UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

ANÁLISIS COMPARATIVO DE LAS RUTAS DE BIOSÍNTESIS DE AMINOÁCIDOS: UNA PERSPECTIVA GENÓMICA

T E S I S

QUE PARA OBTENER EL GRADO DE
DOCTORA EN CIENCIAS BIOQUÍMICAS
PRESENTA

M. EN C. GEORGINA HERNÁNDEZ MONTES

DIRECTOR DE TESIS: DR. LORENZO PATRICK SEGOVIA FORCELLA

CUERNAVACA MORELOS **2009**

AGRADECIMIENTOS

Quiero agradecer muy especialmente a la Universidad Nacional Autónoma de México por abrirme sus puertas desde el bachillerato, por haberme dado la oportunidad de conocer el mundo a través de sus ojos y por permitirme ser miembro de su comunidad en donde se puede crecer no solo como profesionista sino como ser humano.

Al Consejo Nacional de Ciencia y Tecnología por la beca con numero de registro 144886 otorgada para realizar este trabajo.

Quiero a gradecer a las personas de docencia que siempre me ayudaron con toda la" tramitología" Maribel, Gloria y Jalil.

A las personas de cómputo que más de una vez me salvaron de perder todo mi trabajo.

Quiero agradecer al Dr. Lorenzo Segovia por la oportunidad de pertenecer a su grupo, por la confianza que siempre tuvo en mí, por el apoyo y la guía durante todo este tiempo.

También quiero agradecer al Dr. Ernesto Pérez Rueda por toda su ayuda, su paciencia, sus valiosos comentarios que enriquecieron este trabajo y por su amistad.

A los miembros de mi comité tutoral Lorenzo, Ernesto y particularmente a la Dra. Alejandra Covarrubias quien siempre nos puso los pies en la tierra y nos ayudo a enfocar el trabajo.

A los miembros del jurado de mi tutoral ampliado, revisión de tesis y jurado en mi examen de grado: Dr. Enrique Merino, Dr. Guillermo Gosset, Dr. Sergio Encarnación y Dra. Rosa María Gutiérrez.

Quiero agradecer a los chavos del laboratorio: Areli, Viviana, Mariana, Iliana, Adriana, Lorena, Alejandro, Javier, Arcadio y Dago quienes siempre hicieron un ambiente muy agradable en donde trabajar.

Quiero hacer dos agradecimientos especiales uno a Javier D-M quien se involucro en el proyecto y cuya aportación fue muy valiosa y determinante para que el trabajo pudiera concluirse; y a Alex por resolver mis dudas computacionales, por sus observaciones a mi trabajo y por brindarme su amistad.

Finalmente quiero agradecer a todas las personas que de una u otra forma me ayudaron a concluir esta etapa de mi vida.

Dedicatoria

A Javier a quien tengo tantas cosas que agradecer que no me alcanzan las palabras para decirle todo lo que significa en mi vida. Gracias por ser la fortaleza, la nobleza y el amor que me has brindado todo este tiempo.

A mis dos bebes quienes son el regalo más grande, cuando nacieron ustedes yo también volví a nacer.

A mis padres por darme la vida y la oportunidad de crecer y volar y a mis hermanos por enriquecer mi vida. Gracias por todo el apoyo que me han brindado.

A la familia Izquierdo Sánchez, por su apoyo y su solidaridad.

INDICE

RESUMEN

Antecedentes: Las proteínas están constituidas principalmente por 20 aminoácidos que podrían considerarse universales. Sin embargo sus rutas de biosíntesis no parecen ser universales tomando como modelo a la bacteria *Escherichia coli*.
Para entender su origen y evolución es preciso incluir mas especies modelo y las rutas alterativas de cada vía.
En este trabajo se hizo un análisis genómico comparativo para estudiar el origen y evolución de las redes de biosíntesis de aminoácidos.

Resultados: Analizando la distribución taxonómica de las enzimas de la biosíntesis de aminoácidos, se determinó un grupo de redes ampliamente distribuidas para al menos 16 de los 20 aminoácidos estándar. Este resultado sugiere que estas rutas pudieron haber existido en células ancestrales antes de la separación de los tres dominios de la vida. Adicionalmente describimos la distribución de dos tipos de ramas alternas a este grupo: enzimas análogas, que catalizan la misma reacción y que utilizan el mismo metabolito pero que pertenecen a diferentes superfamilias; y alternólogos, que aquí se definen como ramas que proceden de diferentes metabolitos convergiendo en el mismo producto final. Nosotros sugerimos que el origen de rutas alternativas está estrechamente relacionado a las diferentes fuentes de metabolitos y estilos de vida entre las especies.

Conclusión: La estrategia de emplear diferentes semillas de búsqueda que se utilizó en este trabajo mejora la calidad de los datos para determinar las relaciones evolutivas entre las rutas de biosíntesis de aminoácidos. Esta estrategia puede utilizarse para estudiar otras rutas metabólicas y otros procesos biológicos.
Adicionalmente, introdujimos el concepto de alternólogo, el cual no solo juega un papel importante en las relaciones entre estructura y función en redes biológicas, sino que como mostramos aquí, tiene fuertes implicaciones para la evolución al mismo nivel que los conceptos de analogía y paralogia.

ABSTRACT

Background: Twenty amino acids comprise the universal building blocks of proteins. However, their biosynthetic routes do not appear to be universal from an *Escherichia coli*-centric perspective. Nevertheless, it is necessary to understand their origin and evolution in a global context, that is, to include more 'model' species and alternative routes in order to do so. We use a comparative genomics approach to assess the origins and evolution of alternative amino acid biosynthetic network branches.

Results: By tracking the taxonomic distribution of amino acid biosynthetic enzymes, we predicted a core of widely distributed network branches biosynthesizing at least 16 out of the 20 standard amino acids, suggesting that this core occurred in ancient cells, before the separation of the three cellular domains of life. Additionally, we detail the distribution of two types of alternative branches to this core: analogs, enzymes that catalyze the same reaction (using the same metabolites) and belong to different superfamilies; and 'alternologs', herein defined as branches that, proceeding via different metabolites, converge to the same end product. We suggest that the origin of alternative branches is closely related to different environmental metabolite sources and life-styles among species.

Conclusion: The multi-organismal seed strategy employed in this work improves the precision of dating and determining evolutionary relationships among amino acid biosynthetic branches. This strategy could be extended to diverse metabolic routes and even other biological processes. Additionally, we introduce the concept of 'alternolog', which not only plays an important role in the relationships between structure and function in biological networks, but also, as shown here, has strong implications for their evolution, almost equal to paralogy and analogy.

INTRODUCCION

En 1938 Alexander Ivanovich Oparin (Oparin 1938) publicó un trabajo donde propuso una teoría acerca del origen de la vida. En este trabajo el plantea la formación de una sopa primitiva en la cual existen grandes cantidades de compuestos orgánicos sintetizados sin oxígeno a través de la acción de la luz solar a partir de compuestos inorgánicos. Posteriormente estas moléculas se combinarían de una forma cada vez más compleja hasta quedar disueltas en una gota de lo que el llamo coacervados. Estos coacervados crecerían por fusión con otras y se reproducirían mediante fisión en coacervados hijos, y de ese modo podrían haber obtenido un metabolismo primitivo en el que estos factores asegurarían la supervivencia de la "integridad celular", es decir sería la transición de heterotrofía a autotrofía.

Posteriormente en 1953 Stanley Miller retomó las ideas de Oparin y demostró que se pueden formar moléculas orgánicas de forma espontánea a partir de compuestos inorgánicos. Miller junto con su director de tesis Harold Urey (Miller 1952) diseñó un experimento en el que simularon algunas condiciones de la atmósfera de la tierra primitiva y que consistió en mezclar en un matraz metano, amoniaco e hidrógeno con vapor de agua para simular la evaporación de los océanos y sometió esta mezcla a descargas eléctricas. Como resultado se encontraron varios compuestos que eran principalmente aminoácidos, hidroxiácidos, ácidos alifáticos pequeños y urea. La glicina y alanina fueron los principales productos de esta reacción, sin embargo hay algunos otros aminoácidos que fueron sintetizados a muy bajas concentraciones tales como valina, leucina, isoleucina, ácido aspártico, ácido glutámico, serina, prolina y treonina.

Experimentos posteriores, junto con el hallazgo de compuestos orgánicos de origen extraterrestre en el meteorito de Murchinson (Kvenvolden, K. *et al* 1970) dio sustento a la idea de que una síntesis abiótica similar pudo haber ocurrido en el origen de la vida.

Basado en estos estudios se ha propuesto que la síntesis de aminoácidos representa uno de los estados mas tempranos en la evolución molecular, por lo que muchos trabajos se han enfocado a entender y proponer ideas acerca de como pudo haber evolucionado la síntesis de esos compuestos.

ANTECEDENTES

Teorías de evolución metabólica

El origen y la evolución de las rutas metabólicas representaron los pasos cruciales en la evolución de los primeros organismos, ya que les permitió ser independientes de fuentes externas de nutrientes. Para explicar como pudo haberse llevado a cabo este proceso, se han planteado varias teorías. A continuación mencionaré las dos teorías más representativas.

Evolución retrograda

En 1945 Horowitz propuso que las rutas metabólicas evolucionan de una forma retrograda. En esta hipótesis se propone la existencia de un gen "X" que sufre una duplicación y posterior especialización en una nueva reacción conservando la especificidad por el sustrato. Inicialmente este proceso pudo dar como resultado una vía de un solo paso y posteriormente este proceso se repitió hasta formar rutas mucho más grandes. Existen algunos ejemplos en las rutas actuales que parecen haber evolucionado de esta forma; tal es el caso de los últimos pasos de la ruta de biosíntesis de triptófano, dos pasos en la ruta de metionina y dos pasos en la síntesis de histidina.

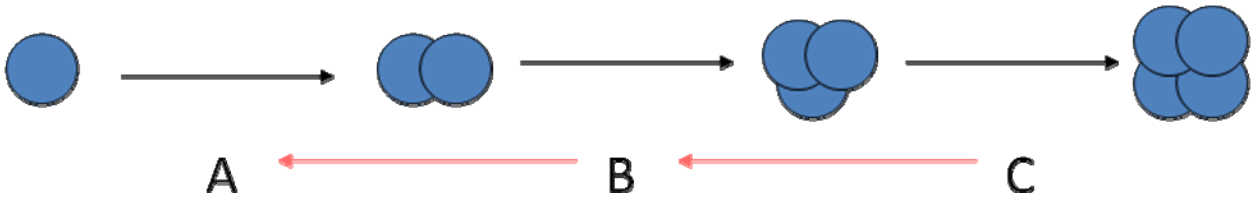

Figura 1. Las figuras representan compuestos que a través de las reacciones se vuelven mas complejos, mientras que las letras representan las enzimas que van depletandose del medio y van dando origen a las nuevas enzimas

Evolución por reclutamiento

En 1976 Jensen propone que las rutas metabólicas son el resultado del reclutamiento de una serie de enzimas que tienen baja especificidad catalítica y que pueden reaccionar con una gran variedad de sustratos químicamente relacionados. Sin embargo, esta hipótesis no se puede aplicar a un estadio temprano de la evolución ya que existían pocas enzimas. Este modelo ha sido

aplicado a varias rutas metabólicas, tales como la glucólisis, el ciclo de Krebs y la degradación de pentaclorofenol.

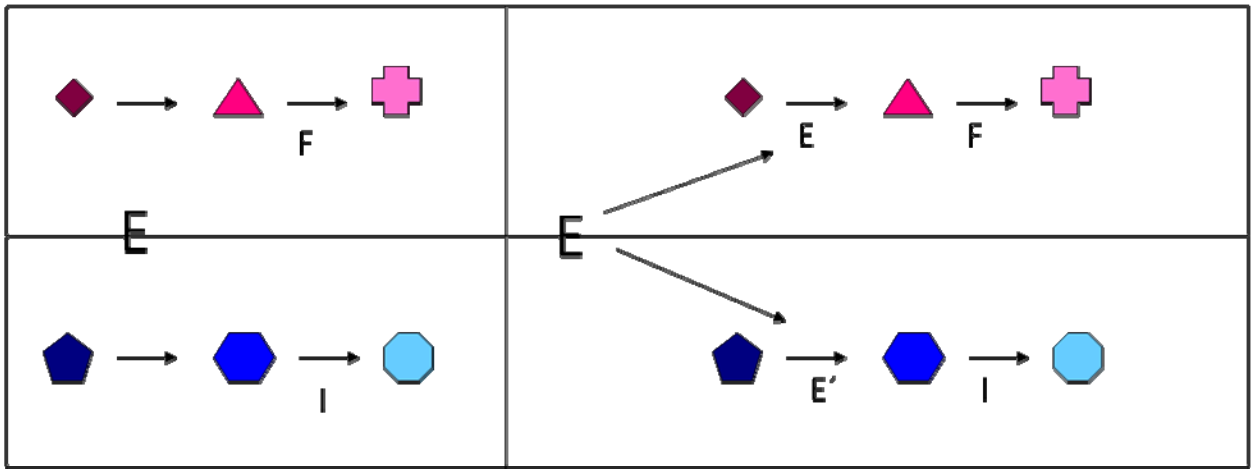

Figura 2. Las figuras representan los compuestos y las letras las enzimas. Se tienen dos escenarios, el primero muestra dos vías diferentes que comparten una misma enzima, el segundo muestra las mismas vías pero ahora hay una adaptación de la enzima a cada una de las vías.

Alternativamente, Lazcano y Miller (1999) propusieron que en un principio las vías metabólicas eran parcial o completamente no enzimáticas, que había reacciones termodinámicamente favorables, sin embargo una vez que se originaron enzimas capaces de llevar a cabo esa reacción, esa enzima podía o no ser reclutada. La biosíntesis de histidina apoya este modelo.

Por mucho tiempo se trato de averiguar a que modelo evolutivo correspondían las vías metabólicas y se encontró que la mayoría de los datos apuntan a que el reclutamiento es el modelo mejor representado, y que este modelo era antagónico al modelo de evolución retrograda. Sin embargo en un análisis reciente sobre la evolución del metabolismo por duplicación desde una perspectiva de redes, Díaz-Mejía J. *et al* 2007 demostraron que usando un enfoque de redes es posible obtener una vision global del mecanismo de evolución del metabolismo, lo que les permite sugerir que en realidad los modelos de reclutamiento y de evolución retrograda no son antagónicos, sino complementarios.

Con base en estos modelos se han realizado diversos estudios de los cuales mencionaremos los más relevantes y que fueron la base para realizar el presente trabajo.

Rutas metabólicas

La mayoría de las vías han sido bien descritas desde el punto de vista bioquímico y genético, principalmente en enterobacterias. En la figura 3 se muestra un mapa donde están representadas las rutas metabólicas como se conocen en *Escherichia coli* y en otras enterobacterias. De acuerdo con esta figura, la síntesis de aminoácidos se puede organizar en varios grupos. A continuación se describirá brevemente la información más relevante de estas vías.

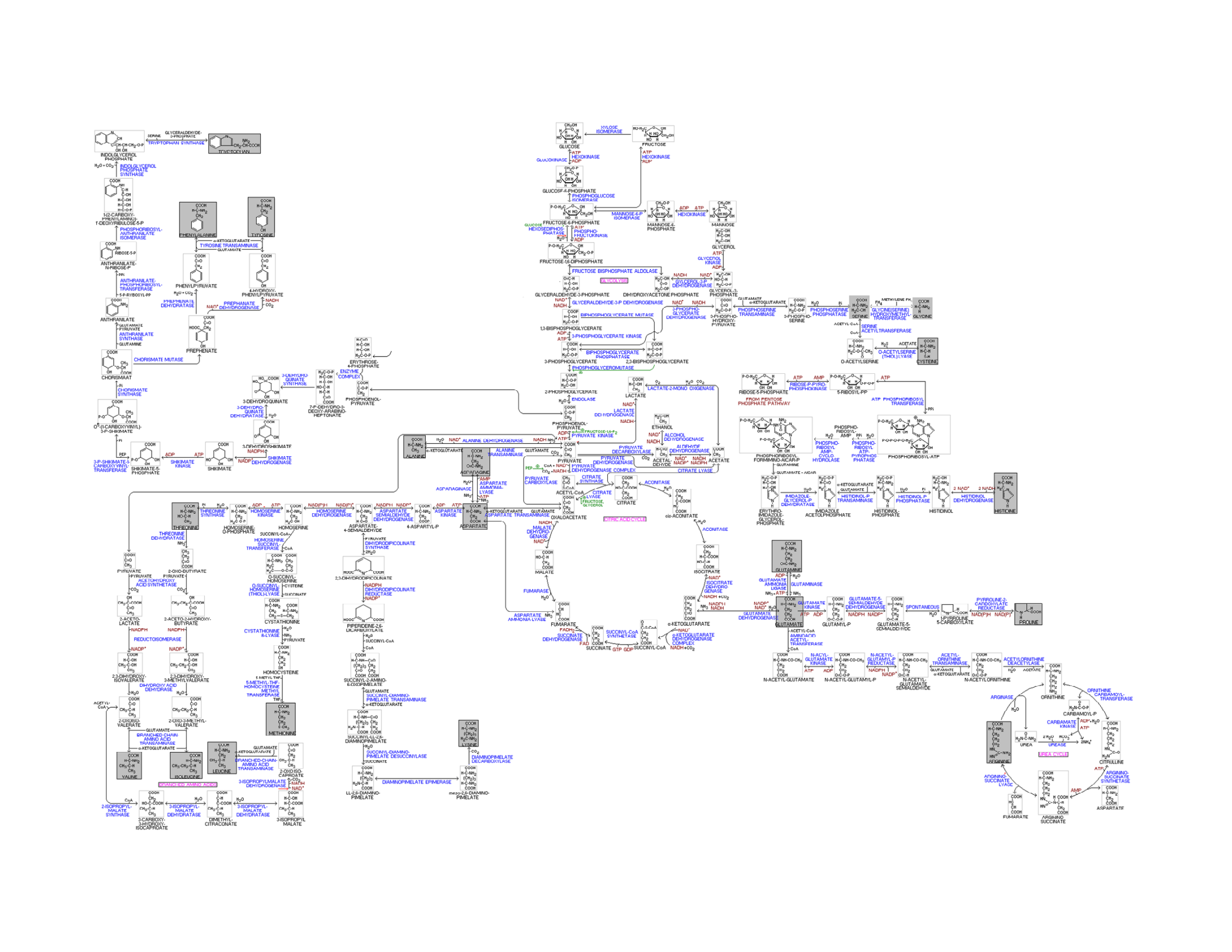

Figura 3. Mapa de las rutas de biosíntesis de aminoácidos conectadas a rutas del metabolismo central. Los 20 aminoácidos se encuentran resaltados en gris. Figura tomada y modificada de internet. http://home.planet.nl/~pvsanten/mmp/mmp.html

Biosíntesis de corismato

Aunque el corismato no es un aminoácido se incluye en el estudio por que es un intermediario muy importante. La vía del corismato, también conocida como ruta del shikimato, une el metabolismo de carbohidratos con la biosíntesis de compuestos aromáticos y se encuentra principalmente en bacterias, plantas, hongos y algunos protozoarios, sin embargo recientemente se reportó la presencia de enzimas de esta vía en algunos metazoarios (Starcevic A, *et al* 2008). Esta vía consta de 7 pasos enzimáticos que requieren la participación de las siguientes enzimas: AroF (EC:2.5.1.54), AroB (EC:4.2.3.4), AroD (EC:4.2.1.10), AroE (EC:1.1.1.25), AroK (EC:2.7.1.71), AroA (EC:2.5.1.19) y AroC (EC:4.2.3.5). Esta vía es de gran importancia, no solo científica sino comercial, ya que se considera que todos los intermediarios sirven como punto de partida para la síntesis de productos secundarios (Herrmann K.M y Weaver L.M 1999).

