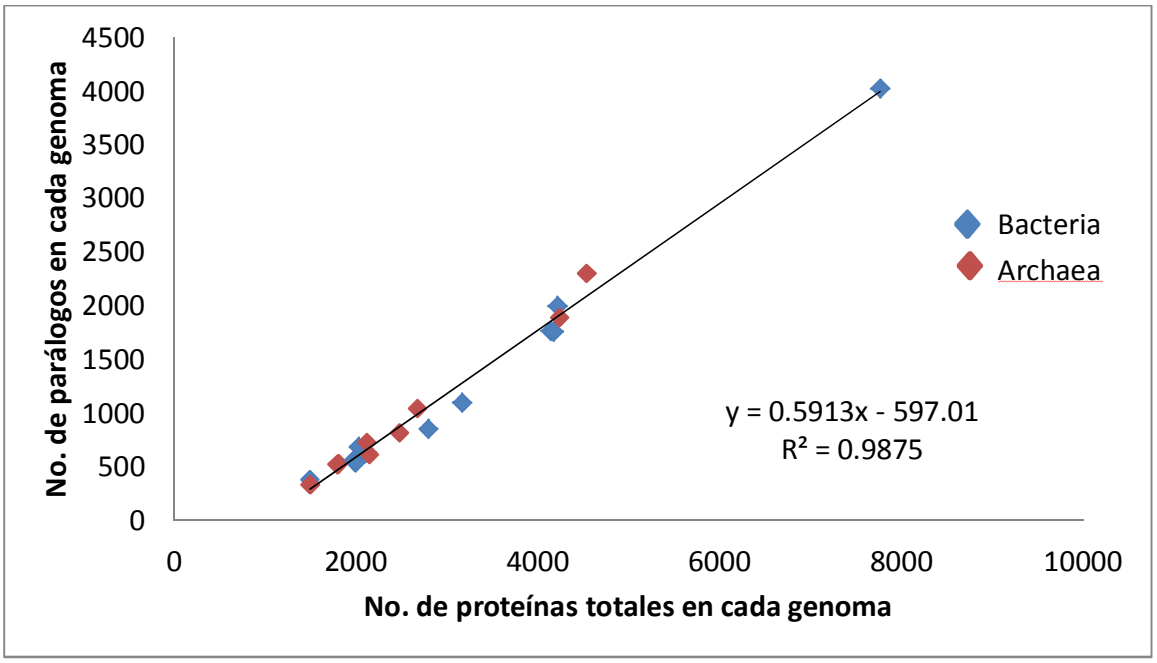


Figura 9. Gráfica en la cual se observa cómo el número de parálogos en cada genoma está correlacionado positivamente con el tamaño del genoma, con un coeficiente de correlación de R = 0.99.

4. DISCUSIÓN

4.1 Las categorías funcionales con mayor proporción de parálogos están asociadas con la interacción ambiental

A pesar de haber utilizado un número menor de organismos, una clasificación funcional distinta, y de haber incluido también al dominio Archaea, existen varias semejanzas con los trabajos reportados por Gevers *et al*. (2004) y Bratlie *et al*., (2010). El primer grupo reportó que las categorías que presentan un mayor número de genes duplicados son aquellas relacionadas con el metabolismo de aminoácidos, metabolismo de carbohidratos, transcripción, transporte y metabolismo de iones inorgánicos, así como la de mecanismos de defensa (Gevers *et al*., 2004). Éstas dos últimas categorías también son reportadas por Bratlie *et al.* (2010) como aquellas que tienen más parálogos, aunque ellos también reportan las de producción y conversión de energía, motilidad celular y transducción de señales. Es importante mencionar que dentro de las categorías de metabolismo de aminoácidos, metabolismo de iones inorgánicos y mecanismos de defensa, muchas de las proteínas parálogas son transportadoras (Gevers *et al*., 2004), las cuales forman familias homólogas (Saier y Paulsen, 1999; Braibant *et al*., 2000; Higgins, 2001; Davidson *et al.*, 2008).

En nuestros resultados (correspondientes al dominio Bacteria), las tres categorías con mayor proporción de parálogos son: metabolismo y degradación de xenobióticos, transporte a través de membrana y transducción de señales. Particularmente en este dominio, la transducción de señales está mediada en su mayoría por sistemas de dos componentes, los cuales constan de una histidina-cinasa sensora, ubicada en la membrana, y un regulador de respuesta afín, el cual se encuentra en el citoplasma (Sheng *et al*., 2012), y en la mayoría de los casos, estas categorías se asocian con procesos de respuesta ante el estrés generado por condiciones ambientales cambiantes (López-Maury *et al*., 2008). La diversificación de este sistema sensorial, se debe a procesos de duplicación génica seguidos por procesos de divergencia y al transporte horizontal. La duplicación de alguno de los dos componentes puede ocasionar que señales diferentes puedan desencadenar la misma respuesta (en caso de que se duplique la cinasa) o bien, que una sola cinasa pueda dirigir efectos de salida diferentes (en caso de que se haya duplicado el regulador de respuesta) (Capra y Laub, 2012). Se ha observado que grupos de bacterias que viven en ambientes constantes poseen

relativamente pocos genes del sistema de dos componentes (Capra y Laub, 2012), lo cual puede ser una razón por la que Gevers *et al.* (2004) no reportaron esta categoría como una de las que poseen mayor proporción de parálogos, debido a que incluyeron una gran cantidad de organismos parásitos. Además, el uso de este sistema parece estar limitado exclusivamente para el dominio Bacteria, y aunque en algunos representantes de Archaea se han encontrado proteínas del sistema de dos componentes, éstas no están distribuidas uniformemente a lo largo del dominio y el número de proteínas es menor que en las bacterias de vida libre (Koretke *et al*., 2000), razón por la cual en nuestro análisis esta categoría no figura entre aquellas con mayor proporción de parálogos.

El transporte a través de membrana es un proceso fundamental en la interacción de los organismos con su ambiente, y muy probablemente, el hecho de poder permear selectivamente ciertas moléculas fue crucial en las primeras células que habitaron el planeta (Davidson *et al*., 2008). Por ejemplo, para aquellos organismos cuya principal fuente de nutrientes es la ingesta de compuestos orgánicos exógenos, la cantidad de transportadores va a ser mayor que para organismos que utilizan solo sustancias inorgánicas como fuente de energía (Saier y Paulsen, 1999). Además, se ha observado que aquellos linajes que se han tenido que adaptar a condiciones muy diversas, presentan un mayor número de transportadores ABC, de los cuales existe prácticamente uno para cada tipo de molécula que tenga que atravesar la membrana celular (Higgins, 2001). Por lo tanto, la naturaleza de procesos como el transporte a través de membrana y la transducción de señales claramente confirma lo propuesto por otros autores respecto a que, por lo menos en procariontes, la duplicación génica es un recurso muy importante para la adaptación a nuevos ambientes (Gevers *et al*., 2004; Bratlie *et al*., 2010; Francino, 2012; Kondrashov, 2012), ya que las categorías mencionadas claramente juegan un papel fundamental en la interacción entre un organismo y su ambiente. De igual forma, la degradación y metabolismo de xenobióticos es un proceso relacionado con la adaptación a nuevas condiciones ambientales generadas por la introducción de sustancias sintetizadas por el hombre, aunque debido a diferencias fundamentales con las dos categorías mencionadas anteriormente, será tratado más a detalle en la siguiente sección.

4.2 El caso del metabolismo y degradación de xenobióticos

En todos los proteomas analizados, existen pocas enzimas asignadas a esta subcategoría, con excepción de *Streptomyces coelicolor*, aunque esto puede deberse al tamaño de su genoma, en comparación con otras subcategorías metabólicas (Tabla 10). Esta subcategoría es de particular interés debido a que muchos de estos compuestos (xenobióticos) se han introducido recientemente al ambiente, y por ende, algunas enzimas que participan en la degradación de estos compuestos probablemente evolucionaron recientemente. Y a pesar de que naturalmente existen compuestos de este tipo, muchos de estos han sido introducidos a partir de la llamada "era industrial", lo cual ha llevado a la modificación de los ambientes originales en los que habitaba un gran número de especies microbianas, particularmente el suelo y el agua. Por lo tanto, así como en los mamíferos interviene el sistema inmune en la protección del hospedero, de una manera análoga los procariontes han optado por una mecanismo similar: aumentar su repertorio de enzimas catabólicas para degradar muchos de los compuestos introducidos por el hombre (Wackett, 2004).

La introducción de compuestos xenobióticos al ambiente ha afectado prácticamente a todos los seres vivos, lo cual ha favorecido la evolución procesos ya sea para adquirir resistencia ante éstos o bien para poder metabolizarlos y utilizarlos como una fuente alternativa de nutrientes (i.e. carbono) (Russell *et al*., 2011). La adquisición de resistencia es una de las mecanismos más utilizados por organismos eucariontes, y básicamente ha consistido en la sobreexpresión o amplificación génica de familias proteínicas muy específicas, tales como esterasas, Cyt P450 y GST (Li *et al*., 2007; Omiecinski *et al*., 2010; Russell *et al.,* 2011; Croom, 2012). La metabolización de xenobióticos es característica de organismos procariontes, y ha sido posible gracias a la evolución de nuevas rutas metabólicas (Russell *et al*., 2011) y al transporte horizontal de genes (Janssen *et al*., 2005). De interés particular resulta el primer caso, debido a que han surgido nuevas enzimas y nuevas rutas en un lapso relativamente corto de tiempo (aprox. dos siglos), lo cual representa un buen modelo para estudiar la evolución enzimática y de nuevas rutas metabólicas por medio de duplicación génica y posterior divergencia funcional (Crawford *et al*., 2007). Éste no es un hecho trivial, sino que es posible gracias a ciertas características de este tipo de organismos, entre las que destacan: (1) tamaños poblacionales muy grandes en comparación con eucariontes, (2) tiempos generacionales muy cortos y (3) un repertorio bioquímico más amplio que en los

