Centro de Investigación Científica de Yucatán, A.C.

Posgrado en Ciencias Biológicas

IDENTIFICACIÓN DE FIRMAS FUNCIONALES EN EL METABOLISMO DE PROCARIOTES Y EUCARIOTES

Tesis que presenta

M. C. PEDRO JAVIER ESCOBAR TURRIZA

En opción al título de

DOCTOR EN CIENCIAS

(Ciencias Biológicas: Opción Biotecnología)

Mérida, Yucatán, México

2021

CENTRO DE INVESTIGACIÓN CIENTÍFICA DE YUCATÁN, A. C.

POSGRADO EN CIENCIAS BIOLÓGICA

RECONOCIMIENTO

Por medio de la presente, hago constar que el trabajo de tesis de **Pedro Javier Escobar Turriza** titulado "**Identificación de funcionales en el metabolismo de procariotes y eucariotes"** fue realizado en la Unidad de Biotecnología del Centro de Investigación Científica de Yucatán, A.C. en conjunto con el Laboratorio de Biología Computacional del Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas de la Universidad Autónoma de México, Sede Mérida, bajo la dirección del Dr. Jorge Humberto Ramírez Prado y el Dr. Ernesto Pérez Rueda, dentro de la opción de Biotecnología, perteneciente al Programa de Posgrado en Ciencias Biológicas de este Centro.

Atentamente.

Dra. Cecilia Hernández Zepeda

Director de Docencia

Mérida, Yucatán, México, a 01 de Octubre de 2021

DECLARACIÓN DE PROPIEDAD

Declaro que la información contenida en la sección de Materiales y Métodos Experimentales, los Resultados y Discusión de este documento proviene de las actividades de experimentación realizadas durante el período que se me asignó para desarrollar mi trabajo de tesis, en las Unidades y Laboratorios del Centro de Investigación Científica de Yucatán, A.C., y que a razón de lo anterior y en contraprestación de los servicios educativos o de apoyo que me fueron brindados, dicha información, en términos de la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, le pertenece patrimonialmente a dicho Centro de Investigación. Por otra parte, en virtud de lo ya manifestado, reconozco que de igual manera los productos intelectuales o desarrollos tecnológicos que deriven o pudieran derivar de lo correspondiente a dicha información, le pertenecen patrimonialmente al Centro de Investigación Científica de Yucatán, A.C., y en el mismo tenor, reconozco que si derivaran de este trabajo productos intelectuales o desarrollos tecnológicos, en lo especial, estos se regirán en todo caso por lo dispuesto por la Ley Federal del Derecho de Autor y la Ley de la Propiedad Industrial, en el tenor de lo expuesto en la presente Declaración.

_____

Pedro Javier Escobar Turriza

Este trabajo se llevó a cabo en la Unidad de Biotecnología del Centro de Investigación Científica de Yucatán en conjunto con el Laboratorio de Biología Computacional del Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas de la Universidad Autónoma de México, Sede Mérida, y forma parte del proyecto titulado “Comparación y predicción de rutas metabólicas utilizando algoritmos genéticos, programación dinámica y cadenas ocultas de Markov”, clave DGAPA-Universidad Nacional Autónoma de México (UNAM) IN209620, en el que participé bajo la dirección del Dr Jorge Humberto Ramírez Prado en conjunto con el Dr. Ernesto Pérez Rueda.

AGRADECIMIENTOS

Al CONACYT por la beca otorgada (CVU/Becario 624129/338189) para realizar el posgrado de Doctorado después de Maestría en Ciencias Biológicas opción Biotecnología del Centro de Investigación Científica de Yucatán A. C. (CICY).

Al financiamiento del proyecto “Comparación y predicción de rutas metabólicas utilizando algoritmos genéticos, programación dinámica y cadenas ocultas de Markov” DGAPA-UNAM IN209620.

Al CICY por todas las instalaciones brindadas para la realización del posgrado, así como los conocimientos científicos ofrecidos para mi formación doctoral.

Al Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), UNAM, Unidad Mérida, por la estancia permanente, cobijarme y ofrecer todas las instalaciones requeridas durante todo el período de mi formación académica.

Al Dr. Ernesto Pérez Rueda, primeramente, por aceptarme para formar parte de su grupo de trabajo en la UNAM, por ser mi guía científico compartiendo todas sus enseñanzas académicas, consejos, experiencias y su visión para afrontar los retos científicos, que son un cimiento para mi desarrollo académico. Gracias por su amistad durante todo este trayecto doctoral.

Al Dr. Jorge Humberto Ramírez Prado, por aceptarme en su grupo de investigación en el CICY, por todos los consejos y asesoramientos académicos brindados durante el posgrado. Gracias por su apoyo y amistad.

Al IMEC. Rafael Hernández Guerrero, por sus valiosos conocimientos en programación y todos los apoyos brindados, siendo un pilar en toda mi estancia de posgrado, incluso haciéndola más amena y enriquecedora. Gracias por tu amistad.

Al Dr. Augusto César Poot Hernández, por todos tus valiosos conocimientos en Biología computacional y programación en Python, fuiste una vía de inspiración para realizar este trabajo doctoral. Te agradezco infinitamente haberme aceptado para una estancia tanto en el IIMAS Sede CU, así como una en el Instituto de Fisiología Celular. Gracias por todo tu apoyo y valiosa amistad.

Al Dr. Edgardo Galán Vázquez, por su sus consejos y conocimientos ofrecidos en torno a las discusiones de los resultados obtenidos en este estudio, permitiendo mejorar mis conocimientos científicos. Así también por la disposición de apoyarme en todas las dudas que fueron surgiendo durante tu estancia posdoctoral. Gracias por tu valiosa amistad.

A la Dra. Katya Rodríguez Vázquez, por sus recomendaciones académicas, por aceptar ser parte de mi comité tutoral. Le agradezco por la estancia de investigación en su laboratorio en el IIMAS Sede CU.

A la Dra. Elsa Góngora Castillo, por sus consejos, cuestionamientos y discusiones en cuanto a los resultados obtenidos en este proceso doctoral, que provocaron a seguir mejorando en la divulgación de la ciencia.

A mi comité predoctoral, por las observaciones, correcciones y consejos brindados en este posgrado.

Al grupo de seminarios de Bioinformática de la UBT, por su consejos, aportaciones y disponibilidad para apoyarme en la divulgación y realización de este proyecto.

Al grupo de seminarios del Laboratorio de Biología Computacional del IIMAS, de la misma forma, les agradezco por su consejos, aportaciones y disponibilidad para apoyarme en la realización de este proyecto.

A mi madre Marcela por su amor sin límites a pesar de la distancia, por su apoyo total en todos los aspectos, por aceptarme como soy, eres una figura ejemplar para mi desarrollo como ser humano.

A mis hermanitos Juan y Nicte-há, por sus apoyos y amor incondicional, la vida con ustedes es maravillosa y especial.

A mi esposa la Dr. Irán Andira Guzmán, por ser mi motivo de superación en los aspectos personales y académicos, por su amor inquebrantable, por decidir acompañarme por el resto de nuestra vida, incentivo para ser un mejor humano cada día.

DEDICATORIAS

Para mis musas,

Océane Danasha e Irán Andira ...

PRODUCTOS ASOCIADOS

Artículo publicado

Escobar-Turriza P, Hernandez-Guerrero R, Poot-Hernández AC, Rodríguez-Vázquez K, Ramírez-Prado J, Pérez-Rueda E (**2019**). Identification of functional signatures in the metabolism of the three cellular domains of life. PLoS ONE 14(5): e0217083. https://doi.org/10.1371/journal.pone.0217083

Martinez-Liu L, Hernandez-Guerrero R, Rivera-Gomez N, Martinez-Nuñez MA, Eveline Peeters, **Escobar-Turriza P,** Perez-Rueda E. (**2021**). Comparative genomics of DNA-binding transcription factors in archaeal and bacterial organisms PLoS ONE 16(7): e0254025.https://doi.org/10.1371/journal.pone.0254025

Participación en Congresos

2019

- Congreso: 4th International Symposium onFunctional Genomics

Título del trabajo: Identification of functional signature in the metabolism of prokaryotes and eukaryotes

Modalidad: Póster

- Congreso: Escuela de Invierno 2019 (IIMAS)

Título del trabajo: Identificación de firmas funcionales en el metabolismo

Modalidad: Ponencia

2018

- Congreso: 3th International Symposium in Functional Genomics and Systems Biology

Título del trabajo: Identification of functional signatures in prokaryote and eukaryote metabolism

Modalidad: Póster

- Congreso: XVIII Congreso de Estudiantes CICY

Título del trabajo: Identificación de firmas funcionales en el metabolismo de procariotes y eucariotes

Modalidad: Póster

Divulgación

2019

- Título del trabajo: Estudio del repertorio enzimático del metabolismo mediante genómica comparativa

Tipo de participación: Seminario; Conferencia

Institución organizadora: IIMAS -Unidad Mérida

Dirigido a: Comunidad científica; Comunidad estudiantil

- Título del trabajo: Estudiando el repertorio enzimático del metabolismo por genómica comparativa

Tipo de participación: Seminario; Conferencia

Institución organizadora: Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco, A.C. (CIATEJ) -SubsedeSureste

Dirigido a: Comunidad científica; Comunidad estudiantil

- Título del trabajo: Comparando las vías metabólicas: un intento por comprender cómo hemos llegado hasta el metabolismo moderno

Tipo de participación: Artículo de divulgación

Dirigido a: Público en general

Revista: QUIU

Tipo de medio: Revista de divulgación en Línea; Internet

Liga:https://quiurevista.com/comparando-las-vias-metabolicas-un-intento-por-comprender-como-hemos-llegado-hasta-el-metabolismo-moderno/

Cursos Impartidos

Enero-Junio 2020

- Curso: Introducción a la Bioinformática

Institución: Universidad Autónoma de Yucatán (UADY)

Profesor responsable: Dr. Ernesto Pérez Rueda

ÍNDICE

LISTADO DE FIGURAS

LISTADO DE TABLAS

RESUMEN

Con el fin de identificar actividades enzimáticas comunes y específicas asociadas con el metabolismo de los dominios celulares de la vida, se evaluó la distribución, las similitudes y las variaciones del repertorio enzimático en organismos que pertenecen a *Bacterias*, *Arqueas* y *Eucariotes*. Para ello, se analizó la información metabólica de 1507 organismos no redundantes, anotados y depositados en la base de datos *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*). De esta manera evaluamos las reacciones enzimáticas, etiquetadas mediante los *EC (Enzyme Commission) numbers* , que están asociadas con cada organismo y sus respectivas vías metabólicas. A partir de esto, hemos encontrado un conjunto de cinco reacciones enzimáticas que se distribuyen ampliamente en todos los organismos de los tres dominios celulares, mediante un perfil taxonómico. Sin embargo, estas reacciones no se distribuyen a lo largo de los mapas metabólicos, lo que sugeriría que no son indispensables en los procesos metabólicos. Finalmente, descubrimos que dichas reacciones están asociadas con una diversidad de dominios estructurales. También, inferimos que estas reacciones poseen dominios ancestrales, como aquellos asociados a grupos que contienen fósforo con un grupo fosfato como aceptor o aquellos relacionados con *barrel of* r*ibulose phosphate binding*, *riosephosphate isomerase* y *D-ribose-5 domain Phosphate isomerase cap* (RpiA), entre otros. Por lo tanto, se considera que este análisis proporciona información sobre las restricciones funcionales asociadas con el repertorio de funciones enzimáticas por organismo.

ABSTRACT

In order to identify common and specific enzymatic activities associated with the metabolism of the three cell domains of life, the distribution, conservation, and variations between enzyme contents of the Bacteria, Archaea, and Eukarya organisms were evaluated. For this, the content of enzymes belonging to a particular pathway in 1507 organisms that have been annotated and deposited in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database was analyzed. In this way, we evaluated the enzymatic reactions, encoded with an EC (Enzyme Commission) number, that are associated with each organism and metabolic map. From this, we found a set of five enzymatic reactions that were widely distributed in all organisms and were considered, in this work, as universal for Bacteria, Archaea, and Eukarya using a taxonomic profile. However, these universal reactions are not widely distributed throughout metabolic maps, suggesting their dispensation to all metabolic processes. Finally, we discover that universal reactions are also associated with a diversity of structural domains; in turn, the reactions are associated with ancestral domains, such as those related to phosphorus-containing groups with a phosphate group as acceptor or those related to barrel of ribulose phosphate binding, triosephosphate isomerase and D-ribose-5 domain Phosphate isomerase cap (RpiA) , among others. Therefore, we consider that this analysis provides clues about the functional restrictions associated with the repertoire of enzymatic functions by organisms.

INTRODUCCIÓN

El metabolismo es un sistema biológico que presenta una amplia diversidad en sus rutas centrales, como por ejemplo el metabolismo central del carbono o la síntesis de nucleótidos (Poot-Hernández *et al.*, 2015; Noor *et al*., 2010). Esta diversidad está asociada a la existencia de una vasta gama de microorganismos que son capaces de sobrevivir en diferentes ambientes, como en el caso de los organismos halófilos, o aquellos que residen en hábitats de bajas temperaturas (Oren, 2008; Boetius *et al.*, 2015), o aquellos organismos identificados en los sedimentos de los océanos, asociados a la producción de metano y otros hidrocarburos (Torsvik *et al*., 2002; Kallmeyer *et al.*, 2012).

El metabolismo consiste en dos fases: la degradación y la biosíntesis. La fase degradativa o catabolismo, se define como la etapa en la que los macronutrientes son transformados en moléculas pequeñas y simples. Los procesos catabólicos se caracterizan por liberar energía, mayormente almacenadas como moléculas de ATP, NADH, NADPH y $FADH_2$. Por otro lado, muchas moléculas pequeñas y simples son requeridas para sintetizar a otras de mayor tamaño, como los oligopolisacáridos, ácidos nucleicos o proteínas. A este proceso se le conoce como anabolismo o biosíntesis. Los procesos anabólicos requieren del consumo de energía para llevar a cabo sus reacciones, refiriéndose a la transferencia de un grupo fosforilo de la molécula energética ATP, y del poder reductor de las moléculas NADH, NADPH y $FADH_2$ (Nelson y Cox, 2017).

En las últimas décadas, el metabolismo se ha estudiado desde una perspectiva de genómica comparativa, utilizando la información derivada de los proyectos de secuenciación masiva. Aunado a los avances en las ciencias computacionales, se han implementado métodos para realizar análisis comparativos entre las vías metabólicas de una misma especie o entre varias especies; tales como el trabajo de comparación del metabolismo de *E. coli*, o la comparación del metabolismo de los aminoácidos. Sin embargo, un análisis exhaustivo de la composición de las reacciones enzimáticas en organismos de todos los dominios celulares no se ha llevado a cabo. En este trabajo, se reporta un conjunto de cinco reacciones enzimáticas que están relacionados con la transferencia de grupos fosfatos y que se distribuyen en la mayoría de los organismos de los tres dominios celulares. Por otro lado, reportamos cinco actividades enzimáticas que

fungen como firmas funcionales, es decir, aquellas reacciones enzimáticas que son comunes a un conjunto de organismos, como la firma funcional 3.1.26 exclusiva de las actinobacterias, o varios grupos taxonómicos. También analizamos la asociación funcional entre las reacciones enzimáticas, donde reportamos 5 pares enzimáticos significativos que, de igual manera, están distribuidos en los tres dominios celulares. Estos resultados, nos permiten asociar dichas reacciones enzimáticas como reacciones ancestrales en la evolución de las vías metabólicas y su papel en la composición estructural de las membranas celulares sintetizando lípidos biológicos, como los fosfatidil fosfolípidos. Por el cual sugerimos que las actividades involucradas en la transferencia de moléculas energéticas se han conservado a lo largo del crecimiento metabólico y posiblemente, sean fundamentales para mantener la maquinaria celular de la vida.

CAPÍTULO I

ANTECEDENTES

1.1 Metabolismo Celular

El metabolismo consiste en dos fases: la degradación y la biosíntesis (Fig. 1.1). La fase degradativa o catabolismo, se define como la etapa en el que los macronutrientes son transformados en moléculas pequeñas y simples. Los procesos catabólicos se caracterizan por liberar energía, mayormente almacenadas como moléculas de ATP, NADH, NADPH y $FADH_2$. Por otro lado, muchas moléculas pequeñas y simples son requeridas para sintetizar a otras de mayor tamaño, como los oligopolisacáridos, ácidos nucleicos o proteínas. A este proceso se le conoce como anabolismo o biosíntesis. Los procesos anabólicos requieren del consumo de energía para llevar a cabo sus reacciones, refiriéndose a la transferencia de un grupo fosforilo de la molécula energética ATP, y del poder reductor de las moléculas NADH, NADPH y $FADH_2$ (Nelson y Cox, 2017).

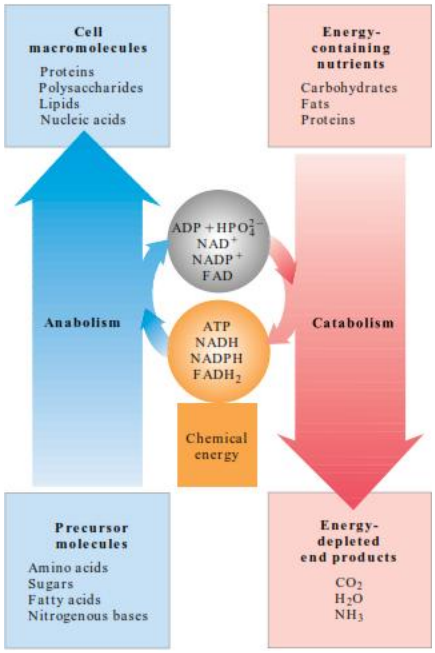

Figura 1.1 Metabolismo. Relación energética entre los procesos catabólicos y anabólicos, en el cual figuran las monedas energéticas ATP, NADH, NADPH y FADH2 (Nelson y Cox, 2017).

1.1.1 Vías metabólicas

Los circuitos multienzimáticos se definen como vías metabólicas que están constituidas por reacciones bioquímicas catalizadas por enzimas de manera sucesiva. Cada una de estas enzimas genera un cambio químico específico, sea transferencia, adición o eliminación de un grupo funcional en particular. A los productos de estos cambios se les denomina metabolitos, siendo productos intermediarios de las reacciones enzimáticas que ocurren en una vía metabólica o entre vías metabólicas (Nelson y Cox, 2017). Los metabolitos son las conexiones que unen a una vía metabólica de otra, ya que los múltiples productos finales de las rutas metabólicas pueden ser los precursores de otras vías, como se ilustra en el metabolismo del piruvato (Fig. 1.2) (Gray *et al.*, 2014).

Las vías metabólicas pueden ser reacciones llevadas en una secuencia lineal o en una ramificada, como las vías de la Glucólisis o la síntesis de los carotenoides (Marini *et al.*, 2016; Cárdenas-Conejo *et al.*, 2015). También existen vías que son cíclicas, es decir, reacciones cuyo precursor principal se regenera a partir de series de reacciones que transforman a un precursor de otra vía en un producto intermedio, como el Ciclo de Krebs (Wu y Minteer, 2015).