Biosíntesis de L-triptófano

La vía mejor estudiada es la vía de la biosíntesis de triptófano. Este aminoácido es esencial en mamíferos, y es sintetizado por procariotes, plantas y algunos eucariotes. Se propone que este aminoácido fue de los últimos que aparecieron durante la síntesis prebiótica de aminoácidos (Trifonov 2000), también se ha propuesto como una ruta presente en el último ancestro común de todos los seres vivos (Xie G. *et al* 2003).

La ruta del triptofano es definida como una ruta no ramificada que empieza con corismato y produce triptofano. Es considerada como muy costosa energéticamente ya que requiere la entrada de eritrosa-4-fosfato, ATP, fosforibosilfosfato, dos moléculas de fosfoenolpiruvato, L-glutamina y L-serina. Consta de 6 pasos enzimáticos y 5 enzimas: antranilato fosforibosil transferasa (TrpD, EC:4.1.3.27 EC:2.4.2.18), antranilato sintasa (TrpE, EC:4.1.3.27), fosforibosilantranilato isomerasa-indol-glicerol fosfato sintasa (TrpC, EC:5.3.1.24, EC:4.1.1.48), triptófano sintasa (TrpB, EC:4.2.1.20), indolglicerol fosfato aldolasa(TrpA, EC:4.2.1.20).

Biosíntesis de L-fenilalanina y L-tirosina

Las vías de biosíntesis de fenilalanina y tirosina requieren de 3 enzimas corismato mutasa-prefenato deshidratasa (PheA, EC:5.4.99.5, EC:4.2.1.51), corismato mutasa-prefenato deshidrogenasa (TyrA, EC:5.4.99.5, EC:1.3.1.12), y aminoácido aromático aminotransferasa (TyrB, EC:2.6.1.57), así como de tres pasos enzimáticos, de los cuales solo dos son compartidos. El primer paso es la conversión de corismato a prefenato por medio de TyrA y PheA. En el segundo paso el prefenato es convertido a fenilpiruvato por PheA en el caso de fenilalanina, y en la síntesis de tirosina el prefenato es convertido a p-hidroxifenilpiruvato por TyrA. El último paso de la vía es una transaminación catalizada por TyrB.

Biosíntesis de L-histidina

La histidina es uno de los aminoácidos que participan en los sitos catalíticos de las proteínas debido a que el grupo imidazol de su cadena lateral es muy reactivo.

La biosíntesis de histidina tiene un papel importante en el metabolismo central ya que está interconectada con las rutas de síntesis de *novo* de purinas y con reacciones del metabolismo de nitrógeno, sin embargo esta ruta también es muy costosa desde el punto de vista energético (Fani R. *et al* 1995). Estudios anteriores a nivel genético sugieren que esta vía pudo haber estado presente en el último ancestro común de todos los seres vivos (Fani R. *et al* 1995). Esta vía consta de 10 reacciones que son llevadas a cabo por 8 enzimas ATP fosforribosiltransferasa (HisG, EC:2.4.2.17), fosforibosil-ATP-pirofosfatasa y fosforibosil-AMP ciclohidrolasa (HisI, EC:3.6.1.31 EC:3.5.4.19), N-(5'-fosfo-L-ribosil-formimino)-5-amino-1-(5'-fosforibosil)-4-imidazolcarboxamida isomerasa (HisA, EC:5.3.1.16), imidazol glicerol fosfato sintasa (HisF e HisH, EC:2.4.2), imidazolglicerol fosfato deshidratasa (HisB, EC:4.2.1.19, EC:3.1.3.15), histidinol fosfato aminotransferasa (HisC, EC:2.6.1.29) e histidinol deshidrogenasa (HisD, EC:1.1.1.23).

Biosíntesis de L-valina, L-leucina e L-isoleucina

La biosíntesis de valina, leucina e isoleucina se lleva a cabo por un conjunto de reacciones paralelas. La vía empieza con piruvato o con 2-α-cetobutirato. Utiliza cuatro enzimas para la biosíntesis de isoleucina y valina. La leucina es formada a partir de un intermediario de la biosíntesis de valina, el oxoisovalerato, y requiere de 4 enzimas. Reportes previos señalan que

estas vías tienen un origen común muy cercano al origen de la vida (Xing R. 1991).

Biosíntesis de L-lisina, L-treonina y L-metionina

La biosíntesis de lisina, treonina y metionina están estrechamente relacionadas en bacterias ya que comparten al aspartato como precursor y también los tres pasos iniciales en la vía.

Para la biosíntesis de lisina se conocen 2 diferentes rutas anabólicas, la ruta del diaminopimelato y la del ácido α aminoadípico. La ruta del diaminopimelato consta de 7 reacciones catalizadas por 7 enzimas (dihidropicolinato sintasa (DapA, EC:4.2.1.52), dihidropicolinato reductasa (DapB, EC:1.3.1.26), tetrahidrodipicolinato succinilasa (DapD, EC:2.3.1.117), N-succinildiaminopimelato transferasa (DapC, EC:2.6.1.17), N-succinil-L-diaminopimlato desuccinilasa (DapE, EC:3.5.1.18), diaminopimelato epimerasa (DapF, EC:5.1.1.7) y diaminopimelato descarboxilasa (LysA, EC:4.1.1.20)) y se ha caracterizado principalmente en enterobacterias. La ruta del ácido aminoadipico ha sido reportada en bacterias termófilas (Velasco A.M. *et al* 2002) y consta de 5 pasos enzimáticos.

La síntesis de treonina a partir de homoserina requiere de dos pasos enzimáticos catalizados por homoserina cinasa (ThrB, EC:2.7.1.39) y treonina sintasa (ThrC, EC:4.2.3.1).

La metionina es el aminoácido universal en el N-terminal de las proteínas. La biosíntesis de metionina parte de homoserina y consiste de cuatro pasos catalizados por enzimas. Los pasos dos y tres están catalizados por la cistationina sintasa (MetB, EC:2.5.1.48) y la cistation liasa (MetC, EC:4.4.1.8), esas dos enzimas muestran una fuerte similitud (Belfaiza 1986), no solo a nivel de secuencia sino en cuanto a su plegamiento (Clausen *et al* 1998). Adicionalmente se han reportado dos rutas alternativas, la ruta de transulfuración donde la cistationina actúa como un intermediario y utiliza la cisteína como fuente de azufre y la segunda ruta donde hay una sulfidrilación donde la cistationina usa azufre inorgánico. El uso de derivado S-adenosilmetionina en una variedad de reacciones sugiere la importancia de la metionina en el metabolismo celular.

Biosíntesis de L-cisteína

La cisteína tiene un papel importante en la estructura de proteínas, así como en la estabilidad, función catalítica y es la principal fuente de azufre para sintetizar otros compuestos. Se han

descrito 3 rutas para la síntesis de este compuesto, sin embargo, parece que las arqueas son muy diversas para esta vía (White *et al* 2003). En eubacterias esta vía consta de 2 pasos enzimáticos y usa 3 enzimas (serina acetiltransferasa (CysE, EC:2.3.1.30), cisteína sintasa (CysM, EC:2.5.1.47) y cisteína sintasa (CysK, EC:2.5.1.47)) cuando se sintetiza a partir de serina.

Biosíntesis de L-prolina

La prolina es un aminoácido cíclico no polar que le confiere una estructura particular a las proteínas. La biosíntesis de prolina tiene como precursor el glutamato, costa de 4 pasos y tres enzimas (glutamato-5-semialdehído deshidrogenasa (ProA, EC:1.5.1.2), glutamil cinasa ProB, EC:2.7.2.11) y pirrolina-5-carboxilato reductasa (ProC, EC:1.2.1.41)).

Biosíntesis de L-arginina

La ruta de la biosíntesis de arginina consta de 8 pasos catalizados por 8 enzimas, inicia con la acetilacion del glutamato por la enzima ArgA y termina con la conversión de la ornitina en L-arginina en un ciclo común que se conoce como el ciclo de la urea. Esta vía ha sido encontrada en Enterobacterias, *Mixococcus xantus* y la arquea *Sulfolobus solfataricus* (Xu Y *et al* 2000).

Debido a que el aspartato, asparagina, glutamato, glutamina y glicina son aminoácidos que requieren uno o dos pasos para sus síntesis, no se han hecho estudios evolutivos sobre su biosíntesis, sino sobre su importancia en el metabolismo ya sea como precursores, reguladores o donadores.

Estudios genómicos

La forma de abordar el estudio del metabolismo cambio de manera radical después de que fue posible obtener la secuencia genómica de una gran cantidad de microorganismos, debido a que con toda la información genómica se podían integrar los aspectos genéticos, metabólicos, estructurales y funcionales así como factores ambientales.

Con esta nueva información se realizaron varios trabajos tratando de buscar las características generales de todos los organismos secuenciados. Uno de los trabajos más sobresaliente fue el que

publicaron Jeon H y colaboradores en el año 2000. En este estudio se analizó la topología de las redes metabólicas de 43 organismos y ellos demostraron que el metabolismo mas allá de ser una red aleatoria, se comporta como una red libre de escala, es decir, tiene nodos altamente conectados. En un estudio posterior del mismo grupo se hicieron análisis complementarios y fue posible identificar claramente módulos organizados de manera jerárquica dentro de esta red de libre escala (Barabasi L. *et al* 2002). Desde el punto de vista genómico, la aportación mas importante fue la noción de que todos los organismos analizados tienen la misma organización independientemente de sus reacciones especie específicas (Jeon H *et al* 2000).

Otro trabajo que también fue importante en este nuevo enfoque del metabolismo es el que realizaron Teichmann S. y colaboradores en 2001. En este estudio se analizan las rutas metabólicas de moléculas pequeñas (SMM por sus siglas en inglés) en *E.coli*. Ellos hacen uso de toda la información bioquímica, estructural y de secuencia para obtener una imagen detallada de las relaciones evolutivas de las proteínas involucradas en las 106 SMM identificadas. En un estudio posterior comparan las proteínas del metabolismo de *E. coli* con las de *Saccharomyces cereviciae* y observan que cerca del 50% de las enzimas son comunes para ambos organismos y que han sido conservadas desde la separación de procariotes y eucariotes incluyendo las modificaciones a nivel regulatorio.

Finalmente en 2003, Cuinchillos y Lecointre hicieron un análisis evolutivo del metabolismo de aminoácidos desde una perspectiva cladista, para ello usan las propiedades de las enzimas tales como especificidad enzimática, función, especificidad de sustrato y familia funcional. Sus resultados dan un panorama general acerca de la aparición de la función y con ello hacen una cronología del orden en que pudo haber aparecido el metabolismo de cada aminoácido. Figura 4.

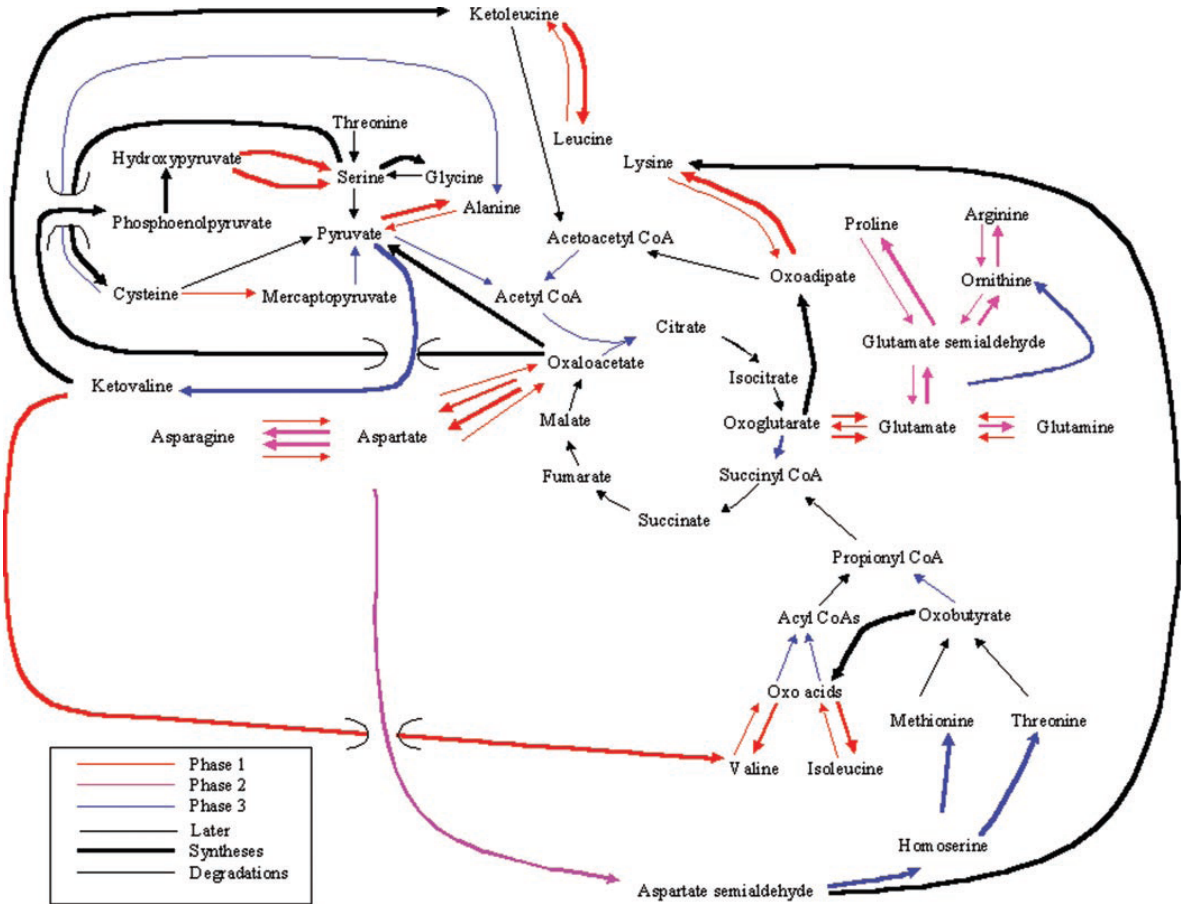

Figura 4. En esta figura se muestra una visión general de las rutas del metabolismo de aminoácidos conectados con el ciclo de krebs. Los colores indican épocas y se obtuvieron de arboles filogenéticos; el color rojo indica primera época, el color rosa segunda, el azul tercera y el negro indica que la aparacion fue posterior a tercera época, pero sin especificar si fueron simultáneas o sucesivas. Figura tomada de Cuinchillos C, y Lecointre G. 2003.

Pese a las aportaciones en el ámbito del metabolismo en general y en la síntesis de aminoácidos en particular, consideramos que las preguntas fundamentales acerca de cómo se originaron y evolucionaron las rutas de la biosíntesis de aminoácidos quedan aun sin resolver, por lo que decidimos llevar a cabo este trabajo de investigación.

PLANTEAMIENTO

Se ha propuesto que las rutas de biosíntesis surgieron como respuesta al agotamiento de los compuestos prebióticos en la fase heterotrófica de la vida, lo que les permitió a los primeros organismos ser menos dependientes de los nutrientes externos (Lazcano A. *et a*l 1999), por lo que se propone que el origen de estas rutas esta muy cercano al origen de la vida. Por otro lado se ha observado que estas rutas tienen la característica de ser más modulares que las rutas de catabolismo debido probablemente a que hay más restricciones termodinámicas o bioquímicas para sintetizar metabolitos que para romperlos (Snel N. *et al* 2004).

Este trabajo se enfoca en analizar la relación evolutiva de las enzimas de las rutas de biosíntesis de aminoácidos ya que consideramos que es un proceso metabólico muy antiguo y bien conservado entre los diferentes dominios celulares

JUSTIFICACION

Consideramos que las rutas de biosíntesis de amino ácidos son muy interesantes dado que

1) Se propone que los aminoácidos fueron de las primeras moléculas que se sintetizaron de manera prebiótica.
2) Se propone una estrecha relación entre la síntesis de aminoácidos y el origen y evolución del código genético. Una vez que ya se encontraban presentes los aminoácidos y se empezaron a “cristalizar” las primeras formas de vida, se propone que la síntesis de aminoácidos debió de acoplarse a otros procesos tales como la síntesis de proteínas.
3) Las rutas de síntesis de aminoácidos están relacionadas con vías del metabolismo central.
4) Se considera que las rutas de biosíntesis pueden organizarse en módulos.
5) Las proteínas de todos los organismos están formadas por aminoácidos, pero también participan en otras funciones tal como la regulación de algunas proteínas.