eucariontes (Russell *et al*., 2011). Además, este mecanismo ocurre preferentemente en procariontes de vida libre (Martínez-Núñez, *et al*., 2015). Aunque existen ciertas rutas y enzimas de esta categoría que han sido adquiridas mediante transporte horizontal (Janssen *et al*., 2005), muchas otras son producto de procesos de duplicación génica y subsecuente divergencia (Russell *et al*., 2011). El hecho de que muchas enzimas metabólicas puedan reconocer compuestos que no son su sustrato nativo (lo que se conoce como "metabolismo subterráneo") es quizá la mejor explicación para el surgimiento de nuevas rutas metabólicas y nuevas enzimas como una forma de interacción con un ambiente cambiante, en el cual un sustrato inicial alternativo para una enzima "A", puede volverse el sustrato primario para una enzima "A1" (surgida gracias a la duplicación de A) (D'Ari y Casadesús, 1998; Notebaart *et al*., 2014). Uno de los casos más estudiados al respecto es el de la ruta degradativa del pentaclorofenol (PCP), un xenobiótico sintetizado artificialmente y que fue introducido al ambiente en 1936 (Copley, 2000; Schmidt *et al*., 2003). Por lo tanto, la ruta para degradar este compuesto debió haber surgido en los últimos 80 años (Teichmann *et al*., 2001b). Ésta consta de tres enzimas, las cuales son homólogas a enzimas de distintas rutas del metabolismo central (específicamente el catabolismo de diclorofenoles y de los aminoácidos fenilalanina y tirosina): (1) pentaclorofenol (PCP) hidroxilasa (surgida a partir de una diclorofenol hidroxilasa), (2) tetraclorohidroquinona (TCHQ) deshalogenasa (a partir de la enzima maleilacetoacetato isomerasa) y (3) 2,6-diclorohidroquinona (DCHQ) dioxigenasa (enzima existente antes de la aparición de la introducción de PCP al ambiente) (Copley, 2000). Sin embargo, esta ruta presenta varias características y desperfectos que pueden ser considerados como evidencia de que ha evolucionado recientemente. En primer lugar, la PCP hidroxilasa tiene una pobre eficiencia catalítica, la cual se debe a su amplia especificidad de sustrato. Además, TCHQ dehalogenasa es inhibida por sustratos aromáticos. Finalmente, la expresión de PCP hidroxilasa y DCHQ dioxigenasa es inducida por PCP, lo cual no ocurre con TCHQ deshalogenasa, que es expresada constitutivamente en presencia de glutamato (Copley, 2000; 2009).

Tabla 10. Proporción de parálogos para la categoría de metabolismo y degradación de xenobióticos. Se indica tanto el número de rutas específicas que entran en esta categoría, identificadas para cada organismo de acuerdo con la base de datos KEGG, así como el número total de enzimas y el número de parálogos para cada uno.

Organismo	Phylum	# de rutas para degradación de xenobióticos	# total de enzimas dentro de estas rutas	# de enzimas en estas rutas con al menos un parálogo	Proporción de parálogos
Bacteria					
Escherichia	Proteobacteria	12	32	25	**0.78**
Bacillus	Firmicutes	5	19	19	**1.00**
Streptomyces	Actinobacteria	15	71	65	**0.92**
Spirochaeta	Spirochaetes	7	16	11	**0.69**
Synechocystis	Cyanobacteria	9	8	4	0.50
Salinibacter	Bacteroidetes	10	22	13	**0.59**
Chlorobium	Chlorobi	2	6	6	**1.00**
Thermus	Deinococcus-Thermus	9	19	13	**0.68**
Aquifex	Aquificae	6	15	14	**0.93**
Deferribacter	Deferribacteres	3	8	4	0.50
Thermodesulfovibrio	Nitrospirae	4	15	6	0.40
Archaea					
Methanococcus	Euryarchaeota	1	6	5	**0.83**
Methanosarcina	Euryarchaeota	8	33	21	**0.64**
Ferroglobus	Euryarchaeota	4	38	34	**0.89**
Haloarcula	Euryarchaeota	7	42	32	**0.76**
Pyrococcus	Euryarchaeota	2	7	4	**0.57**
Sulfolobus	Crenarchaeota	9	40	35	**0.88**
Pyrobaculum	Crenarchaeota	4	28	17	**0.61**
Acidilobus	Crenarchaeota	0	0	0	0.00
Nitrosopumilus	Thaumarchaeota	5	6	6	**1.00**

Aharoni *et al*, (2005) demostraron experimentalmente que una enzima especializada en una función, aunque con presencia de una o más actividades secundarias, puede sufrir mutaciones que la lleven a incrementar la eficiencia sin disminuir la actividad nativa. Con el tiempo, dicha enzima podría especializarse más (Fig. 10) (Aharoni *et al*., 2005; Khersonsky y Tawfik, 2010). A pesar de que ha habido procesos de duplicación y divergencia, algunos de estos parálogos retienen la capacidad de llevar a cabo más de una sola reacción y de aceptar sustratos diferentes, por lo que se conocen como intermediarios generalistas (Aharoni *et al*., 2005).

Aunque no se identificó que, individualmente, cada una de las enzimas de esta subcategoría metabolice más de un sustrato, existen varias evidencias que apoyan este punto: (1) son muchos los compuestos xenobióticos que han sido vertidos al ambiente y relativamente pocas las enzimas encargadas de metabolizarlos (Tabla 10) (Agrawal y Shahi, 2015), (2) algunos de estos compuestos tienen estructuras químicas similares a la de ciertos metabolitos que ocurren naturalmente y (3) el hecho de que esta categoría sea

la que más parálogos tiene implica que la mayoría de sus enzimas han surgido por duplicación y divergencia de enzimas pertenecientes a otras rutas del metabolismo central (específicamente, a partir de la ruta degradativa de diclorofenoles y del metabolismo de fenilalanina y tirosina) las cuales probablemente poseían una actividad secundaria para metabolizar muchos de estos compuestos.

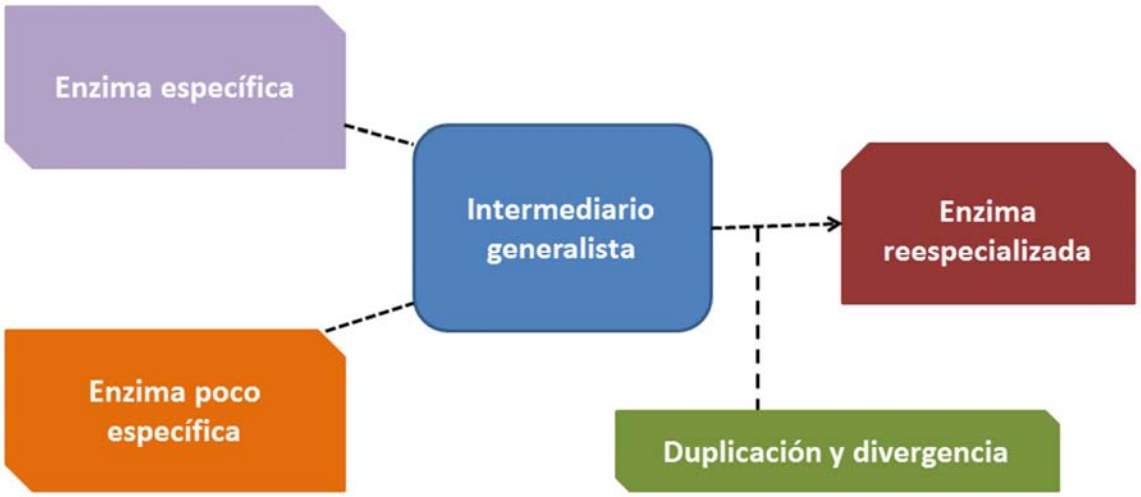


Figura 10. Esquema general que explica la aparición de nuevas enzimas y nuevas funciones. Una enzima totalmente específica hacia una función, o con actividad(es) secundaria(s) leve puede sufrir mutaciones que le confieran la capacidad de llevar a cabo una nueva función (para el caso de que la enzima fuera específica) o de incrementar la o las actividades secundarias que posean pero sin disminuir considerablemente (o no disminuir del todo) la actividad nativa. Esto dará lugar a un intermediario generalista, el cual llevará a cabo más de una actividad. Finalmente, si una de estas nuevas actividades resulta benéfica para la adecuación del organismo, los procesos de duplicación y divergencia pueden generar una enzima reespecializada, la cual será óptima para una de las actividades secundarias, mientras que la copia original retendrá la función original.

4.3 La proporción de parálogos para la clase de las óxidorreductasas difiere significativamente del resto de las clases enzimáticas

La clase de las óxidorreductasas (EC 1) agrupa a todas aquellas enzimas que intervienen en la catálisis de reacciones de oxidación-reducción, en las cuales siempre hay un grupo que se oxida y otro que se reduce. Muchas de las enzimas de esta clase reciben también el nombre de deshidrogenasas debido a que la oxidación de su respectivo sustrato ocurre mediante la sustracción de protones (iones H^+), mientras que en aquellas en las que el O_2

es el aceptor se utiliza el término oxigenasa (Boyce y Tipton, 2001). Como se observa en la Figura 6, esta es la única clase enzimática que presenta una retención preferencial de sus parálogos.

Después de que ocurre la duplicación de un gen, la prevalencia de las dos copias depende de que, o bien una de ellas adquiera una nueva función, o se lleve a cabo una reducción de funciones (en caso de que el producto génico duplicado tuviera más de una), de manera que una de las copias mantuviera una de las funciones y su parálogo la otra. Debido a que este último escenario es el más probable (Force *et al.*, 1999; Lynch y Force, 2000) puedo pensar que la mayoría de los parálogos que pertenecen a las óxidorreductasas ha logrado permanecer debido a: (1) la enzima ancestral poseía baja especificidad y por lo tanto podía aceptar más de un sustrato (promiscuidad catalítica); (2) alguna característica inherente a la enzima ayudó a que pudiera llevar a cabo una nueva función. Para este último caso, lo más probable es que la nueva actividad esté relacionada con la actividad original y no que sea una totalmente diferente (Zhang, 2003), lo cual no se limita solamente a las enzimas de esta clase.

De acuerdo con O'Brien y Herschlag (1999), las actividades secundarias o reacciones alternativas pueden ser facilitadas gracias a grupos catalíticos dentro del sitio activo de una enzima, tales como (1) iones metálicos, (2) cofactores, (3) nucleófilos, (4) aceptores y donadores de protones, así como (5) ácidos y bases generales. Las propiedades químicas de las óxidorreductasas muestran que éstas constituyen la clase enzimática que emplea una mayor diversidad de aceptores y donadores de protones. Además, de entre todas las clases enzimáticas, ésta presenta una mayor proporción de enzimas que requieren de un cofactor orgánico (80.4%), superando por más del doble a la siguiente clase enzimática que más requiere de cofactores (EC2: 36.5%) y por más del triple al resto de las clases (EC 3: 4%; EC 4: 27.2%; EC 5: 23.4% y EC 6: 29.6%). Asimismo, es la clase enzimática que utiliza una mayor diversidad de cofactores orgánicos (19 en total) (Fischer *et al*., 2010) (Tabla 11), en particular el NAD^+(P) y el FAD, los cuales son dos de los metabolitos más utilizados en todo el metabolismo (Alves *et al.*, 2002). También, esta clase enzimática también es la que emplea una mayor diversidad de iones metálicos para la catálisis óxido-reducción (Andreini *et* al., 2008; Ji *et al.*, 2008; Kim *et al*., 2013). Esta característica es muy importante, ya que gracias a esto las óxidorreductasas son la clase enzimática que tiene una mayor intervención en procesos del metabolismo energético (Ji *et al.*, 2008) y ciclos biogeoquímicos (Falkowski *et al*., 2008; Kim *et al*., 2013); procesos que se basan en reacciones de óxidorreducción. Por lo

tanto, la gran diversidad de estos grupos catalíticos presentes a lo largo de esta clase enzimática pudo haber sido el factor determinante para la adquisición de una nueva función después del proceso de duplicación, o bien, para la optimización de alguna de las subfunciones presentes previamente a la duplicación, ya que como señalan O'Brien y Herschlag (1999): "No hay razón para esperar que en las reacciones alternativas se utilicen estos grupos del sitio activo de la misma manera en que se utilizan en la reacción nativa."