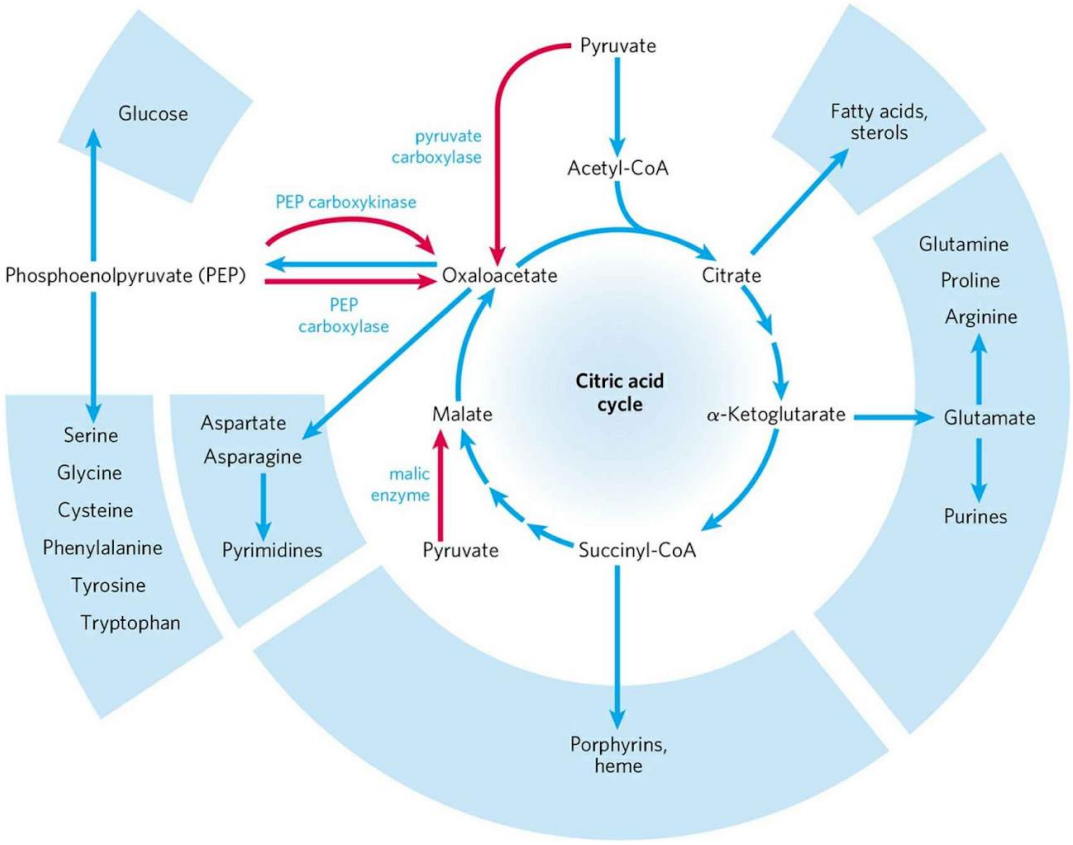

Figura 1.2 Metabolismo del piruvato. Reacciones enzimáticas conectadas mediante intermediarios metabólicos, en la cual un metabolito final es la molécula precursora para otra vía.

1.1.2 Reacción enzimática

Una reacción enzimática es una reacción química que es mediada por un catalizador de origen proteico denominado enzima. Las enzimas son macromoléculas compuestas de polímeros de aminoácidos conectados por enlaces amino. El sitio activo de las enzimas a menudo está rodeado de bolsas hidrofóbicas, lo que proporciona el poder de la especificidad de su sustrato (Singh *et al*., 2016).

1.1.2.1 Número de Clasificación Enzimática (*EC number*)

La reacción enzimática es una de las funciones biológicas proteicas mejor descritas, y se han clasificado con base en el tipo de reacción por la *Enzyme Commission Number* (Rahman *et al.* 2014), en siete clases (Tabla 1.1) (Tipton y McDonald, 2018). Cada reacción enzimática está representada por un código jerárquico de cuatro niveles. El primer nivel corresponde a siete clases diferentes según el tipo de química que se realiza (Cuesta *et al.*, 2015; Tipton y McDonald, 2018).

Tabla 1.1 Clases de reacciones enzimáticas

No. de Clase	***Clase de enzima***	***Función***
1.-.-.-	Oxidorreductasa	Catalizan reacciones de oxidación / reducción
2.-.-.-	Transferasa	Transfieren un grupo químico, por ejemplo, un residuo metilo o glicosilo
3.-.-.-	Hidrolasa	Realizan la hidrólisis de los enlaces químicos
4.-.-.-	Liasa	Rompen los enlaces químicos por otros medios que no sean la oxidación o la hidrólisis.
5.-.-.-	Isomerasa	Catalizan cambios geométricos y estructurales entre los isómeros.
6.-.-.-	Ligasa	Unen dos compuestos con hidrólisis de una molécula de nucleósido trifosfato
7.-.-.-	Translocasa	Catalizan el movimiento de iones o moléculas a través de las membranas o su separación dentro de las membranas

Estas clases de reacciones se dividen en subclases y sub-sub-clases (segundo y tercer nivel, respectivamente) con base en una variedad de criterios, como el enlace químico escindido o formado, el centro de reacción, el grupo químico transferido y el cofactor utilizado para la catálisis. El nivel final de clasificación define la especificidad del sustrato. Por ejemplo, la alanina racemasa es una isomerasa (EC 5.-.-.-), en particular una racemasa (EC 5.1.-.-) que actúa sobre el aminoácido (EC 5.1.1.-) alanina (EC 5.1.1.1) (Cuesta *et al.*, 2015).

1.2. Bases de datos biológicos

En la era Post-Genoma surge la necesidad de analizar millones de secuencias provenientes de la secuenciación de genomas completos (Kanehisa, 1997). En la actualidad existen, principalmente, cinco tipos de datos biológicos que son masivos y que se utilizan en gran medida en la investigación bioinformática: los datos de expresión génica, datos de secuencia de DNA, RNA y proteínas, datos de interacción proteína-proteína (PPI), datos metabólicos, y la ontología de genes (GO) (Kashyap *et al.*, 2015). De la mano de las nuevas tecnologías informáticas y aplicadas a la biología molecular, se han desarrollado proyectos enfocados en el desarrollo de bases de datos, con el objetivo de almacenar, organizar, recuperar, modificar y actualizar los datos (Kanehisa, 2003). En ese sentido, las bases de datos biológicas se clasifican en tres categorías: las bases de datos primarias, las bases de datos secundarias y las bases de datos especializadas.

Las bases de datos primarias son repositorios públicos que almacenan y catalogan secuencias de DNA, RNA y de proteínas (secuencias primarias y estructurales). Las bases de datos secundarias derivan de los datos referenciados de la información depositada de las primarias, y sobre ellos se llevan a cabo análisis computacionales, por ejemplo, un banco de datos de conjuntos de familias de secuencias de proteínas y/o una clasificación jerárquica de patrones de plegamiento de proteínas (Lesk, 2019). La tercera categoría corresponde a las bases de datos especializadas, que es la integración de información de las bases de datos primarias y secundarias asociadas a un organismo en particular o a un tipo de molécula determinada.

1.2.1. Bases de datos metabólicos.

Las bases de datos metabólicos nos proporcionan información de vías metabólicas referentes, con la finalidad de predecir las vías metabólicas que un organismo pueda presentar a partir de la anotación de su genoma completo. Existen bases de datos, como *KEGG* (Kanehisa *et al.,* 2018) y *MetaCyc* (Caspi *et al.,* 2018), que generan modelos de flujos metabólicos que dependen en gran medida de las reacciones metabólicas referentes. Desde un punto de vista cuantitativo, *KEGG* y *MetaCyc* son las principales bases de datos con información colectada y de buena calidad respecto a una vía metabólica y sus reacciones enzimáticas. Si comparamos a *KEGG* y *MetaCyc*, podemos visualizar la manera en cómo estructuran la información. En el 2018, *KEGG* analizó el doble de información genómica con respecto a *MetaCyc*; por otro lado, en *KEGG* se generaron un 39% más de vías metabólicas, se describieron 41% menos de reacciones enzimáticas y un 20 % más de *EC numbers* con respecto a *MetaCyc* (ver Tabla 1.2). Sin embargo, pese al contraste de información, un estudio sobre la retención de enzimas duplicadas en las rutas metabólicas, reveló resultados similares al analizar diferentes redes metabólicas provenientes de diversas fuentes biológicas (Díaz-Mejía et al., 2007).

Adicionalmente, se han descrito otras bases de datos como *Rhea*, *BiGG*, *UniPathway*, *BioPath* y *Reactome*, que estructuran la información metabólica en menor proporción; *The SEED* y *BRENDA* contienen un número comparable de reacciones, aunque el contenido metabólico de *The SEED* se deriva en gran parte de KEGG, mientras que BRENDA no incluye vías metabólicas (Altman *et al.*, 2013).

Tabla 1.2 Tabla comparativa entre las bases de datos KEGG y MetaCyc

Características	***KEGG (*Kanehisa *et al.*, 2018*)***	***MetaCyc (*Caspi *et al.*, 2018*)***
Método de Anotación	Semiautomático (curación manual)	Automática (machine learning)
Número de. de genomas	6233	3045
Número de vías	537 (*mapas*[1])	385 (*superpathways*[2])

metabólicas		
Número de reacciones enzimáticas	11324	16034
Número de EC numbers	7672	6349

*1. *Mapas*: son una integración de reacciones y vías metabólicas que se encuentran en múltiples especies. No se encuentran en su totalidad en ninguna especie.

*2.- *Superpathways*: similares a los *mapas* de KEGG. Son una integración de reacciones que comprenden a múltiples sub-vías metabólicas. A diferencia de los mapas de KEGG, la mayoría de los *superpathways* ocurren en un solo organismo.

1.2.1.1 Base de datos *Kyoto Encyclopedia of Genes and Genomes*

En 1995 se creó la *Kyoto Encyclopedia of Genes and Genomes* (*KEGG*), que funge como un recurso de referencia para la asignación de funciones biológicas a genes y proteínas asociados a un organismo (Kanehisa *et al.*, 2013; Kanehisa *et al.*, 2016). Actualmente, *KEGG* ha sufrido una expansión significativa, en los que se destacan 4 secciones principales: *PATHWAYS*, *GENES*, *COMPOUNDS* y *ENZYMES* (ver Tabla 1.3). Y no menos importante, la adición de herramientas para el análisis de datos transcriptómicos proteómicos, metabolómicos y metagenómicos, entre otros (Kanehisa *et al.*, 2016).

Tabla 1.3 Descripción de la colección de datos de *KEGG*

Fecha de actualización	2015/01	2015/04	2015/07
Número de pathways[1]	470	475	478
Número de módulos[2]	652	685	707
Número de genomas	3,495	3,712	3,936

Número de genes	15,346,261	16,400,093	17,427,876
Número de compuestos[3]	17,343	17,402	17,421
Número de reacciones[4]	9,775	9,862	9,889
Número de enzimas distintas	6,415	6,463	6,51

*1.-pathways: son representaciones gráficas de un diagrama de interacción/reacción molecular de todas las vías metabólicas presentes en múltiples especies.

*2.-módulos: son unidades funcionales (definidas manualmente) de conjuntos de genes y conjuntos de reacciones. Los módulos de vías son aquellos conjuntos de genes que caracterizan a las vías metabólicas, por ejemplo, a los complejos moleculares. Los módulos de firmas son aquellos conjuntos de genes que caracterizan rasgos fenotípicos. Los módulos de reacción son unidades funcionales de pasos consecutivos de reacciones presentes en las vías metabólicas.

*3.-compuestos: son moléculas, biopolímeros y otras sustancias químicas relevantes para los sistemas biológicos, que contiene información de sus estructuras químicas utilizada para inferir con repertorios químicos de diversas sustancias a partir de la información genómica.

*4.- reacciones: son todas las reacciones químicas (en su mayoría reacciones enzimáticas) que se presentan en los *pathways* y aquellas reacciones adicionales que solo están descritas en la Nomenclatura Enzimática.

1.2.1.1.1. Mapas metabólicos

Un mapa metabólico es la integración de 2 redes de información: la red química por el cual las moléculas son transformadas, degradadas o sintetizadas y, la red genómica que consiste en cómo las enzimas (codificadas por un genoma) están conectadas para realizar actividades catalíticas consecutivas (Fig. 1.3) (Kanehisa, 2013). Con base en lo anterior,

los datos, la información, el conocimiento y los principios se usan para mejorar la arquitectura y el contenido de la base de datos KEGG.

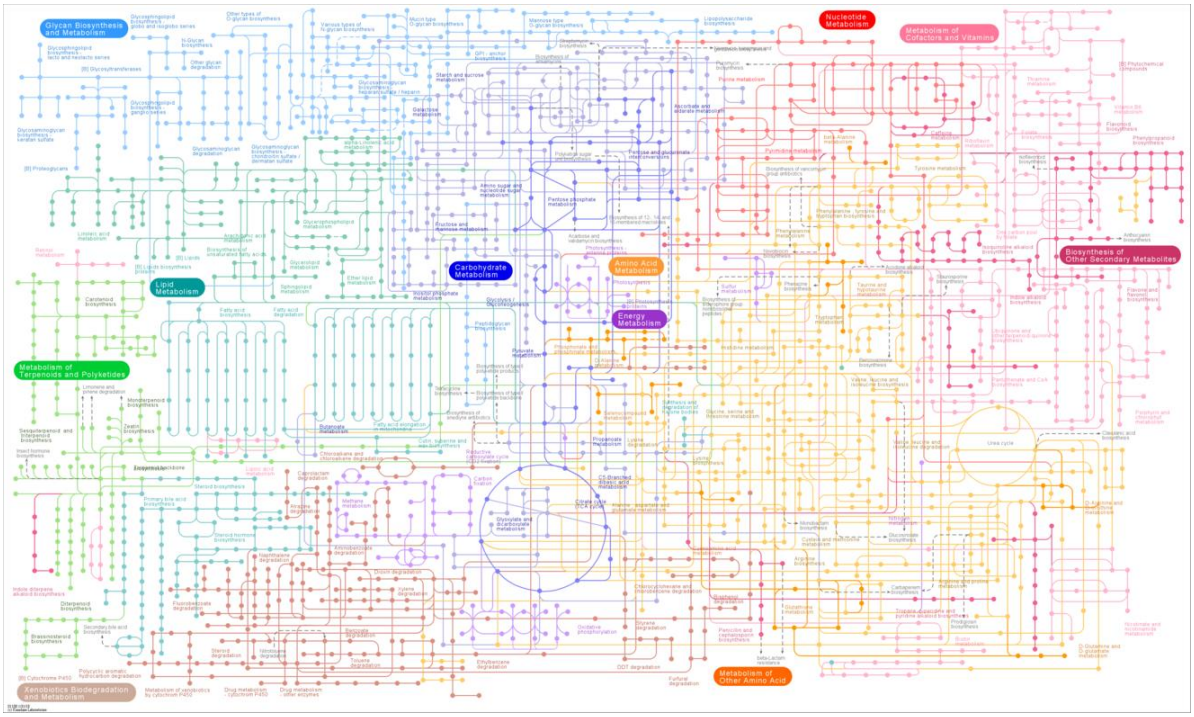

Figura 1.3 Representación gráfica de los mapas metabólicos existentes en la base de datos KEGG. Los puntos (nodos) representan a sustratos y las aristas a la reacción enzimática o *EC number*. Los colores representan al tipo de metabolismo. Por ejemplo, el Metabolismo de carbohidratos está representado por el color *Azul*; mientras que el Metabolismo energético se identifica mediante el color *Púrpura*; en color *Verde* también podemos apreciar aquéllos que pertenecen al Metabolismo de lípidos; y en color rojo a aquéllos asociados al Metabolismo de nucleótidos (más detalles en Tabla S1).

1.3. El Metabolismo moderno y la Diversificación de la Vida

1.3.1 Expansión metabólica

El metabolismo ha despertado un interés en cómo se ha originado tras la posible existencia de genomas primitivos que estaban compuestos de un puñado de genes (Scossa & Fernie, 2020). Aunque, en los inicios de la vida, el metabolismo podría haber involucrado un conjunto de pocas reacciones químicas simplificadas, y ante la necesidad selectiva, las células primitivas sufren mecanismos de replicación para su supervivencia y reproducción (Caetano-Anollés et al. 2009). En realidad, existen diferentes mecanismos moleculares que sin duda son partícipes de la expansión metabólica, es decir, de la estructuración de los genomas ancestrales y de sus vías metabólicas (Scossa & Fernie, 2020). Los mecanismos más aceptados son, la duplicación de genes, la fusión de genes y la transferencia horizontal de genes. Se ha evidenciado que en múltiples organismos la mayoría de la información genética provienen de los eventos de duplicación (Qiao *et al*. 2019). Bajo esta perspectiva, se han planteado diversas teorías hipótesis en las que destacan la retrógrada, la de Granik, la de *patchwork* y la *shell*.

1.3.1.1 Hipótesis de evolución, modelo retrógada

En 1945, Horowitz plantea que las reacciones enzimáticas de las vías metabólicas emergieron por duplicación de genes en un orden inverso al que se encuentra actualmente. en las vías actuales (Horowitz, 1945). Es decir, en la Tierra primitiva, se produjo una escasez de compuestos claves para subsistir, provocando una presión selectiva donde surge la necesidad de duplicar la información genética de una enzima clave para generar nuevas reacciones enzimáticas capaces de transformar otros compuestos existentes en el entorno para obtener al precursor en desabasto; de esta manera, se fue construyendo una vía de nuevas reacciones enzimáticas desde el producto final hacia el precursor inicial (Muto-Fujita, 2019).

1.3.1.2 Hipótesis de evolución, modelo *patchwork*

La hipótesis *patchwork* propone que las vías metabólicas se expandieron con el reclutamiento de enzimas cuyas capacidades metabólicas eran muy diversas, ya que podrían transformar a una amplia variedad de compuestos similares químicamente (Jensen,

1976). Reclutar este tipo de enzimas, les permitió que las células primitivas pudieran aumentar sus capacidades de codificación limitadas (Scossa & Fernie, 2020). Los mecanismos de duplicación génica y la subfuncionalización sustentan a que el reclutamiento de una enzima promiscua se incline hacia la especificidad de un sustrato para cumplir funciones nuevas en vías emergentes, como la enzima ligasa presente en las biosíntesis de peptidoglicano (Muto-Fujita, 2019; Díaz-Mejía *et al.*, 2009).

1.3.2 La ancestralidad de las arquitecturas proteicas como parte de un estudio sobre la diversificación de los Dominios Celulares

El mundo proteico contemporáneo nos permite recorrer el pasado a través de sus estructuras espaciales, de manera que dicha información se integra en un sistema de clasificación de pliegues arquitectónicos, sus aspectos termodinámicos y su función biológica (Caetano-Anollés *et al.*, 2009a). La Bioinformática evolutiva y estructural se ha esforzado en descifrar el comportamiento de la arquitectura proteica por medio de estudios filogenómicos sobre los dominios estructurales en los proteomas de una amplia diversidad de organismos (Caetano-Anollés *et al.*, 2018). De esta forma, se ha propuesto una línea de tiempo a partir de un censo filogenético en el cual se trazan las arquitecturas proteicas frente a un conjunto de genomas anotados como arqueas, bacterias y eucariotas (Wang, *et al.*, 2006). Basado en un método cladístico, la edad relativa de los dominios proteicos se calculó como el número de eventos de ramificación que se conservaron en un árbol reconstruido, y se proporcionó una escala relativa de 0 a 1 (Figura 1.4).