HIPÓTESIS

La aparición de las rutas de biosíntesis de aminoácidos se considera cercana al origen de la vida, y aunque se propone que algunas están bien conservadas, es posible encontrar una gran variabilidad en los diferentes organismos debido a sus propios procesos de divergencia, por lo que empleando múltiples secuencias semillas podremos inferir los probables eventos evolutivos asociados a ellas.

OBJETIVO

Estudiar la evolución de las vías de la biosíntesis de aminoácidos combinando la genómica comparativa y una perspectiva de redes.

MATERIAL Y METODOS

Distribución taxonómica

Se utilizó la secuencia de 537 dominios funcionales de las enzimas que participan en la biosíntesis de aminoácidos de diferentes organismos. Estas secuencias fueron buscadas en 410 genomas completamente secuenciados, para obtener la secuencia que se usó como semilla.

Para llevar a cabo al búsqueda de las semillas se usó BLASTP con un valor de corte $E=10^{-20}$ y un porcentaje de identidad >95%.

Cada secuencia genómica fue usada para la detección de ortólogos en 410 genomas, de los cuales 30 son arqueas, 363 son eubacterias, y 17 son eucariotes. Para el análisis se usó el criterio de mejor hit reciproco MHR (BRH) utilizando BLASTP con un valor de corte de 10^{-5} y una cobertura en el alineamiento de >50%. En este estudio no se incluyeron los genomas con menos de 1500 marcos abiertos de lectura que en su mayoria son parasitos obligados, ya que estos organismos han experimentado pérdidas secundarias de enzimas anabólicas y consideramos que introducirían ruido en la distribución taxonómica.

Secuencias en el mismo genoma con >95% de identidad estimada con CD-HIT (Li, W *et al* 2002) fueron organizadas en grupos. Como se ha reportado anteriormente, este procedimiento reduce la frecuencia de resultados falsos negativos causados por resultados cruzados entre secuencias muy similares dentro de un genoma. Debido a la redundancia de cepas secuenciadas para algunas especies bacterianas se depuraron de manera sistemática el grupo original de genomas, para obtener una medida normalizada de la distribución de ortólogos de acuerdo a los siguientes pasos.

En el paso 1 se construyó una distribución taxonómica (DT) para cada una de las enzimas *versus* la especie del organismo analizado, asignando el valor de 1 a las enzimas con ortologos (MHR) en $\geq$ 50% de cepas de cada especie. Por otro lado se asignó el valor de 0 a las enzimas con ortólogos en < 50. En el paso 2 se construyó una DT de enzimas versus género. En esta DT cada vector representa el porcentaje de especies que tienen el valor de 1 (asignado en el paso1) para cada género. En el paso 3 se construyo una DT de enzimas versus clados. En esta TD cada vector representa un promedio del porcentaje de género obtenido en el paso 2 para cada clado. Las clados corresponden a las categorías taxonómicas del KEGG (Kyoto Enciclopedia of Genes and

Genomes) Este procedimiento provee una promedio normalizado de distribución de enzimas a través de genomas.

Reconstrucción de redes

En las redes metabólicas bipartita hay dos tipos de nodos: enzimas y compuestos tales como sustratos, productos y cofactores. Los ejes relacionan a las enzimas con los compuestos que ocurren en una misma reacción. Por ejemplo, en una reacción R1 que consume el compuesto C1 y produce C2 y C3 y que es catalizada por la enzima E1, el proceso se establece de la siguiente manera: C1 → E1, E1 →C2 y E1 → C3. En reacciones reversibles se adiciona un segundo grupo de ligas de productos a enzimas y de regreso de enzimas a sustratos. En este trabajo se reconstruyó una red bipartita derivada de tres bases de datos metabólicos: EcoCyc v8.0 (Karp PD *et al* 2007) para *E.coli*, MjCyc (Tsoka S *et al* 2004) de *Methanococcus jannaschii* y MetaCyc v8.0 (Caspi R *et al* 2008) para asignaciones de múltiples organismos. Para obtener información concerniente a los nodos y a los ejes para cada reacción se usaron los archivos de EcoCyc y MetaCyc: reaction.dat (sustrato/producto), enzrxns.dat (reversibilidad) y reaction-links.dat (números EC). De MjCyc, la información correspondiente fue adquirida de forma manual de la base de datos de la página web. Redes derivadas de las bases de datos fueron preparadas para presentación con Cytoscape v2.5.2 (Shannon P *et al* 2003). Los aminoácidos se resaltaron (triángulos rojos en la figura) para denotar los puntos terminales de las ramas y los puntos terminales de las rutas. Para mayor claridad en la presentación los compuestos mas conectados (principalmente cofactores) y los compuestos terminales no aminoácidos fueron removidos de la red.

Las secuencias de enzimas multifuncionales fueron divididas manualmente de acuerdo a la asignación de su dominio funcional en la base de datos Swiss-Prot (Boeckmann B *et al* 2003).

En la figura cada nodo representa una reacción catalizada por un dominio funcional. Las enzimas análogas que catalizan la misma reacción pero poseen diferente plegamiento fueron identificadas por comparar el contenido del dominio estructural entre proteínas de acuerdo a la base de datos Superfamily v1.69 (Gough J *et al* 2001) usando el paquete HMMer (Eddy SR 1996). Los alternologos fueron identificados por inspección manual de la red de la figura 5, identificando ramas que proceden de diferentes metabolitos y que convergen en un mismo compuesto, generalmente en aminoácidos.

Bases de datos

Ecocy. Es una base de datos para la bacteria *Escherichia coli* K12 MG1655. (Karp PD. *et al* 2007). El grupo que participa en el proyecto Ecocyc realiza revisiones constantes de la información que se encuentra en la literatura acerca del genoma, de la regulación transcripcional, los transportadores y las rutas metabólicas. Esta base de datos también contiene una descripción de las redes celulares de *E.coli* así como resúmenes de sus genes.

MetaCyc. Es una base de datos de rutas metabólicas no redundantes y enzimas elucidadas experimentalmente. Contiene más de 1,100 rutas de más de 1500 organismos diferentes y es curada de la literatura experimental. (Caspi R. *et al* 2008). Esta base de datos también sirve como referencia para hacer predicciones de rutas metabólicas de organismos a partir de sus secuencias genómicas.

MjCyc. Base de datos de *Methanococcus jannaschii* (Tsoka S *et al* 2004)

Superfamily 1.69. Es una base de datos de anotaciones funcionales y estructurales de las proteínas de todos los genomas secuenciados. (Gough J. *et al* 2001)

KEGG. Es un conjunto de bases de datos de sistemas biológicos. Contiene información acerca genes y proteínas (KEGG GENES), de información acerca de sustancias endógenas y exógenas (KEGG LIGAND) de información acerca de redes de reacciones e interaccion (KEGG PATHWAY), y finalmente también tiene información de jerarquías funcionales que representan el conocimiento de los sistemas biológicos (KEGG BRITE)(Kaneshida M y Goto S. 2000).

Swiss-Prot. Fue creada en 1986 y es una base de datos de secuencias de proteínas que es curada manualmente. Esta base de datos provee información muy completa acerca de las secuencias de proteínas tal como descripción de función, dominio estructural, modificaciones post traduccionales, variantes etc. También tiene un nivel mínimo de redundancia y un alto nivel de integración con otras bases de datos (Boeckmann B *et al* 2003).

RESULTADOS Y DISCUSION

Distribución de las redes de biosíntesis de aminoácidos.

Para evaluar el origen y evolución de las rutas metabólicas de biosíntesis de aminoácidos se analizó la distribución taxonomica (DT) de las enzimas que catalizan las reacciones. La DT de cada enzima es un vector de distribución de ortólogos (presencias/ausencias) en un set de genomas o cladas (ver material y métodos). La base es que la DT provee información concerniente a la aparición relativa de enzimas, ramas y rutas durante la evolución del metabolismo.

Se determinó la DT para 537 dominios funcionales de enzimas que catalizan 188 reacciones en las rutas de biosíntesis de aminoácido de diferentes especies en un set de 410 genomas (30 arqueas, 363 bacteria y 17 eucarias).

Para obtener la DT se uso la siguiente estrategia:

A) Primero se obtuvieron 113 enzimas de la biosíntesis de aminoácidos de la bacteria *E. coli* K12 reportados en la base de datos EcoCyc(Karp PD. *et al* 2007),

B) Segundo, se analizaron los genomas para identificar los ortólogos (mejor hit reciproco) de las 113 enzimas de *E. coli*.

C) Tercero, se uso un segundo grupo de enzimas ortólogas, parálogas, análogas y ramas alternologas de diferentes especies definidas en las bases de datos MetaCyc (Caspi R. *et al* 2008) y MjCyc (Tsoka S *et al* 2004) para llenar los espacios en los DT basados en *E. coli*.

La figura 5 se muestra una red formada por 188 reacciones analizadas en este trabajo y el promedio de la distribución de ortólogos de las enzimas catalizantes. Se consideraron dos amplias categorías para la distribución de ortólogos: enzimas ampliamente distribuidas, cuya distribución de ortólogos es ≥ 50% a través de los clados analizados aquí y enzimas parcialmente distribuidas cuya distribución de ortólogos es < 50% a través de las cladas. La amplia distribución de enzimas, ramas y rutas sugiere su ocurrencia en el último ancestro común (LUCA), sin embargo esas categorías son simplemente una herramienta para los propósitos de presentación. Aun cuando una ruta muestra una baja distribución de ortólogos, algunas de esas ramas, pueden estar distribuidas a través de los tres dominios celulares y quizá esas ramas pudieron estar presentes en LUCA. El escenario opuesto puede tener lugar también, esto es, algunas enzimas pueden exhibir una amplia distribución pero

pueden estar restringidos a dominios celulares específicos o divisiones tales como Bacteria o las γ-proteobacterias que están sobrerepresentadas en secuencias genómicas.

Así una amplia distribución no necesariamente significa su ocurrencia en el LUCA. Por esas razones examinamos exhaustivamente las DT de las enzimas formando cada rama dentro de las rutas de biosíntesis de aminoácidos.

En las siguientes secciones, se describirán los resultados en orden decreciente de distribución ortóloga, enfatizando la posible existencia de algunas ramas en el LUCA.

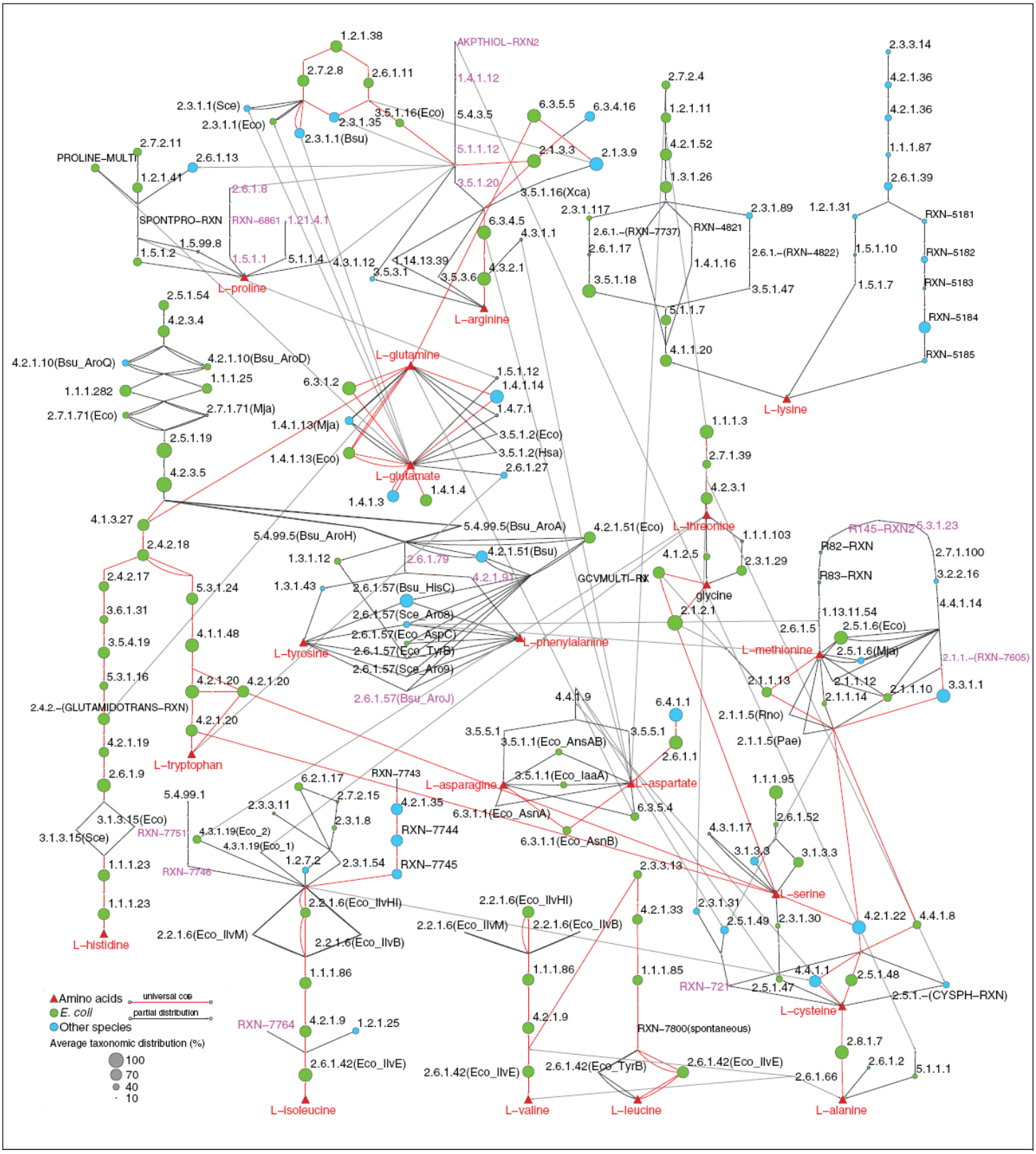
1.2.1.38
AKPTHIOL–RXN2
2.7.2.8
2.6.1.11
1.4.1.12
2.3.1.1(Sce)
2.3.1.1(Eco)
2.3.1.35
3.5.1.16(Eco)
5.4.3.5
6.3.5.5
6.3.4.16
2.3.1.1(Bsu)
2.7.2.11
PROLINE–MULTI
2.6.1.13
1.2.1.41
5.1.1.12
2.1.3.3
2.1.3.9
3.5.1.20
2.6.1.8
3.5.1.16(Xca)
SPONTPRO–RXN
RXN–6861
1.21.4.1
1.5.99.8
1.5.1.2
1.5.1.1
5.1.1.4
4.3.1.12
1.14.13.39
3.5.3.1
6.3.4.5
4.3.1.1
3.5.3.6
4.3.2.1
L–proline
L–arginine
2.5.1.54
4.2.3.4
4.2.1.10(Bsu_AroQ)
4.2.1.10(Bsu_AroD)
1.1.1.282
1.1.1.25
2.7.1.71(Eco)
2.7.1.71(Mja)
2.5.1.19
4.2.3.5
L–glutamine
6.3.1.2
1.5.1.12
1.4.1.14
1.4.7.1
1.4.1.13(Mja)
1.4.1.13(Eco)
3.5.1.2(Eco)
3.5.1.2(Hsa)
2.6.1.27
L–glutamate
1.4.1.3
1.4.1.4
4.1.3.27
2.4.2.18
5.4.99.5(Bsu_AroH)
5.4.99.5(Bsu_AroA)
4.2.1.51(Eco)
4.2.1.51(Bsu)
2.6.1.79
4.2.1.91
1.3.1.12
1.3.1.43
2.6.1.57(Bsu_HisC)
2.6.1.57(Sce_Aro8)
2.6.1.57(Eco_AspC)
2.6.1.57(Eco_TyrB)
2.6.1.57(Sce_Aro9)
2.6.1.57(Bsu_AroJ)
L–tyrosine
L–phenylalanine
2.4.2.17
5.3.1.24
3.6.1.31
3.5.4.19
4.1.1.48
5.3.1.16
4.2.1.20
4.2.1.20
2.4.2.–(GLUTAMIDOTRANS–RXN)
4.2.1.20
4.2.1.19
2.6.1.9
L–tryptophan
3.1.3.15(Eco)
3.1.3.15(Sce)
RXN–7751
5.4.99.1
4.3.1.19(Eco_2)
4.3.1.19(Eco_1)
1.1.1.23
RXN–7746
1.1.1.23
L–histidine
6.2.1.17
RXN–7743
2.7.2.15
2.3.3.11
2.3.1.8
4.2.1.35
RXN–7744
1.2.7.2
2.3.1.54
RXN–7745
2.2.1.6(Eco_IlvHI)
2.2.1.6(Eco_IlvM)
2.2.1.6(Eco_IlvB)
1.1.1.86
RXN–7764
4.2.1.9
1.2.1.25
2.6.1.42(Eco_IlvE)
L–isoleucine
4.4.1.9
3.5.5.1
3.5.1.1(Eco_AnsAB)
3.5.5.1
6.4.1.1
2.6.1.1
L–asparagine
3.5.1.1(Eco_IaaA)
L–aspartate
6.3.1.1(Eco_AsnA)
6.3.5.4
6.3.1.1(Eco_AsnB)
2.2.1.6(Eco_IlvHI)
2.2.1.6(Eco_IlvM)
2.2.1.6(Eco_IlvB)
1.1.1.86
4.2.1.9
2.6.1.42(Eco_IlvE)
L–valine
2.6.1.42(Eco_TyrB)
RXN–7800(spontaneous)
L–leucine
2.3.3.13
4.2.1.33
2.3.1.31
2.5.1.49
2.3.1.30
4.2.1.22
4.4.1.8
1.1.1.85
4.4.1.1
2.5.1.48
RXN–721
2.5.1.47
2.5.1.–(CYSPH–RXN)
L–cysteine
2.6.1.42(Eco_IlvE)
2.6.1.66
2.8.1.7
2.6.1.2
5.1.1.1
L–alanine
1.1.1.95
2.6.1.52
4.3.1.17
3.1.3.3
3.1.3.3
L–serine
2.7.2.4
1.2.1.11
4.2.1.52
1.3.1.26
2.3.1.117
2.3.1.89
2.6.1.–(RXN–7737)
RXN–4821
2.6.1.17
1.4.1.16
2.6.1.–(RXN–4822)
3.5.1.18
3.5.1.47
5.1.1.7
4.1.1.20
L–lysine
2.3.3.14
4.2.1.36
4.2.1.36
1.1.1.87
2.6.1.39
1.2.1.31
RXN–5181
1.5.1.10
RXN–5182
1.5.1.7
RXN–5183
RXN–5184
RXN–5185
1.1.1.3
2.7.1.39
4.2.3.1
L–threonine
4.1.2.5
1.1.1.103
2.3.1.29
GCVMULTI–RX
glycine
2.1.2.1
R145–RXN2
5.3.1.23
R82–RXN
2.7.1.100
R83–RXN
3.2.2.16
4.4.1.14
1.13.11.54
2.6.1.5
2.5.1.6(Eco)
L–methionine
2.5.1.6(Mja)
2.1.1.–(RXN–7605)
2.1.1.13
2.1.1.12
2.1.1.14
2.1.1.10
3.3.1.1
2.1.1.5(Rno)
2.1.1.5(Pae)
Amino acids
E. coli
Other species
universal core
partial distribution
Average taxonomic distribution (%)
100
70
40
10