Tabla 11. Comparación entre el total de números EC diferentes con respecto al número de enzimas que requieren algún cofactor orgánico para llevar a cabo la reacción que catalizan. El número total de enzimas fue obtenido a partir de la base de datos ENZYME, del Instituto Suizo de Bioinformática (Bairoch, 2000), mientras que la lista de enzimas que utilizan cofactores orgánicos fue obtenida de la base de datos CoFactor Database, del Instituto Europeo de Bioinformática (Fischer *et al*., 2010). Los resultados se comparan con los porcentajes obtenidos por Fischer *et al.* en 2010.

EC Number	No. total de enzimas	No. de enzimas que utilizan algún cofactor orgánico	Porcentaje (%)	Porcentaje (Fischer *et al*., 2010) (%)	# cofactores distintos
EC 1	1567	1204	76.8	80.4	19
EC 2	1654	460	27.8	36.5	16
EC 3	1303	34	2.6	4.0	11
EC 4	582	109	18.7	27.2	15
EC 5	253	34	13.4	23.4	9
EC 6	187	37	19.8	29.6	3

El hecho de que una enzima sea poco específica no implica necesariamente un efecto negativo para la adecuación del organismo, sobre todo si dicha actividad raramente ocurre o si no implica gasto energético alguno (Copley, 2003). Lo que sí se requiere es que la promiscuidad catalítica se encuentre por encima de un umbral de ventaja selectiva para que pueda fijarse en el genoma, o lo que es lo mismo, un gen con una escasa actividad para una reacción secundaria, requeriría un gran número de mutaciones para optimizar dicha actividad y por lo tanto tendría una menor probabilidad de ser fijada (O'Brien y Herschlag, 1999). Si suponemos que una deshidrogenasa ancestral catalizaba dos reacciones similares pero utilizando el mismo grupo catalítico, esto implica que ambas funciones no pueden llevarse a cabo de la manera más óptima debido a que ocupan el mismo sitio activo. Si ambas funciones son importantes para una célula y se encuentran por encima del umbral de ventaja selectiva, una vez que ocurra el proceso de duplicación

lo más probable sería que ocurrieran mutaciones degenerativas en ambas copias para que así cada una solo lleve a cabo una de las dos funciones (Force *et al.*, 1999; Hittinger y Carroll, 2007; Des Marais y Rausher, 2008).

Como se mencionó anteriormente, los grupos catalíticos presentes en el sitio activo de una enzima pueden ser la clave para la evolución de nuevas funciones. Particularmente destaca el caso de los cofactores orgánicos, los cuales son metabolitos muy importantes para el correcto funcionamiento de las enzimas, y de los cuales se piensa que han sido un factor muy importante en la evolución de nuevas enzimas (Alves *et al.*, 2002) y de nuevas rutas metabólicas (Schmidt *et al.*, 2003) (Fig. 11).

La mayoría de las oxidorreductasas son capaces de utilizar cofactores orgánicos como el NAD^+ y el FAD gracias a la presencia de un dominio de unión a nucleótidos (conocido como plegamiento Rossmann), el cual consta de un conjunto de hojas β paralelas a las cuales se unen dos o más α-hélices, mismas que están unidas a un loop (conocido como "P loop") que interacciona con los fosfatos de los nucleótidos (Kuriyan *et al*., 2013). Se piensa que fue a partir de un dominio ancestral de unión a la porción piridínica de ciertos nucleótidos que evolucionó el dominio de unión a NAD (Buehner *et al.*, 1973), el cual posteriormente se fusionó con una diversidad de otros genes (Eventoff *et al.*, 1975), dando lugar así a familias de proteínas que tienen en común este dominio de unión a nucleótido (Teichmann *et al.*, 2001a; 2001b).

De acuerdo con la base de datos CoFactor (Fischer *et al.*, 2010), tanto el NAD^+ como el FAD están involucrados en más del 80% de las reacciones llevadas a cabo por deshidrogenasas. Además, muchas de las enzimas que requieren de estos dos cofactores para su funcionamiento son las que a su vez poseen un número mayor de parálogos (datos no mostrados) en comparación con aquellas que no requieren o que utilizan alguno distinto (como el heme). Debido a los valores de corte utilizados en el análisis del BLAST, particularmente el porcentaje de cobertura (query coverage), así como al tamaño del dominio de unión a nucleótidos, creemos que estos hits constituyen verdaderos parálogos y no únicamente proteínas cuya única región homóloga sea el dominio Rossmann. Además, el hecho de que muchas proteínas con este dominio posean a su vez algún dominio adicional (Teichmann *et al.*, 2001a; 2001b), implica que muchos de los homólogos encontrados dentro de esta clase enzimática comparten otros dominios además del dominio de unión a dinucleótido.

Tomando en cuenta lo descrito anteriormente, propongo que en etapas tempranas de la evolución biológica, enzimas con baja especificidad y un dominio de unión a

nucleótidos podían haber catalizado un gran número de reacciones en las que nucleótidos como el NAD y el FAD eran los principales donadores y/o aceptores de protones. La versatilidad de estos cofactores, así como la naturaleza de las reacciones en las que intervienen, fueron quizá el sustrato que propició la evolución de enzimas a través de duplicación génica, y una vez ocurrido esto, los parálogos pudieron ser retenidos gracias a que: 1) pudieron llevar a cabo reacciones diferentes (aunque en esencia basadas en la misma química), o 2) llevaron a cabo más eficientemente una de las subfunciones ancestrales. A esto se le puede añadir el hecho de que la mayoría de los procariontes participa en procesos metabólicos cuya química se basa principalmente en reacciones de óxido/reducción, por lo cual, muy probablemente no encontraríamos esta retención preferencial de oxidorreductasas dentro de los genomas eucariontes.

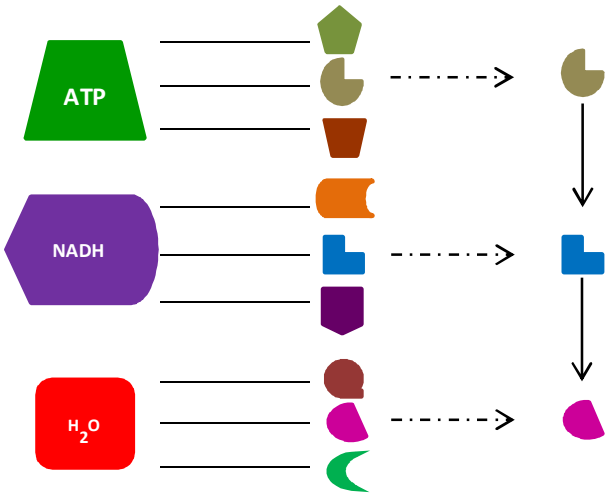


Figura 11. Modelo de evolución en mosaico del metabolismo, con énfasis en los metabolitos involucrados. Éste expresa la evolución de nuevas rutas metabólicas a partir del reclutamiento de enzimas preexistentes que participan en rutas diferentes. Sin embargo, este modelo es ligeramente diferente debido a que considera a los metabolitos muy utilizados como el agente director del reclutamiento de enzimas. En la parte izquierda se encuentran tres de los metabolitos que son más ampliamente utilizados (H_2O, ATP y NADH). En la parte del centro está representada una serie de enzimas, las cuales están unidas con una línea al metabolito que utilizan en la reacción que llevan a cabo, y que no necesariamente participan en la misma ruta metabólica. Finalmente, en la parte derecha se representa una ruta metabólica ensamblada mediante el modelo de mosaico (Waley, 1969; Ycas, 1974; Jensen, 1976), la cual está compuesta por enzimas que utilizan alguno de los metabolitos más ampliamente usados.

4.4 La estructura química y función de los metabolitos en los organismos influye en la retención de enzimas parálogas

Los resultados de este trabajo concuerdan con aquellos obtenidos por Gevers *et al*. (2004) y Bratlie *et al*. (2010) en cuanto a las categorías con una menor proporción de parálogos, tales como el metabolismo de nucleótidos, metabolismo de cofactores y vitaminas y el proceso de traducción. Sin embargo, tanto Gevers *et al*. (2004) como Bratlie *et al* (2010). únicamente proponen una explicación para las categorías con más parálogos, la cual tiene que ver con la adaptación a nuevos ambientes, y no dan una explicación convincente para aquellas categorías con una baja proporción (básicamente, argumentan que su proporción de genes parálogos es baja debido a que son procesos que no están implicados en la interacción organismo-ambiente). Debido a esto, en este trabajo trato de proporcionar algunas otras razones para este hecho, las cuales están asociadas con las características bioquímicas de estas subcategorías.

A pesar de que el proceso de duplicación génica puede ocurrir de manera aleatoria, sí existen restricciones para la retención de los genes duplicados (Papp *et al*., 2003; Davis y Petrov, 2004) por lo cual no todos tendrán la misma probabilidad de ser retenidos (Pap *et al*., 2003; Marland *et al*., 2004; He y Zhang, 2005). Si el gen duplicado no mejora en forma alguna la eficiencia de la función original, la no funcionalización puede ser el escenario más probable. Pero, ¿qué ocurre cuando se duplica más de un gen en el mismo evento? Suponiendo que todos los genes duplicados codifican para enzimas, la presión de selección podría ser la capacidad para metabolizar un sustrato nuevo. Si el hecho de metabolizarlo representa una ventaja para el organismo, los candidatos más probables para llevar a cabo esto serían las enzimas recién duplicadas, mediante la acumulación de mutaciones (lo cual representaría un escenario de neofuncionalización) (Fig. 12a-c), prevaleciendo aquellas que requieran un menor número de mutaciones, mientras que las otras serían eliminadas (Khersonsky y Tawfik, 2010). Incluso, se ha demostrado que en ocasiones una sola sustitución es más que suficiente para que la especificidad de una enzima por un sustrato cambie hacia uno nuevo (Umeno *et* al., 2005; Varadarajan *et al*., 2005; Watts *et al*., 2006), como ocurre con las enzimas MDH y LDH, las cuales son homólogas y un simple cambio de R102 en MDH por Q102 en LDH es suficiente para que la especificidad cambie de malato a lactato (Goward y Nicholls, 1994).