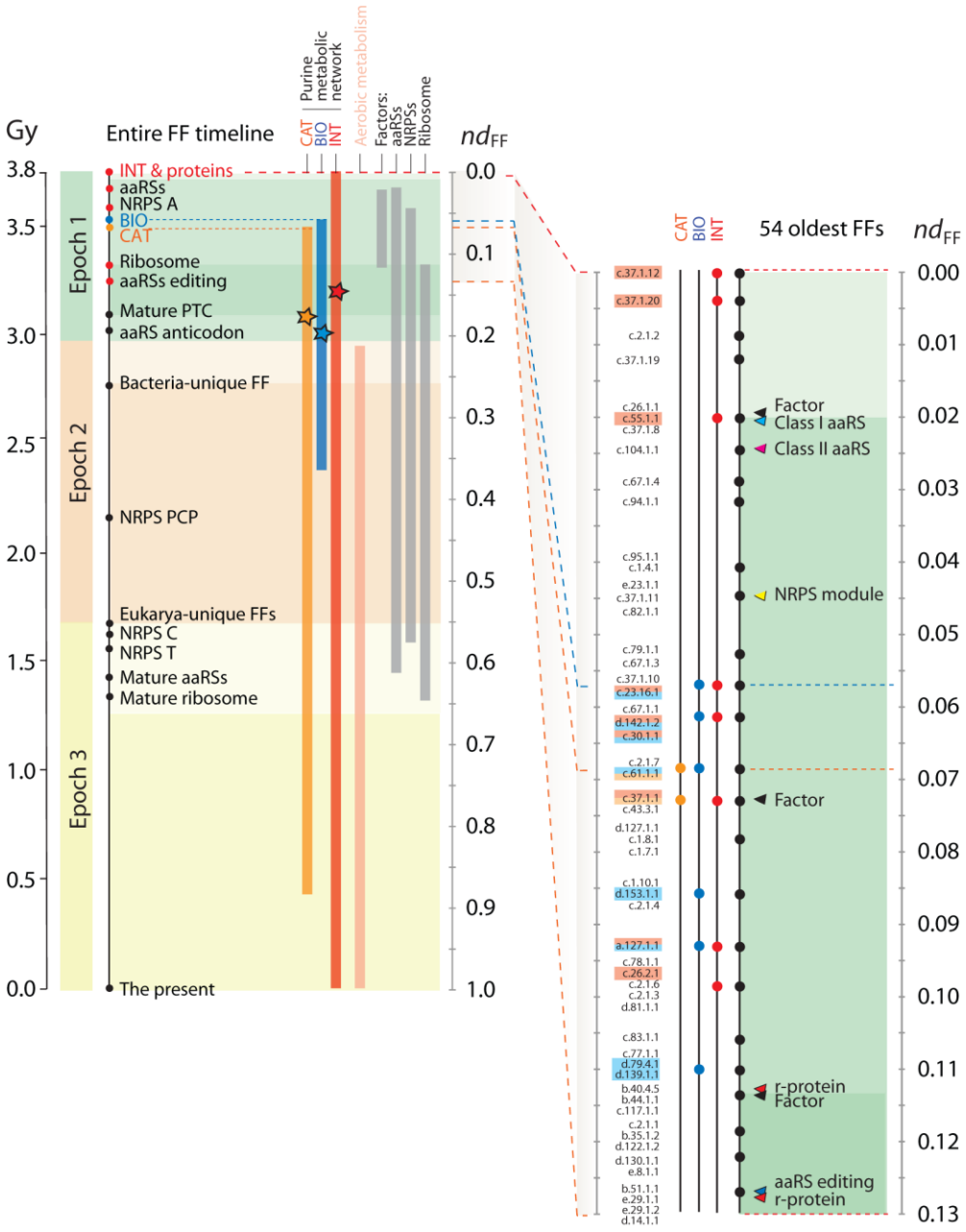

Figura 1.4 Línea de tiempo que describe la evolución de las estructuras del dominio FF y la evolución de las principales vías del metabolismo de las purinas. La línea de tiempo se derivó directamente del árbol de FF reconstruido a partir de organismos de vida libre. Las edades se dan como distancias de nodo (ndFF) y tiempo geológico (Gy). El tiempo fluye de arriba hacia abajo. Gráfico tomado de Caetano-Anollés K. y Caetano-Anollés G. (2013).

Posteriormente, se asociaron datos geológicos en el cual consideran que la línea de tiempo de la evolución de los dominios proteicos abarca ~ 3.8 mil millones de años (Kim, *et al.*, 2013) (Figura 1.4). Aunque bajo ese universo de arquitecturas proteicas, no se aprecia algún cambio con respecto a la diversidad de los organismos y se puntualiza que la evolución proteica actúa de forma independiente a los procesos de convergencia y de transferencia horizontal de genes (Wang et *al.*, 2006). Recientemente, se ha descrito la historia evolutiva de los dominios de la vida, donde las Arqueas son el dominio más antiguo y del cual se emprende la diversificación de todos los organismos (Staley y Caetano-Anollés, 2018). Ellos trazaron diversos análisis filogenómicos de las distribuciones de las familias de dominios proteicos en comparación de los proteomas de organismos de vida libre, donde se aprecia una co-evolución Arqueas-Eucariotes al compartir ocho familias de estructuras proteicas (Staley & Caetano-Anollés, 2018).

1.4. Genómica comparativa aplicada al metabolismo

En las últimas décadas, el metabolismo se ha estudiado desde una perspectiva de genómica comparativa. Es decir, que, gracias a los grandes avances tecnológicos aplicados a la biología molecular, como la secuenciación de nueva generación (*NGS*, siglas en inglés) se han podido obtener grandes volúmenes de datos asociados con el metabolismo de muchos organismos. A la par, también se han implementado diversos métodos computacionales con el objetivo de realizar análisis comparativos entre las vías metabólicas de una misma especie o entre varias especies. En la mayoría de estos métodos, el metabolismo es representado como un grafo, cuyos nodos suelen ser los metabolitos, enzimas, y los genes que codifican a estas, y las aristas, que muestra la relación que existe entre ellos (Yamada y Peer, 2009). Por una parte, la finalidad de los estudios comparativos es proporcionar información acerca de las relaciones evolutivas entre las especies y las limitaciones funcionales de estas; por ejemplo, con base en estos enfoques se ha identificado la versatilidad de la vía de la asimilación de la glucosa o glucólisis en diversas especies (Dandekar et al., 1999); y, por otro lado, permite proponer las aplicaciones biotecnológicas a partir de estos hallazgos. Por ejemplo, las reconstrucciones metabólicas permitirían generar vías con un número mínimo de reacciones para fungir como una unidad funcional ó redirigir y rediseñar a las rutas metabólicas (Papin et al., 2004; Ron et al., 2005).

Las perspectivas actuales sobre los procesos metabólicos nos indican que el metabolismo tiene la capacidad para contrarrestar fallos (tales como mutaciones que desbalancean al flujo metabólico) utilizando rutas y enzimas alternativas que proceden de diferentes vías metabólicas pero que convergen a los mismos productos (Hernández-Montes *et al.*, 2008). Estas rutas alternativas podrían correlacionarse a los cambios ambientales, pues se ha observado una divergencia en los niveles enzimáticos del metabolismo de nucleótidos relacionada a los cambios en la arquitectura celular durante la evolución (Armenta-Medina *et al.*, 2014). En este contexto, un análisis comparativo del metabolismo en las *Proteobacterias* de la división *Gamma* evidenció que en diferentes mapas metabólicos se conserva un alto contenido del repertorio enzimático similar, como en la biosíntesis de ácidos grasos y lisina, así como también en las vías metabólicas del metabolismo de nucleótidos, lo cual refuerza el modelo *patchwork* en la evolución del metabolismo, ya que probablemente pueda ocurrir una transferencia de actividades enzimáticas en diferentes rutas metabólicas de las *Gammaproteobacterias* (Poot-Hernández *et al.*, 2015).

Uno de los métodos de comparación de vías metabólicas es el algoritmo de alineamientos de secuencias y que que se basa en la similitud entre las reacciones enzimáticas, que son clasificadas jerárquicamente mediante un *Enzyme commission number* (*EC number*) (Tohsato *et al.*, 2000). Posteriormente, el uso de algoritmos genéticos para realizar alineamientos múltiples en el repertorio metabólico de la bacteria *E. coli* fue propuesto; así como algoritmos de programación dinámica para realizar alineamientos en pares entre las vías metabólicas de las *Gammaproteobacterias* (Ortegón Cano *et al.*, 2015; Poot Hernández *et al.*, 2015). Con base en estos estudios, se identificó que las *Gammaproteobacterias* tienen un contenido similar a nivel de reacciones enzimáticas al interior de un mapa metabólico y entre mapas metabólicos, con múltiples eventos de transferencia de actividades enzimáticas entre las diferentes vías metabólicas (Ortegón Cano *et al.*, 2015; Poot Hernández *et al.*, 2015). La particularidad de estos estudios, es que a partir de la información de la base de datos de KEGG, los mapas metabólicos se transformaron en secuencias lineales de reacciones enzimáticas. Posteriormente, dichas secuencias de pasos pueden ser comparadas, ya que cada paso en la secuencia representa a un *EC number*. En este contexto, actualmente se está intentando dilucidar si la observación que presentan las Gammaproteobacterias también puede ser generalizada

a todos los organismos, con el objetivo de comprender cómo el metabolismo ha llegado a ser lo que es en la actualidad.

JUSTIFICACIÓN

Entender los orígenes y la expansión del metabolismo, sigue siendo una pregunta abierta. Sin embargo, los grandes avances tecnológicos nos han permitido obtener datos biológicos masivos asociados a las vías metabólicas de diversos organismos, lo que nos permite estudiar al metabolismo desde una perspectiva de genómica comparativa. Por ello, se han implementado diversos métodos computacionales con el objetivo de realizar análisis comparativos entre las vías metabólicas de una misma especie o entre varias especies. Y es que a partir de la información de la base de datos de KEGG se ha observado que en la bacteria *E. coli* y otras *Gammaproteobacterias*, a nivel funcional, las vías metabólicas poseen una alta similitud de reacciones enzimáticas al interior de un mapa metabólico y entre mapas metabólicos. Sin embargo, en los trabajos anteriores no se realizaron estudios de abundancia genómica a nivel de reacciones enzimáticas con respecto a las especies analizadas. En este trabajo evaluamos el repertorio enzimático de organismos que pertenecen a los tres dominios celulares: *Arqueas*, *Bacterias* y *Eucariotes*, para detectar qué tipo de reacciones enzimáticas trascienden en el metabolismo contemporáneo.

HIPÓTESIS

El estudio del repertorio enzimático permite identificar reacciones enzimáticas comunes en los organismos *Arqueas*, *Bacterias* y *Eucariotes*, así como aquellas que son específicas en un grupo taxonómico determinado. Por consiguiente, las reacciones enzimáticas con mayor presencia genómica y distribución taxonómica son antiguas y necesarias para realizar los procesos metabólicos.

OBJETIVO GENERAL

Identificar a las reacciones enzimáticas abundantes, ampliamente distribuidas y antiguas del repertorio enzimático de los organismos de los tres dominios celulares: Bacterias, Arqueas y Eucariotes.

OBJETIVOS ESPECÍFICOS

I. Determinar la abundancia y distribución de las actividades enzimáticas en los organismos de los tres dominios celulares.

II. Identificar reacciones enzimáticas comunes y específicas a los organismos analizados.

III. Determinar la diversidad y antigüedad de dominios proteicos asociados a las reacciones enzimáticas.

ESTRATEGIA EXPERIMENTAL

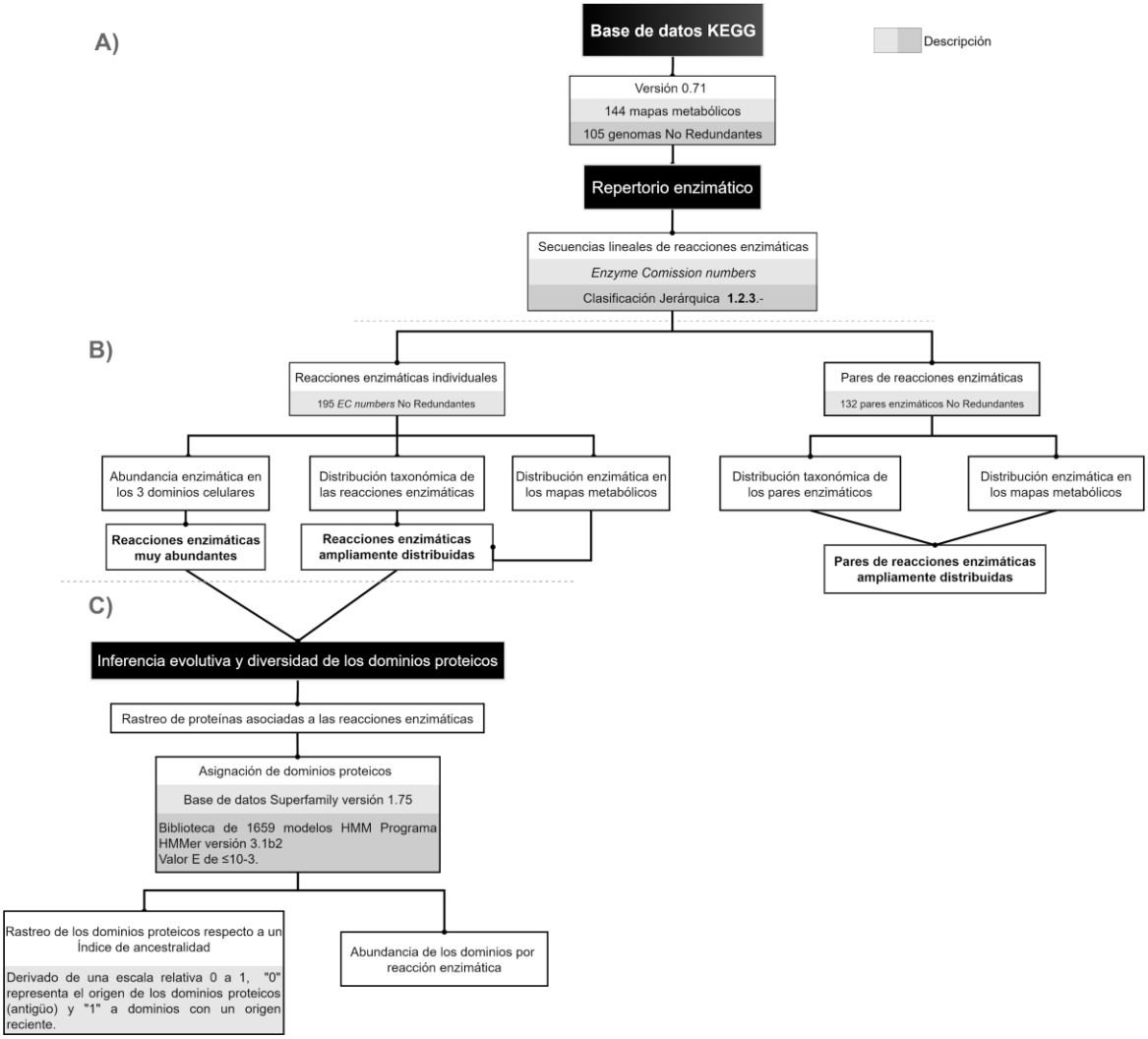

Figura 1.5 Estrategia general experimental. Esquema resumido de la estrategia experimental para abordar la hipótesis de la tesis. A) Obtención de los datos biológicos asociados a mapas metabólicos, que se estructuraron como secuencias lineales de reacciones enzimáticas. B) Para determinar la abundancia y distribución de las actividades enzimáticas, se trazaron dos caminos: primero se evaluó a las reacciones enzimáticas de forma individual, y segundo mediante pares de reacciones enzimáticas. C) De los resultados obtenidos, procedemos a comprobar si las reacciones enzimáticas abundantes y/o distribuidas, son antiguas. Para ello, los dominios proteicos pertenecientes a proteínas asociadas de cada reacción enzimática fueron rastreados en un índice de ancestralidad (Wang, *et al.,* 2006).

CAPÍTULO II

Título: *Identification of functional signatures in the metabolism of the three cellular domains of life*

Autores: Pedro Escobar-Turriza, Rafael Hernandez-Guerrero, Augusto Cesar Poot-Hernández, Katya Rodríguez-Vázquez, Jorge Ramírez-Prado, Ernesto Pérez-Rueda

Estado: Publicado

Publicado: 28 de Mayo del 2019

Cita: Escobar-Turriza P, Hernandez-Guerrero R, Poot-Hernández AC, Rodríguez-Vázquez K, Ramírez-Prado J, Pérez-Rueda E (2019) Identification of functional signatures in the metabolism of the three cellular domains of life. PLoS ONE 14(5): e0217083. https://doi.org/10.1371/journal.pone.0217083

IDENTIFICACIÓN DE FIRMAS FUNCIONALES EN LOS GENOMAS DE PROCARIOTES Y EUCARIOTES

2.1. INTRODUCCIÓN

Actualmente, las bases de datos KEGG y MetaCyc organizan los datos metabólicos para contribuir en la comprensión de los procesos de adaptación de la vida celular, la diversidad de la organización celular y la complejidad del mundo de la vida (Okuda *et al.,* 2008; Caspi *et al.*, 2018; Caetano-Anolles *et al.*, 2018). El análisis comparativo del metabolismo ha proporcionado información sobre la identificación del reclutamiento enzimático y los eventos de duplicación génica. Por ejemplo, se ha identificado que las rutas metabólicas presentan una alta retención de enzimas duplicadas dentro de los módulos funcionales,como en el caso de 4 enzimas homólogas *EC number* 6.3.-.- (ligasas carbón-nitrógeno) que catalizan de manera consecutivas la vía de la síntesis de peptidoglicanos (Light *et al.*, 2005, Díaz-Mejía *et al.*, 2007, Hernández-Montes *et al.*, 2008, Armenta-Medina *et al.*, 2011). En este trabajo, evaluamos cómo se distribuyen los pares de reacciones enzimáticas individuales y consecutivas (mediante el uso de los números de la Comisión Enzimática (*EC numbers*) a lo largo del metabolismo de los tres dominios de la vida, Bacterias, Arqueas y Eucariotes, y cómo esta distribución ha influido en las vías metabólicas en su forma actual. Con este fin, se evaluó, en términos de su composición enzimática, la información de los mapas metabólicos de 1507 organismos no redundantes depositados en la base de datos KEGG. Adicionalmente, se evalúo, la composición de los dominios estructurales a partir de sus asignaciones con la base de datos Superfamily, permitiendo identificar reacciones mayormente distribuidas que están asociadas con dominios “evolutivamente antiguos” , como los relacionados con los grupos que contienen un grupo de fosfato como aceptor o los relacionados al *ribulose-phosphate binding barrel*, *triosephosphate isomerase*, and *D-ribose-5-phosphate isomerase (RpiA) lid domain*, entre otros. Por lo tanto, consideramos que este análisis proporciona pistas sobre las restricciones funcionales asociadas con el repertorio de funciones enzimáticas en los tres dominios celulares.

2.2. MATERIALES Y MÉTODOS

2.2.1 Repertorio enzimático

A partir de la base de datos KEGG versión 0.71, se obtuvo información de 144 mapas metabólicos asociados a 1507 genomas de bacterias, arqueas y eucariotes , extraídos y depositados en un archivo *Structured Query Language* (`.*sql*`). Cada reacción enzimática perteneciente a los mapas metabólicos se identificó mediante los primeros tres niveles del *Enzyme Commission Number* (EC number) (TablaS1). En ese contexto, la información se estructuró de acuerdo con el trabajo anterior de Poot-Hernández y colaboradores (2015), generando cadenas o conjuntos de pasos enzimáticos consecutivos (*ESS*, siglas en inglés), en donde cada *EC number* representa a un paso enzimático. Para eliminar la redundancia asociada a las *ESS,* se aplicaron dos filtros: a) si dos *ESS* de diferentes organismos pero de un mismo mapa metabólico eran idénticos, entonces solo se consideró a uno de ellos; y b) si dos secuencias idénticas del mismo mapa metabólico y organismo tenían diferentes longitudes, sólo se consideró la secuencia más larga, quedando solamente un conjunto de *ESS* representativas o no redundantes totales (*nrESS*). A partir de estos *nrESS*, se obtuvieron 195 EC numbers individuales y 3151 posibles pares consecutivos de reacciones enzimáticas.