Figura 5. Red de la biosíntesis de aminoácidos de múltiples especies. Los 20 aminoácidos están representados con triángulos rojos y se muestran al final de las rutas. Los círculos verdes representan las enzimas canónicas de *E.coli* . Los círculos azules representan enzimas alternativas (análogas o alternologas) de otras especies. El tamaño de los nodos corresponde al promedio normalizado de la distribución taxonómica de ortólogos para cada dominio enzimático (dominios en enzimas multimericas) que catalizan la reacción correspondiente. Un nodo grande implica una enzima de amplia distribución en los genomas. Los ejes rojos indican pasos que podrían ocurrir en LUCA de acuerdo con el DT de sus enzimas catalizantes. Los números EC de color púrpura corresponden a reacciones sin genes o enzimas conocidos. Figura tomada de artículo Hernández-Montes *et al* 2008.

Nueve rutas ampliamente distribuidas a través de los tres dominios de la vida.

L-arginina

Hay cuatro rutas de síntesis de arginina que interactúan con la conversión de L-ornitina y citrulina, sin embargo pueden ser agrupados dentro de dos supervías (figura 5). La primer supervía, que involucra carbamoil fosfato y N-acetyl-L-citrullina puede proceder mediante dos ramas alternólogas: la primer rama es la canónica de *E. coli* catalizada por dos enzimas ampliamente distribuidas, carbamoil fosfato sintetiza (EC:6.3.5.5) y ornitina carbamoil transferasa (EC:2.1.3.3). La segunda rama usa tres enzimas: EC:6.3.4.16, EC:2.1.39 y EC:3.5.1.16 de las cuales dos están ampliamente distribuidas. De manera interesante EC:6.3.5.5 y EC:6.3.4.16 son parálogos así como EC:2.1.3.3 y EC:2.1.39, representando un evento de retención de genes duplicados como grupos en lugar de entidades independientes. Se ha sugerido que la retención de grupos de duplicados tiene un papel importante en la evolución del metabolismo (Diaz-Mejía JJ *et al* 2007). Alternativamente, la segunda supervía se lleva a cabo vía N-acetyl-L-ornitina y también esta ampliamente distribuida a través de los tres dominios, con la excepción de animales. Esta supervía muestra tres DTs interesantes, primero cuando se usaron las enzimas de *E. coli* como semillas para el MHR en esta supervía, solamente se detectaron ortólogos en algunos clados, sin embargo cuando se usaron las secuencias de *S. cerevisae*, *M. jannaschii* y *Bacillus subtilis* como semillas, se detectaron los ortólogos faltantes en los demás grupos filogenéticos, demostrando así la importancia de usar la secuencia de diferentes organismos en lugar de usar estrategias *E.coli*-céntricas. Segundo, hay dos N-acetilglutamato sintasa. La que es tipo *E.coli* es una enzima monomérica y monofuncional, mientras que la del tipo de *B.subtilis* es una enzima heterodimérica bifuncional cuyos constituyentes son proteolíticamente autoprocesados a partir de un solo precursor. Ambos tipos de enzimas están ampliamente distribuidas a través

de los tres dominios, sin embargo la de *E. coli* no fue identificada en firmicutes, sugiriendo que este ha sido un desplazamiento por la del tipo de *B. subtilis*. Tercero, ocurre otra retención de genes duplicados como grupos, en lugar de cómo entidades independientes en tres pasos consecutivos en la biosíntesis de L-arginina/L-lisisna. EC:2.7.2.8/EC:2.7.2.4, EC:1.2.1.38/EC:1.2.1.11, EC:2.6.1.11/EC:2.6.1.17 y EC:3.5.1.16/EC:3.5.1.18 respectivamente.

A partir de estos resultados, proponemos que no todas las rutas para sintetizar L-arginina estuvieron presentes en LUCA, solamente los que proceden de la vía de N-acetyl-L-ornitina y citrulina.

L-glicina

Existen cuatro rutas para sintetizar L-glicina. Dos de ellas, involucran la degradación de L-treonina (figura 5) y están parcialmente distribuidas en Bacteria y Eucaria (figura 6). En contraste, las otras dos vías están interconectada a través de 5,10-metilen-tetrahidrofolato e involucran el sistema de corte de glicina o serina hidroximetiltransferasa (EC:2.1.2.1). Ambas ramas están ampliamente distribuidas en los tres dominios celulares. De hecho EC:2.1.2.1 es una de las enzimas mas ampliamente distribuidas a través de todas las especies, probablemente por que también participa en la síntesis de folato, otra ruta ampliamente distribuida. Colectivamente la distribución de esas enzimas sugiere que el LUCA sintetizo glicina vía la rama de 5,10-metilene-tetrahidrofolato.

L-triptofano

Las cinco enzimas de la ruta de biosíntesis de triprofano están ampliamente distribuidas en los tres dominios, confirmando reportes previos (Xie G. *et al* 2003). Sin embargo no se detectaron ortólogos para estas enzimas en los animales, con excepción de *Nematostella vectensis*, que es una cnidaria representativa de los estadios tempranos en la evolución de los animales (Putnam NH *et al* 2007). Esto indica que algunos animales tuvieron perdidas secundarias de las enzimas para la biosíntesis de L-triptofano y también explica por que este aminoácido es esencial para los humanos. Así, el LUCA probablemente fue capaz de sintetizar L-triptofano en forma similar a las especies contemporáneas.

L-prolina

Hay por lo menos seis ramas biosintéticas para la L-prolina. Tres de ellas convergen en L-glutamato γ–semialdehido y juzgando por su DT, la ornitina δ-aminotransferasa EC:2.6.1.13 es la enzima mas distribuida dentro de esta ruta, aun en los genomas de arqueas. Las otras 2 ramas han sido caracterizadas bioquímicamente, a pesar de que sus enzimas son desconocidas. La sexta rama la cual convierte directamente L-ornitina a L-prolina vía ornitina ciclodeaminasa EC:4.3.1.12, fue encontrada en algunas Arquea y escasamente en Bacteria y Eucaria. Sin embargo son necesarios mas análisis para corroborar experimentalmente las actividades de esos marcos abiertos de lectura arqueales, por que la enzima putativa EC:2.6.1.13 no tiene los residuos catalíticos canónicos involucrados en esta actividad, y se conoce poca información acerca de la actividad de EC:4.3.1.12. Es por ellos que la biosíntesis de L-prolina en arqueas permanece enigmática y hace difícil inferir las capacidad de síntesis de este aminoácido en el LUCA.

L-leucina

La biosíntesis de L-leucina es una ruta lineal que consiste de 5 reacciones (figura 5). Usando como modelo las secuencias de *E. coli* y *M. jannaschii*, se detectó que las enzimas putativas que catalizan las tres primeras reacciones están ampliamente distribuidas. Esas tres enzimas pertenecen a un grupo de genes duplicados que catalizan pasos consecutivos en la ruta de biosíntesis de L- lisina, L- leucina y L-isoleucina. Las relaciones evolutivas entre lisina y leucina han sido documentadas previamente (Nishida H. *et al* 1999, Irving SD. *et al* 1998 Velasco AM. *et al* 2002). Nosotros encontramos que la L-isoleucina también está involucrada en este fenómeno. Estos genes duplicados junto con los genes duplicados de arginina/lisina respaldan nuestro reporte previo sobre la importancia de la retención de grupos de genes duplicados en lugar de cómo entidades individuales en la evolución del metabolismo (Díaz-Mejia JJ *et al* 2007). La cuarta reacción se lleva a cabo de manera espontanea y no requiere enzima. De manera complementaria, el quinto paso de la ruta en *E. coli* es catalizado por una de las dos enzimas análogas de la aminotransferasa EC:2.6.1.42, una de ellas pertenece a la superfamilia de D-amino acido aminotransfersasa-like PLP-dependiente y está ampliamente distribuida a través de los tres dominios incluyendo los animales. En contraste la segunda enzima EC: 2.6.1.42 pertenece a las

transferasas PLP-dependientes y está escasamente distribuida en los genomas. De manera colectiva, esas observaciones sugieren que el LUCA fue capaz de sintetizar L-leucina como las especies contemporáneas. Sin embargo es necesario caracterizar los marcos abiertos de lectura ya que la leucina es un aminoácido esencial en humanos.

L-histidina

Estructuralmente hablando la histidina y el triptofano son similares, ambas rutas son principalmente lineares y divergieron del antranilato usando EC: 2.4.2.18 y dada su amplia distribución, estas rutas han sido propuestos como rutas ancestrales. La enzima histidinol-fosfatasa EC:3.1.3.15 es la única enzima de está ruta que esta parcialmente distribuida en los genomas. Esto es probablemente debido a la existencia de 2 enzimas análogas EC:3.1.3.15 (*S. cerevisae* y *E. coli*). Ambos tipos son altamente divergentes en secuencia, y cuando los parámetros de los análisis MHR se relajan (se incrementa el umbral de E de 10^{-6} a 10^{-1}) se detectaron ortólogos en el 84% y 40% de los genomas analizados con las secuencias de *S. cerevisae* y *E. coli*, respectivamente. Las otras enzimas analizadas en este estudio no son afectadas por los parámetros del análisis de MHR. Adicionalmente, se encontró que los animales con excepción de *N. vectensis* han experimentado una perdida secundaria de la maquinaria de síntesis de la L-histidina. Tomando esos resultados juntos, sugerimos que el LUCA tuvo la misma ruta de síntesis de L-histidina que tienen los organismos actualmente.

L-treonina

Dos de las tres enzimas de la síntesis de L-treonina fueron encontrados en los tres dominios de la vida. Cuando se usó la treonina sintasa de *E. coli* como semilla no se encontró ortólogo en Arquea, sin embargo cuando se usó como semilla un parálogo de *M. janaschii* con la misma función fue posible identificar ortólogos en Arquea. De nuevo esta información refuerza la importancia de usar enzimas de múltiples especies como semillas. Algunos animales aparentemente han perdido la maquinaria para este aminoácido, pero *N. vectensis* lo ha retenido. Nosotros sugerimos que LUCA podría sintetizar L-treonina como especies contemporáneas.

L-glutamina y L-glutamato

Como se muestra en la figura 5, la interconversion de L-glutamina en L-glutamato se puede llevar a cabo por muchas enzimas alternólogas. Tanto la glutamato sintasa NADH dependiente EC:1.4.1.14 y la glutamato sintasa NADPH dependiente EC:1.4.1.13 producen L-glutamato a partir de L-glutamina y están ampliamente distribuidas en los tres dominios de la vida. En la dirección opuesta, de L-glutamato a L-glutamina, encontramos que la glutamina sintasa EC: 6.3.1.2 también está altamente distribuida en los tres dominios. Esto sugiere que el LUCA fue capaz de interconvertir L-glutamina y L-glutamato. Pero esto deja abierta una nueva pregunta: ¿el LUCA fue capaz de producir esos aminoácidos de manera independiente unos de otros? De forma similar a la glutamato sintasa, los parálogos de la glutamato deshidrogenasa, el NAD (P)+-dependiente EC:1.4.1.3 y el NADP+-dependiente EC:1.4.1.4 que producen L-glutamato a partir de 2-oxoglutarato y amonio también están ampliamente distribuidos. Finalmente, todas las otras reacciones que sintetízan L-glutamina usan L-glutamato como sustrato y están escasamente distribuidas. A partir de estos datos es posible sugerir que el LUCA fue capaz de sintetizar L-glutamato de 2-oxoglutarato e interconvertir este con L-glutamina, pero es difícil determinar si el LUCA fue capaz de producir este último aminoácido de forma independiente al L-glutamato.

L-cisteína

Hay por lo menos cuatro formas de sintetizar L-cisteína. La ruta mas distribuida procede vía cistation usando cistation β-sintasa EC:4.2.1.22 y cistation γ-liasa EC:4.4.1.1 y aunque está documentada como una vía eucarionte, nosotros los hemos encontrado en los tres dominios. Alternativamente la cistationina β-liasa EC:4.4.1.8 cistation γ-sintasa EC:2.5.1.48 y la O-succinilhomoserina(tiol)-liasa EC:2.5.1.48 catalizan la reacción equivalente y están ampliamente distribuido en Bacteria y Eucaria. En contraste una rama alternóloga que usa EC:2.5.1.47 vía O-acetil-L-serina está escasamente distribuida en los genomas, mientras que otra rama sin enzima asignada (ni genes) usa O-acetil-L-homoserina. Esos resultados sugieren que no todas las vías reportadas pudieron haber existido en LUCA pero el tipo eucarionte si pudo haber estado en LUCA.