Sin embargo, este no es el único escenario posible para la preservación de una enzima duplicada. Supóngase ahora que al aparecer un nuevo sustrato, alguna enzima

posee una ligera afinidad hacia éste. Si el hecho de metabolizarlo representa una ventaja para el organismo, es evidente que este hecho no podría ocurrir de la manera más eficiente porque la enzima en cuestión posee una mayor especialización hacia su función original. Pero si ocurriera una duplicación para el gen que codifica a la enzima en cuestión, este hecho podría ocasionar que cada copia se especializara hacia una función diferente y por ende, la eficiencia de ambas podría alcanzarse bajo un escenario de subfuncionalización (D'Ari y Casadesús, 1998; Bergthorsson *et al*., 2007) (Fig. 12d-f).

Un factor importante en las suposiciones anteriores es la existencia de sustratos similares. Tomando en cuenta esto, una posible explicación al hecho de que ciertas categorías metabólicas (particularmente metabolismo de nucleótidos y metabolismo de cofactores y vitaminas) presenten una proporción de parálogos muy baja puede encontrarse precisamente en los sustratos (metabolitos) involucrados. Por ejemplo, muchos de los metabolitos que participan en la biosíntesis y degradación de cofactores orgánicos son casi exclusivos para esta categoría funcional y no suelen fungir como intermediarios en otras categorías (Kuper *et al*., 2000; Begley *et al*., 2008; Bettendorf y Wins, 2009; Braakman y Smith, 2013; Walsh y Wencewicz, 2013). Debido a esto, si alguna enzima que participe en este proceso se duplicara, su retención dependería de la utilidad de su función. Podría coincidir con la aparición de un nuevo sustrato, pero tomando en cuenta las características únicas de muchos de los metabolitos del metabolismo de cofactores, otra enzima que participase en un proceso distinto podría ser una mejor opción para metabolizar el nuevo sustrato y por lo tanto, tendría mayor probabilidad de retener a su parálogo. En pocas palabras, sin la presencia de sustratos similares, la adaptación de una enzima duplicada a una función nueva se vuelve más difícil y lo más probable es que se vuelva una copia no funcional.

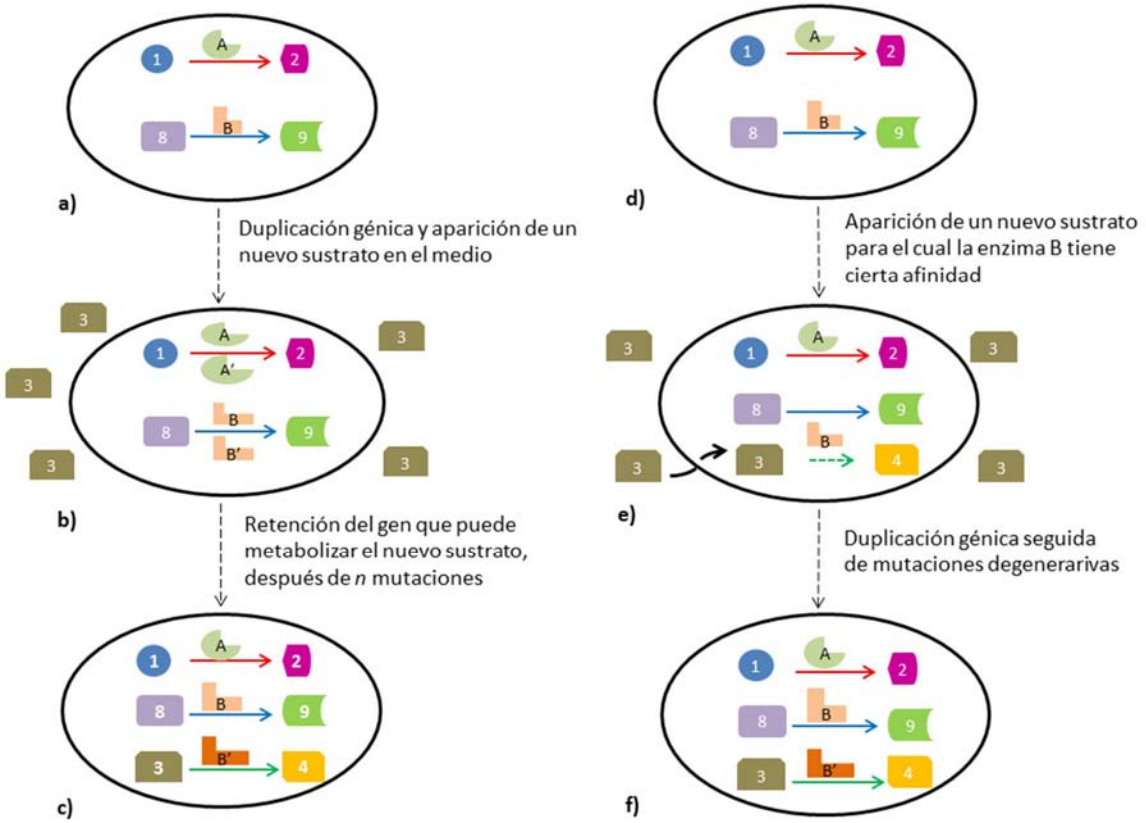


Figura 12. Escenarios distintos mediante los cuales podría preservarse un gen parálogo como consecuencia de la aparición de nuevos sustratos. Las condiciones iniciales son las mismas para ambos casos (a y d), sin embargo, en el primero ocurre un evento de duplicación génica coincidente con la aparición de un nuevo sustrato (D3b), el cual no puede ser metabolizado por ninguna de las enzimas del organismo en cuestión. Sin embargo, dicho sustrato es químicamente similar al sustrato 8, por lo que lo más probable es que la enzima B' requiera un número menor de mutaciones para poder metabolizar al sustrato 3 (D3c), lo cual puede ser visto como la adquisición de una nueva función (neofuncionalización). Por el contrario, también existe la posibilidad de que una de las enzimas presente cierta afinidad al nuevo sustrato (D3e), y si esto resulta benéfico para el organismo (por ejemplo, si el nuevo sustrato representa una buena fuente de carbono), será muy probable que ocurra un evento de duplicación para que de esta manera cada copia (B y B') pueda especializarse hacia una función.

Adicionalmente, la promiscuidad enzimática parece estar determinada por la ruta metabólica en la que participa cada una. No todas las rutas metabólicas son tan versátiles y promiscuas (Khersonsky y Tawfik, 2010), y parece ser que mientras más versátil sea determinada ruta metabólica y más promiscuas las enzimas que la conforman, más fácil será su especialización hacia una nueva función (Umeno *et al*., 2005; Khersonsky y

Tawfik, 2010), tal como ocurre con las enzimas de la ruta biosintética de carotenoides, en la cual se ha demostrado experimentalmente una gran diversificación en cuanto a la especificidad de sustrato gracias a cambios genéticos (mutaciones) limitados (Umeno *et al*., 2005).

Otra posible explicación tiene que ver con la idea de "subsistemas cristalizados" (Woese, 1987; 1998; Koonin, 2014), la cual propone que existen ciertos procesos celulares tan universales y conservados (y por ende muy antiguos) gracias a que se "cristalizaron" tempranamente en el curso de la evolución biológica. Se cree que esto ocurrió debido a que la forma de llevar a cabo dichos procesos era quizá la mejor (o la única) forma de hacerlo, y el hecho de que ocurrieran mutaciones en los genes involucrados podría modificar de manera drástica la adecuación del organismo. El ejemplo más típico es el proceso de traducción (Woese, 1998), seguido quizá por otros procesos informacionales como la transcripción y la replicación (Koonin, 2014). Incluso se ha propuesto la biosíntesis de cofactores como uno de los eventos más antiguos en la evolución del metabolismo debido a que ninguna ruta metabólica funciona sin éstos (White, 1982; Braakman y Smith, 2013). Precisamente, estas cuatro categorías aparecen como aquellas con menor proporción de parálogos en nuestro análisis (Tablas R2 y R3) y apoyan la idea de que los genes más esenciales rara vez presentan muchos parálogos (He y Zhang, 2006).

A partir de las dos consideraciones anteriores, propongo que en procesos muy conservados y/o muy antiguos, así como en aquellos que utilizan intermediarios metabólicos cuya naturaleza química es casi exclusiva para estas rutas, no habrá casi parálogos debido a que se vuelve menos probable que encuentren una función útil.

Finalmente, cabe la posibilidad de que la proporción tan baja de parálogos en estas categorías tenga que ver más con cuestiones metodológicas. Ésta apela a la mayor antigüedad de estos procesos con respecto a otros, lo cual podría provocar que con los métodos y parámetros utilizados sea imposible detectar secuencias que guarden cierto parecido. Probablemente al relajar los parámetros se podría encontrar un mayor número de "hits" para estas categorías, pero debido a que en los otros trabajos publicados estas categorías son la regla más que la excepción, se podrían realizar comparaciones a nivel estructura terciaria para confirmar si en efecto estas categorías casi no poseen parálogos, o si la misma antigüedad de dichos procesos ha provocado que no se detecten parálogos mediante la simple comparación de las estructuras primarias (Mayr *et al*., 2007).

5. Conclusiones

El análisis de las diferentes categorías funcionales en las cuales se agrupan las rutas metabólicas (y por ende las enzimas) de los microorganismos permitió corroborar la idea de que por lo menos dentro de los procariontes, aquellas categorías con una mayor proporción de enzimas parálogas están relacionadas con la interacción entre los organismos y su ambiente, lo cual a su vez les confiere la capacidad de hacer frente a nuevas condiciones ambientales o a sustancias ajenas al ambiente original. El uso de la base de datos KEGG en vez de COG como recurso para clasificar a las rutas metabólicas de los organismos de la muestra nos permitió identificar a la categoría de metabolismo y degradación de xenobióticos como aquella con la mayor proporción de parálogos y para la cual no existe un equivalente directo en la base de datos COG. Aunque probablemente no fueron consideradas algunas enzimas de esta categoría (particularmente aquellas que son producto del transporte horizontal, las cuales se hallan en plásmidos) el estudio posterior de las rutas y enzimas involucradas en este proceso puede ser un buen modelo para conocer mejor el proceso de evolución por medio de duplicación génica, gracias a que muchas de estas enzimas son quizá de las más recientes en haber surgido y en su mayoría a partir de la duplicación de enzimas del metabolismo central.