2.2.2 Distribución de las reacciones enzimáticas

Para determinar la distribución de las reacciones enzimáticas en los organismos analizados, se rastrearon 195 diferentes *EC numbers,* de los cuales solo se consideraron los primeros tres niveles de clasificación, en 105 genomas de arqueas, 1264 genomas bacterianos y 138 de eucariotes. La tasa de ocurrencia de cada *EC number* por organismo y por división taxonómica se calculó, considerando la presencia (un valor de 1) y la ausencia (valor de 0), utilizando la siguiente fórmula:

$$RA = \frac{Ni}{ODiv}$$

Donde,

RA= Abundancia relativa;

i = 1… n es una división taxonómica;

N = Ocurrencia total de cada *EC number* por división taxonómica;

ODiv = Total de organismos por división taxonómica.

En esta normalización se consideraron 50 divisiones taxonómicas a nivel *phylum*. Para b**acterias**: *Acidobacteria, Actinobacteria, Alphaproteobacteria, Aquificae, Bacteroidetes, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, Epsilonproteobacteria,* Otras *Proteobacterias*, *Chlamydiae, Chlorobi, Chloroflexi, Chrysiogenetes, Cyanobacteria, Deferribacteres, Deinococcus-Thermus, Dictyoglomi, Elusimicrobia, Fibrobacteres, Firmicutes-Bacilli, Firmicutes-Clostridia, Firmicutes-Others, Fusobacteria, Gemmatimonadetes, Nitrospirae, Planctomycetes, Spirochaetes, Synergistetes, Tenericutes, Thermotogae,* Unclassified *Terrabacteria* group, *Verrucomicrobia*; **Arqueas**: *Crenarchaeota*, *Euryarchaeota*, *Korarchaeota*, *Nanoarchaeota* and *Thaumarchaeota*; y **Eucariotes**: *Alveolata*, *Amoebozoa*, *Choanoflagellida*, *Diplomonadida*, *Euglenozoa*, Fungi, *Heterolobosea*, *Metazoa*, *Parabasalia*, *Rhodophyta*, *Stramenopiles* y *Viridiplantae*.

Para finalizar, la Abundancia relativa (RA), que representa a los *EC numbers* por cada división, se evalúo con un método de agrupamiento jerárquico (HCA) utilizando un algoritmo de enlace completo con la correlación de Pearson como una medida de similitud, mediante el programa *Mev4* (Saeed *et al.,* 2003).

2.2.3 Distribución de los EC numbers por mapa metabólico

La distribución de las reacciones enzimáticas individuales y pares se realizó en los 144 mapas metabólicos depositados en la base de datos KEGG. Se construyó una matriz de presencia y ausencia de los *EC numbers*, y se calculó su distribución. Dicho cálculo, se basa en la tasa de aparición de cada *EC number* por mapa metabólico, en función de su presencia (un valor de 1) o ausencia (valor de 0).

2.2.4 Asignación de dominios a *EC numbers*

Cada reacción enzimática se asoció a su proteína, así como su dominio estructural por medio de las asignaciones de la base de datos Superfamily versión 1.75 (Wilson *et al.*, 2009). Para ello, las proteínas de los 1,507 genomas se analizaron con una biblioteca de 1659 modelos HMM de Superfamily mediante el programa HMMer versión 3.1b2 (Finn *et al*., 2011), con un valor E de $\leq 10^{-3}$.

2.2.5 Inferencia evolutiva recurriendo a un índice de ancestralidad

Cada dominio proteico se asoció a un índice de ancestralidad propuesto por Wang *et al.* (2009). El índice que representa una línea de tiempo, va de una escala relativa 0 a 1, donde "0" representa el origen de los dominios (antigüo) y "1" representa a los dominios con un origen reciente (Caetano-Anollés K. y Caetano-Anollés G., 2013). De esta forma, el valor de ancestralidad se define por el nivel de conservación de los componentes estructurales que conforman a los dominios proteicos, obtenido a través de un censo filogenómico reconstruido a partir de organismos de vida libre (Wang *et al.*, 2009).

2.2.6 Procesamientos y análisis de los datos metabólicos

Los datos metabólicos y los correspondientes análisis se desarrollaron en *scripts ad hoc* de archivos *Jupyter notebook* (*.ipynb*), *python3.6* (*.py*) y *bash* (*.sh*). Los scripts se generaron y ejecutaron a través de una *terminal linux*. Toda la información está resguardada en un disco duro externo.

2.3. RESULTADOS Y DISCUSIÓN

2.3.1 Abundancia de las reacciones enzimáticas en los tres dominios celulares.

Las enzimas relacionadas con cada reacción metabólica en los 1507 organismos se obtuvieron de los mapas metabólicos depositados en la base de datos KEGG y se representaron mediante el uso de los tres primeros niveles de la clasificación de la Comisión de Enzimas (*EC numbers*) para describir su tipo general de reacción química (Klein *et al.*, 2012). Del total de las reacciones enzimáticas, 43.87% están anotadas como transferasas (EC: 2.-.-), el 21.93% como oxidorreductasas (EC: 1.-.-), el 17.22% como liasas (EC: 4.-.-),

13.44% como hidrolasas (EC: 3.-.-), 12.75% como ligasas (EC: 6.-.-), 8.32% como isomerasas (EC: 5.-.-), y finalmente, 0.10% como translocasas (EC: 7.-.-) (Tabla S1). Esta distribución sugiere que las reacciones de oxidorreducción son muy abundantes en el metabolismo, probablemente debido a que los procesos metabólicos pueden verse como electrones en movimiento entre las moléculas, y que a menudo capturan parte de la energía liberada cuando los electrones pasan de estados de alta energía a estados de baja energía, como ocurre en la glucólisis o en la respiración (González y Quiñones, 2000).

Así, para determinar la abundancia de reacciones enzimáticas específicas, 195 *EC numbers* diferentes (considerando los tres niveles de información) se rastrearon en todos los genomas divididos en sus respectivos dominios celulares. Para este fin, los valores superiores en la intersección entre una frecuencia relativa y una frecuencia relativa acumulativa se consideraron como umbral, para identificar a los *EC numbers* más abundantes. A partir de este análisis, encontramos que 15 *EC numbers* altamente abundantes y que representan el 55.2% del total de *EC numbers* de las arqueas (Figura 2.1, TablaS2). En contraste, 14 *EC numbers* representan el 49.2% de las bacterias, y 13 *EC numbers* representan el 44.6% del total de los *EC numbers* de eucariotes (Figura 2.2, Tabla S3; Figura 2.3, Tabla S4).

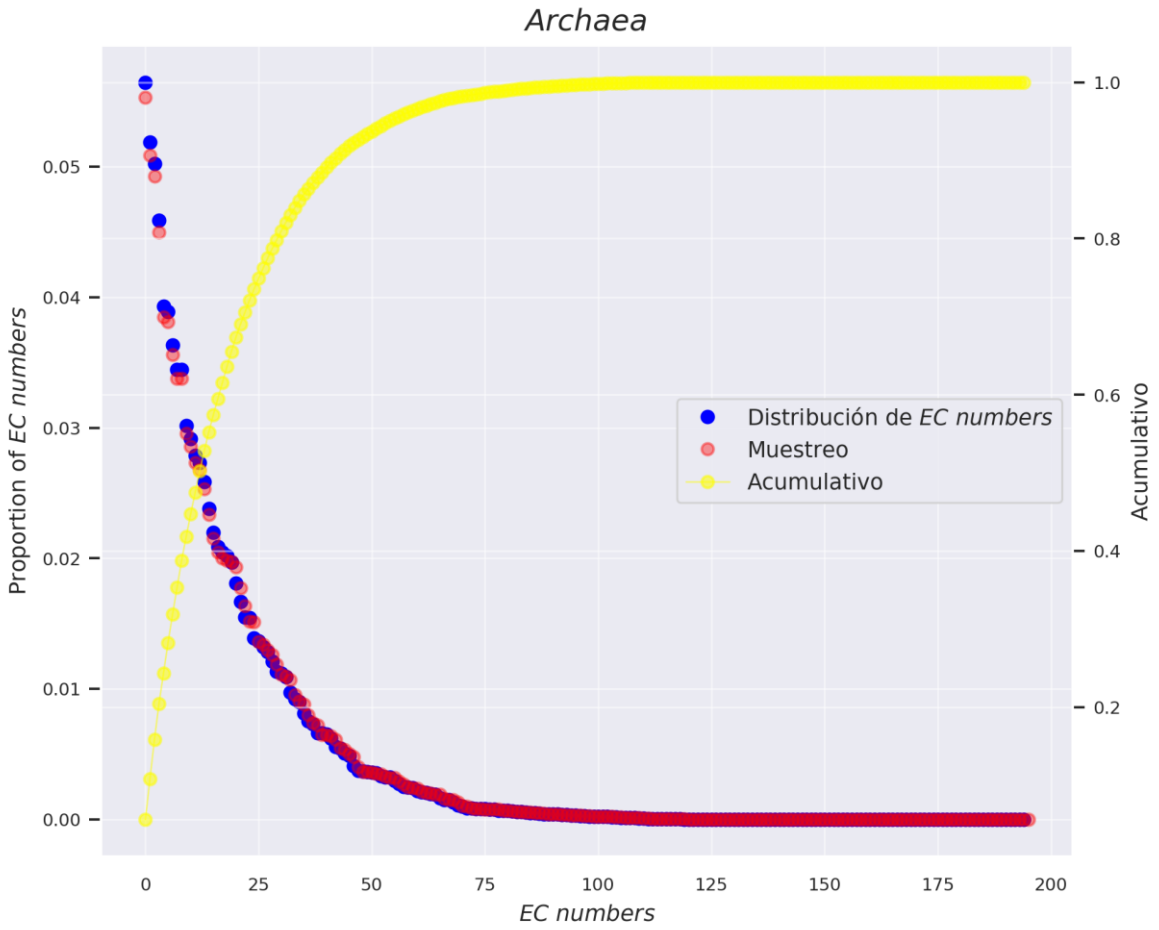

Figura 2.1 Abundancia de los EC number en Arqueas. En el eje Y se indica la proporción de *EC numbers* en el Dominio celular; En el eje X se indican los *EC numbers*. Cada punto corresponde a un *EC number*. La intersección con el gráfico amarillo indica la parte superior de lo *EC numbers* más abundantes y su porcentaje. El eje secundario indica la proporción acumulada de los *EC numbers*.

A partir de estos *EC numbers* abundantes por dominio celular, ocho actividades enzimáticas (EC 1.1.1, 2.4.2, 2.5.1, 2.6.1, 2.7.1, 2.7.7, 4.1.1 y 4.2.1) fueron también identificadas altamente abundantes en el tres dominios celulares (Figura 2.4). Es decir, son abundantes en todos los dominios celulares. Estos grupos se asocian principalmente a las transferasas (Figura. 2.4). Se identificó un *EC number* como abundante en *Arqueas* y *Bacterias* (6.3.4) pero no en *Eucariotes*; cuatro *EC numbers* (1.2.1, 2.3.1, 3.1.3 y 3.5.1) son abundantes en *Bacterias* y *Eucariotes* pero no en *Arqueas*; una ligasa (6.3.2) fue identificada como

abundante en bacterias pero no en arqueas y eucariotas; un *EC number* (2.4.1) fue abundante en *Eucariotes*; y finalmente, seis actividades (1.2.7, 2.7.4, 4.1.2, 4.3.2, 5.3.1 y 6.3.5) fueron identificadas como altamente abundantes solamente en Arqueas (Fig. 2.4).

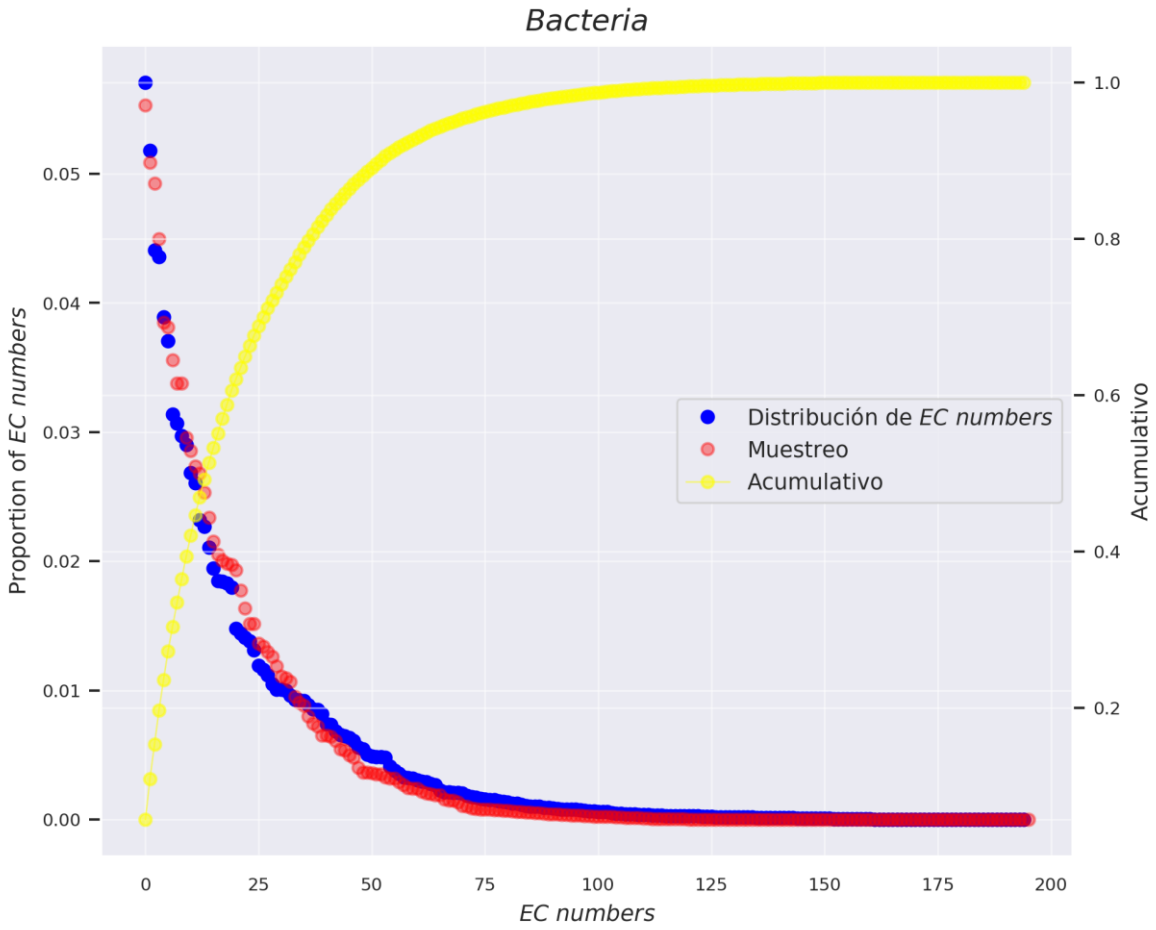

Figura 2.2 Abundancia de los EC number en Bacterias. En el eje Y se indica la proporción de *EC numbers* en el Dominio celular; En el eje X se indican los *EC numbers*. Cada punto corresponde a un *EC number*. La intersección con el gráfico amarillo indica la parte superior de lo *EC numbers* más abundantes y su porcentaje. El eje secundario indica la proporción acumulada de los *EC numbers*.

Con el fin de excluir un sesgo como consecuencia de la sobre representación de genomas, es decir, asociado a un mayor número de genomas bacterianos, que arqueales o eucariotas en los resultados previamente descritos, se realizó un análisis que consideraba muestreos aleatorios de organismos para los tres dominios celulares. En el proceso, seleccionamos

aleatoriamente 100 genomas por dominio 1000 veces, obtuvimos el promedio de cada uno y comparamos el resultado con la distribución original (considerando el conjunto completo de genomas).A partir de estos análisis, identificamos una consistencia entre el muestreo y los datos observados, lo que sugiere que nuestros resultados son lo suficientemente sólidos y confirman que 15 actividades enzimáticas son abundantes en Arqueas, 14 en Bacterias y 13 en Eucariotes, es decir, también los encontramos cuando consideramos el conjunto de datos completo (Tabla S2, Tabla S3, Tabla S4).

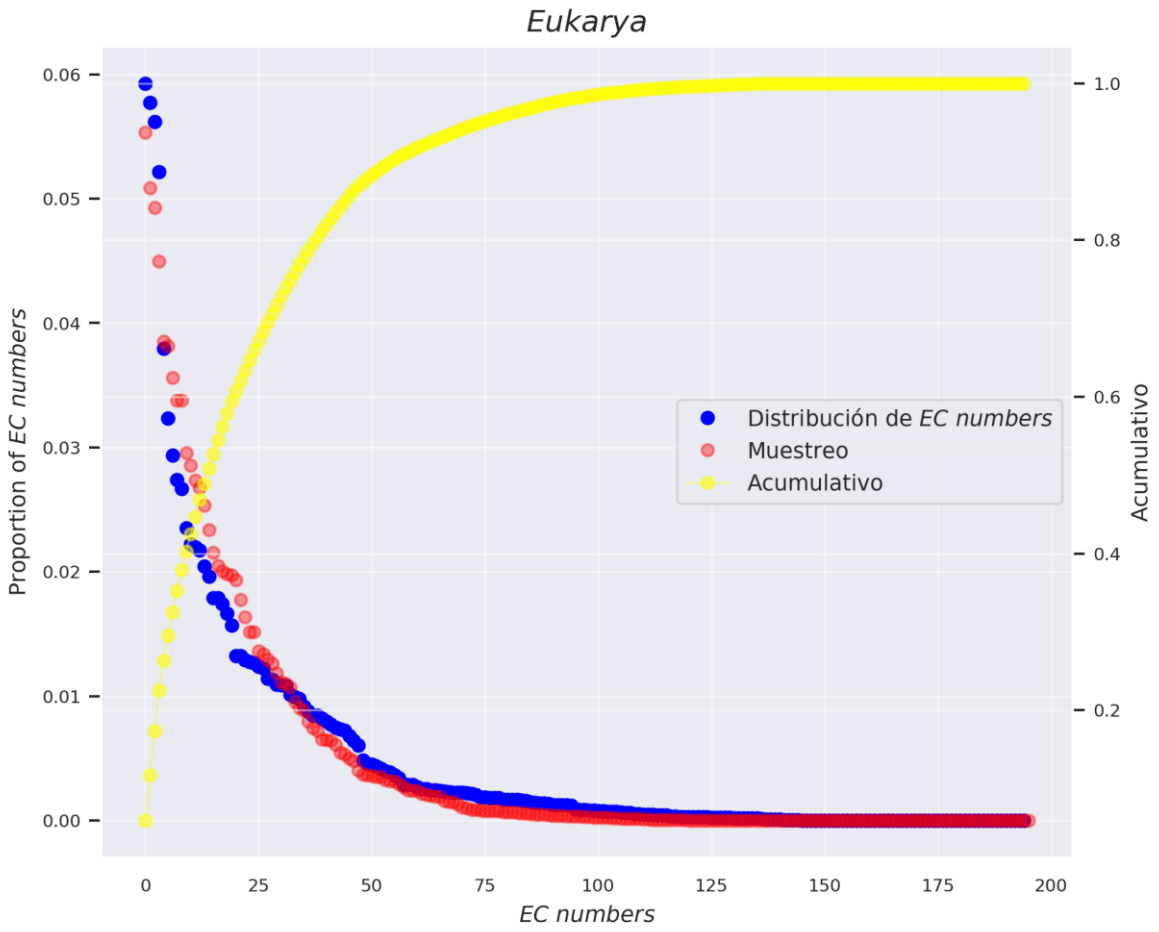

Figura 2.3 Abundancia de los EC number en Eucariotes. En el eje Y se indica la proporción de *EC numbers* en el Dominio celular; En el eje X se indican los *EC numbers*. Cada punto corresponde a un *EC number*. La intersección con el gráfico amarillo indica la parte superior de lo *EC numbers* más abundantes y su porcentaje. El eje secundario indica la proporción acumulada de los *EC numbers*.