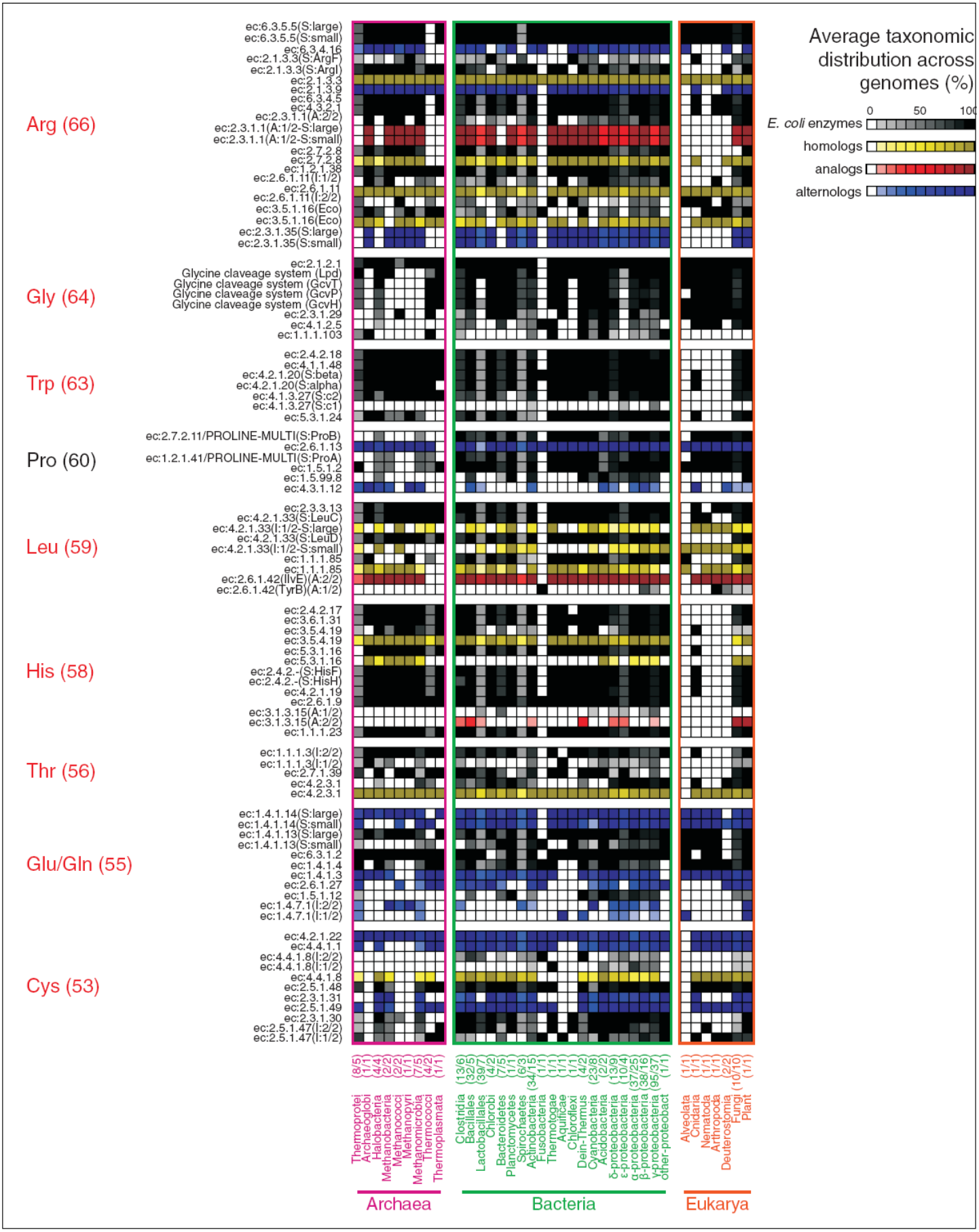
Arg (66)
ec:6.3.5.5(S:large)
ec:6.3.5.5(S:small)
ec:6.3.4.16
ec:2.1.3.3(S:ArgF)
ec:2.1.3.3(S:ArgI)
ec:2.1.3.3
ec:2.1.3.9
ec:6.3.4.5
ec:4.3.2.1
ec:2.3.1.1(A:2/2)
ec:2.3.1.1(A:1/2-S:large)
ec:2.3.1.1(A:1/2-S:small)
ec:2.7.2.8
ec:2.7.2.8
ec:1.2.1.38
ec:2.6.1.11(I:1/2)
ec:2.6.1.11
ec:2.6.1.11(I:2/2)
ec:3.5.1.16(Eco)
ec:3.5.1.16(Eco)
ec:2.3.1.35(S:large)
ec:2.3.1.35(S:small)
Gly (64)
ec:2.1.2.1
Glycine claveage system (Lpd)
Glycine claveage system (GcvT)
Glycine claveage system (GcvP)
Glycine claveage system (GcvH)
ec:2.3.1.29
ec:4.1.2.5
ec:1.1.1.103
Trp (63)
ec:2.4.2.18
ec:4.1.1.48
ec:4.2.1.20(S:beta)
ec:4.2.1.20(S:alpha)
ec:4.1.3.27(S:c2)
ec:4.1.3.27(S:c1)
ec:5.3.1.24
Pro (60)
ec:2.7.2.11/PROLINE-MULTI(S:ProB)
ec:2.6.1.13
ec:1.2.1.41/PROLINE-MULTI(S:ProA)
ec:1.5.1.2
ec:1.5.99.8
ec:4.3.1.12
Leu (59)
ec:2.3.3.13
ec:4.2.1.33(S:LeuC)
ec:4.2.1.33(I:1/2-S:large)
ec:4.2.1.33(S:LeuD)
ec:4.2.1.33(I:1/2-S:small)
ec:1.1.1.85
ec:1.1.1.85
ec:2.6.1.42(IlvE)(A:2/2)
ec:2.6.1.42(TyrB)(A:1/2)
His (58)
ec:2.4.2.17
ec:3.6.1.31
ec:3.5.4.19
ec:3.5.4.19
ec:5.3.1.16
ec:5.3.1.16
ec:2.4.2.-(S:HisF)
ec:2.4.2.-(S:HisH)
ec:4.2.1.19
ec:2.6.1.9
ec:3.1.3.15(A:1/2)
ec:3.1.3.15(A:2/2)
ec:1.1.1.23
Thr (56)
ec:1.1.1.3(I:2/2)
ec:1.1.1.3(I:1/2)
ec:2.7.1.39
ec:4.2.3.1
ec:4.2.3.1
Glu/Gln (55)
ec:1.4.1.14(S:large)
ec:1.4.1.14(S:small)
ec:1.4.1.13(S:large)
ec:1.4.1.13(S:small)
ec:6.3.1.2
ec:1.4.1.4
ec:1.4.1.3
ec:2.6.1.27
ec:1.5.1.12
ec:1.4.7.1(I:2/2)
ec:1.4.7.1(I:1/2)
Cys (53)
ec:4.2.1.22
ec:4.4.1.1
ec:4.4.1.8(I:2/2)
ec:4.4.1.8(I:1/2)
ec:4.4.1.8
ec:2.5.1.48
ec:2.3.1.31
ec:2.5.1.49
ec:2.3.1.30
ec:2.5.1.47(I:2/2)
ec:2.5.1.47(I:1/2)
Average taxonomic distribution across genomes (%)
0
50
100
E. coli enzymes
homologs
analogs
alternologs
(8/5)
(1/1)
(4/4)
(2/2)
(2/2)
(1/1)
(7/5)
(4/2)
(1/1)
Thermoprotei
Archaeoglobi
Halobacteria
Methanobacteria
Methanococci
Methanopyri
Methanomicrobia
Thermococci
Thermoplasmata
(13/6)
(32/5)
(39/7)
(4/2)
(7/5)
(1/1)
(6/3)
(34/15)
(1/1)
(1/1)
(1/1)
(1/1)
(4/2)
(23/8)
(2/2)
(13/9)
(10/4)
(37/25)
(38/16)
(95/37)
(1/1)
Clostridia
Bacillales
Lactobacillales
Chlorobi
Bacteroidetes
Planctomycetes
Spirochaetes
Actinobacteria
Fusobacteria
Thermotogae
Aquificae
Chloroflexi
Dein-Thermus
Cyanobacteria
Acidobacteria
δ-proteobacteria
ε-proteobacteria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
other-proteobact
(1/1)
(1/1)
(1/1)
(1/1)
(2/2)
(10/10)
(1/1)
Alveolata
Cnidaria
Nematoda
Arthropoda
Deuterostomia
Fungi
Plant
Archaea
Bacteria
Eukarya

Figura 6. Distribución taxonómica promedio de las enzimas de las rutas de biosíntesis de aminoácidos ampliamente distribuidos a través de los tres dominios de la vida. La DT de cada una de las enzimas de la ruta de biosíntesis de aminoácidos (vertical) fue calculada buscando su distribución ortóloga a través de diversos grupos taxonómicos (horizontal). En la grafica se muestran las enzimas con una distribución normalizada promedio ≥ 50%. Los aminoácidos están indicados en un código de tres letras y los que están en color rojo indican que esta vía pudo haber estado en el último ancestro común. Para este análisis se usaron 4 tipos de semillas: las enzimas canónicas de *E.coli* (en gris); enzimas homologas- paralogos y ortologos- de otras especies que mostraron una mayor distribución que las enzimas de *E.coli* (en amarillo); las enzimas análogas-que catalizan la misma reacción pero que proviene de una superfamilia diferente- se muestran en rojo; y las enzimas y ramas alternologas-convergen en el mismo producto final, pero proceden de diferentes metabolitos- en otras especies (en azul). En el eje vertical las subunidades de enzimas multimericas se indican como 'S', las enzimas análogas se indican como 'A' y las isoenzimas se indican como 'I'. El promedio de la distribución de ortólogos para cada ruta se muestra en paréntesis seguido del código de tres letras para cada aminoácido. Las enzimas están organizadas de acuerdo a su aparición dentro del vía. Figura tomada de artículo Hernández-Montes *et al* 2008

Ocho rutas parcialmente distribuidas en los tres dominios de la vida.

L-lisina

En la síntesis de L-lisina se pueden reconocer seis rutas alternativas agrupadas en dos supervías que proceden vía tanto de L,L-diaminopimelato o alfa-aminoadipato.

La supervía que involucra L, L-diaminopimelato tiene cuatro ramas alternólogas que corresponden a la síntesis de lisina tipo I, II, III, y IV en MetaCyc: estas rutas comparten un set común de 6 reacciones catalizadas por enzimas ampliamente distribuidos. Cuatro de esas enzimas catalizan los primeros pasos de la supervía, de aspartato cinasa (EC: 2.7.2.4) a dihidropicolinato reductasa (EC:1.3.1.26) y forman los pares de genes duplicados entre la biosíntesis de L-arginina y L-lisina. Las otras dos enzimas EC:5.1.17 y EC:4.1.120 los últimos pasos de la supervía.

A continuación se describe la DT de las enzimas que catalizan los pasos intermedios en las ramas alternólogas. En la ruta del tipo I (tipo *E. coli*), la cual es catalizada por tres enzimas, solamente N-succinil-L, L-diaminopimelato desuccinilasa EC:3.5.1.18 está ampliamente distribuida en los tres dominios. En la ruta del tipo II (tipo *B. subtilis*), está catalizada por otras tres enzimas, de las cuales solamente tetrahidropicolinato acetiltransferasa EC:2.3.1.89 está ampliamente distribuida en Bacteria y está ausente en Arquea y Eucaria. La ruta del tipo III de *Corynebacterium glutamicum* EC:1.4.1.16 está restringida a algunas actinobacterias y firmicutes, mientras que la ruta VI, recientemente descubierta, formada por una sola enzima denominada L,Ldiaminopimelato aminotransferasa EC:2.6.1 parece ser específica de plantas. Estos resultados ilustran una observación general de este trabajo: las

rutas lineares parecen estar más distribuidas que las que están bifurcadas. Como se describió anteriormente, las rutas de L-histidina, L-triptofano y L-leucina respaldan esta observación y correlacionan con estudios previos que muestran que dentro de la biosíntesis de aminoácidos las rutas mas largas tienden a tener bajas tasas de cambio en su estructura que las rutas cortas (Rutter MT y Zufall RA 2004). Sin embargo, estudios posteriores sobre redes metabólicas completas son necesarios para evaluar la generalidad de esta propiedad en la evolución del metabolismo. Por otro lado, la segunda supervía que procede de la degradación de α-aminoadipato, está formada por rutas linaje específico tipo IV y V que comparten un grupo de cinco reacciones de homocitrato sintasa EC:2.3.3.14 a α-aminoadipato aminotransferasa EC:2.6.1.39. Este grupo contiene las cuatro enzimas que forman pares de genes duplicados entre la biosíntesis de L-leucina y L-lisina. La ruta de tipo V usa N-2-acetil-L-lisina y fue caracterizada en el linaje Thermus-Deinococcus y sus representantes se encuentran en Arquea y algunas Bacteria, mientras que la ruta del tipo IV, que se lleva a cabo vía sacaropina EC:1.2.1.31 y EC:1.5.1.7 aparece restringido a Eucaria y algunas Bacteria. Colectivamente, las DT de esas dos supervías muestran que rutas alternativas llevaron al origen de la biosíntesis de lisina. Ninguna de esas alternologas parece estar distribuida universalmente, así que probablemente el LUCA llevaba a cabo la síntesis de lisina con alguno otro grupo de enzimas diferentes a las analizadas aquí. De manera interesante, las dos supervías retienen grupos de genes duplicados de la biosíntesis de leucina y de arginina, los cuales como mencionamos anteriormente probablemente existieron en LUCA. Así, existe la posibilidad que la lisina fue incorporada al metabolismo de las rutas de leucina y arginina.

L-metionina

La biosíntesis de metionina puede llevarse a cabo por tres diferentes supervías. Una involucra la degradación de cistation vía homoserina usando cistation β-sintasa EC:4.2.1.22 o cistation β-liasa EC:4.4.1.8, seguida por la metionina sintasa EC:2.1.1.13. Esas tres enzimas están ampliamente distribuidas en los tres dominios y probablemente esa rama pudo haber ocurrido en LUCA. Alternativamente la segunda supervía, también llamada el ciclo de salvamento de L-metionina, el cual empieza con EC:4.4.1.14 vía S-adenosil-

metionina y termina en L-metionina usando EC:2.6.1.5 vía 2-oxo-4-metiltiobutanoato está ampliamente distribuida en Eucaria y casi ausente en Arquea y Bacteria.

Una excepción a esta distribución es el paso de L-metionina a S-adenosil-L-metionina el cual puede ser catalizado por uno de los dos análogos metionin adenosiltransferasa EC:2.5.1.6. Esos análogos muestran una anti-correlación casi perfecta en sus DT; una es restringida a Arquea mientras la otra se encuentra en Bacteria y Eucaria. De manera complementaria, una tercera supervía, caracterizada en plantas es llamada ciclo S-adenosil-L-metionina convierte S-adenosil-L-metionina a metionina vía S-adenosil-L-homocisteina. Se encontró que una de las enzimas de este ciclo, la S-adenosilhomocisteina hidrolasa EC:3.3.1.1 está ampliamente distribuida en los tres dominios. En resumen, nosotros sugerimos que el LUCA fue capaz de producir L-metionina degradando cistation vía homocisteina.

L-valina y L-isoleucina

Los cuatro pasos terminales en la biosíntesis de valina e isoleucina emplean un grupo común de enzimas ampliamente distribuidas. De EC:2.2.1.6 a la amino ácido cadena ramificada aminotransferasa EC:2.6.1.42. Este grupo no fue encontrado en animales, con excepción de *N. vectensis*. Complementariamente, 5 ramas alternólogas pueden catalizar el paso inicial de la síntesis de isoleucina, convirgiendo en 2-oxobutanoato, el cual es a su vez un sustrato de acetolactato sintasa EC:2.2.1.6. Nosotros encontramos que la rama canónica de *E. coli* que lleva a cabo esos pasos vía propionato usa EC:2.7.2.15 y EC:2.3.1.8 y está aleatoriamente distribuida entre genomas bacterianos. En contraste la rama alternóloga caracterizada en espiroquetas, que procede vía (R)-cit-ramalato, usa isopropilmalato isomerasa EC:4.2.1.35 y β-isopropilmalato deshidrogenasa (sin EC) y ambas enzimas están ampliamente distribuidas en los tres dominios.

Esos resultados ejemplifican claramente que las rutas canónicas de *E. coli* no son necesariamente las más ampliamente distribuidas y así las rutas alternólogas deben ser incluidas en el análisis evolutivo. Adicionalmente, esta rama participa en la retención de un grupo de genes duplicados que catalizan reacciones consecutivas en la síntesis de lisina, leucina e isoleucina. Tomando en cuenta la amplia distribución de la rama tipo espiroqueta y las enzimas compartidas entre la biosíntesis de L-valina y L-isoleucina nosotros

sugerimos que el LUCA y aún especies contemporáneas pudieron combinar esas ramas para sintetizar ambos aminoácidos.

Corismato

Debido a la importancia de este compuesto en la biosíntesis de los aminoácidos aromáticos se decidió considerar la vía del corismato de manera independiente. La biosíntesis de corismato comprende 7 pasos, los dos últimos son catalizados por dos enzima ampliamente distribuidas, la 3-fosfoshikimato-1-carboxyviniltransferasa EC:2.5.1.9 y la corismato sintasa EC:4.2.3.5. Por otro lado, los primeros dos pasos son catalizados por enzimas ampliamente distribuidas en Bacteria y en algunas Eucaria, pero ausentes en Arquea. Un reporte reciente que sugiere una ruta nueva para la biosíntesis de aminoácidos aromáticos y acido p-aminobenzoico en *Methanococcus maripaludis* nos ayuda a entender su distribución (Porat I et al 2006). Adicionalmente, tres pasos intermedios son catalizados por enzimas análogas y alternólogas escasamente distribuidas como sigue. Primero, la transformación de 3-dehidroquinato a 3-dehidro-shikimato puede ser catalizada por 2 análogas de 3-dehidroquinato dehidratasa EC:4.2.1.10. *B. subtilis* posee ambas análogas, mientras que Arquea, algunas Eucaria y unas pocas Bacteria llevan solamente la enzima tipo II que pertenece a la superfamilia de la aldolasa (TIM barrel). En contraste, la mayoría de las Bacteria, incluyendo *E. coli*, usa la enzima de tipo I perteneciente a la superfamilia de la 3-dehidroquinato dehidratasa. Segundo, en *E. coli* hay dos parálogos catalizando la conversión de 3-dehidro-shikimato a shikimato. Uno de ellos, la NADP+-dependiente EC:1.1.1.25, está ampliamente distribuido, mientras EC:1.1.1.282 (usando NAD+ o NADP+, y quinato o shikimato) está escasamente distribuido. En contraste, *B. subtilis* tiene solamente la shikimato dehidrogenasa NADP+ dependiente, y cuando su secuencia se usa como semilla para obtener el MHR se encuentran mas ortólogos que con la contraparte de *E. coli*. Este resultado es probablemente causado por resultados cruzados entre los parálogos de *E. coli* durante la construcción de los DT. Tercero, la transformación de shikimato a shikimato-3-fosfato puede ser catalizada por 2 shikimato cinasas análogas EC:2.7.1.71. El tipo arqueal pertenece a la superfamilia de GHMP cinasa, mientras que la del tipo Bacteria/Eucaria pertenece a la superfamilia de P-loop que contienen nucleósido trifosfato hidrolasas. De manera interesante, hay una casi perfecta anti-correlación entre las

DT de esas enzimas. Los animales, incluyendo *N. vectensis*, han perdido todas las enzimas que catalizan las reacciones intermediarias en la síntesis de corismato, respaldando el hecho que los aminoácidos aromáticos son esenciales para humanos. Resumiendo, nosotros encontramos que la porción baja de la biosíntesis de corismato que convierte 3-dehidro-shikimato a corismato está ampliamente distribuida en los tres dominios, sugiriendo que probablemente existió en el LUCA. Por otro lado la porción superior y los pasos intermedios de esta ruta parecen haberse originado de manera independiente en cada linaje específico durante la evolución.

L-aspartato y L-asparagina

La biosíntesis e interconversión de L-aspartato y L-asparagina son mediadas por un set diverso de enzimas alternólogas, la mayoría de las cuales han sido caracterizadas en *E. coli* y están aleatoriamente distribuidas. Sin embargo, la aspartato aminotransferasa EC:2.6.1.1 y piruvato carboxilasa EC:6.4.1.1 son capaces de producir L-asparato a partir de piruvato, vía oxaloacetato y ambas enzimas están ampliamente distribuidas en los tres dominios. Complementariamente, la conversión de L-asparato a L-asparagina puede llevarse a cabo por tres asparagina sintetasas, dos de las cuales son glutamina dependientes EC:6.3.5.4 mientras que la otra es amonio dependiente EC:6.3.1.1. Tanto la EC:6.3.1.1 tipo I como la EC:6.3.5.4 pertenecen a la superfamilia de la adenina nucleotido alfa hidrolasa-like y está ampliamente distribuida en los tres dominios. En contraste, la producción de L-aspartato a L-asparagina vía 3-ciano-L-alanina, la cual es mediada por la β-ciano-L-alanina sintasa EC:4.4.1.9 y dos paralogos de nitrilasas EC:3.5.5.1 parecen estar restringidas a las plantas, cianobacterias y α-proteobacteria. Esta distribución podría ser el producto de transferencia horizontal entre esos clados, probablemente por simbiosis - como algunas α-proteobacterias son simbiontes y parásitos de plantas o por endosimbiosis - porque las cianobacterias son consideradas descendientes de plástidos ancestrales en plantas. Nosotros no detectamos algún otro posible evento de transferencia horizontal de genes en esas rutas usando bases de datos de genes putativos transferidos horizontales a genomas completos de procariontes. Finalmente, las dos asparaginas análogas EC:3.5.1.1, que convierten L-asparagina a L-aspartato muestran DT anti-correlacionados. Una de ellas de la superfamilia de la glutaminasa/asparaginasa fue encontrada en Arquea, algunas Bacteria, hongos y animales,

mientras que la segunda, de la superfamilia de las nucleófilo amino-terminal aminohidrolasas muestran una distribución similar a la de EC:4.4.1.9 y EC:3.5.5.1. En resumen, el LUCA probablemente no fue capaz de producir L-aspartato o L-asparagina vía los alternólogos modernos canónicos (nitrilasa y asparaginasa) pero pudo haberlo hecho vía la degradación de oxalacetato usando las ramas descritas arriba.