A su vez, para aquellas categorías metabólicas con una proporción muy baja de parálogos podrían utilizarse bases de datos como BRENDA (Schomburg *et al.*, 2002) en las que se indican los sustratos alternativos que han sido reportados para la mayoría de las enzimas. Si mi propuesta (respecto a que la química orgánica y la bioquímica de las enzimas es una limitante para la retención de sus parálogos) es correcta, entonces esperaríamos ver que las enzimas de estas rutas casi no aceptaran sustratos alternativos. Las comparaciones a nivel de estructura terciaria podrían ayudar a determinar si en efecto las enzimas de estas rutas casi no tienen parálogos, o si estos han divergido tanto que ya es imposible detectar la homología a nivel de estructura primaria.

Por otra parte, la gran diversidad metabólica presente en los dominios Archaea y Bacteria (particularmente en el caso de los ciclos biogeoquímicos), sumado al hecho de que en la mayoría de las óxidorreductasas existen grupos potencialmente catalíticos (principalmente metabolitos tan ampliamente utilizados, como el FAD y el NAD^+) que pueden favorecer que dichas enzimas lleven a cabo reacciones alternativas, son quizá los dos factores más importantes que propician la retención de una gran proporción de

enzimas de esta clase después de eventos de duplicación génica, y es la misma razón por la que en el resto de las clases enzimáticas no observamos el mismo fenómeno.

6. Referencias

Agrawal, N. y S. Shahi. (2015). An environmental cleanup strategy – Microbial transformation of xenobiotic compounds. *International Journal of Current Microbiology and Applied Sciences* 4(4), 429-461

Aharoni, A., L. Gaidukov, O. Khersonsky, S. McQ Gould, C. Roodveldt, y D. Tawfik. (2005). The "evolvability" of promiscuous protein functions. *Nature Genetics* 37(1), 73-76

Altschul, S., W. Gish, W. Miller, E. Myers y D. Lipman. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215(3), 403-410

Andreini, C., I. Bertini, G. Cavallaro, G. Holliday y J. Thornton. (2008). Metal ions in biological catalysis: from enzyme databases to general principles. *Journal of Biological Inorganic Chemistry* 13, 1205-1218

Alves, R., R. Chaleil y M. Sternberg. (2002). Evolution of enzymes in metabolism: a network perspective. *Journal of Molecular Biology* 320, 751-770

Ashburner, M., C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, H. Rubin y G. Sherlock. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics* 25, 25-29

Babcock, E. B. y J. T. Collins. (1922). A case of duplicate genes in Crepis capillaris (L.) Wallr. *Science* 56(1449), 392.

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research* 28(1), 304-305

Begley, T., A. Chatterjee, J. Hanes, A. Hazra y S. Ealick. (2008). Cofactor biosynthesis—still yielding fascinating new biological chemistry. *Current Opinion in Chemical Biology* 12(2), 118-125

Bergthorsson, U., D. Andersson y J. Roth. (2007). Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences* 104 (43), 17004-17009

Bettendorf, L. y P. Wins. (2009). Thiamin diphosphate in biological chemistry: new aspects of thiamin metabolism, especially triphosphate derivatives acting other than as cofactors. *FEBS Journal* 276, 2917-2925

Blakeslee, A. F., J. Belling y M. E. Farnham. (1920). Chromosomal duplication and mendelian phenomena in Datura Mutans. *Science* 52(1347), 388-390

Boyce, S. y K. Tipton. (2001). Enzyme classification and nomenclature. En *Encyclopedia of Life Sciences, Nature Publishing Group*

Braakman, R. y E. Smith. (2013). The compositional and evolutionary logic of metabolism. *Physical Biology* 10, 1-62

Braibant, M., P. Gilot y J. Content. (2000). The ATP binding cassette (ABC) transport systems of *Mycobacterium tuberculosis. FEMS Microbiology Reviews* 24, 449-467

Bratlie, M., J. Johansen, B. Sherman, D. Huang, R. Lempicki y F. Drablos. (2010). Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics* 11, 588

Bridges, C. B. (1936). The bar "gene" a duplication. *Science* 83(2148), 210-211

Buehner, M., G. Ford, D. Moras, K. Olsen y M. Rossmann. (1973). D-Glyceraldehyde-3-Phosphate Dehydrogenase: three-dimensional structure and evolutionary significance. *Proceedings of the National Academy of Sciences* 70(11), 3052-3054

Caetano-Anollés, G., L. Yafremana, H. Gee, D. Caetano-Anollés y J. Mittenthal. (2009). The origin and evolution of modern metabolism. *The International Journal of Biochemistry and Cell Biology* 41, 285-297

Capra, E. y M. Laub. (2012). Evolution of two-component signal transduction systems. *Annual Review of Microbiology* 66, 325-347

Chang, A., I. Schomburg, S. Placzek, L. Jeske, M. Ulbrich, M. Xiao, C. Sensen y D. Schomburg. (2014). BRENDA in 2015: exciting developments in its 25th year of existence. *Nucleic Acids Research* 43(Database issue), D439-D446

Conant, G. y A. Wagner. (2002). GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30(15), 3378-3386

Conant, G. y K. Wolfe. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* 9, 938-950

Copley, S. (2000). Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *TRENDS in Biochemical Sciences* 25(6), 261-265

Copley, S. (2003). Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology* 7, 265-272

Copley, S. (2009). Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nature Chemical Biology* 5(8): 559-566

Crawford, R., C. Jung y J. Strap. (2007). The recent evolution of pentachlorophenol (PCP)-4-monooxygenase (PcpB) and associated pathways for bacterial degradation of PCP. *Biodegradation* 18, 525-539

Croom, E. (2012). Metabolism of xenobiotics of human environments. En *Progress in Molecular Biology and Translational Science* 112, 32-88

Cruz-González, C. (2015). La reconstrucción del genoma ancestral de proteobacterias: un control metodológico para el estudio del último ancestro común (LUCA) (tesis de licenciatura). Universidad Nacional Autónoma de México, México D.F., México.

D'Ari, R. y J. Casadesús. (1998). Underground Metabolism. *BioEssays* 20: 181-186

Davis, J. y D. Petrov. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biology* 2(3), 0318-0326

Davidson, A., E. Dassa, C. Orelle y J. Chen. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiology and Molecular Biology Reviews* 72(2), 317-364

Des Marais, D. y M. Rausher. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454, 762-766

Eventoff, W., M. Rossmann y C. Brändén. (1975). The evolution of dehydrogenases and kinase. *CRC Critical Reviews in Biochemistry* 3(2), 111-140

Falkowski, P., T. Fenchel y E. Delong. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science* 320. 1034-1039

Fani, R. y M. Fondi. (2009). Origin and evolution of metabolic pathways. *Physics of Life Reviews* 6, 23-52

Fischer, J., G. Holliday, S. Rahman y J. Thornton. (2010). The structures and physicochemical properties of organic cofactors in biocatalysis. *Journal of Molecular Biology* 403: 803-824

Fischer, J., G. Holliday y J. Thornton. (2010). The CoFactor database: Organic cofactors in enzyme catalysis. *Bioinformatics* 26(19), 2496-2497

Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology* 19(2), 99-113

Force, A., M. Lynch, F. Pickett, A. Amores, Y. Yan y J. Pstlethwait. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4), 1531-1545

Francino, MP. (2005). An adaptive radiation model for the origin of new gene functions. *Nature Genetics* 37(6), 573-577

Francino, MP. (2012). The ecology of bacterial genes and the survival of the new. *International Journal kanehisaof Evolutionary Biology* 1-14

Gevers, D., K. Vandepoele, C. Simillion e Y. Van de Peer. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *TRENDS in Microbiology* 12(4), 148-154

Gogarten, J. y O. Olendzenski. (1999). Orthologs, paralogs and genome comparisons. *Current Opinion in Genetics & Development* 9(6), 630-636

Gout, J., L. Duret y D. Kahn. (2009). Differential retention of metabolic genes following whole-genome duplication. *Molecular Biology and Evolution* 26(5), 1067-1072

Goward, C. y D. Nicholls. (1994). Malate dehydrogenase: a model for structure, evolution, and catalysis. *Protein Science* 3: 1883-1888

Hahn, M. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity* 100(5), 605-617

Haldane, J. B. S. (1933). The part played by recurrent mutation in evolution. *The American Naturalist* 67(708), 5-19

He, X. y J. Zhang. (2005). Gene complexity and gene duplicability. *Current Biology* 15, 1016-1021

He, X. y J. Zhang. (2006). Higher duplicability of less important genes in yeast genomes. *Molecular Biology and Evolution* 23(1), 144-151

Higgins, C. (2001). ABC transporters: physiology, structure and mechanism – an overview. *Research in Microbiology* 152, 205-210

Hittinger, C. y S. Carroll. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449, 677-682

Horowitz, N. (1945). On the evolution of biochemical syntheses. *Proceedings of the National Academy of Sciences* 31, 153-157

Hughes, A. (1994). The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London B* 256, 119-124

International Union of Biochemistry and Molecular Biology. Nomenclature Committee, Edwin Clifford Webb. (1992). Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. *Academic Press* 862 p.

Janssen, D., I. Dinkla, G. Poelarends y P. Terpstra. (2005). Bacterial degradation of xenobiotic compounds: evolution and distribution of novel enzyme activities. *Environmental Microbiology* 7(12), 1868-1882

Jensen, R. (1976). Enzyme recruitment in evolution of new function. *Annual Review of Microbiology* 30, 409-425

Ji, H., L. Chen y H. Zhang. (2008). Organic cofactors participated more frequently than transition metals in redox reactions of primitive proteins. *BioEssays* 30: 766-771

Kanehisa, M., y S. Goto. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28(1), 27-30

Karp, P., M. Riley, S. Paley y A. Pellegrini-Toole. (2002). The MetaCyc Database. *Nucleic Acids Research* 30(1), 59-61

Khersonsky, O. y D. Tawfik. (2010). Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry* 79, 471-505

Kim, J., S. Senn, A. Harel, B. Jelen, y P. Falkowski. (2013). Discovering the electronic circuit diagram of life: structural relationships among transition metal binding sites in oxidoreductases. *Philosophical Transactions of the Royal Society B* 368, 1622, 20120257

Kondrashov, F. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. R. Soc. B* 279, 5048-5057

Koonin, E., K. Makarova y L. Aravind. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. *Annual Review of Microbiology* 55, 709-742

Koonin, E. (2014). Carl Woese's vision of cellular evolution and the domains of life. *RNA Biology* 11(3), 197-204

Koretke, K., A. Lupas, P. Warren, M. Rosenberg y J. Brown. (2000). Evolution of two-component signal transduction. *Molecular Biology and Evolution* 17(12), 1956-1970