Una de las actividades enzimáticas más recurrentes identificadas en todos los organismos correspondió a las transferasas de los grupos que contienen fósforo (2.7.-), en particular, las nucleotidil fosfotransferasas (2.7.7) involucradas en la transferencia de acilo, glicosilo, amino y fosfato (incluye difosfato, residuos de nucleotidilo, y otros). En contraste, las fosfotransferasas (2.7.4) fueron abundantes en Arqueas; dichas enzimas están involucradas en la adición de fosfato a las moléculas de UMP y CMP, entre otras moléculas. Este resultado concuerda con las simulaciones de las redes metabólicas donde, se encontró que las actividades de transferasa estaban asociadas con nuevas vías metabólicas, en particular, con enzimas multifuncionales como consecuencia de la dependencia hacia el metabolito donador o aceptor (Pfeiffer *et al.*, 2005; Caetano-Anolles *et al*., 2009).

En resumen, hemos identificado ocho reacciones enzimáticas como las más abundantes en todos los organismos analizados en este trabajo, sugiriendo un conjunto recurrente de funciones utilizadas en todos los organismos, probablemente como consecuencia de duplicación y reclutamiento de eventos en varias ocasiones a lo largo de la evolución para abastecer a las vías metabólicas en todos los organismos (Figura S2.1).

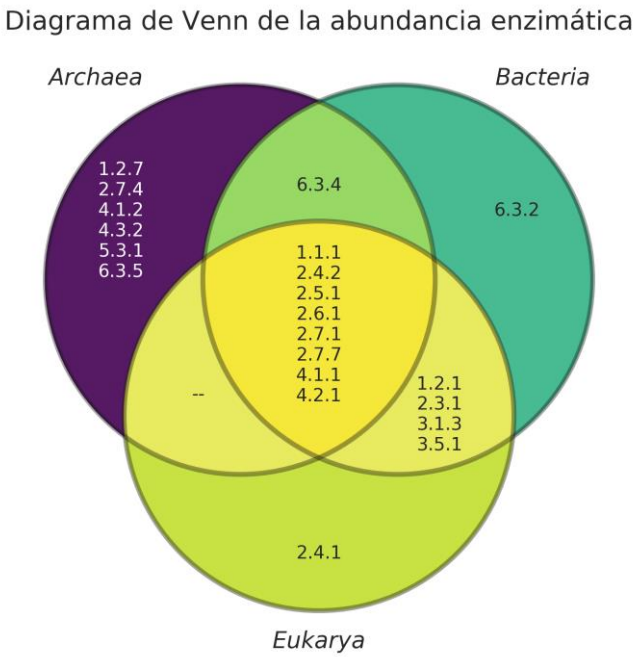

Figura 2.4 Reacciones enzimáticas identificadas como abundantes en Arqueas, Bacterias y Eucariotes. Diagrama de Venn de la abundancia enzimática, que revela los *EC numbers* 1.1.1, 2.4.2, 2.5.1, 2.6.1, 2.7.1, 2.7.7, 4.1.1 y 4.2.1 como abundantes en los tres dominios celulares. Se identificó el *EC number* 6.3.4 como abundante en arqueas y bacterias, pero no en eucariotes; los EC numbers 1.2.1, 2.3.1, 3.1.3 y 3.5.1 son abundantes en bacterias y eucariotes pero no en arqueas; el *EC number* 6.3.2 fue identificado como abundante en bacterias pero no en arqueas y eucariotas; el *EC number* 2.4.1 fue abundante en eucariotes; y finalmente, los *EC numbers* 1.2.7, 2.7.4, 4.1.2, 4.3.2, 5.3.1 y 6.3.5 fueron identificadas como altamente abundantes en arqueas.

2.3.2 Distribución de los *EC numbers* en todos los organismos.

Se determinó la distribución de los EC numbers en los tres dominios celulares mediante un agrupamiento jerárquico (HCA) (Figura S2.2). La pregunta a responder es si los EC numbers más abundantes son también los más ampliamente distribuidos en todos los organismos, lo que sugeriría una distribución universal (Tabla S5 [archivo: ecdivtax.txt]). De acuerdo con esta distribución, cinco reacciones enzimáticas (tres transferasas, 2.7.4, 2.7.7 y 2.7.1; y dos isomerasas, 5.3.1 y 5.4.2) se agruparon y se encontraron distribuidas en todos los organismos con una RA ≥ 0,95 (Figura 2.5), lo que sugiere una actividad catalítica ancestral (Tabla 2.1). De estas, las reacciones 2.7.7 y 2.7.1 se identificaron como altamente abundantes en todos los organismos como se describió anteriormente, mientras que las 2.7.4 se identificaron como abundantes en las eucariotas. Las dos reacciones catalíticas de isomerasa, 5.3.1 y 5.4.2 no se identificaron como abundantes, lo que sugiere que las reacciones catalíticas ubicuas no son necesariamente abundantes en todos los organismos.

En contraparte, se encontraron diversos *EC numbers* con menores proporciones en diversas divisiones celulares, como *Chlamydia* y *Tenericutes* (Bacterias), *Nanoarchaeum* (Arqueas), y *Parabasalia* y *Diplomonadida* (Eucariotes). Esta disminución en la proporción de reacciones enzimáticas se correlaciona con los pocos mapas metabólicos asociados. Probablemente porque estos organismos están asociados a entornos específicos y

restringidos, como *Nanoarchaeum equitans,* un endosimbionte de *Ignococcus* sp. Seis actividades enzimáticas (3.1.26, 5.4.4, 4.2.99, 1.16.1, 5.1.2 y 1.3.7) se presentan en *Actinobacteria, Cyanobacteria, Gammaproteobacteria* y *Viridiplantae*. Por un lado, una actividad enzimática 3.1.26, definida como una endoribunucleasa que produce 5'-fosfo monoésteres, se aprecia sólo en las actinobacterias. Se han reportado que las endoribunucleasas se concentran en mayor proporción en las actinobacterias, como en las especies *Frankia* y *Salinispora* que contienen enzimas descritas como RNasa *J* y la RNasa *Y* (Even *et al.*, 2005; Shahbabian *et al*., 2009). También, una actividad enzimática 1.3.7, definida como una oxidorreductasa que actúa sobre el grupo de donantes CH-CH teniendo una proteína de hierro-azufre como aceptor, es exclusiva de las cianobacterias. Se ha reportado que algunas enzima atípica ficocianobilina:ferredoxina oxidoreductasa (EC 1.3.7.5) de la familia de la bilina reductasa dependiente de ferredoxina, pues cataliza las transferencias directas de electrones sin iones metálicos ni cofactores orgánicos, participan en la biosíntesis de ficocianobilina, que es el pigmento precursor de los cromóforos de fitocromo y ficobiliproteína exclusivas de las cianobacterias (Tu *et al.*, 2006). La actividad enzimática 5.4.4, es una isomerasa que realiza reacciones transferencias intramoleculares de grupos hidroxi, se puede apreciar en tres grupos taxonómicos: *Cyanobacteria, Gammaproteobacteria* y *Viridiplantae*. La actividad enzimática 4.2.99, una liasa de carbono-oxígeno, se presenta en las cianobacterias y las gammaproteobacterias. Por último, un conjunto de cuatro reacciones enzimáticas (5.4.4, 4.2.99, 1.16.1 y 5.1.2) se atribuyen a las gammaproteobacterias, donde la actividad enzimática 5.1.2 es una isomerasa cuya función es una racemasa y/o epimerasa que actúa sobre hidroxiácidos y derivados.

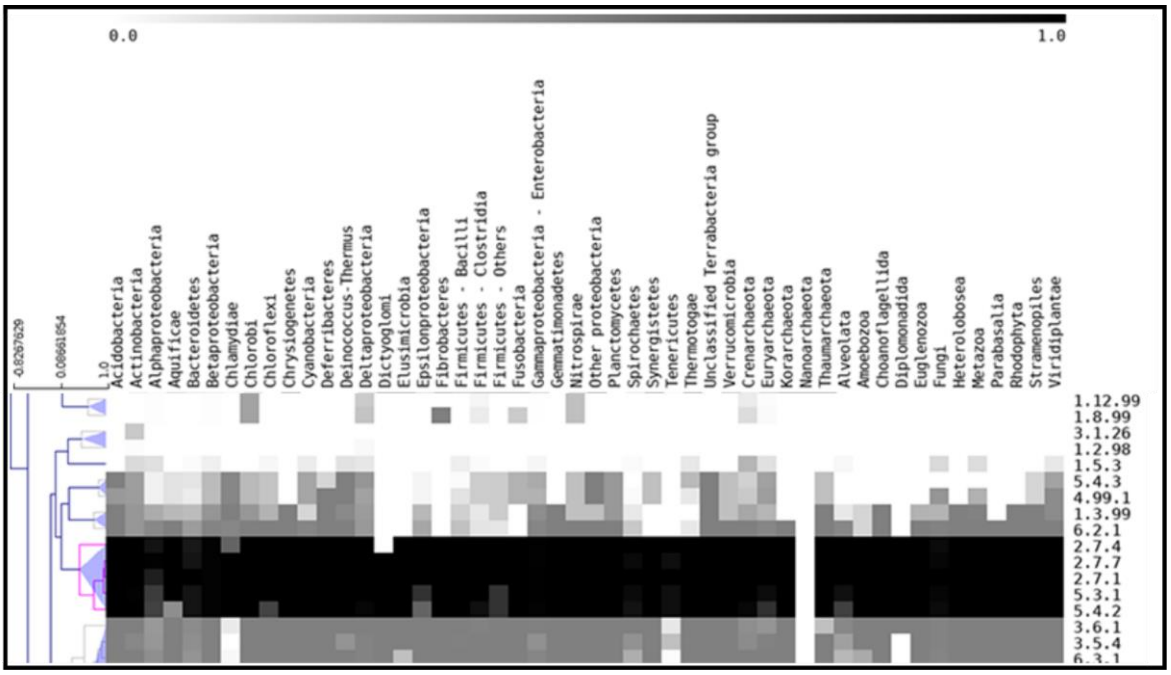

Figura 2.5 Análisis de agrupamiento de los *EC numbers* que muestra la presencia de un conjunto de actividades enzimáticas en todos los organismos. Un grupo de cinco *EC numbers* (2.7.4, 2.7.7, 2.7.1, 5.3.1 y 5.4.2) se distribuyen ampliamente en 50 divisiones taxonómicas con un RA ≥ 0.95 (ramas en rosa y resaltadas en negro). Los *EC numbers* se agruparon con un HCA utilizando la correlación al cuadrado de Pearson como métrica de distancia.

2.3.3 ¿Qué tan antiguos son los dominios estructurales de las enzimas asociadas al metabolismo?

Para determinar si las reacciones enzimáticas ampliamente distribuidas y específicas, están asociadas a dominios proteicos antiguos, los 195 *EC numbers* se evaluaron en términos de las asignaciones de la base de datos Superfamily (Oates et *al.*, 2015). Esta información es relevante, pues planteamos la siguiente hipótesis: los *EC numbers* más abundantes y más ampliamente distribuidos, podrían estar asociados a dominios proteicos antiguos. Para evaluar esta hipótesis, los dominios de proteínas identificados por las asignaciones de Supfam se anotaron con base en el índice de ancestralidad propuesto por Wang *et al.* (2009). En resumen, el enfoque considera que la línea de tiempo de la evolución de la enzima abarca ~ 3,8 mil millones de años de evolución, donde "0" representa el origen de las enzimas y "1" el presente (Figura 1.4). Por lo tanto, la ancestralidad está definida por los ancestros de los componentes del dominio de la proteína derivados de un censo filogenómico estructural (Caetano-Anollés K. y Caetano-Anollés G., 2013).

En detalle, la enzima asociada a la actividad 2.7.1, identificada como una de las más distribuidas en todos los organismos, transfiere grupos que contienen fósforo con un grupo de alcohol como aceptor. En general, se identificaron 261 dominios estructurales diferentes en proteínas asociadas a esta función enzimática, principalmente dedicadas a actividades de fosfato. De estos dominios, el *Actin-like ATPase domain* (SF:53067), el *P-loop containing nucleoside triphosphate hydrolases* (SF:52540) y el *Ribosomal protein S5 domain 2-like* (SF: 54211), representan el 40 % de su repertorio de dominios (Tabla 2.1; Figura 2.6), sugiriendo un uso preferencial de los mismos. Es interesante observar que la alta diversidad de dominios de proteínas asociados a esta actividad sugiere múltiples eventos de reclutamiento de dominios de proteínas a lo largo de la historia de la vida, lo que refuerza la idea de que el reclutamiento de funciones catalíticas es muy importante para aumentar el tamaño de los mapas metabólicos o para mantener Integridad de las funciones metabólicas.

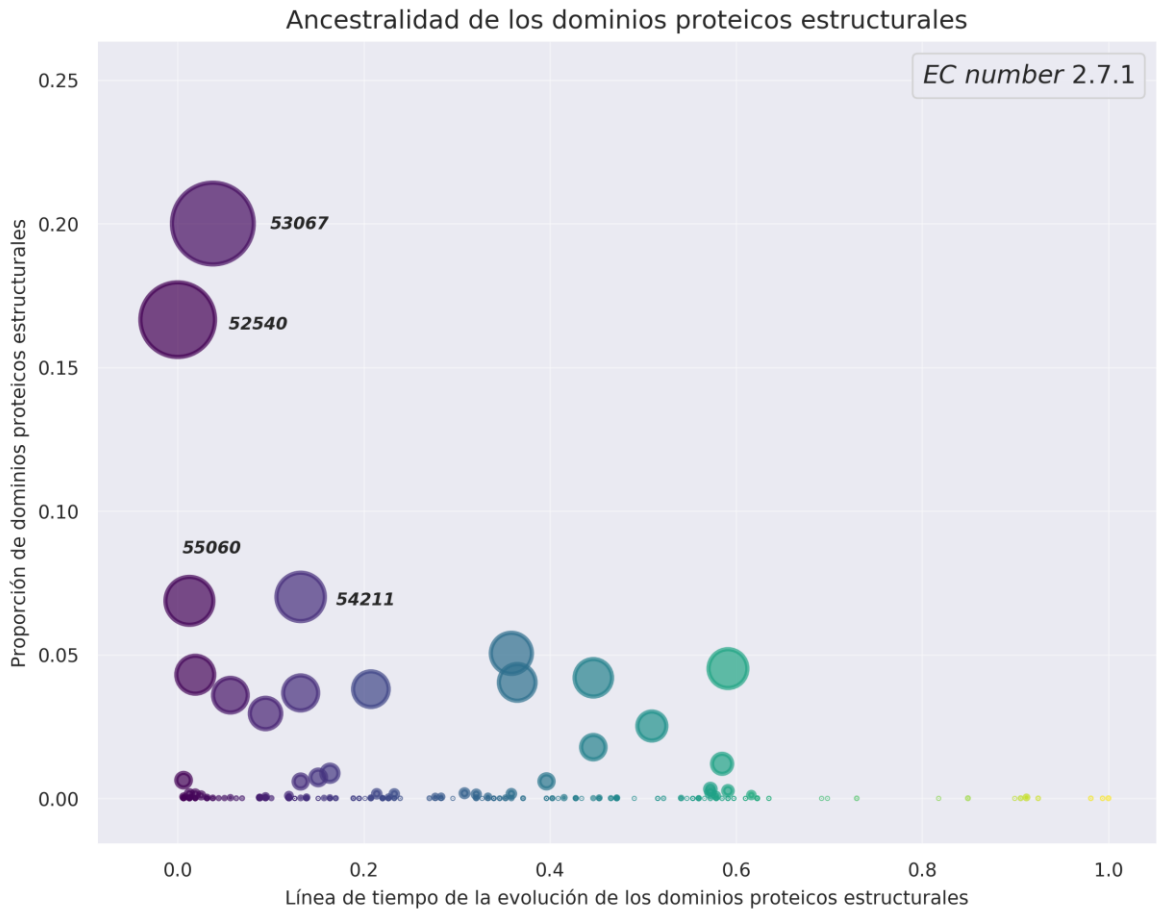

Figura 2.6 Ancestralidad y abundancia de los dominios estructurales de la reacción enzimática 2.7.1. La línea de tiempo asigna la antigüedad de cada dominio estructural presente en una de las reacciones enzimáticas con mayor distribución, *EC number* 2.7.4, como lo sugiere Wang *et al.*, (2009). "0" representa los dominios proteicos antiguos y "1" a los dominios contemporáneos. Los ID de cada superfamilia más abundante está resaltada en *Negritas*.

El segundo grupo de actividad enzimática corresponde a las fosfotransferasas con un grupo fosfato como aceptor (EC 2.7.4), dicha actividad presenta una amplia distribución entre los organismos. Las proteínas que llevan a cabo esta actividad se han relacionado con 48 dominios diferentes, principalmente dedicados a actividades de unión a grupos fosfato, tales como el *Carbamate kinase-like; el Phospholipase D-nuclease, el Ribokinase-like,* y *el Nucleoside diphosphate kinase NDK*. De estos, el dominio más abundante está relacionado *con el P-loop containing nucleoside triphosphate hydrolase* (SF: 52540), considerado como

cercano al último ancestro común de todos los organismos y que representa el 37% del total de dominios proteicos identificados en esta actividad (Figura 2.7) (Caetano-Anolles *et al.*, 2007). De hecho, Alva, *et al* (2015), identificaron el *P-loop* como uno de los 40 fragmentos estructurales cuya similitud y función sugieren un papel primordial más cercano al mundo del ARN.

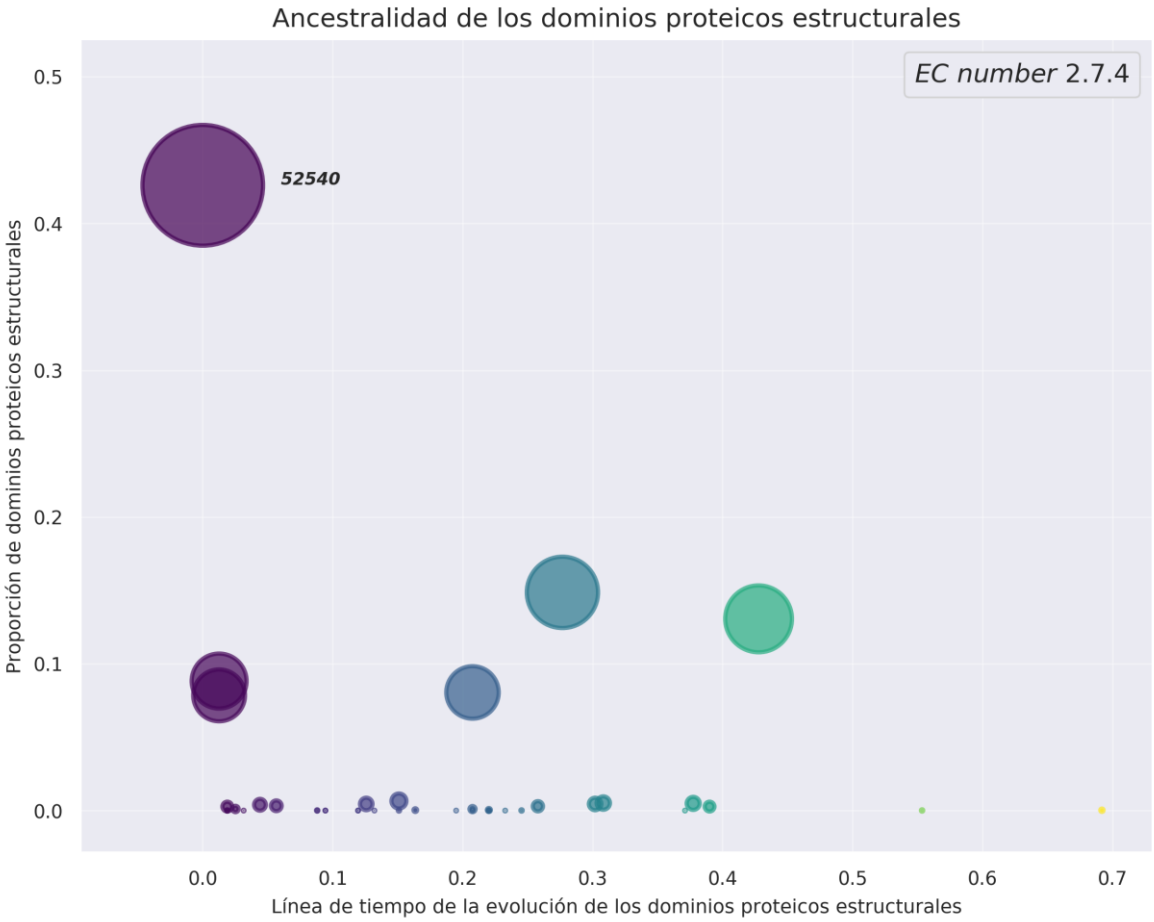

Figura 2.7 Ancestralidad y abundancia de los dominios estructurales de la reacción enzimática 2.7.4. La línea de tiempo asigna la antigüedad de cada dominio estructural presente en una de las reacciones enzimáticas con mayor distribución, *EC number* 2.7.4, como lo sugiere Wang *et al.*, (2009). "0" representa los dominios proteicos antiguos y "1" a los dominios contemporáneos. Los ID de cada superfamilia más abundante está resaltada en *Negritas*.