L-tirosina y L-fenilalanina

Hay por lo menos 5 ramas que divergen del prefenato para la biosíntesis de L-tirosina y L-fenilalanina. Dos de ellas proceden vía fenilpiruvato y uno de los dos análogos de prefenato dehidratasa EC:4.2.1.51 está ampliamente distribuido. Otras dos ramas proceden vía L-arogenato y usan a la arogenato dehidrogenasa EC:1.3.1.42 para sintetizar L-tirosina o a la arogenato dehidratasa EC:4.2.1.91 para sintetizar L-fenilalanina. EC:1.3.1.43 se encuentra en Bacteria y algunas Arquea mientras que EC:4.2.1.91 no tiene asignada una enzima ni secuencia genética. La quinta rama usa la prefenato dehidrogenasa EC:1.3.1.12 seguida por una aminoácido aromatico aminotransferasa EC:2.6.1.57. *E. coli*, *B. subtilis* y *S. cerevisae* tienen dos EC:2.6.1.57 y todas ellas pueden ser clasificadas en la familia de las transferasas PLP-dependientes, con excepción de AroJ en *B. subtilis* cuya secuencia es desconocida.

Sin embargo es difícil establecer relaciones de ortología entre esas enzímas porque ninguna de ellas son MHR y así solo son parálogos putativos. Aparentemente, esta alta diversidad se mantiene por una expresión diferencial y multifuncional de las propiedades de esas enzimas. Por ejemplo TyrB en *E. coli* es aproximadamente 1000 veces más activa sobre sustratos aromáticos que el parálogo AspC, el cual es más específico para L-aspartato. De manera similar, en *B. subtilis*, HisC es más activa que AroJ sobre L-fenilalanina y L-tirosina, en lugar de su actividad primaria que es sobre histidinolfosfato y *S. cerevisae* usa Aro8 preferencialmente en el anabolismo y Aro9 en catabolismo (Nester EW y Montoya AL 1976, Weigent DA y Nester EW 1976). HisC pudo representar uno de los linajes más ancestrales en esta familia, por que es el único miembro ampliamente distribuido en los tres dominios. Esto concuerda con el hecho que la biosíntesis de L-hisidina, la ruta en la cual participa HisC de manera preferencial es una ruta ancestral. La amplia distribución de HisC y dos enzimas análogas EC:4.2.1.51 sugiere que la síntesis de tirosina y fenilalanina también es antigua. Sin embargo el paso que precede a esas enzimas, de corismato a

prefenato, puede ser catalizada por los análogos de la corismato mutasa EC:5.4.99.5; estos muestran una distribución aleatoria con algunos representantes en firmicutes, proteobacterias y plantas. Por ejemplo, *E. coli* posee dos enzimas parálogas EC:5.4.99.5 pertenecientes a la superfamilia de la corismato mutasa II y que están fusionadas a dominios que catalizan EC:1.3.1.12 y EC:4.2.1.51 actividades, mientras *B. subtilis* también tiene dos EC:5.4.99.5, una de la superfamilia corismato mutasa II y otra de la superfamilia de YjgF-like. Ninguno de esos dominios EC:5.4.99.5 está ampliamente distribuido, por lo tanto difícil establecer si LUCA fue capaz de sintetizar L-fenilalanina y L-tirosina.

L-alanina

Hay cuatro pasos alternólogos independientes para sintetizar L-alanina. Uno de ellos usa cisteína desulfurasa EC:2.8.1.7, esta enzima degrada L-cisteína y está ampliamente distribuida en los tres dominios. Dado que la biosíntesis de L-cisteína probablemente ocurrió en LUCA, nosotros sugerimos que la biosíntesis de L-alanina pudo haber ocurrido en LUCA usando este paso. En contraste la alanina racemasa EC:5.1.1.1 isomerisa D-alanina a L-alanina y está restringida a Bacteria. La alanina aminotransferasa EC:2.6.1.2 convierte L-glutamato y piruvato a L-alanina y está ampliamente distribuida en Eucaria pero pobremente representada en Bacteria y Arquea, mientras que la valina piruvato aminotransferasa EC:2.6.1.66 degrada L-valina a L-alanina y fue detectada solo en pocas bacterias.

L-serina

Finalmente, hay 4 ramas para sintetizar L-serina. La primera procede vía 3-fosfo-hidroxypiruvato, es catalizada por tres enzimas, dos de ellas 3-fosfoglicerato dehidrogenasa EC: 1.1.1.95 y 3-fosfoserina fosfato EC:3.1.3.3 están ampliamente distribuidas, pero la tercera fosfoserina aminotransferasa EC:2.6.1.52 está restringida a Eucaria y algunas Bacteria. La segunda rama es un solo paso que convierte amonia y piruvato a L-serina por L-serina aminoa liasa EC:4.3.1.17 y esta restringida a algunas bacterias. La tercera y cuarta rama están estrechamente relacionadas a la biosíntesis de L-cisteína y L-metionina, interconvierten cistationina y homocisteina por EC:4.2.1.22 o EC:4.4.1.8, las cuales están ampliamente distribuidas en los tres dominios. Dado que L-cisteína y L metionina pudieron

haber existido en LUCA, la biosíntesis de serina pudo haber existido por medio de esas enzimas. De hecho, esta cadena de enzimas ampliamente distribuidas puede ser extendida a la biosíntesis de alanina.

En resumen, nuestros resultados han descubierto un grupo de 64 dominios enzimáticos participando en la biosíntesis de por lo menos 16 de las 20 rutas de biosíntesis de aminoácidos que tentativamente estuvieron presentes en LUCA. La figura 6 muestra un sesgo marcado en la DT de este grupo de dominios con respecto a la tendencia general de todo el metabolismo y otras vías menos conservadas de la biosíntesis de aminoácidos sugiriendo que las ramas en otros procesos metabólicos podrían tener también una universalidad oculta.

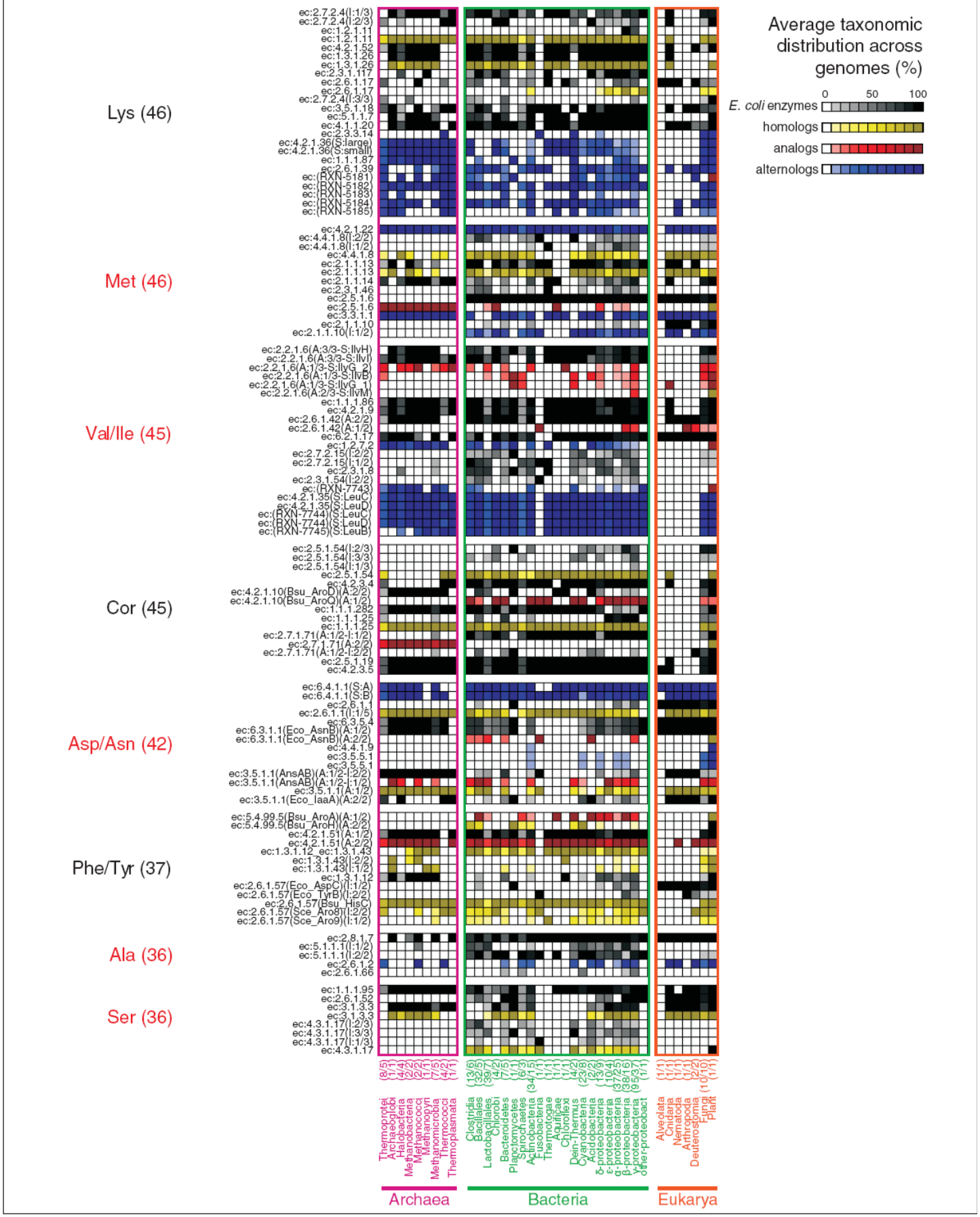
Lys (46)
Met (46)
Val/Ile (45)
Cor (45)
Asp/Asn (42)
Phe/Tyr (37)
Ala (36)
Ser (36)
ec:2.7.2.4(I:1/3)
ec:2.7.2.4(I:2/3)
ec:1.2.1.11
ec:1.2.1.11
ec:4.2.1.52
ec:1.3.1.26
ec:1.3.1.26
ec:2.3.1.117
ec:2.6.1.17
ec:2.6.1.17
ec:2.7.2.4(I:3/3)
ec:3.5.1.18
ec:5.1.1.7
ec:4.1.1.20
ec:2.3.3.14
ec:4.2.1.36(S:large)
ec:4.2.1.36(S:small)
ec:1.1.1.87
ec:2.6.1.39
ec:(RXN-5181)
ec:(RXN-5182)
ec:(RXN-5183)
ec:(RXN-5184)
ec:(RXN-5185)
ec:4.2.1.22
ec:4.4.1.8(I:2/2)
ec:4.4.1.8(I:1/2)
ec:4.4.1.8
ec:2.1.1.13
ec:2.1.1.13
ec:2.1.1.14
ec:2.3.1.46
ec:2.5.1.6
ec:2.5.1.6
ec:3.3.1.1
ec:2.1.1.10
ec:2.1.1.10(I:1/2)
ec:2.2.1.6(A:3/3-S:IlvH)
ec:2.2.1.6(A:3/3-S:IlvI)
ec:2.2.1.6(A:1/3-S:IlvG_2)
ec:2.2.1.6(A:1/3-S:IlvB)
ec:2.2.1.6(A:1/3-S:IlvG_1)
ec:2.2.1.6(A:2/3-S:IlvM)
ec:1.1.1.86
ec:4.2.1.9
ec:2.6.1.42(A:2/2)
ec:2.6.1.42(A:1/2)
ec:6.2.1.17
ec:1.2.7.2
ec:2.7.2.15(I:2/2)
ec:2.7.2.15(I:1/2)
ec:2.3.1.8
ec:2.3.1.54(I:2/2)
ec:(RXN-7743)
ec:4.2.1.35(S:LeuC)
ec:4.2.1.35(S:LeuD)
ec:(RXN-7744)(S:LeuC)
ec:(RXN-7744)(S:LeuD)
ec:(RXN-7745)(S:LeuB)
ec:2.5.1.54(I:2/3)
ec:2.5.1.54(I:3/3)
ec:2.5.1.54(I:1/3)
ec:2.5.1.54
ec:4.2.3.4
ec:4.2.1.10(Bsu_AroD)(A:2/2)
ec:4.2.1.10(Bsu_AroQ)(A:1/2)
ec:1.1.1.282
ec:1.1.1.25
ec:1.1.1.25
ec:2.7.1.71(A:1/2-I:1/2)
ec:2.7.1.71(A:2/2)
ec:2.7.1.71(A:1/2-I:2/2)
ec:2.5.1.19
ec:4.2.3.5
ec:6.4.1.1(S:A)
ec:6.4.1.1(S:B)
ec:2.6.1.1
ec:2.6.1.1(I:1/5)
ec:6.3.5.4
ec:6.3.1.1(Eco_AsnB)(A:1/2)
ec:6.3.1.1(Eco_AsnB)(A:2/2)
ec:4.4.1.9
ec:3.5.5.1
ec:3.5.5.1
ec:3.5.1.1(AnsAB)(A:1/2-I:2/2)
ec:3.5.1.1(AnsAB)(A:1/2-I:1/2)
ec:3.5.1.1(A:1/2)
ec:3.5.1.1(Eco_IaaA)(A:2/2)
ec:5.4.99.5(Bsu_AroA)(A:1/2)
ec:5.4.99.5(Bsu_AroH)(A:2/2)
ec:4.2.1.51(A:1/2)
ec:4.2.1.51(A:2/2)
ec:1.3.1.12_ec:1.3.1.43
ec:1.3.1.43(I:2/2)
ec:1.3.1.43(I:1/2)
ec:1.3.1.12
ec:2.6.1.57(Eco_AspC)(I:1/2)
ec:2.6.1.57(Eco_TyrB)(I:2/2)
ec:2.6.1.57(Bsu_HisC)
ec:2.6.1.57(Sce_Aro8)(I:2/2)
ec:2.6.1.57(Sce_Aro9)(I:1/2)
ec:2.8.1.7
ec:5.1.1.1(I:1/2)
ec:5.1.1.1(I:2/2)
ec:2.6.1.2
ec:2.6.1.66
ec:1.1.1.95
ec:2.6.1.52
ec:3.1.3.3
ec:3.1.3.3
ec:4.3.1.17(I:2/3)
ec:4.3.1.17(I:3/3)
ec:4.3.1.17(I:1/3)
ec:4.3.1.17
(8/5)
(1/1)
(4/4)
(2/2)
(2/2)
(1/1)
(7/5)
(4/2)
(1/1)
(13/6)
(32/5)
(39/7)
(4/2)
(4/5)
(1/1)
(6/3)
(34/15)
(1/1)
(1/1)
(1/1)
(4/2)
(23/8)
(2/2)
(12/2)
(15/4)
(10/4)
(37/25)
(38/16)
(95/37)
(1/1)
(1/1)
(1/1)
(1/1)
(2/2)
(1/1)
(10/10)
(1/1)
Thermoprotei
Archaeoglobi
Halobacteria
Methanobacteria
Methanococci
Methanopyri
Methanomicrobia
Thermococci
Thermoplasmata
Clostridia
Bacillales
Lactobacillales
Chlorobi
Bacteroidetes
Planctomycetes
Spirochaetes
Actinobacteria
Fusobacteria
Thermotogae
Aquificae
Chlorioflexi
Dein-Thermus
Cyanobacteria
Acidobacteria
δ-proteobacteria
ε-proteobacteria
α-proteobacteria
β-proteobacteria
γ-proteobacteria
other-proteobact
Alveolata
Cnidaria
Nematoda
Arthropoda
Deuterostomia
Fungi
Plant
Archaea
Bacteria
Eukarya
Average taxonomic distribution across genomes (%)
0
50
100
E. coli enzymes
homologs
analogs
alternologs

Figura 7. Distribución taxonómica promedio de las enzimas de las rutas de biosíntesis de aminoácidos que se encuentran parcialmente distribuidas a través de los tres dominios de la vida. Las DTs para cada enzima tienen un promedio de distribución normalizado < 50%. Se usaron los mismos códigos que en la figura 6**.** Figura tomada de artículo Hernández-Montes *et al* 2008.

CONCLUSIONES

En este trabajo se llevo a cabo un análisis sobre el origen y la evolución de las rutas de biosíntesis de aminoácidos.

La estrategia de utilizar diferentes organismos modelo, así como combinar la comparación entre genomas con una perspectiva de redes nos permitió identificar un grupo de enzimas ampliamente distribuidas y que probablemente estuvieron presentes en el último ancestro común. Se encontró que por lo menos 16 de los 20 aminoácidos estándares pudieron haber sido sintetizados en el último ancestro común. Esto no implica sin embargo que las 16 rutas hayan aparecido en estadios tempranos, sino que sus ramas podrían satisfacer los mínimos requerimientos bioquímicos y estructurales.