Kuper, J., T. Palmer, R. Mendel y G. Schwarz. (2000). Mutations in the molybdenum cofactor biosynthetic protein Cnx1G from Arabidopsis thaliana define functions for molybdopterin

binding, molybdenum insertion, and molybdenum cofactor stabilization. *Proceedings of the National Academy of Sciences* 97(12), 6475-6480

Kuriyan, J., B. Konforti y D. Wemmer. (2013). The molecules of life. 1ª Edición. *Garland Science, Taylor & Francis Group, LLC*. pp. 228-230

Li, X., M. Schuler y M. Berenbaum. (2007). Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. *Annual Review of Entomology* 52, 231-253

Lawrence, J. y H. Hendrickson. (2003). Lateral gene transfer: when will adolescence end? *Molecular Microbiology* 50(3), 739-749

Lazcano, A. y S. Miller. (1999). On the origin of metabolic pathways. *Journal of Molecular Evolution* 49, 424-431

Liao, D. (2008). Concerted evolution. En *Encyclopedia of Life Sciences, John Wiley and Sons, Ltd*

Liberles, D., G. Kolesov y K. Dittmar. (2010). Understanding gene duplication through biochemistry and population genetics. En *Evolution after gene duplication*. K. Dittmar y D. Liberles (Eds.) 1-21

López-Maury, L., S. Marguerat y J. Bähler. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nature Reviews Genetics* 9, 583-593

Lynch, M. y A. Force. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459-473

Lynch, M. y J. Conery. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151-1155

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. van Montagu, M. Kuiper e Y. van de Peer. (2005). Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences* 102(15), 5454-5459

Maere, S. e Y. van de Peer. (2010). Duplicated retention after small- and large-scale duplication. En *Evolution after gene duplication*. K. Dittmar y D. Liberles (Eds.) 31-56

Marland, E., A. Prachumwat, N. Maltsev, Z. Gu y WH Li. (2004). Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *Journal of Molecular Evolution* 59, 806-814

Martínez-Núñez, M., K. Rodríguez-Vázquez y E. Pérez-Rueda. (2015). The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. *Proteins* 83, 1625-1631

Mayr, G., F. Domingues y P. Lackner. (2007). Comparative analysis of protein structure alignements. *BMC Structural Biology* 7, 50

Michal, G. y D. Schomburg (Eds.) Biochemical pathways: An atlas of biochemistry and molecular biology. 2a Edición. *John Wiley & Sons, Inc*. Singapur. pp. 23-24

Mudunuri, U., A. Che, M. Yi y RM Stephens. (2009). bioDBnet: the biological database network. *Bioinformatics* 25(4), 555-556

Müller, H. J. (1934). The origination of chromatin deficiencies as minute deletions subject to insertions elsewhere. *Genetica* 17(3-4), 237-252.

Näsvall, J., L. Sun, J. Roth y D. Andersson. (2012). Real-time evolution of new genes by innovation, amplification, and divergence. *Science* 338 (6105), 3854-387

Notebaart, R., B. Szappanos, B. Kintses, F. Pál, A. Gyorkei, B. Bogos, V. Lázar, R. Spohn, B. Csorgo, A. Wagner, E. Ruppin, S. Pál y B. Papp. (2014). Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences* 111(32), 11762-11767

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

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono y M. Kanehisa. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 27(1), 29-34

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag. Berlín. 160 pp.

Olendzenski, L., O. Zhaxbayeva y JP Gogarten. (2006). Orthologs, paralogs and xenologs in human and other genomes. En *Encyclopedia of Life Sciences, John Wiley and Sons, Ltd*

Omiecinski, C., J. Vanden-Heuvel, G. Perdew y J. Peters. (2011). Xenobiotic metabolism, disposition, and regulation by receptors: from biochemical phenomenon to predictors of major toxicities. *Toxicological Sciences* 120(S1), S49-S75

Papp, B., C. Pál y L. Hurst. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194-197

Piskur, J., E. Rozpedowska, S. Polakova, A. Merico y C. Compagno. (2006). How did *Saccharomyces* evolve to become a good brewer? *TRENDS in Genetics* 22(4), 183-186

Riley, M. (1993). Functions of the gene-products of *Escherichia coli*. *Microbiological Reviews* 57(4), 862-952

Roth, C., S. Rastogi, L. Arvestad, K. Dittmar, S. Light, D. Ekman y D. Liberles. (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *Journal of Experimental Zoology (Mol Dev Evol)* 308B, 58-73

Russell, R., C. Scott, C. Jackson, R. Pandey, G. Pandey, M. Taylor, C. Coppin, J. Liu y J. Oakeshott. (2011). The evolution of new enzyme function: lessons from xenobiotic metabolizing bacteria versus insecticide-resistant insects. *Evolutionary Applications* 4(2), 225-248

Saier, M. e I. Paulsen. (1999). Paralogous genes encoding transport proteins in microbial genomes. *Research in Microbiology* 150, 689-699

Schmidt, S., S. Sunyaev, P. Bork y T. Dandekar. (2003). Metabolites: a helping hand for pathway evolution? *TRENDS in Biochemical Sciences* 28(6), 336-341

Schomburg, I., A. Chang y D. Schomburg. (2002). BRENDA, enzyme data and metabolic information. *Nucleic Acids Research* 30(1), 47-49

Serres, M., A. Kerr, T. McCormack y M. Riley. (2009). Evolution by leaps: gene duplication in bacteria. *Biology Direct* 4, 46

Sheng, X., M. Huvet, J. Pinney y M. Stumpf. (2012). Evolutionary characteristics of bacterial two-component systems. *Advances in Experimental Medicine and Biology* 751, 121-137

Tatusov, R., E. Koonin y D. Lipman. (1997). A genomic perspective on protein families. *Science* 278 (5338), 631-637

Tatusov, R. N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, B. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin y D. Natale. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41

Taylor, J. y J. Raes. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annual Review of Genetics* 38, 615-643

Teichmann, S., S. Rison, J. Thornton, M. Riley, J. Gough y C. Chothia. (2001a). The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *Journal of Molecular Biology* 311, 693-708

Teichmann, S., S. Rison, J. Thornton, M. Riley, J. Gough y C. Chothia. (2001b). Small-molecule metabolism: an enzyme mosaic. *TRENDS in Biotechnology* 19(12), 482-486

Thomson, M., E. Gaucher, M. Burgan, D. De Kee, T. Li, J. Aris y S. Benner. (2005). Resurrecting ancestral alcohol dehydrogenases from yeast. *Nature Genetics* 37(6), 630-635

Umeno, D., A. Tobias y F. Arnold. (2005). Diversifying carothenoid biosynthetic pathways by directed evolution. *Microbiology and Molecular Biology Reviews* 69(1), 51-78

Varadarajan, N., J. Gam, M. Olsen, G. Georgiou y B. Iverson. (2005). Enngineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity. *Proceedings of the National Academy of Sciences* 102(19), 6855-6860

Wackett, L. (2004). Evolution of enzymes for the metabolism of new chemical inputs into the environment. *The Journal of Biological Chemistry* 279(40), 41259-41262

Waley, S. (1969). Some aspects of the evolution of metabolic pathways. *Comparative Biochemistry and Phisiology* 30, 1-11

Walsh, C. y T. Wencewicz. (2013). Flavoenzymes: Versatile catalysts in biosynthetic pathways. *Natural Product Reports* 30, 175-200

Watts, K., B. Mijts, P. Lee, A. Manning y C. C. Schmidt-Dannert. (2006). Discovery of a substrate selectivity switch in tyrosine ammonia-lyase, a member of the aromatic amino acid lyase family. *Chemistry & Biology* 13, 1317-1326

White, H. (1982). Evolution of coenzymes and the origin of pyridine nucleotides. En *The Pyridine Nucleotide Coenzymes*, eds. Everse, J., B. Anderson y K. You. Academic Press. Pp. 1-17

Woese, C. (1987). Bacterial evolution. *Microbiological Reviews* 51(2), 221-271

Woese, C. (1998). The universal ancestor. *Proceedings of the National Academy of Sciences*. 95: 6854-6859

Ycas, M. (1974). On earlier states of the biochemical system. *Journal of Theoretical Biology* 44(1), 145-160

Zhang, J., Y. Zhang y H. Rosenberg. (2002). Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nature Genetics* 30, 411-415

Zhang, J. (2003). Evolution by gene duplication: an update. *TRENDS in Ecology and Evolution* 18(6), 292-298