Las proteínas que llevan la actividad de las nucleotidil transferasas (EC 2.7.7), cuya distribución entre los organismos fue muy amplia, se han relacionado con 692 dominios diferentes, principalmente dedicados a las actividades de fosfato, tales como las *Nucleotide-diphospho-sugar transferases* (SF:53448), *Nucleotidylyl transferase* (SF:81301) y *Nucleotidyltransferases* (SF:52374), entre otros. Ocho dominios representan el 39.2% del conjunto total de dominios identificados en esta actividad enzimática, siendo el *Nucleotide-diphospho-sugar transferases* (SF: 53448) el dominio más abundante asociado a esta actividad catalítica, seguido por el *P-loop containing nucleoside triphosphate hydrolase* (SF: 52540) (Caetano-Anolles *et al.*, 2007) (Tabla 2.1; Figura 2.8).

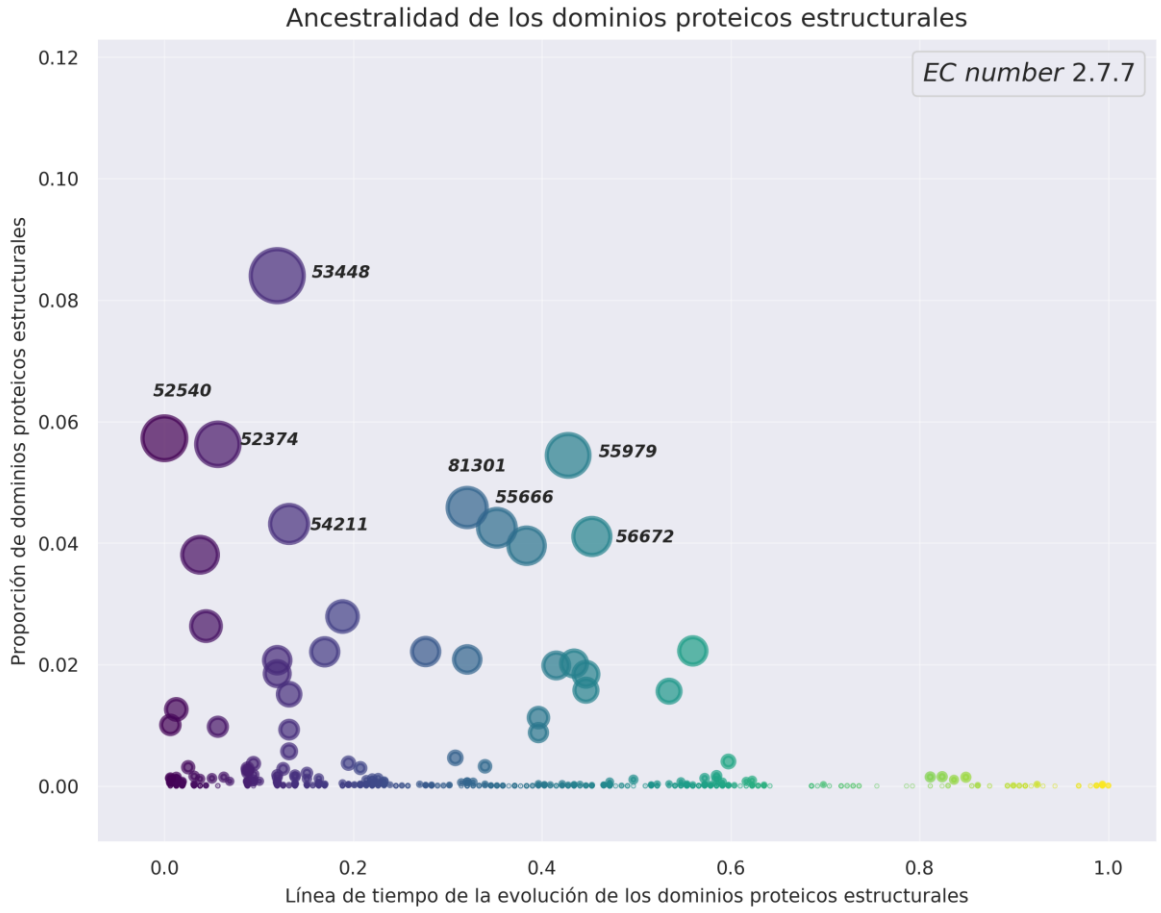

Figura 2.8 Ancestralidad y abundancia de los dominios estructurales de la reacción enzimática 2.7.7. La línea de tiempo asigna la antigüedad de cada dominio estructural presente en una de las reacciones enzimáticas con mayor distribución, *EC number* 2.7.4, como lo sugiere Wang *et al.*, (2009). "0"

representa los dominios proteicos antiguos y "1" a los dominios contemporáneos.
Los ID de cada superfamilia más abundante está resaltada en *Negritas*.

Las isomerasas que interconvierten las aldosas y las cetosas (EC 5.3.1) son proteínas relacionadas con 326 dominios diferentes, dedicados principalmente a las actividades del fosfato, como el *Ribulose-phosphate binding barrel* (SF:51366)*, Triosephosphate isomerase* (SF:51351) y *D-ribose-5-phosphate isomerase (RpiA) lid domain*, entre otros. De estos, seis dominios representan el 67% del total de los 326 dominios identificados. De hecho, el *Ribulose-phosphate binding barrel* es el dominio más abundante asociado a esta actividad y también se considera como uno de los dominios más antiguos (Tabla 2.1; Figura 2.9).

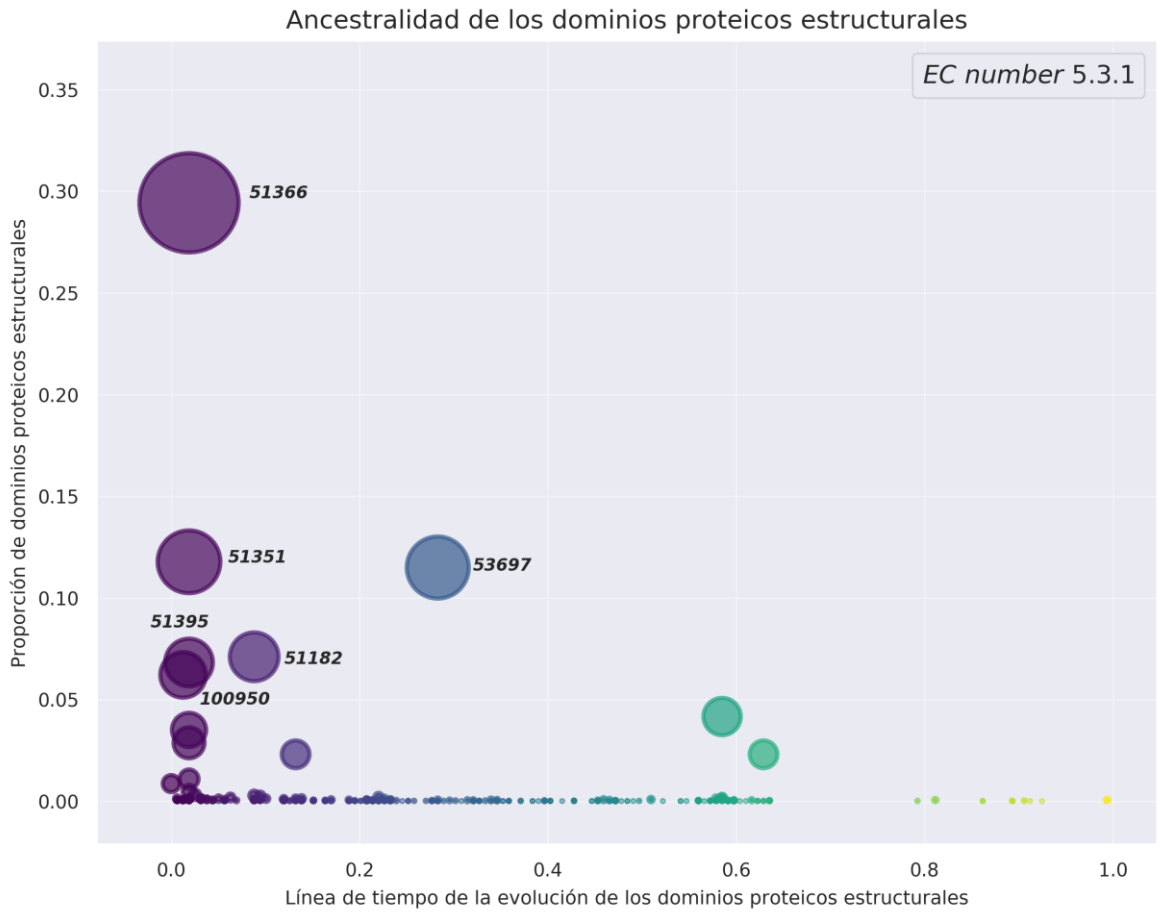

Figura 2.9 Ancestralidad y abundancia de los dominios estructurales de la reacción enzimática 5.3.1. La línea de tiempo asigna la antigüedad de cada dominio estructural presente en una de las reacciones enzimáticas con mayor distribución, *EC number* 2.7.4, como lo sugiere Wang *et al.*, (2009). "0" representa los dominios proteicos antiguos y "1" a los dominios contemporáneos. Los ID de cada superfamilia más abundante está resaltada en *Negritas*.

Finalmente, las fosfotransferasas (fosfomutasa o EC 5.4.2) están asociadas a 12 dominios diferentes, principalmente dedicados a las actividades de fosfato, como *Ribonuclease H-like* o *P-loop containing nucleoside triphosphate hydrolases*. Solo dos dominios representan el 61,5% de los dominios de proteínas totales, siendo la *Phosphoglucomutase* (SF: 53738) más abundante, seguida de *Phosphoglycerate mutase-like* (SF: 53254) (Tabla 2.1; Figura 2.10).

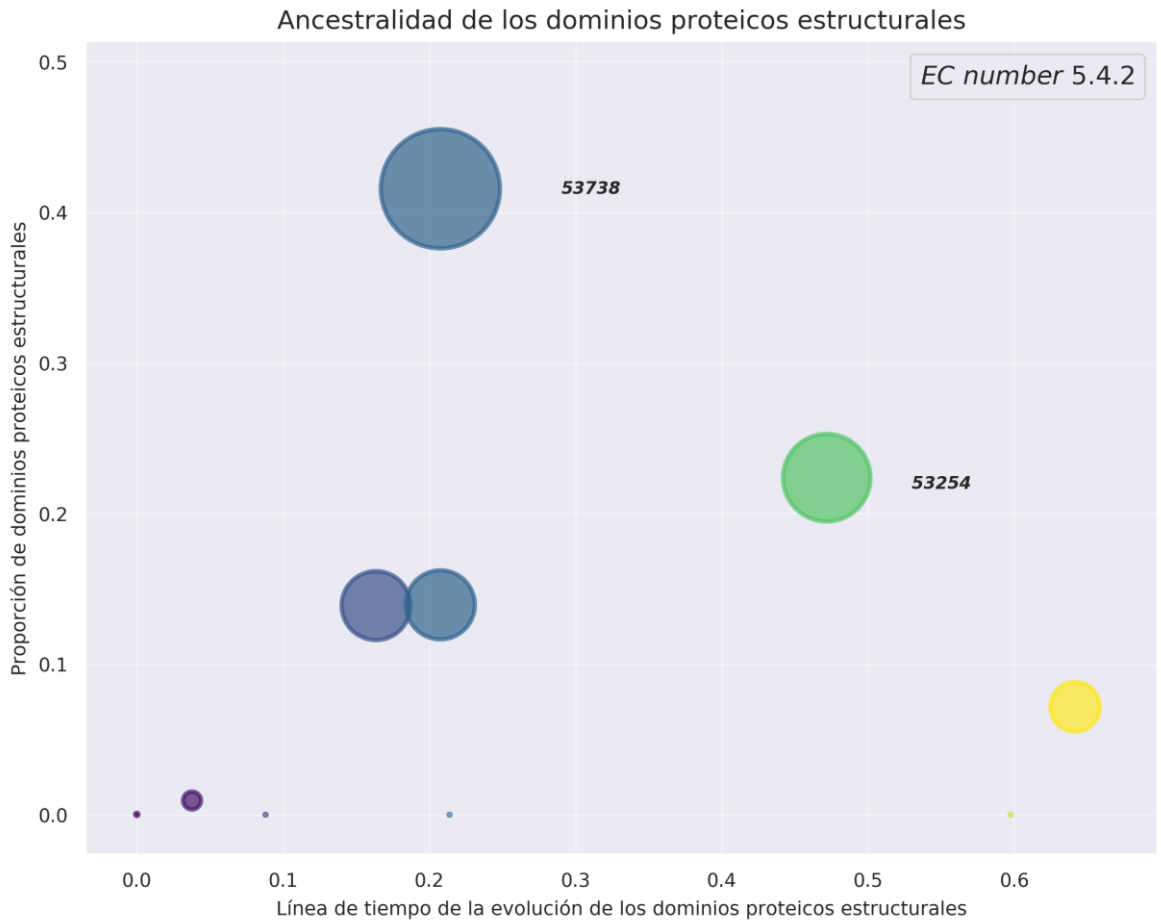

Figura 2.10 Ancestralidad y abundancia de los dominios estructurales de la reacción enzimática 5.4.2. La línea de tiempo asigna la antigüedad de cada dominio estructural presente en una de las reacciones enzimáticas con mayor distribución, *EC number* 2.7.4, como lo sugiere Wang *et al.*, (2009). "0" representa los dominios proteicos antiguos y "1" a los dominios contemporáneos. Los ID de cada superfamilia más abundante está resaltada en *Negritas*.

En resumen, las cinco actividades enzimáticas identificadas como ampliamente distribuidas a lo largo de las bacterias, arqueas y eucariotas, se relacionaron con los dominios asociados a las funciones relacionadas con el fosfato (transferasas e isomerizaciones) que pueden resaltar la importancia del metabolismo del fósforo en el mantenimiento global de la función celular. El dominio *P-loop containing nucleoside triphosphate hydrolases* (SF: 52540), es el más recurrente y ancestral de las actividades enzimáticas identificadas (Tabla 2.1). En contraste, en el resto de *EC numbers*, que presentan menor distribución en los organismos,

el dominio más abundante corresponde al antiguo *NAD(P)-binding Rossmann-fold domain* (Schaeffer *et al.*, 2017); asociados a procesos funcionales fundamentales, como el enlace FAD, NAD o NADP (Hanukoglu, 2015; Laurino *et al.*, 2016).

Tabla 2.1 *EC numbers* ampliamente distribuidos en los tres dominios celulares

EC number	***Descripción***	***Función***	***Total de dominios estructurales***	**Dominios estructurales más abundantes (ID Supfam / Descripción) / %**
2.7.1	Transferasa	Fosfotransferasa con un grupo alcohol como aceptor	261	53067 / Actin-like ATPase domain/ 0.183274 52540 / P-loop containing nucleoside triphosphate hydrolases/ 0.152635 54211 / Ribosomal protein S5 domain 2-like/ 0.064262
2.7.4	Transferasa	Fosfotransferasa con un grupo fosfato como aceptor	48	52540 / P-loop containing nucleoside triphosphate hydrolases / 0.375094
2.7.7	Transferasa	Nucleotidiltransferasa	692	53448 / Nucleotide-diphospho-sugar transferases / 0.077639 52540 / P-loop containing nucleoside triphosphate hydrolases / 0.052910 52374 / Nucleotidylyl transferase / 0.051996 55979 / DNA clamp / 0.050308 81301 / Nucleotidyltransferase/ 0.042375 54211 / Ribosomal protein S5 domain 2-like / 0.039876 55666 / Ribonuclease PH domain 2-like/ 0.039296 56672 / DNA/RNA polymerases/ 0.037970
5.3.1	Isomerasa	Interconvierte aldosas y ketosas	326	51366 / Ribulose-phoshate binding barrel / 0.271909 51351 / Triosephosphate isomerase (TIM) / 0.108794 53697 / SIS domain / 0.106178 51182 / RmlC-like cupins / 0.065631 51395 / FMN-linked oxidoreductases / 0.063015 100950 / NagB/RpiA/CoA transferase-like / 0.059398
5.4.2	Isomerasa	Fosfotransferasa (fosfomutasa)	12	53738 / Phosphoglucomutase, first 3 domains / 0.400197 53254 / Phosphoglycerate mutase-like / 0.215499

2.3.4 Las relaciones funcionales de los pares enzimáticos consecutivos proyectan grupos taxonómicos conservados y variables

Para determinar la relación de los EC numbers consecutivos y su distribución en los genomas completos, se evaluaron las asociaciones funcionales entre las reacciones enzimáticas a través de la distribución de pares enzimáticos consecutivos no redundantes. Para determinar los pares de reacciones enzimáticas más significativos, se construyeron miniseries de dos EC numbers consecutivos (EC: a.b.c → EC: w.x.y). Se compararon las frecuencias de pares enzimáticos consecutivos con valores esperados sobre *N* conjuntos de ESS aleatorios. En ese contexto, se construyeron 10 bases de datos aleatorias mezclando el nrESS real, manteniendo la composición y longitudes de los *EC numbers*, de manera similar a la ESS aleatoria construida para comparaciones de proteobacterias (Poot-Hernández *et al.,* 2015) . A partir de ellos, obtuvimos los pares enzimáticos utilizando un valor de Z a través de la siguiente fórmula: valor Z (Zi) = (Nreal i — <Nrandi>) / std (Nrandi). A partir de esto, un valor de Z ≥ 5 sugiere que la frecuencia del par en el ESS real es significativamente mayor de lo esperado por azar, dejando un conjunto de 132 pares de EC como significativo, lo que sugiere que están involucrados en una gran cantidad de pares consecutivos. reacciones en los organismos considerados en este análisis.

Basándonos en el patrón de distribución asociado con los pares enzimáticos en todos los genomas, identificamos cinco pares (EC 4.2.1: 5.4.2; 5.4.2: 4.2.1; 2.7.7: 2.7.1; 2.7.4: 3.6.1; y 2.7.7: 2.7.8) distribuidos ampliamente entre los organismos o pares enzimáticos "universales". (Tabla 2.2). Estas reacciones están involucradas principalmente en funciones relacionadas con el fosfato (transferasas e isomerasas) y también relacionadas con el metabolismo del fósforo.