Cabe notar que algunas especies tales como parásitos y animales, incluyendo los mamíferos, carecen de partes significativas de este grupo "universal" debido a que pueden importar aminoácidos de su hospedero o adquirirlos a través de su dieta, por lo que se propone pudieron haber tenido perdidas secundarias.

En este sentido, consideramos que usar tantas especies como estén disponibles en las bases de datos y no solo un modelo tradicional para llevar a cabo los análisis genomicos fue fundamental. De las 20 rutas de biosíntesis de aminoácidos 8 están pobremente distribuidos desde un punto de vista de *E. coli*, pero ampliamente distribuidas cuando se utilizan los ortólogos, parálogos análogos y alternólogos de otras especies, lo que nos permite obtener un resultado mas representativo y por lo tanto mas fiel a lo que pudo haber sucedido en el origen del metabolismo.

El principal papel biológico de los aminoácidos es que son los constituyentes de las proteínas, por lo que una pregunta lógica sería si este grupo podría ser suficiente para el repertorio de proteínas del ancestro común. Recientemente, Atchley *et al* agruparon a los aminoácidos de acuerdo a por lo menos 500 propiedades o atributos que van desde propiedades biofísicas hasta bioquímicas y estructurales produciendo una representación multifuncional de la variabilidad de los aminoácidos. Cuando se localiza este grupo de los 16 aminoácidos sobre la gráfica de Atchley, el resultado sugiere que el ancestro común fue capaz de poblar todas la regiones del espacio de variabilidad de los aminoácidos, es decir que este grupo pudo haber sido suficiente para cubrir las funciones de las proteínas en sistemas tempranos.

La mayoría de las rutas universales encontradas en este trabajo están conectadas unas a otras, permitiendo la posibilidad de que retroalimenten y completen un grupo mínimo de enzimas que dirigen reacciones para la biosíntesis de aminoácidos. La incorporación de nuevos aminoácidos a este grupo universal pudo ser el resultado de la combinación entre la variabilidad y la presión de selección que se originó al incrementarse el tamaño del genoma y la complejidad estructural de las proteínas.

Por otro lado la estrategia de utilizar semillas de diferentes organismos para buscar ortologos, permitió identificar ramas y rutas alternas, lo que refleja que los organismos han generado estrategias específicas para la síntesis de aminoácidos, probablemente debido a sus diferentes estilos de vida.

Finalmente también pudimos corroborar que en la biosíntesis de aminoácidos la retención de genes duplicados como grupos en lugar de entidades individuales ha sido un factor importante en la evolución de estas vías ya que encontramos que 11 de las 20 rutas revelan una importante contribución de la paralogía a la generación de diversidad.

También encontramos que las enzimas análogas contribuyen en 8 de los 20 aminoácidos estándar, mientras que las rutas alternólogas participan en nueve. Esto implica que enzimas análogas y ramas alternólogas tiene una participación casi tan importante como la duplicación genética en la evolución de la biosíntesis de aminoácidos.

En conclusión, nosotros sugerimos que las rutas de biosíntesis de aminoácidos actuales más que ser heredadas de sistemas ancestrales, han sido originadas de manera independiente por cada linaje de acuerdo a sus fuentes ambientales como se ve reflejado por la gran diversidad de las ramas anabólicas.

PERSPECTIVAS

Consideramos que es interesante hacer un análisis similar con las rutas de biosíntesis de nucleótidos y determinar la generalidad de estas observaciones. Por otro lado es interesante saber como se comportan los homólogos dentro de cada grupo de rutas y entre rutas de diferentes procesos metabólicos.

BIBLIOGRAFIA

Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS: **The 'evolvability' of promiscuous protein functions.** *Nat Genet* 2005, **37:**73-76.

Alves R, Chaleil RA, Sternberg MJ: **Evolution of enzymes in metabolism: a network perspective.** *J Mol Biol* 2002, **320:**751-770.

Atchley WR, Zhao J, Fernandes AD, Druke T: **Solving the protein sequence metric problem.** *Proc Natl Acad Sci USA* 2005, **102:**6395-6400.

Benner SA, Ellington AD, Tauer A: **Modern metabolism as a palimpsest of the RNA world.** *Proc Natl Acad Sci USA* 1989, **86:**7054-7058

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.

Cunchillos C, Lecointre G: **Evolution of amino acid metabolism inferred through cladistic analysis** *J Biol Chem* 2003, **278**(48):47960-70**.**

Díaz-Mejía JJ, Perez-Rueda E, Segovia L: **A network perspective on the evolution of metabolism by gene duplication.** *Genome Biol* 2007, **8:**R26.

Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6:**361-365.

Fani R, Lio P, Lazcano A: **Molecular evolution of the histidine biosynthetic pathway.** *J Mol Evol* 1995, **41:**760-774.

Galperin MY, Walker DR, Koonin EV: **Analogous enzymes: independent inventions in enzyme evolution.** *Genome Res* 1998, **8:**779-790.

Garcia-Vallve S, Guzman E, Montero MA, Romeu A: **HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes.** *Nucleic Acids Res* 2003, **31:**187-189.

Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313:**903-919.

Herrmann MK and Weaver LM. **The shikimate pathway**. *Ann. Rev. Plant. Physiol.Plant.Mol.Biol* 1999, 50:473-503.

Hernandez-Montes G, Diaz-Mejia JJ, Perez-Rueda E, Segovia L: **The hidden universal distribution of amino acid biosynthetic networks: a genomic pespective on their origin and evolution.** *Genome Biology* 2008, 9:R95

Horowitz NH: **On the evolution of biochemical synthesis.** *Proc Natl Acad Sci USA* 1945, **31:**153-157.

Hudson AO, Bless C, Macedo P, Chatterjee SP, Singh BK, Gilvarg C, Leustek T: **Biosynthesis of lysine in plants: evidence for a variant of the known bacterial pathways.** *Biochim Biophys Acta* 2005, **1721:**27-36.

Irvin SD, Bhattacharjee JK: **A unique fungal lysine biosynthesis enzyme shares a common ancestor with tricarboxylic acid cycle and leucine biosynthetic enzymes found in diverse organisms.** *J Mol Evol* 1998, **46:**401-408.

Jensen RA: **Enzyme recruitment in the evolution of new function.** *Annu Rev Microbiol* 1976, **30:**409-425.

Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654.

Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: **The EcoCyc Database.** *Nucleic Acids Res* 2002, **30:**56-58.

Kvenvolden K, Lawless J, Pering K, Peterson E, Flores J, Ponnamperuma C, Kaplan IR, Moore C: **Evidence for extraterrestrial amino-acids and hydrocarbons in the Murchison meteorite**. *Nature* 1970, Dec 5;228(5275):923-6.

Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32:**D438-442.

Lazcano A. and Miller S.L. **On the origin of Metabolic pathways**. *J Mol Evol* 1999, **49**:424-431.

Li W, Jaroszewski L, Godzik A: **Tolerating some redundancy significantly speeds up clustering of large protein databases.** *Bioinformatics* 2002, **18:**77-82.

Light S, Kraulis P: **Network analysis of metabolic enzyme evolution in *Escherichia coli*.** *BMC Bioinformatics* 2004, **5:**15.

Miyazaki J, Kobashi N, Nishiyama M, Yamane H: **Functional and evolutionary relationship between arginine biosynthesis and prokaryotic lysine biosynthesis through alpha-aminoadipate.** *J Bacteriol* 2001, **183:**5067-5073.

Nester EW, Montoya AL: **An enzyme common to histidine and aromatic amino acid biosynthesis in *Bacillus subtilis*.** *J Bacteriol* 1976, **126:**699-705.

Nishida H, Nishiyama M, Kobashi N, Kosuge T, Hoshino T, Yamane H: **A prokaryotic gene cluster involved in synthesis of lysine through the amino adipate pathway: a key to the evolution of amino acid biosynthesis.** *Genome Res* 1999, **9:**1175-1183.

Nishida H: **Evolution of amino acid biosynthesis and enzymes with broad substrate specificity.** *Bioinformatics* 2001, **17:**1224-1225.

Ohno S: *Evolution by Gene Duplication* New York: Springer; 1970.

Oparin, A. I. El Origen de la Vida. New York: Dover (1952) (1ª publicación en 1938).

Porat I, Sieprawska-Lupa M, Teng Q, Bohanon FJ, White RH, Whitman WB: **Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p-aminobenzoic acid in the archaeon *Methanococcus maripaludis*.** *Mol Microbiol* 2006, **62:**1117-1131.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization.** *Science* 2007, **317:**86-94.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.

Rutter MT, Zufall RA: **Pathway length and evolutionary constraint in amino acid biosynthesis.** *J Mol Evol* 2004, **58:**218-224.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.

Starcevic A, Akthar S, Dunlap WC, Shick JM, Hranueli D, Cullum J, and Long PF: **Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, Nematostella vectensis, have microbial origins**
PNAS 2008**,105**: **2533–**2537.

Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C:
The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J Mol Biol* 2001, **311:**693-708.

Tsoka S, Simon D, Ouzounis CA: **Automated metabolic reconstruction for *Methanococcus jannaschii*.** *Archaea* 2004

Velasco AM, Leguina JI, Lazcano A: **Molecular evolution of the lysine biosynthetic pathways.** *J Mol Evol* 2002, **55:**445-459.

Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268:**1803-1810.

Weigent DA, Nester EW: **Purification and properties of two aromatic aminotransferases in *Bacillus subtilis*.** *J Biol Chem* 1976, **251:**6974-6980.

Xie G, Keyhani NO, Bonner CA, Jensen RA: **Ancient origin of the tryptophan operon and the dynamics of evolutionary change.** *Microbiol Mol Biol Rev* 2003, **67:**303-342.

Xu Y, Liang Z, Legrain C, Ruger HJ, Glansdorff N: **Evolution of arginine biosynthesis in the bacterial domain: novel gene-enzyme relationships from psychrophilic Moritella strains (Vibrionaceae) and evolutionary significance of N-alpha-acetyl ornithinase**. *J Bacteriol* 2000, **182**:1609-15.

Research

The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution

Georgina Hernández-Montes[¤*], J Javier Díaz-Mejía[¤*†], Ernesto Pérez-Rueda[*] and Lorenzo Segovia[*]

Addresses: [*]Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México. Av. Universidad, Col. Chamilpa, Cuernavaca, Morelos, México, CP 62210. [†]Department of Biology, Wilfrid Laurier University, University Av. Waterloo, ON N2L 3C5, Canada; and Donnelly Centre for Cellular and Biomolecular Research, University of Toronto. College St., Toronto, ON M5S 3E1, Canada.

¤ These authors contributed equally to this work.

Correspondence: Lorenzo Segovia. Email: lorenzo@ibt.unam.mx

Published: 9 June 2008

Genome ***Biology*** 2008, **9:**R95 (doi:10.1186/gb-2008-9-6-r95)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/6/R95

Received: 4 December 2007
Revised: 6 May 2008
Accepted: 9 June 2008

Abstract

Background: Twenty amino acids comprise the universal building blocks of proteins. However, their biosynthetic routes do not appear to be universal from an *Escherichia coli*-centric perspective. Nevertheless, it is necessary to understand their origin and evolution in a global context, that is, to include more 'model' species and alternative routes in order to do so. We use a comparative genomics approach to assess the origins and evolution of alternative amino acid biosynthetic network branches.

Results: By tracking the taxonomic distribution of amino acid biosynthetic enzymes, we predicted a core of widely distributed network branches biosynthesizing at least 16 out of the 20 standard amino acids, suggesting that this core occurred in ancient cells, before the separation of the three cellular domains of life. Additionally, we detail the distribution of two types of alternative branches to this core: analogs, enzymes that catalyze the same reaction (using the same metabolites) and belong to different superfamilies; and 'alternologs', herein defined as branches that, proceeding via different metabolites, converge to the same end product. We suggest that the origin of alternative branches is closely related to different environmental metabolite sources and life-styles among species.

Conclusion: The multi-organismal seed strategy employed in this work improves the precision of dating and determining evolutionary relationships among amino acid biosynthetic branches. This strategy could be extended to diverse metabolic routes and even other biological processes. Additionally, we introduce the concept of 'alternolog', which not only plays an important role in the relationships between structure and function in biological networks, but also, as shown here, has strong implications for their evolution, almost equal to paralogy and analogy.

Background

Metabolism represents an intricate set of enzyme-catalyzed reactions synthesizing and degrading compounds within cells. It is likely that a small number of enzymes with broad specificity existed in early stages of metabolic evolution. Genes encoding these enzymes probably have been duplicated, generating paralog enzymes that, through sequence divergence, became more specialized, giving rise, for instance, to the isomerases HisA (EC:5.3.1.16) and TrpC (EC:5.3.1.24), which act in histidine and tryptophan biosynthesis, respectively [1-4]. Additionally, gene duplication can promote innovations, generating enzymes catalyzing functionally different reactions, such as HisA, HisF (EC:2.4.2.-) and TrpA (EC:4.2.1.10). The classic view of metabolism is that relatively isolated sets of reactions or pathways are enough for the synthesis and degradation of compounds. The new perspective views metabolic components (substrates, products, cofactors, and enzymes) as nodes forming branches within a single network [5,6].

In the past few years, an increasing amount of information on metabolic networks from different species has become available [7-10], allowing for comparative genomic-scale studies on the evolution of both specific pathways [11,12] and whole metabolic networks [13-16]. Collectively, these studies highlight the contribution of gene duplication in the evolution of metabolism. Nevertheless, analog enzymes - those catalyzing the same reaction, even belonging to different evolutionary families - have been suggested to play an important role on this process as well [17]. This results, for instance, in three different types of acetolactate synthases (EC:2.2.1.6) acting in the biosynthesis of L-valine and L-leucine in *Escherichia coli*. Additionally, the modern perspective of metabolic processes has shown that evolutionary studies must include not only phylogenetic relationships among enzymes, but also the influence of some topological properties of metabolic networks [5,6,18-20]. One of these properties is the capability of metabolism to circumvent failures - for example, mutations promoting unbalanced fluxes - using alternative network branches and enzymes. Here, we introduce the term 'alternolog' to refer to these alternative branches and enzymes that, proceeding via different metabolites, converge in a common product. Some authors have suggested that alternative branches can contribute to genetic buffering in eukaryotes to a degree similar to gene duplication [18], but the role of these alternologs in the evolution of metabolism in other phylogenetic groups remains to be solved. In evolutionary terms, one can assume that the universal occurrence of some pathways and branches in modern species suggests that they existed in the last common ancestor (LCA). The evolution of these pathways and the emergence of paralogs, analogs and alternologs reflect an increased metabolic diversity as a consequence of increasing genome size, protein structural complexity and selective pressures in changing environments. In the evolution of amino acid biosynthesis, for instance, alternative pathways synthesizing L-lysine via either L,L-diaminopimelate or alpha-aminoadipate have been suggested to have developed independently in diverse clades [21-23]. The evolution of these pathways is closely related to the biosynthesis of L-arginine and L-leucine [22-24] and even to the Krebs cycle [24], but the origin of all these pathways is still under discussion. Diverse studies [6,25,26] have suggested that amino acids could be among the earliest metabolic compounds. However, two main questions have emerged from these studies: from what did their biosynthetic networks originate and how did they evolve? And how did gene duplication (paralogs), functional convergence (analogs) and network structural alternatives (alternologs) contribute to these processes? The purpose of this work is to broach these questions, combining both a network perspective and a comparative genomics approach. For this purpose we consider that the architecture of proteins preserves structural information that can be used to identify their relative emergence during the evolution of metabolism. Specifically, we identified a set of enzymes and branches that originated closer to the existence of the LCA, delimiting a core of enzyme-driven reactions that putatively catalyzed the biosynthesis of at least 16 out of the 20 amino acids in early stages of evolution. Additionally, we determined the contributions of biochemical functional alternatives to this core (paralogs, analogs, and alternologs) during the evolution of amino acid biosynthesis in diverse species.

Results and discussion

Biological distribution of amino acid biosynthetic networks

The origins and evolution of amino acid biosynthesis were assessed by analyzing the taxonomic distributions (TDs) of its catalyzing enzymes. Each enzyme's TD is a vector of ortholog distribution (presences/absences) in a set of genomes or clades (see Materials and methods). The rationale is that TDs provide clues concerning the relative appearance of enzymes, branches and pathways during the evolution of metabolism. We determined the TDs for 537 enzyme functional domains, catalyzing 188 reactions in the biosynthesis of amino acids from diverse species, in a set of 410 genomes (30 Archaea, 363 Bacteria and 17 Eukarya). To this end, we followed a two step strategy: first, we scanned the genomes to identify orthologs (best reciprocal hits (BRHs)) for the 113 amino acid biosynthetic enzymes from *E. coli* K12 defined in the EcoCyc database [8]: and second, a second set of ortholog, paralog, analog and alternolog enzymes and branches from different species, defined in the MetaCyc [9] and MjCyc [9] databases, was used to fill out the gaps in the *E. coli*-based TDs. Figure 1 shows a network formed by the 188 reactions analyzed in this work and the average distribution of orthologs for their catalyzing enzymes (see Materials and methods). We considered two broad categories for ortholog distribution: widely distributed enzymes, whose ortholog distribution is ≥ 50% across the clades analyzed here; and partially distributed enzymes, whose ortholog distribution is <50% across these clades. The

wide distribution of enzymes, branches and pathways suggests their occurrence in the LCA, although these categories are simply a tool for presentation purposes. Even when a pathway shows a low average distribution of orthologs, some of its branches can be widely distributed across the three cellular domains (Archaea, Bacteria and Eukarya), and hence these branches might be present in the LCA. The opposite scenario can also take place, that is, some enzymes can exhibit a high average distribution, but they could be restricted to specific cellular domains or divisions, such as Bacteria or γ-proteobacteria, that are overrepresented in sequenced genomes. Thus, their distribution does not necessarily signify their occurrence in the LCA. For these reasons, we exhaustively examined the TDs of enzymes forming each branch within amino acid biosynthetic pathways. In the following sections we describe our main findings in decreasing order of average ortholog distribution, emphasizing the possible existence of some branches in the LCA.