Anexos I

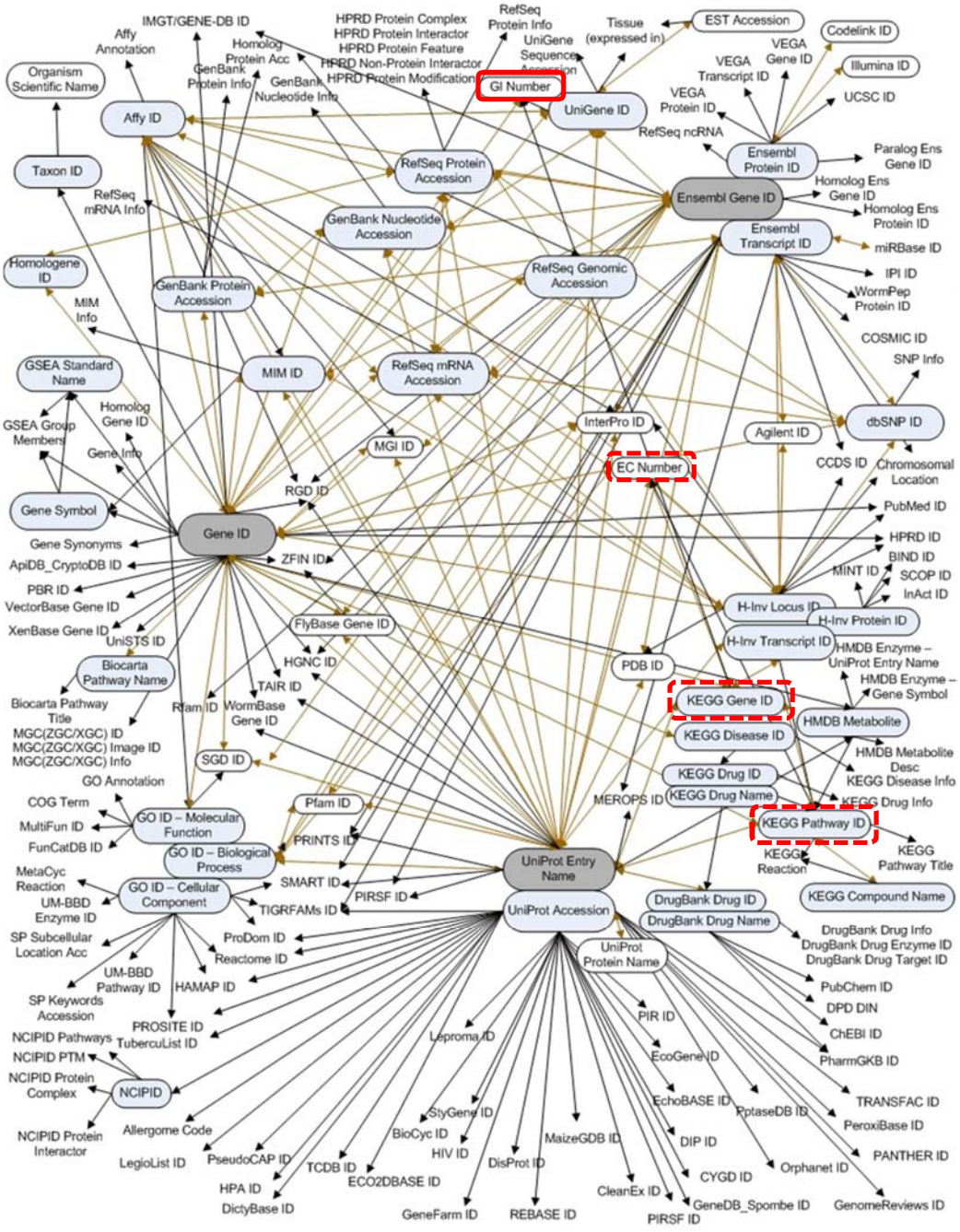


Figura S1. Esquema en el que se ilustra, mediante nodos y conectores, el conjunto de las bases de datos a las cuales tiene acceso la herramienta db2db del sitio bioDBnet. El formato de entrada de la lista de genes fue el GI Number (rectángulo con línea sólida), a partir del cual se obtuvieron los respectivos identificadores para KEGG Gene ID, EC Number y KEGG Pathway ID (rectángulos con línea discontínua). Modificado a partir de https://biodbnet-abcc.ncifcrf.gov/dbInfo/netGraph.php

Tabla S1. Proporción de enzimas dentro de cada clase enzimática para los organismos del dominio Bacteria, con respecto al número total de enzimas que presenta cada uno. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG.

EC Class	eco	bsu	sco	ssm	syn	sru	cch	tth	aae	ddf	tye
EC 1	19.89	18.76	22.73	16.39	20.63	20.09	19.97	19.66	21.55	18.20	19.97
EC 2	32.75	31.70	29.30	32.11	31.75	30.13	32.43	31.27	32.66	35.32	35.97
EC 3	26.74	31.31	24.62	25.64	20.50	23.73	19.97	19.66	16.33	17.89	15.51
EC 4	9.40	7.71	9.30	11.48	10.75	10.19	9.74	11.30	10.44	10.70	10.23
EC 5	6.06	4.90	6.22	6.80	7.13	6.11	6.71	6.66	6.90	6.57	6.77
EC 6	5.15	5.62	7.83	7.58	9.25	9.75	11.18	11.46	12.12	11.31	11.55
Total (%)	100	100	100	100	100	100	100	100	100	100	100

Tabla S2. Proporción de enzimas dentro de cada clase enzimática para los organismos del dominio Archaea, con respecto al número total de enzimas que presenta cada uno. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG.

EC Class	mmp	mac	fpl	hma	pfu	sol	pcl	asc	nmr
EC 1	18.30	20.00	20.83	23.07	19.71	23.21	22.80	19.81	15.75
EC 2	34.94	34.39	33.67	29.02	33.63	31.05	33.83	35.75	35.89
EC 3	14.97	18.22	14.50	19.39	18.44	15.86	16.07	16.67	12.04
EC 4	10.91	10.32	11.67	12.67	10.67	12.02	11.03	9.90	12.91
EC 5	7.95	6.50	7.83	5.70	6.51	5.84	5.05	6.28	8.53
EC 6	12.94	10.57	11.50	10.14	11.03	12.02	11.21	11.59	14.88
Total (%)	100	100	100	100	100	100	100	100	100

Tabla S3. Proporción de enzimas dentro de cada clase enzimática para los organismos del dominio Bacteria, con respecto al número total de enzimas parálogas que presenta cada uno. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG.

EC Class	eco	bsu	sco	ssm	syn	sru	cch	tth	aae	ddf	tye
EC 1	25.40	24.12	27.38	20.14	29.60	26.75	24.40	28.04	34.27	28.66	30.88
EC 2	28.99	30.43	28.15	28.20	26.40	27.63	27.38	25.40	24.16	29.30	32.35
EC 3	25.27	31.44	22.49	31.28	21.20	21.93	20.83	19.05	11.80	14.65	13.24
EC 4	9.18	5.30	8.74	10.90	8.00	11.40	8.93	8.99	10.67	7.64	10.78
EC 5	6.78	3.91	6.17	6.16	9.20	6.14	4.76	7.41	7.87	5.73	5.39
EC 6	4.39	4.80	7.07	3.32	5.60	6.14	13.69	11.11	11.24	14.01	7.35
Total (%)	100	100	100	100	100	100	100	100	100	100	100

Tabla S4. Proporción de enzimas dentro de cada clase enzimática para los organismos del dominio Archaea, con respecto al número total de enzimas parálogas que presenta cada uno. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG.

EC Class	mmp	mac	fpl	hma	pfu	sol	pcl	asc	nmr
EC 1	36.36	24.06	30.81	33.33	34.63	31.98	38.46	35.04	21.05
EC 2	24.68	33.42	32.70	22.85	27.80	28.38	27.22	35.90	29.82
EC 3	11.69	19.52	11.37	20.43	14.15	13.51	12.43	9.40	8.77
EC 4	7.14	9.36	9.95	11.83	9.27	8.11	8.88	3.42	14.04
EC 5	10.39	5.08	6.64	4.57	4.88	4.95	2.37	4.27	7.89
EC 6	9.74	8.56	8.53	6.99	9.27	13.06	10.65	11.97	18.42
Total (%)	100	100	100	100	100	100	100	100	100

Tabla S5. Proporción de enzimas parálogas dentro de cada clase enzimática para los organismos del dominio Bacteria. Cada valor se obtuvo al dividir el número de enzimas parálogas en cada clase enzimática sobre el total de enzimas presentes en la misma clase. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG.

EC Class	eco	bsu	sco	ssm	syn	sru	cch	tth	aae	ddf	tye
EC 1	0.58	0.67	0.66	0.58	0.45	0.44	0.33	0.42	0.48	0.38	0.52
EC 2	0.40	0.50	0.52	0.41	0.26	0.30	0.23	0.24	0.22	0.20	0.30
EC 3	0.43	0.52	0.50	0.57	0.32	0.31	0.28	0.28	0.22	0.20	0.29
EC 4	0.45	0.36	0.51	0.45	0.23	0.37	0.25	0.23	0.31	0.17	0.35
EC 5	0.51	0.41	0.54	0.43	0.40	0.33	0.19	0.33	0.34	0.21	0.27
EC6	0.39	0.44	0.49	0.21	0.19	0.21	0.33	0.28	0.28	0.30	0.21

Tabla S6. Proporción de enzimas parálogas dentro de cada clase enzimática para los organismos del dominio Archaea. Cada valor se obtuvo al dividir el número de enzimas parálogas en cada clase enzimática sobre el total de enzimas presentes en la misma clase. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG.

EC Class	mmp	mac	fpl	hma	pfu	sol	pcl	asc	nmr
EC 1	0.57	0.57	0.52	0.68	0.65	0.51	0.53	0.50	0.33
EC 2	0.20	0.46	0.34	0.37	0.31	0.34	0.25	0.28	0.21
EC 3	0.22	0.51	0.28	0.50	0.28	0.32	0.24	0.16	0.18
EC 4	0.19	0.43	0.30	0.44	0.32	0.25	0.25	0.10	0.27
EC 5	0.37	0.37	0.30	0.38	0.28	0.31	0.15	0.19	0.23
EC6	0.21	0.39	0.26	0.33	0.31	0.40	0.30	0.29	0.31

Tabla S7. Proporción de enzimas parálogas dentro de cada categoría funcional (de acuerdo con la base de datos KEGG) para los organismos del dominio Archaea. Cada valor se obtuvo al dividir el número de enzimas parálogas en cada clase enzimática sobre el total de enzimas presentes en la misma clase. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG**.** Se resalta en gris los cinco valores promedio más altos del total de categorías consideradas. Se omiten los valores de las categorías de biosíntesis y metabolismos de glucanos, así como de motilidad celular debido a que solo están reportadas en unos cuantos organismos.

	mmp	**mac**	**fpl**	**hma**	**pfu**	**sol**	**pcl**	**asc**	**nmr**	**Prom.**
Metabolismo de carbohidratos	0.43	0.57	0.62	0.65	0.54	0.51	0.49	0.41	0.38	0.51
Metabolismo energético	0.46	0.57	0.38	0.41	0.19	0.52	0.49	0.21	0.19	0.38
Metabolismo de lípidos	0.00	0.36	0.64	1.00	0.67	0.89	0.74	0.73	0.25	0.59
Metabolismo de nucleótidos	0.11	0.25	0.19	0.28	0.20	0.25	0.15	0.17	0.16	0.20
Metabolismo de aminoácidos	0.33	0.51	0.48	0.60	0.46	0.48	0.40	0.36	0.35	0.44
Metabolismo de otros aminoácidos	0.13	0.55	0.75	0.74	0.73	0.56	0.33	0.25	0.17	0.47
Biosíntesis y metabolismo de glicanos	-	-	-	1.00	1.00	-	-	0.00	-	0.67
Metabolismo de cofactors y vitaminas	0.18	0.40	0.21	0.35	0.25	0.33	0.30	0.19	0.35	0.28
Metabolismo de terpenos y policétidos	0.09	0.47	0.61	0.71	0.20	0.54	0.50	0.47	0.23	0.42
Biosíntesis de otros metabolitos secundarios	0.50	0.67	0.54	0.57	0.60	0.75	0.42	0.44	0.50	0.55
Metabolismo y degradación de xenobióticos	0.83	0.64	0.89	0.76	0.57	0.88	0.61	-	1.00	0.77
Transcripción	0.06	0.17	0.11	0.42	0.12	0.12	0.12	0.00	0.32	0.16
Traducción	0.05	0.11	0.08	0.11	0.14	0.14	0.11	0.10	0.06	0.10
Plegamiento, clasificación y degradación	0.10	0.40	0.34	0.29	0.45	0.59	0.42	0.35	0.44	0.38
Replicación y reparación	0.00	0.28	0.07	0.37	0.00	0.31	0.27	0.23	0.00	0.17
Transporte membranal	0.63	0.86	0.72	0.68	0.69	0.64	0.58	0.65	0.52	0.66
Transducción de señales	0.57	0.50	0.58	0.60	0.00	0.67	0.36	0.58	0.59	0.50
Motilidad celular	0.60	0.79	0.50	0.69	-	-		-	-	0.65