Tabla 2.2. Pares de *EC numbers* significativos y ampliamente distribuidos en los tres Dominios celulares

Par *EC number* (A:B)	***Descripción***	**Función *EC A***	**Función *EC B***
4.2.1:5.4.2	Liasa:Isomerasa	Hidro-liasa	Fosfotransferasa (Fosfomutasa)
5.4.2:4.2.1	Isomerasa:Liasa	Fosfotransferasa (Fosfomutasa)	Hidro-liasa
2.7.7:2.7.1	Transferasa:Transferas	Nucleotidiltransferasa	Fosfotransferasa con un grupo alcohol como aceptor
2.7.4:3.6.1	Transferasa:Hidrolasa	Fosfotransferasa con un grupo fosfato como aceptor	En anhídridos que contienen fósforo
2.7.7:2.7.8	Transferasa:Transferasa	Fosfotransferasa (Fosfomutasa)	Transferasa para otros grupos fosfatos sustituidos

Para evaluar los roles de estos pares enzimáticos en todos los mapas metabólicos, estas reacciones "universales" se rastrearon a lo largo del metabolismo completo de Bacterias, Arqueas y Eucariotes. Por tanto, las cinco reacciones se identificaron en el metabolismo de los glicerolípidos, probablemente porque esta vía es una vía fundamental asociada con el origen y evolución de las membranas celulares y vinculada al componente estructural central de las principales clases de lípidos biológicos, triglicéridos y fosfatidil fosfolípidos que participan en la composición de las membranas (Peretó *et al.*, 2004). En este sentido, se han identificado diversas estructuras lipídicas en los tres dominios celulares, como el enlace éster en ácidos grasos de cadena larga en Bacteria y Eucaria o éter lípidos con isoprenoides en Arqueas; hay lípidos polares comunes con una columna vertebral de glicerol en todos los organismos, con la excepción de sus estereoestructuras (Yokobori *et al.*, 2016). Por lo tanto, esta columna vertebral común se asocia a los organismos analizados en este trabajo, sin embargo se requieren más análisis.

Finalmente, dos mapas metabólicos antiguos, para la glucólisis y el metano, contienen dos y tres pares de reacciones, respectivamente, 5.4.2: 4.2.1 y 4.2.1: 5.4.2 y 5.4.2: 4.2.1, 4.2.1: 5.4 .2 y 2.7.7: 2.7.8; mientras que el par 2.7.7: 2.7.1 se asocia preferentemente con ocho mapas metabólicos, entre los que destacan los mapas de metabolismo de aminoazúcares y nucleótidos, metabolismo de fructosa y manosa, entre otros.

2.4 CONCLUSIONES

Las actividades enzimáticas reflejan la organización del metabolismo en todos los organismos y su análisis podría proporcionar información valiosa acerca de cómo las reacciones se han asociado a su metabolismo y a sus dominios estructurales. En este trabajo, evaluamos la abundancia y distribución de las reacciones enzimáticas en organismos de los tres dominios celulares, identificando cinco *EC numbers* (en los tres primeros niveles de clasificación) que están ampliamente distribuidos Bacterias, Arqueas y Eucariotes, aunque están limitados a mapas metabólicos específicos (es decir, no están asociados con todos los mapas metabólicos). Además, identificamos que esas reacciones están asociadas a dominios estructurales propuestos como antiguos, como el *P-loop containing nucleoside triphosphate hydrolases* (SF: 52540). Cuando analizamos la asociación funcional entre las reacciones enzimáticas, identificamos 132 pares enzimáticos como significativos; sin embargo, 5 de ellas se definen como universales para los dominios celulares. Esta asociación sugiere que dichas reacciones también podrían ser ancestrales en la evolución de las vías metabólicas o partícipes en la evolución de las membranas celulares sintetizando lípidos biológicos como los fosfatidil fosfolípidos. También, identificamos firmas funcionales para diversos grupos taxonómicos. En el caso de las cianobacterias la firma funcional es la actividad enzimática 1.3.7. La firma funcional 3.1.26 se presenta para las actinobacterias. Un conjunto de reacciones enzimáticas (5.4.4, 4.2.99, 1.16.1 y 5.1.2) se denominan firma funcional para el grupo taxonómico Gammaproteobacteria. En resumen, encontramos que las reacciones enzimáticas conservadas están relacionadas principalmente con las reacciones de fosforilación, que son esenciales en el metabolismo moderno.

CAPÍTULO III

DISCUSIÓN, CONCLUSIONES GENERALES Y PERSPECTIVAS

3.1 DISCUSIÓN GENERAL

El metabolismo en un contexto amplio, es una gran colección de reacciones químicas entrelazadas. Desde el punto de vista funcional, las reacciones bioquímicas están reguladas por enzimas específicas, promiscuas y/o paquetes multienzimáticos con un alto grado de complejidad, que se codifican a partir de secuencias de DNA y RNA altamente conservadas, en un nivel evolutivo (Smith y Morowitz, 2004). Dicha conservación ha despertado un interés para determinar su posible origen, razón por la cual se han propuesto diversas hipótesis acerca de su expansión, tales como aquellos que se rigen bajo los mecanismos moleculares de la duplicación de genes; es decir, en los inicios de la vida, las células primitivas tendrían un conjunto de pocas reacciones químicas simplificadas, y ante la necesidad selectiva, estas lograron la estructuración de sus genomas y por ende, las vías metabólicas (Scossa & Fernie, 2020). En este contexto, nuestra pregunta biológica es, desde una perspectiva funcional, ¿qué tipo de reacciones enzimáticas repercuten directamente en el metabolismo que actualmente conocemos? Para ello, en este trabajo evaluamos el repertorio enzimático de 1507 organismos pertenecientes a los tres dominios celulares (Bacterias, Arqueas y Eucariotes), cuya información se encuentra depositada en KEGG.

Con este análisis, hemos identificado que las reacciones enzimáticas son, en mayor proporción, actividades ligadas al transporte de electrones y moléculas, entre ellos, los grupos fosfatos. Algunos trabajos sugieren que las actividades transferasas (2.x.x.x) están asociadas con nuevas vías metabólicas, en particular, con enzimas multifuncionales como consecuencia de la dependencia hacia el metabolito donador o aceptor (Pfeiffer *et al.*, 2005). Por ejemplo, si hablamos de comunidades entre bacterias y arqueas en ambientes hostiles, como entornos metanogénicos, la transferencia de electrones entre este tipo especies es un proceso vital, ya que se aprovechan de las capacidades metabólicas de su pareja sintrófica para obtener energía a partir de la descomposición de los compuestos que no pueden digerir por sí mismos (Stams y Plugge, 2009). De esta manera, se sitúan cinco reacciones enzimáticas 2.7.1, 2.7.4, 2.7.7, 5.3.1, 5.4.2 que se conservan en los tres

dominios celulares. Este resultado sugiere tener relación en la aparición primordial transferasas como ATPasa, GTPasa y helicasa, que fueron cruciales para la unión y el transporte, la aparición de ácidos nucleicos y polímeros de proteínas y la comunicación de las células primordiales con el medio ambiente (Kim y Caetano Anollés, 2010). Por otro lado, estas cinco reacciones enzimáticas se encuentran limitadas a mapas metabólicos como la vía de la Pentosa fosfato, o el metabolismo de azúcares, entre otros. Se ha observado la importancia de la relación de la molécula del inositol con el fosfato. En las arqueas, se conserva una única reacción de isomerización irreversible que convierte la glucosa en la forma mucho más estable de inositol, dando a un azúcar metabólicamente inerte y versátil, el lienzo ideal para decorar con fosfatos; esto permite que las arqueas, puedan adaptarse a ambientes hostiles (Livermore *et al.*, 2016). Cuando analizamos la diversidad de dominios asociados a dichas actividades enzimáticas, identificamos diferentes eventos de reclutamiento de dominios a lo largo de su historia evolutiva. El *P-loop containing nucleoside triphosphate hydrolases* (SF: 52540) es el dominio más antiguo presente y recurrente en las actividades enzimáticas conservadas. Los *P-loops*, así como los *Rossmanns folds,* se describen como dominios de unión a nucleótidos porque ambos utilizan ribonucleósidos fosforilados como ATP o NAD, así como otros cofactores pre-LUCA como SAM (Longo *et al.*, 2020). Por ello, se ha visto que la arquitectura de algunas proteínas anunciadas como antiguas, la mayoría de ellas representada bajo el *EC number* 2.7.x.x, recurren principalmente a los dominios *P-loop containing nucleoside triphosphate hydrolases*, aunque también están asociadas a los dominios *TIM beta/alpha-barrel*, *NAD(P)-binding Rossmann-fold domains*, entre otros (Ma *et al.*, 2008). En ese contexto, sugerimos que dichas reacciones enzimáticas, 2.7.1, 2.7.4, 2.7.7, 5.3.1, 5.4.2, podrían ser ancestrales en la evolución de las vías metabólicas.

Por otra parte, cuando se analizan las reacciones enzimáticas consecutivas (a manera de pares), identificamos 5 pares enzimáticos significativos (4.2.1: 5.4.2; 5.4.2: 4.2.1; 2.7.7: 2.7.1; 2.7.4: 3.6.1; y 2.7.7: 2.7.8) se conservan ampliamente entre todos los organismos analizados, cuyas funciones rigen ser transferasas e isomerasas, así como su participación en el metabolismo del fósforo. La mayoría de estas reacciones enzimáticas de los pares enzimáticos conservados, están asociadas a enzimas con *P-loops* como son las adenilato quinasas, la ATPasa transportadora de arsenito y la ATPasa de dos sectores transportadora de H + (Ma *et al.*, 2008). Por otra parte, estas mismas reacciones de los pares enzimáticos

conservados se presentan en el metabolismo de los glicerolípidos, vía metabólica que está asociada con el origen y evolución de las membranas celulares, siendo partícipe en el componente estructural central de las principales clases de lípidos biológicos que participan en la composición de las membranas (Peretó *et al.*, 2004).

Finalmente, seis actividades enzimáticas (3.1.26, 5.4.4, 4.2.99, 1.16.1, 5.1.2 y 1.3.7) se presentan como firmas funcionales en *Actinobacteria, Cyanobacteria, Gammaproteobacteria* y *Viridiplantae*. En concreto, la actividad enzimática 3.1.26 es la firma funcional del grupo taxonómico de las actinobacterias. Mientras que la actividad enzimática 1.3.7 funge como firma funcional para las cianobacterias. También un conjunto de cuatro actividades enzimáticas (5.4.4, 4.2.99, 1.16.1 y 5.1.2) se denomina firma funcional para las gammaproteobacterias.

Considero que este trabajo permite la posibilidad de entender la historia evolutiva del metabolismo a través de sus reacciones enzimáticas, en la cual se pueden asociar a la conservación y reclutamiento de diversos dominios proteicos que propician a las actividades catalíticas de las mismas. Como consecuencia, podemos determinar que la capacidad energética del metabolismo moderno es llevada a cabo por reacciones asociadas al ión fósforo y grupos fosfatos como NAD, NADH, FAD, FADH, así como los ácidos nucleicos DNA y ARN.

3.2 CONCLUSIONES GENERALES

- Cinco reacciones enzimáticas, (2.7.1, 2.7.4, 2.7.7, 5.3.1, 5.4.2) se conservan en los tres Dominios celulares, involucradas en los mecanismos de fosforilación esenciales en el metabolismo moderno.

- Las reacciones enzimáticas conservadas, están limitadas a mapas metabólicos como la vía de la Pentosa fosfato, El metabolismo de azúcares, etc.

- La diversidad de dominios asociados a las actividades enzimáticas conservadas, sugieren que existieron diferentes eventos de reclutamiento de dominios a lo largo de su historia evolutiva.

- El *P-loop containing nucleoside triphosphate hydrolases* (SF: 52540), es el dominio más antiguo presente en las actividades enzimáticas conservadas, que sugiere que dichas reacciones también podrían ser ancestrales en la evolución de las vías metabólicas.

- 5 pares enzimáticos significativos (4.2.1: 5.4.2; 5.4.2: 4.2.1; 2.7.7: 2.7.1; 2.7.4: 3.6 .1; y 2.7.7: 2.7.8) se conservan ampliamente entre los organismos de los tres Dominios celulares, cuyas funciones rigen ser transferasas e isomerasas, así como su participación en el metabolismo del fósforo.

- Las reacciones enzimáticas de los pares enzimáticos conservados se presentan en el metabolismo de los glicerolípidos,vía metabólica que está asociada con el origen y evolución de las membranas celulares, siendo partícipe en el componente estructural central de las principales clases de lípidos biológicos que participan en la composición de las membranas.

- Seis actividades enzimáticas (3.1.26, 5.4.4, 4.2.99, 1.16.1, 5.1.2 y 1.3.7) se presentan como firmas funcionales en *Actinobacteria, Cyanobacteria, Gammaproteobacteria* y *Viridiplantae*. En concreto, la actividad enzimática 3.1.26 es la firma funcional del grupo taxonómico de las actinobacterias. Mientras que la actividad enzimática 1.3.7 funge como firma funcional para las cianobacterias.

También un conjunto de cuatro actividades enzimáticas (5.4.4, 4.2.99, 1.16.1 y 5.1.2) se denomina como firma funcional para las gammaproteobacterias.

3.3 PERSPECTIVAS

Partiendo de nuestros resultados y las conclusiones obtenidas en este trabajo doctoral en proceso, las perspectivas pueden orientarse hacia las siguientes mociones.

En primera instancia, profundizar la distribución de las reacciones enzimáticas con respecto a los organismos proyectados en este estudio. Así, realizar el análisis del repertorio enzimático en función de su primer nivel de clasificación enzimática ahora con respecto a los mapas metabólicos, nos permitirá visualizar qué actividades son las que han predominado (y con qué frecuencia) en las vías metabólicas y además, si existe alguna correlación con la diversidad de tamaños de los genomas completos. También, analizaremos las firmas funcionales que están particularmente asociadas a una especie, a división taxonómica o a algún dominio celular y, desde una perspectiva aplicada, se puedan ofrecer como marcadores funcionales para futuros análisis experimentales.

Por otro lado, se debe continuar con las comparaciones entre vías metabólicas de diferentes especies, abarcando el metabolismo de arqueas, bacterias y eucariotes. Las comparaciones partirán de secuencias lineales de actividades enzimáticas que conforman a los mapas metabólicos y, mediante algoritmos de programación dinámica, nos enfocaremos en alineamientos de secuencias metabólicas mediante un score definido. Esto nos permitirá conocer cuál es la limitante funcional del crecimiento del metabolismo. También, la similitud entre mapas metabólicos se podría correlacionar con la expresión global de los genomas completos analizados, es decir, podemos inferir la expresión de los genes con respecto a las actividades conservadas entre vías metabólicas, aplicando una perspectiva de teoría de redes. Un enfoque aplicado de los alineamientos de los mapas metabólicos es tener la posibilidad de predecir rutas alternas o rutas nuevas con respecto a las ya existentes. De esta manera, sugerimos mapear las firmas funcionales que se conservan en las vías metabólicas para después tomarlas como base y generar secuencias metabólicas mediante estrategias estadísticas como los modelos ocultos de Markov.

ANEXOS

El anexo se encuentra en el siguiente enlace:

https://docs.google.com/document/d/1bQDgi1L7vg9W2kT1-BICwhdNo_CxCGCfUgqVahvZBYg/edit?usp=sharing

Figuras

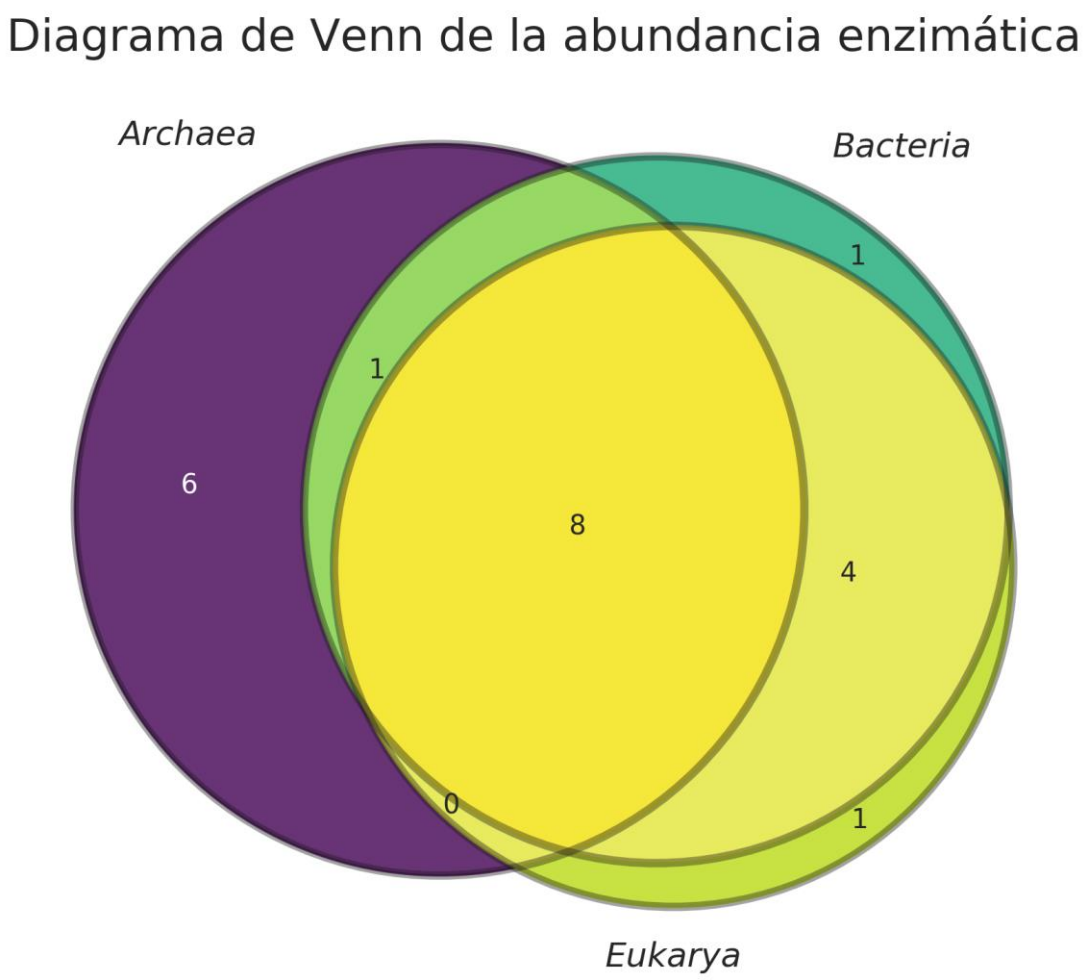

Figura Suplementaria 2.1 Diagrama de Venn de la abundancia enzimática de los tres Dominios celulares. Diagrama de Venn de la abundancia enzimática, que revela los 8 *EC numbers* como abundantes en los tres dominios celulares. Se identificó un *EC number* como abundante en arqueas y bacterias pero no en eucariotes; cuatro *EC numbers* son abundantes en bacterias y eucariotes pero no en arqueas; un *EC number* fue identificado como abundante en bacterias pero no en arqueas y eucariotas; un *EC number* fue abundante en eucariotes; y finalmente, seis *EC numbers* 1.2.7, 2.7.4, 4.1.2, 4.3.2, 5.3.1 y 6.3.5 fueron identificadas como altamente abundantes en arqueas.