Tabla S8. Proporción de enzimas parálogas dentro de cada categoría funcional (de acuerdo con la base de datos KEGG) para los organismos del dominio Bacteria. Cada valor se obtuvo al dividir el número de enzimas parálogas en cada clase enzimática sobre el total de enzimas presentes en la misma clase. En la primera fila se indica el código de tres letras usado para identificar a cada organismo en la base de datos KEGG**.** Se resalta en gris los cinco valores promedio más altos del total de categorías consideradas.

	eco	**bsu**	**sco**	**ssm**	**syn**	**sru**	**cch**	**tth**	**aae**	**ddf**	**tye**	**Prom.**
Metabolismo de carbohidratos	0.57	0.60	0.70	0.60	0.36	0.42	0.36	0.39	0.46	0.35	0.48	0.48
Metabolismo energético	0.57	0.50	0.59	0.60	0.32	0.28	0.32	0.29	0.44	0.29	0.54	0.43
Metabolismo de lípidos	0.45	0.75	0.79	0.61	0.29	0.57	0.35	0.57	0.39	0.38	0.52	0.52
Metabolismo de nucleótidos	0.34	0.40	0.38	0.24	0.14	0.14	0.15	0.22	0.19	0.18	0.27	0.24
Metabolismo de aminoácidos	0.48	0.54	0.58	0.50	0.33	0.40	0.28	0.45	0.37	0.28	0.36	0.41
Metabolismo de otros aminoácidos	0.52	0.75	0.73	0.56	0.20	0.50	0.22	0.46	0.40	0.10	0.16	0.42
Biosíntesis y metabolismo de glicanos	0.36	0.50	0.67	0.10	0.00	0.29	0.24	0.00	0.24	0.27	0.15	0.26
Metabolismo de cofactors y vitaminas	0.23	0.35	0.40	0.27	0.26	0.24	0.25	0.14	0.21	0.14	0.19	0.24
Metabolismo de terpenos y policétidos	0.48	0.66	0.82	0.42	0.37	0.55	0.35	0.45	0.20	0.20	0.33	0.44
Biosíntesis de otros metabolitos secundarios	0.47	0.44	0.44	0.76	0.56	0.57	0.43	0.63	0.00	0.29	0.62	0.47
Metabolismo y degradación de xenobióticos	0.78	1.00	0.92	0.69	0.50	0.59	1.00	0.68	1.00	0.50	0.40	0.73
Transcripción	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Traducción	0.14	0.19	0.27	0.11	0.14	0.08	0.04	0.10	0.09	0.08	0.09	0.12
Plegamiento, clasificación y degradación	0.27	0.49	0.27	0.19	0.28	0.26	0.11	0.21	0.19	0.18	0.24	0.24
Replicación y reparación	0.14	0.18	0.42	0.18	0.05	0.21	0.16	0.08	0.13	0.15	0.14	0.17
Transporte membranal	0.72	0.81	0.74	0.84	0.63	0.53	0.41	0.67	0.21	0.46	0.39	0.58
Transducción de señales	0.67	0.67	0.66	0.75	0.49	0.59	0.21	0.23	0.35	0.62	0.76	0.55
Motilidad celular	0.23	0.46	0.00	0.70	0.00	0.37	0.00	0.00	0.22	0.51	0.44	0.27

Tabla S9. Diferencias significativas entre cada una de las categorías funcionales con respecto a las demás, correspondientes al dominio Archaea y obtenidas a partir de la tabla S7. Éstas se identificaron por medio de pruebas de t pareada y con un valor de P = 0.05. Las celdas en color negro representan la intersección de una categoría con ella misma y aquellos casos en donde existen diferencias significativas se indican con una X. Las columnas correspondientes a las categorías 1.7 y 4.1 se omitieron de este análisis debido a que en la mayoría de los organismos no se encuentra representada y su comparación con el resto de las categorías podría sesgar los resultados.

	1.1	1.2	1.3	1.4	1.5	1.6	1.8	1.9	1.10	1.11	2.1	2.2	2.3	2.4	3.1	3.2
1.1		X		X			X				X	X	X	X		
1.2	X			X					X	X	X	X		X	X	
1.3				X			X				X	X		X		
1.4	X	X	X		X	X	X	X	X	X		X	X		X	X
1.5				X			X		X	X	X	X		X	X	
1.6				X							X	X		X		
1.8	X		X	X	X				X	X	X	X			X	X
1.9				X						X	X	X		X	X	
1.10		X		X	X		X				X	X	X	X	X	
1.11		X		X	X		X	X			X	X	X	X		
2.1	X	X	X		X	X	X	X	X	X			X		X	X
2.2	X	X	X	X	X	X	X	X	X	X			X		X	X
2.3	X			X					X	X	X	X		X	X	
2.4	X	X	X		X	X		X	X	X			X		X	X
3.1		X		X	X		X	X	X		X	X	X	X		X
3.2				X			X				X	X		X	X	

Tabla S10. Diferencias significativas entre cada una de las categorías funcionales con respecto a las demás, correspondientes al dominio Bacteria y obtenidas a partir de la tabla S8. Éstas se identificaron por medio de pruebas de t pareada y con un valor de P = 0.05. Las celdas en color negro representan la intersección de una categoría con ella misma y aquellos casos en donde existen diferencias significativas se indican con una X. Los encabezados de cada fila y columna corresponden a los códigos de las categorías funcionales de KEGG PATHWAY (ver Tabla 2 en Metodología)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	1.10	1.11	2.1	2.2	2.3	2.4	3.1	3.2	4.1
1.1				X			X	X			X	X	X	X	X			
1.2				X			X	X			X	X	X	X	X	X		
1.3				X			X	X			X	X	X	X	X			
1.4	X	X	X		X	X			X	X	X	X	X			X	X	X
1.5				X		X			X	X	X	X	X			X	X	X
1.6				X				X			X	X	X	X	X			
1.7	X	X	X		X				X	X	X	X				X	X	
1.8	X	X	X		X	X			X	X	X	X	X			X	X	X
1.9				X			X	X			X	X	X	X	X			
1.10				X			X	X			X	X	X	X	X			
1.11	X	X	X	X	X	X	X	X	X	X		X	X	X	X			X
2.1	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
2.2	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X
2.3	X	X	X		X	X			X	X	X	X	X			X	X	X
2.4	X	X	X		X	X			X	X	X	X				X	X	X
3.1		X		X	X		X	X				X	X	X	X			
3.2				X			X	X				X	X	X	X			
4.1				X				X			X	X	X	X	X			

Anexos II

Script 1: parseblast1.pl

```
#!/usr/bin/perl

# This scripts takes two parameters

print ("Dame el nombre del archivo (salidad del blast)\n");
$blastfile = <STDIN>;
chomp ($blastfile);

print ("Dame el valor de corte\n");
$cutoff = <STDIN>;
chomp ($cutoff);

open(BLAST, "$blastfile") || die " Unable to load $blastfile.\n";
open(RESULTS1, ">>Results1");

while (<BLAST>) {
        $totallines++;
        # using the split function we populate the @fields array
        @fields = split(/\t/, $_);

        # Compare alignment length with user supplied cutoff
        if ($fields[2] < $cutoff)  {
                $printedlines++;
                # we print the whole line ($_) in case
                print RESULTS1 "$_";
        }
}
$percentage = int($printedlines*100/$totallines);
print RESULTS1 "Printed $printedlines/$totallines ($percentage%).\n";
```

Script 2: parseblast2.pl

```
#!/usr/bin/perl

# This scripts takes two parameters

print ("Dame el nombre del archivo (salidad del blast)\n");
$blastfile = <STDIN>;
chomp ($blastfile);

print ("Dame el valor de corte\n");
$cutoff = <STDIN>;
chomp ($cutoff);

open(BLAST, "$blastfile") || die " Unable to load $blastfile.\n";
open(RESULTS, ">>Results");

while (<BLAST>) {
        $totallines++;
        # using the split function we populate the @fields array
        @fields = split(/\t/, $_);

        # Compare alignment length with user supplied cutoff
        if ($fields[6] > $cutoff)  {
                $printedlines++;
                # we print the whole line ($_) in case
                print RESULTS "$_";
        }
}
$percentage = int($printedlines*100/$totallines);
print RESULTS "Printed $printedlines/$totallines ($percentage%).\n";
```

Script 3: blast2adjLists_ed1.pl

```
#!/usr/local/bin/perl -w

# blast2adjLists_ed1.pl, version ed1.
# copyright Manuel J. Gómez, CNB-CSIC

# USAGE: blast2adjLists_ed1.pl blast_output.file e-value
# EXAMPLE: blast2adjLists_ed1.pl Myco.blast 1e-80


# This script parses a BLAST output in table format, and constructs adjacency
# lists by considering as neighbours those pairs of sequences that
# have been identified as similar with e-values under a certain value.

# For each line parsed that fulfils the two following conditions: i) that the
# line does not refer to a self-match, and, ii) that the e-value is under a given
# value, the two entries in the line are entered as first and second values in a
# hash of hashes, forcing the construction of a simmetrical matrix. The advantage
# of doing this in Perl is that even in this situation the result is not a complete
# matrix, because only the existing pairs of neighbours are stored. The e-values
# are also stored, and if there are several hits for the same pair of sequences,
# with different e-values, only the one that has the lowest e-value is kept.


use strict;


my ($threshold);
my (@column);
my ($prot1,$prot2);
my (%adMat);
my ($first,$second);

$threshold = $ARGV[1];          # Maximum e-value allowed.


# This part parses the blast output

open (FILE , "$ARGV[0]") or die "Can not open $ARGV[0]\n";

while (<FILE>){
        chomp;
        @column = split (/\t/,$_);
        if ($column[0] ne $column[1]) {
                if ($column[10]<$threshold) {
                        $column[0] =~ /gi\|(\d*)\|/;
                        $prot1 = $1;
                        $column[1] =~ /gi\|(\d*)\|/;
                        $prot2 = $1;
                        if (not exists $adMat{$prot1}{$prot2}) {
```

```
                        $adMat{$prot1}{$prot2} = $column[10];
                        $adMat{$prot2}{$prot1} = $column[10];
                    } elsif ($adMat{$prot1}{$prot2} > $column[10]){
                        $adMat{$prot1}{$prot2} = $column[10];
                        $adMat{$prot2}{$prot1} = $column[10];
                    }
                }
            }
}


# This part prints the adjacency list


print "Nodes\t\tNeighbours\n";

foreach $first (sort keys %adMat) {
        print "$first\t";
        foreach $second (sort keys %{$adMat{$first}}) {
                print "$second ";
        }
        print "\n";
}

##########################################################################
#####
```