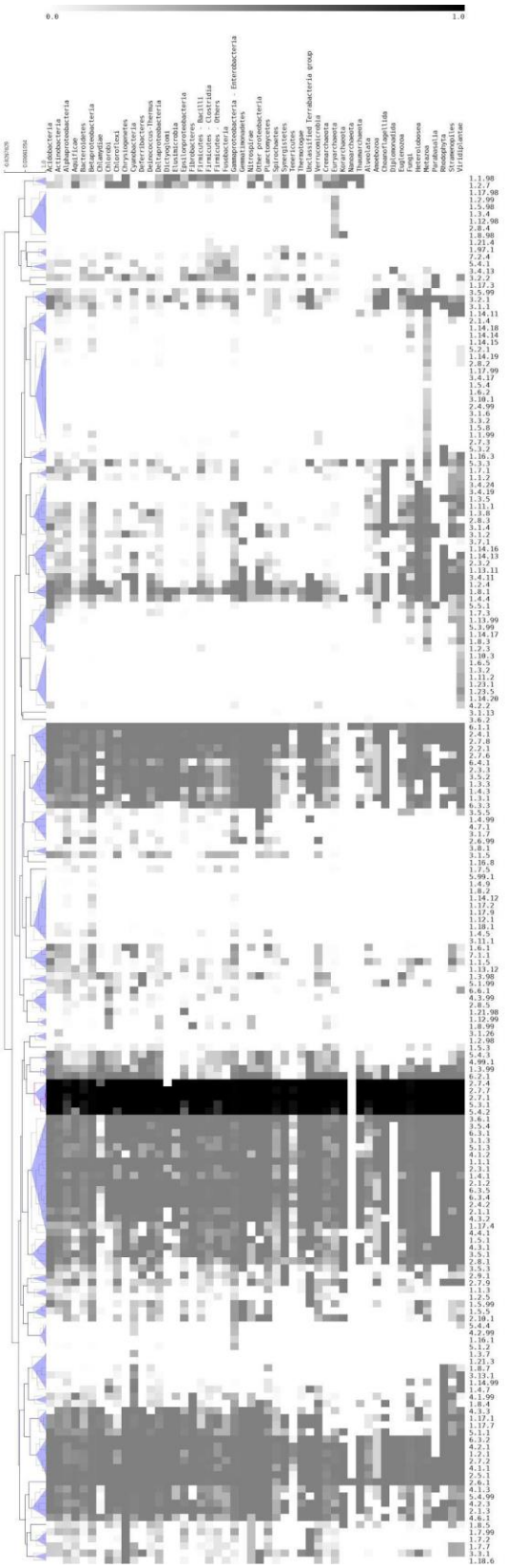

Figura Suplementaria 2.2 Análisis de agrupamiento de los *EC numbers* que muestra la presencia de un conjunto de actividades enzimáticas en todos los organismos. 195 *EC numbers* se agruparon con un HCA utilizando la correlación al cuadrado de Pearson como métrica de distancia. Se obtuvieron 50 agrupamientos utilizando un umbral de distancia de 0.668.

Tablas

Tabla Suplementaria 1

Código	Categoría funcional (utilizado en mapas metabólicos globales)	Código HEX color	Nombre color
09101	Metabolismo de carbohidratos	#0000ee	*Blue*
09102	Metabolismo energético	#9933cc	*Purple Heart*
09103	Metabolismo de lípidos	#009999	*Persian Green*
09104	Metabolismo de nucleótidos	#ff0000	*red*
09105	Metabolismo de aminoácidos	#ff9933	*Neon Carrot*
09106	Metabolismo de otros aminoácidos	#ff6600	*Blaze Orange*
09107	Biosíntesis y metabolismo de glicanos	#3399ff	*Dodger Blue*
09108	Metabolismo de cofactores y vitaminas	#ff6699	*Hot Pink*
09109	Metabolismo de terpenoides y policétidos	#00cc33	*Malachite*
09110	Biosíntesis de otros metabolitos secundarios	#cc3366	*Hibiscus*
09111	Biodegradación y metabolismo de xenobióticos	#ccaa99	*Eunry*

Tabla Suplementaria 2

Se encuentra en el siguiente enlace:

https://drive.google.com/file/d/1OwulCFn-ZshJqZpP8GXCk905ffT8wLWI/view?usp=sharing

Tabla Suplementaria 3

Distribución proporcional de la abundancia en *Arqueas*				**Distribución proporcional del muestreo aleatorio en *Arqueas***		
EC number	***Proporción***	**Acumulativo**		***EC number***	**Sesgo (media)**	**Proporción de la media**

6.3.4	0.056478241584939	0.056478241584939		6.3.4	3271.066	0.055319865999757
4.2.1	0.051878501141721	0.10835674272666		4.2.1	3007.319	0.050859409164634
2.7.4	0.050252164342155	0.158608907068815		2.7.4	2914.011	0.049281395741272
2.7.1	0.045882410921098	0.204491317989913		2.7.1	2659.701	0.044980536289827
1.1.1	0.039294925500633	0.243786243490546		1.1.1	2277.189	0.038511540377394
2.4.2	0.038917089678511	0.282703333169057		2.4.2	2255.35	0.038142201894597
4.1.2	0.036321521856981	0.319024855026038		4.1.2	2104.816	0.035596389395428
4.3.2	0.034481625679694	0.353506480705732		1.2.7	1998.188	0.033793109769819
1.2.7	0.034481625679694	0.387988106385425		4.3.2	1997.401	0.033779800122584
2.6.1	0.030161155191957	0.418149261577382		2.6.1	1748.384	0.029568455236342
2.5.1	0.029142641236673	0.447291902814056		2.5.1	1689.206	0.028567644176542
2.7.7	0.027894140259228	0.475186043073284		2.7.7	1617.45	0.027354115527265
6.3.5	0.027335600348266	0.50252164342155		6.3.5	1583.946	0.026787499998732
5.3.1	0.02585711234866	0.52837875577021		5.3.1	1498.474	0.025342008043898

4.1.1	0.023836512082533	0.552215267852743		4.1.1	1381.813	0.023369051555891

Tabla Suplementaria 4

Distribución proporcional de la abundancia en *Bacterias*				**Distribución proporcional del muestreo aleatorio**		
EC number	**Proporción**	**Acumulativo**		***EC number***	**Sesgo (media)**	**Proporción de la media**
2.7.1	0.057096170954447	0.057096170954447		2.7.1	7495.817	0.056178814276557
1.1.1	0.051786827405484	0.108882998359931		1.1.1	6794.558	0.050923096438093
2.5.1	0.044076689845495	0.152959688205426		2.5.1	5773.9	0.043273582552965
2.6.1	0.043592536780503	0.196552224985929		2.6.1	5714.953	0.042831793143597
2.3.1	0.038916828555348	0.235469053541277		2.3.1	5096.832	0.038199168726613
4.2.1	0.037040735428506	0.272509788969783		4.2.1	4860.484	0.036427814063521
6.3.2	0.031389458777393	0.303899247747175		6.3.2	4114.017	0.030833276342472
2.7.7	0.030668675901886	0.334567923649062		2.7.7	4018.499	0.030117398676026
2.4.2	0.029734260486453	0.364302184135514		2.4.2	3900.555	0.02923344512286

6.3.4	0.029008030888966	0.39331021502448	6.3.4	3798.433	0.028468072532847
4.1.1	0.026869889915697	0.420180104940177	4.1.1	3530.088	0.026456910318369
1.2.1	0.026082535993755	0.446262640933931	1.2.1	3422.943	0.025653891907479
3.5.1	0.023217560231667	0.469480201165599	3.5.1	3048.475	0.022847370853868
3.1.3	0.022694069730145	0.492174270895744	3.1.3	2976.302	0.022306456692972

Tabla Suplementaria 5

Distribución proporcional de la abundancia en *Eucariotes*			**Distribución proporcional del muestreo aleatorio**		
EC numbers	**Proporción**	**Acumulativo**	***EC number***	**Sesgo (media)**	**Proporción de la media**
2.7.1	0.059289099832524	0.059289099832524	2.7.1	13253.522	0.057869975970606
2.3.1	0.057737198429993	0.117026298262517	2.3.1	12908.332	0.056362743628494
1.1.1	0.056217628306682	0.173243926569199	1.1.1	12563.686	0.054857886599671
2.6.1	0.052169752148414	0.225413678717612	2.6.1	11661.925	0.050920451146571
4.1.1	0.037953688675646	0.263367367393258	4.1.1	8484.998	0.037048765631554

1.2.1	0.032321579835628	0.295688947228886		1.2.1	7228.268	0.031561398960148
4.2.1	0.029379433426663	0.325068380655549		4.2.1	6571.404	0.028693278026259
2.7.7	0.027403992266358	0.352472372921907		2.7.7	6126.258	0.026749599332897
3.1.3	0.026666839100156	0.379139212022063		3.1.3	5963.434	0.026038647106957
2.4.1	0.023517772504187	0.40265698452625		2.4.1	5256.42	0.022951551979271
2.4.2	0.022240686975021	0.424897671501271		2.4.2	4972.578	0.021712188607071
3.5.1	0.021975570485422	0.446873241986692		3.5.1	4908.155	0.02143089300414
2.5.1	0.021742785275042	0.468616027261735		2.5.1	4858.268	0.021213067169524

Tabla Suplementaria 6

Se encuentra en el siguiente enlace:

https://drive.google.com/file/d/1bntH9IPx1Z_QJKVK_bilMMNMnbGBHu5f/view?usp=sharing

REFERENCIAS

Altman, T., Travers, M., Kothari, A., Caspi, R., & Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. BMC bioinformatics, 14(1), 112.

Alva V, Soding J., & Lupas AN (2015). A vocabulary of ancient peptides at the origin of folded proteins. Elife , 4:e09410

Armenta-Medina, D., Pérez-Rueda, E., & Segovia, L. (2011). Identification of functional motions in the adenylate kinase (ADK) protein family by computational hybrid approaches. *Proteins: Structure, Function, and Bioinformatics*, *79*(5), 1662-1671.

Armenta-Medina, D., Segovia, L., & Perez-Rueda, E. (2014). Comparative genomics of nucleotide metabolism: a tour to the past of the three cellular domains of life. *BMC genomics*, *15*(1), 800.

Caetano-Anollés, G., Kim, H. S., & Mittenthal, J. E. (2007). The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences*, *104*(22), 9358-9363.

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D., & Mittenthal, J. E. (2009a). The origin, evolution and structure of the protein world. Biochemical Journal, 417(3), 621-637.

Caetano-Anollés, G., Yafremava, L. S., Gee, H., Caetano-Anollés, D., Kim, H. S., & Mittenthal, J. E. (2009b). The origin and evolution of modern metabolism. *The international journal of biochemistry & cell biology*, *41*(2), 285-297.

Caetano-Anollés, K., & Caetano-Anollés, G. (2013). Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism. PLoS One, 8(3), e59300.

Cárdenas-Conejo, Y., Carballo-Uicab, V., Lieberman, M., Aguilar-Espinosa, M., Comai, L., & Rivera-Madrid, R. (2015). De novo transcriptome sequencing in Bixa orellana to

identify genes involved in methylerythritol phosphate, carotenoid and bixin biosynthesis. *BMC genomics*, *16*(1), 877.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., & Walk, T. C. (2007). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, *36*, D623-D631.

Caspi, R., Billington, R., Keseler, I. M., Kothari, A., Krummenacker, M., Midford, P. E., & Karp, P. D. (2019). The MetaCyc database of metabolic pathways and enzymes-a 2019 update. Nucleic acids research, 48(D1), D445-D453.

Cuesta, S. M., Rahman, S. A., Furnham, N., & Thornton, J. M. (2015). The classification and evolution of enzyme function. Biophysical journal, 109(6), 1082-1086.

Dandekar, T., Schuster, S., Berend, S. N. E. L., Huynen, M., & Peer, B. O. R. K. (1999). Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochemical Journal, 343(1), 115-124.

de la Osa, J. L., Bateman, D. A., Ho, S., González, C., Chakrabartty, A., & Laurents, D. V. (2007). Getting specificity from simplicity in putative proteins from the prebiotic Earth. *Proceedings of the National Academy of Sciences*, *104*(38), 14941-14946.

Díaz-Mejía, J. J., Pérez-Rueda, E., & Segovia, L. (2007). A network perspective on the evolution of metabolism by gene duplication. Genome biology, 8(2), R26.

Even, S., O. Pellegrini, L. Zig, V. Labas, J. Vinh, D. Brechemmier-Baey & H. Putzer, (2005). Ribonucleases J1 and J2: two novel endoribonucleases in B. subtilis with functional homology to E. coli RNase E. Nucleic Acids Res 33: 2141-2152.

Fani, R., & Fondi, M. (2009). Origin and evolution of metabolic pathways. Physics of Life Reviews, 6(1), 23–52.3

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, *39*(suppl_2), W29-W37.

Gray, L. R., Tompkins, S. C., & Taylor, E. B. (2014). Regulation of pyruvate metabolism and human disease. *Cellular and molecular life sciences*, *71*(14), 2577-2604.

Hernández-Montes, G., Díaz-Mejía, J. J., Pérez-Rueda, E., & Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome biology*, *9*(6), R95.

Horowitz, N. H. (1945). On the evolution of biochemical synthesis. *Proceedings of the National Academy of Sciences*, *31*(6), 153-157.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology*, *30*(1), 409-425.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. Proceedings of the National Academy of Sciences, 109(40), 16213-16216.

Kanehisa, M. (1997). A database for post-genome analysis. *Trends in Genetics*, *13*(9), 375-376.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2013). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic acids research*, *42*(D1), D199-D205.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research*, *45*(D1), D353-D361.

Kashyap, H., Ahmed, H. A., Hoque, N., Roy, S., & Bhattacharyya, D. K. (2015). Big data analytics in bioinformatics: A machine learning perspective. arXiv preprint arXiv:1506.05101.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., & Tanabe, M. (2018). New approach for understanding genome variations in KEGG. Nucleic acids research, 47(D1), D590-D595.

Kim, K. M., & Caetano-Anollés, G. (2010). Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. Molecular Biology and Evolution, 27(7), 1710-1733.

Lesk, A. (2019). Introduction to bioinformatics. Oxford university press.

Light, S., Kraulis, P., & Elofsson, A. (2005). Preferential attachment in the evolution of metabolic networks. *Bmc Genomics*, *6*(1), 159.

Livermore, T. M., Azevedo, C., Kolozsvari, B., Wilson, M. S., & Saiardi, A. (2016). Phosphate, inositol and polyphosphates. Biochemical Society Transactions, 44(1), 253-259.

Longo, L. M., Jabłońska, J., Vyas, P., Kanade, M., Kolodny, R., Ben-Tal, N., & Tawfik, D. S. (2020). On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. Elife, 9, e64415.

Ma, B. G., Chen, L., Ji, H. F., Chen, Z. H., Yang, F. R., Wang, L., & Zhang, H. Y. (2008). Characters of very ancient proteins. Biochemical and biophysical research communications, 366(3), 607-611.

Marini, C., Bianchi, G., Buschiazzo, A., Ravera, S., Martella, R., Bottoni, G., & Inglese, E. (2016). Divergent targets of glycolysis and oxidative phosphorylation result in additive effects of metformin and starvation in colon and breast cancer. *Scientific reports*, *6*, 19569.

Muto-Fujita, A. (2019). A Novel model for the Chemical Evolution of Metabolic Networks. 日本化学会情報化学部会誌, 37(3), 57.

Nelson, D. L., & Cox, M. M. (2017). Lehninger Principles of Biochemistry. W. H. Freeman. 7ta Edición, 528-542.

Noor, E., Eden, E., Milo, R., & Alon, U. (2010). Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Molecular cell*, *39*(5), 809-820.

Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J., Sardar, A. J., & Gough, J. (2014). The SUPERFAMILY 1.75 database in 2014: a doubling of data. *Nucleic acids research*, *43*(D1), D227-D233.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., & Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic acids research*, *36*(suppl_2), W423-W426.

Oren, A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline systems, 4(1), 2.

Ortegon, P., Poot-Hernández, A. C., Perez-Rueda, E., & Rodriguez-Vazquez, K. (2015). Comparison of Metabolic Pathways in Escherichia coli by Using Genetic Algorithms. *Computational and structural biotechnology journal*, *13*, 277-285.

Pál, C., Papp, B., & Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature genetics*, *37*(12), 1372.

Papin, J. A., Stelling, J., Price, N. D., Klamt, S., Schuster, S., & Palsson, B. O. (2004). Comparison of network-based pathway analysis methods. Trends in biotechnology, 22(8), 400-405.

Pereto J, Lopez-Garcia P, Moreira D: Ancestral lipid biosynthesis and early membrane evolution. Trends Biochem Sci 2004, 29(9):469–477.

Poot-Hernandez, A. C., Rodriguez-Vazquez, K., & Perez-Rueda, E. (2015). The alignment of enzymatic steps reveals similar metabolic pathways and probable recruitment events in Gammaproteobacteria. *BMC genomics*, *16*(1), 957.

Poot-Hernández, A. C., Rodríguez-Vázquez, K., Hernández-Guerrero R., & Pérez-Rueda, (2019). De los genomas al metabolismo celular: el estudio sistemático y comparativo del metabolismo. SMBB. 335-358.

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome biology, 20(1), 1-23.

Rahman, S. A., Cuesta, S. M., Furnham, N., Holliday, G. L., & Thornton, J. M. (2014). EC-BLAST: a tool to automatically search and compare enzyme reactions. Nature Methods, 11(2), 171–174. doi:10.1038/nmeth.2803

Ron Y. Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, Michal Ziv-Ukelson (2005). Alignment of metabolic pathways, Bioinformatics, Volume 21, Issue 16, , Pages 3401–3408, https://doi.org/10.1093/bioinformatics/bti554

Tipton, K., & McDonald, A. (2018). A Brief Guide to Enzyme Nomenclature and Classification.

Torsvik, V., Øvreås, L., & Thingstad, T. F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. Science, 296(5570), 1064-1066.

Tohsato, Y., Matsuda, H., & Hashimoto, A. (2000, August). A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In ISMB (Vol. 8, pp. 376-383).

Tu, S. L., Sughrue, W., Britt, R. D., & Lagarias, J. C. (2006). A conserved histidine-aspartate pair is required for exovinyl reduction of biliverdin by a cyanobacterial phycocyanobilin: ferredoxin oxidoreductase. Journal of Biological Chemistry, 281(6), 3127-3136.

Scossa, F., & Fernie, A. R. (2020). The evolution of metabolism: How to test evolutionary hypotheses at the genomic level. Computational and Structural Biotechnology Journal.

Shahbabian, K., A. Jamalli, L. Zig & H. Putzer, (2009). RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis. EMBO J 28: 3523-3533.

Silvela, J., & Portillo, J. (2001). Breadth-first search and its application to image processing problems. *IEEE Transactions on Image Processing*, *10*(8), 1194-1199.

Singh, R., Kumar, M., Mittal, A., & Mehta, P. K. (2016). Microbial enzymes: industrial progress in 21st century. 3 Biotech, 6(2), 174.

Staley, J. T., & Caetano-Anollés, G. (2018). Archaea-First and the Co-Evolutionary Diversification of Domains of Life. BioEssays, 40(8), 1800036.

Stams, A. J., & Plugge, C. M. (2009). Electron transfer in syntrophic communities of anaerobic bacteria and archaea. Nature Reviews Microbiology, 7(8), 568-577.

Wang, M., Boca, S. M., Kalelkar, R., Mittenthal, J. E., & Caetano-Anollés, G. (2006). A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. Complexity, 12(1), 27–40.

Weng, J. K. (2014). The evolutionary paths towards complexity: a metabolic perspective. *New Phytologist*, *201*(4), 1141-1149.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., & Gough, J. (2009). SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic acids research*, *37*(suppl_1), D380-D386.

Wu, F., & Minteer, S. (2015). Krebs cycle metabolon: structural evidence of substrate channeling revealed by cross-linking and mass spectrometry. *Angewandte Chemie International Edition*, *54*(6), 1851-1854.

Yamada, T., & Bork, P. (2009). Evolution of biomolecular networks—lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*, *10*(11), nrm2787.

Yokobori, S. I., Nakajima, Y., Akanuma, S., & Yamagishi, A. (2016). Birth of archaeal cells: molecular phylogenetic analyses of G1P dehydrogenase, G3P dehydrogenases, and glycerol kinase suggest derived features of archaeal membranes having G1P polar lipids. Archaea, 2016.