Nine amino acid biosynthetic pathways are widely distributed across the three domains of life, and eight of their branches probably occurred in the LCA

L-arginine

There are at least four L-arginine synthesis pathways, interplaying with the conversion of L-ornithine and citrulline, although they can be grouped in two superpathways (Figure 1). The first superpathway, involving carbamoyl-phosphate and N-acetyl-L-citrulline, can proceed via two alternolog branches: the first branch is the canonical *E. coli* pathway, catalyzed by two widely distributed enzymes, carbamoyl phosphate synthetase (EC:6.3.5.5) and ornithine carbamoyltransferase (EC:2.1.3.3). The second branch uses three enzymes (EC:6.3.4.16, EC:2.1.39 and EC:3.5.1.16), of which two are also widely distributed (Figure 2). Interestingly, EC:6.3.5.5 and EC:6.3.4.16 enzymes are paralogs, and EC:2.1.3.3 and EC: 2.1.39 are paralogs as well (Figure 3), representing an event of retention of duplicated genes as groups, instead of single entities. The retention of groups of duplicates has been suggested to play a significant role in the evolution of metabolism [16]. Alternatively, the second superpathway occurring via N-acetyl-L-ornithine is also widely distributed across the three domains, with the exception of animals, and shows three interesting TDs. First, using the *E. coli* enzymes as seeds for BRHs in this superpathway, we detected a small amount of orthologs in some clades, but using the ortholog sequences from *Saccharomyces cerevisiae*, *Methanocaldococcus jannaschii* and *Bacillus subtilis*, the gaps were filled in their respective phylogenetic groups (yellow squares in Figure 2), showing the importance of using enzymes from multiple species as queries instead of the simpler *E. coli*-centric strategies. Second, there are two analog N-acetylglutamate synthases (EC:2.3.1.1). The *E. coli*-type is a monomeric monofunctional enzyme, while the *B. subtilis*-type is a heterodimeric bifunctional enzyme (EC:2.3.1.1/2.3.1.35) whose constituents are proteolytically self-processed from a single precursor protein. Both types of enzymes are widely distributed across the three domains (Figure 2), although the *E. coli*-type was not identified in firmicutes, suggesting its displacement by the *B. subtilis*-type. Third, another retention of duplicated genes as groups, instead of as single entities, occurs between three consecutive steps in the biosynthesis of L-arginine/L-lysine [22]: EC:2.7.2.8/EC:2.7.2.4, EC:1.2.1.38/EC:1.2.1.11, EC:2.6.1.11/EC:2.6.1.17 and EC:3.5.1.16/EC:3.5.1.18 (Figure 3). In summary, we propose that not all pathways to synthesize L-arginine occurred in the LCA, only those proceeding via N-acetyl-L-ornithine and citrulline.

L-glycine

There are four branches to synthesize L-glycine. Two of them, involving the degradation of L-threonine (Figure 1), are partially distributed in Bacteria and Eukarya (Figure 2). In contrast, the other two branches, interconnected through 5,10-methylene-tetrahydrofolate, involve either the glycine-cleavage system or serine hydroxymethyltransferase (EC:2.1.2.1). Both branches are widely distributed across the three cellular domains (Figure 2). Indeed, EC:2.1.2.1 is one of the most widely distributed enzymes across all the species, probably as it also participates in folate biosynthesis, another broadly distributed pathway. Collectively, the distribution of these enzymes suggests that the LCA synthesized glycine via the branch of 5,10-methylene-tetrahydrofolate.

L-tryptophan

We found the five L-tryptophan biosynthetic enzymes widely distributed across the three domains of life, confirming previous reports [27]. Nevertheless, we did not identify orthologs for these enzymes in animals (Figure 2), with the exception of *Nematostella vectensis*, a cnidaria representative of early stages in animal evolution [28]. This indicates that some animals had a secondary loss of the L-tryptophan biosynthetic enzymes and also explains why this amino acid is essential for humans. Thus, the LCA probably was able to synthesize L-tryptophan in a similar fashion to contemporary species.

L-proline

There are at least six L-proline biosynthetic branches (Figure 1). Three of them converge in L-glutamate γ-semialdehyde and, judging from their TDs, ornithine-δ-aminotransferase (EC:2.6.1.13) is the most widely distributed enzyme within this pathway, even in some archaeal genomes (Figure 2). The other two branches have been biochemically characterized, although their catalyzing enzymes are unknown. The sixth branch, which directly converts L-ornithine to L-proline via ornithine cyclodeaminase (EC:4.3.1.12), was found in some Archaea and scarcely in Bacteria and Eukarya (Figure 2). Further analyses are necessary to corroborate experimentally the activities of these archaeal open reading frames, because the putative EC:2.6.1.13 enzymes do not have the canonical catalytic residues involved in this activity, and little information is known about the EC:4.3.1.12 activity. Thus, the archaeal biosynthesis of L-proline remains enigmatic and makes it

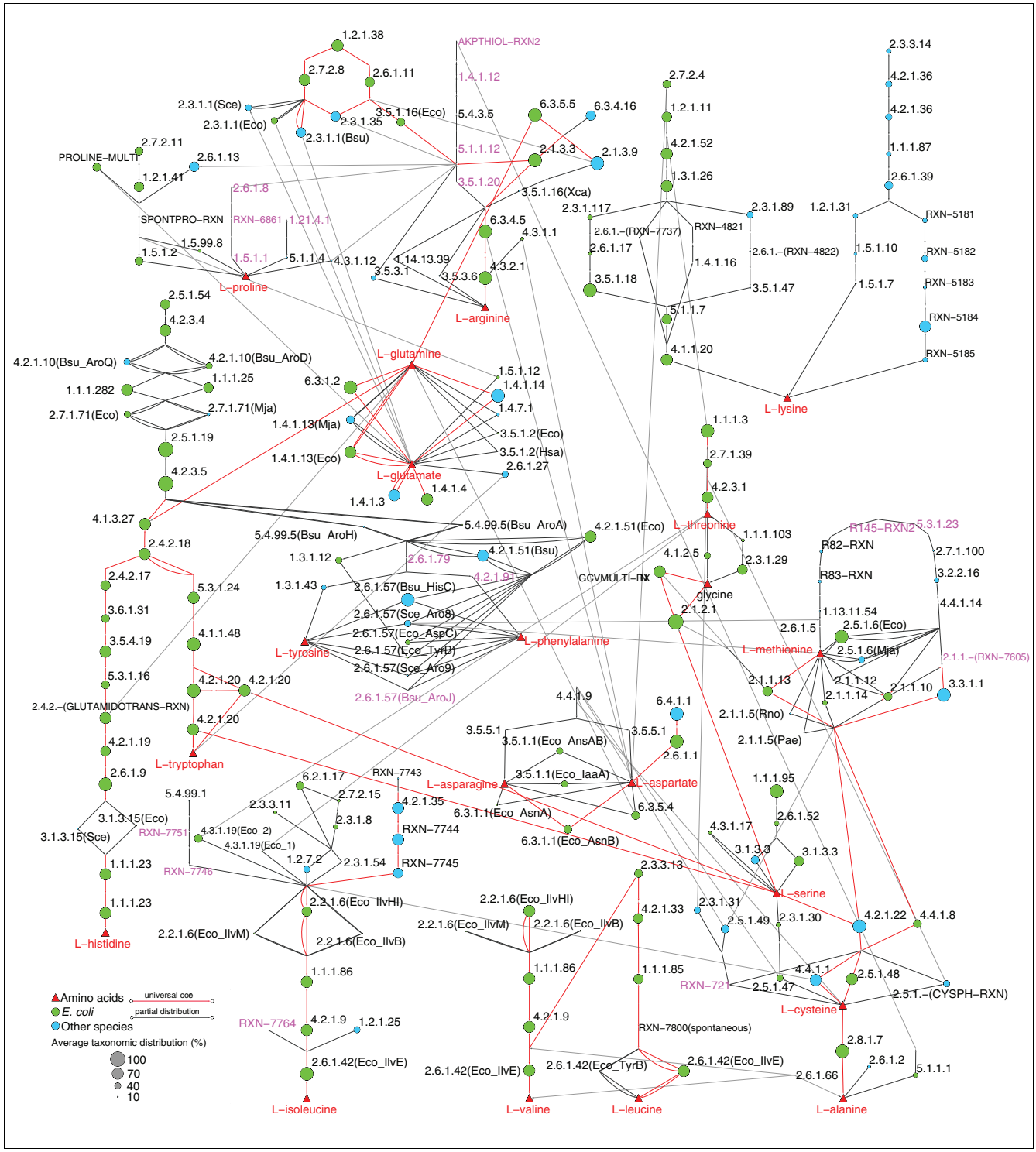

Figure 1

The amino acid biosynthetic network analyzed in this work. Bipartite amino acid biosynthetic network from multiple species. The 20 standard amino acids (red triangles) are shown as the ends of pathways. Green circles represent the canonical *E. coli* enzymes. Blue circles represent alternative enzymes (analogs and alternologs) from other species. The size of nodes corresponds to the normalized average taxonomic distribution of orthologs for each enzyme domain (domains in multimeric enzymes) catalyzing the corresponding reaction. The larger a node is the wider the distribution of orthologs for the corresponding enzyme across genomes. Red edges denote steps that could occur in the LCA based on the TDs of their catalyzing enzymes (Figures 2 and 4). Purple EC numbers correspond to reactions without known gene/enzymes. A detailed view of this network, including substrates and products, is provided in Additional data files 1 and 3, and the data for its construction are provided in Additional data files 2 and 4.

difficult to infer if the LCA was capable of synthesizing L-proline.

L-leucine

The biosynthesis of L-leucine consists of five reactions following a mainly linear pathway (Figure 1). Using the *E. coli* and *M. jannaschii* sequences for BRHs, we detected that putative enzymes catalyzing the first three reactions are widely distributed (Figure 2). These three enzymes belong to a group of duplicated genes catalyzing consecutive steps in the biosynthesis of three amino acids, L-lysine, L-leucine and L-isoleucine (Figure 3). The evolutionary relationships between L-lysine and L-leucine biosynthesis have been documented previously [23,24,29]: we found that L-isoleucine biosynthesis is also implied in this phenomenon. These duplicates together with those from L-arginine/L-lysine biosynthesis support our previous report on the importance of the retention of duplicated genes as groups, instead of as single entities, in the evolution of metabolism [16]. The fourth reaction occurs spontaneously and does not require a catalyzing enzyme. Complementarily, the fifth step in *E. coli* is catalyzed by one out of the two analog branched-chain amino acid transferases (EC:2.6.1.42); one of them belongs to the D-amino acid aminotransferase-like PLP-dependent superfamily and is widely distributed across the three domains, including some animals. In contrast, the second EC:2.6.1.42 belongs to the PLP-dependent transferases superfamily and is sparsely distributed across genomes. Collectively, these observations suggest that the LCA was able to synthesize L-leucine-like contemporary species. Further biochemical characterization of animal open reading frames is necessary, as L-leucine is an essential amino acid for humans.

L-histidine

Structurally speaking, L-histidine and L-tryptophan biosynthesis are similar; both are mainly linear pathways diverging from anthranilate using EC:2.4.2.18 (Figure 1) and, given their wide distribution, they have been proposed to be ancient pathways. The L-histidine biosynthesis enzyme histidinol-phosphatase (EC:3.1.3.15) is the only enzyme from this pathway partially distributed across genomes (Figure 2). This is probably due to the existence of two analog EC:3.1.3.15 enzymes (*S. cerevisiae*- and *E. coli*-types). Both types are highly divergent in sequence, and when we relaxed the stringency of BRH analysis (increasing the threshold E-value from 10^{-6} to 10^{-1}), we detected orthologs in 84% and 40% of the analyzed genomes for the *S. cerevisiae* and *E. coli* types, respectively. The other enzymes analyzed in this study are not affected by the stringency of BRHs. Additionally, we found that animals, with the exception of *N. vectensis*, have experienced a secondary loss of the L-histidine biosynthetic machinery (Figure 2). Taking these results together, we suggest that the LCA had the same L-histidine synthesis pathway as extant species.

L-threonine

Two out of the three L-threonine biosynthetic enzymes from *E. coli* were found across the three domains. We did not find any orthologs in Archaea when we performed a genome scan with the *E. coli* threonine synthase (EC:4.2.3.1) as seed. Alternatively, when we used as seed an *M. jannaschii* paralog with the same function, we identified orthologs in Archaea (Figure 2). Again, this finding reinforces the importance of using enzymes from multiple species as seeds. Some animals apparently lost the biosynthetic machinery for this amino acid, but *N. vectensis* retained it. We suggest that the LCA could synthesize L-threonine like contemporary species.

L-glutamine and L-glutamate

As depicted in Figure 1, the inter-conversion of L-glutamine and L-glutamate can be performed by many alternolog enzymes. Both paralog glutamate synthases, the NADH dependent (EC:1.4.1.14) and the NADPH dependent (EC:1.4.1.13), produce L-glutamate from L-glutamine, and are widely distributed across the three domains (Figure 2). In the reverse direction, from L-glutamate to L-glutamine, we found that glutamine synthetase (EC:6.3.1.2), which is ATP dependent, is also widely distributed across the three domains. This suggests that the LCA was able to inter-convert L-glutamine and L-glutamate. But it leaves one open question: was the LCA capable of producing these amino acids independently of each other? Similarly to glutamate synthases, both paralog glutamate dehydrogenases, the $NAD(P)^+$-dependent (EC:1.4.1.3) and the $NADP^+$-dependent (EC:1.4.1.4) enzymes, produce L-glutamate from 2-oxoglutarate and ammonia, and are also widely distributed across the three domains. On the other hand, all other reactions synthesizing L-glutamine use L-glutamate as substrate and are sparsely distributed. In summary, we suggest that the LCA was able to synthesize L-glutamate from 2-oxoglutarate and inter-convert it with L-glutamine, but it is difficult to determine if the LCA was able to produce this last amino acid independently of the former one.

L-cysteine

There are at least four ways to synthesize L-cysteine (Figure 1). The most widely distributed, proceeding via cystathionine, uses cystathionine β-synthase (EC:4.2.1.22) and cystathionine γ-lyase (EC:4.4.1.1) and is documented as being eukaryotic-type, yet we found it distributed across the three domains (Figure 2). Alternatively, cystathionine-β-lyase (EC:4.4.1.8), cystathionine γ-synthase (EC:2.5.1.-) and O-succinylhomoserine(thiol)-lyase (EC:2.5.1.48) catalyze equivalent reactions and they are widely distributed in Bacteria and Eukarya. In contrast, an alternolog branch using EC:2.5.1.47 via O-acetyl-L-serine is sparsely distributed across genomes (Figure 2), while another branch without assigned enzymes (nor genes) uses O-acetyl-L-homoserine. These findings suggest that not all the L-cysteine biosynthetic pathways occurred in the LCA, but that the contemporary eukaryotic-like type could.

Eight amino acid biosynthetic pathways are partially distributed across the three domains of life, and five of their branches probably occurred in the LCA

L-lysine

L-lysine biosynthesis has been used largely to exemplify the existence of alternolog branches in amino acid biosynthesis [21-23]. Six alternative pathways can be recognized for the biosynthesis of L-lysine (Figure 1), grouped in two superpathways proceeding via either L,L-diaminopimelate or alpha-aminoadipate. The superpathway involving L,L-diaminopimelate has four alternolog branches, corresponding to L-lysine biosynthesis types I, II, III and VI in MetaCyc; they share a common set of six reactions catalyzed by widely distributed enzymes. Four of these enzymes catalyze the upper steps of the superpathway, from aspartate kinase (EC:2.7.2.4) to dihydrodipicolinate reductase (EC:1.3.1.26), and form the pairs of duplicated genes between the biosynthesis of L-arginine/L-lysine (Figure 3). The other two enzymes (EC:5.1.17 and EC:4.1.120) catalyze the lower portion of the superpathway. The TDs of enzymes catalyzing intermediate steps in these alternologs are as follow. In the type I pathway (*E. coli*-type), which is catalyzed by three enzymes, only N-succinyl-L,L-diaminopimelate desuccinylase (EC:3.5.1.18) is widely distributed across the three domains. In the type II pathway (*B. subtilis*-type), catalyzed by the other three enzymes, only tetrahydrodipicolinate acetyltransferase (EC:2.3.1.89) is widely distributed in Bacteria, while it is absent in Archaea and Eukarya. The type III pathway of *Corynebacterium glutamicum* (EC:1.4.1.16) appears constrained to some actinobacteria and firmicutes, while the recently discovered type VI pathway, formed by a single enzyme, namely L,L-diaminopimelate aminotransferase (EC:2.6.1.-), seems to be specific for plants. These results illustrate a general finding of this work: linear pathways seem to be more widely distributed than bifurcating ones. As described above, L-histidine, L-tryptophan and L-leucine pathways support this observation, and correlate with previous studies showing that within amino acid biosynthesis, larger pathways tend to have lower rates of change in their structure than shorter pathways [31]. However, further studies on whole metabolic networks are necessary to assess the generality of this property in the evolution of metabolism. On the other hand, the second superpathway, proceeding via the degradation of alpha-aminoadipate, is formed by lineage specific type IV and V pathways that share a core of five reactions from homocitrate synthase (EC:2.3.3.14) to α-aminoadipate aminotransferase (EC:2.6.1.39). This core contains the four enzymes forming pairs of duplicated genes between the biosynthesis of L-leucine/L-lysine (Figure 3). The type V pathway, using N-2-acetyl-L-lysine (RXN-5181 to RXN-5185), was characterized in the Thermus-Deinococcus lineage, and its representatives were found in Archaea and some Bacteria, while the type IV pathway, proceeding via saccharopine (EC:1.2.1.31 to EC:1.5.1.7), appears restricted to Eukarya and some Bacteria. Collectively, the TDs of these two superpathways show that alternative pathways have led the origin of the biosynthesis of L-lysine. None of these alternologs appears to be universally distributed and, thus, the LCA probably was not able to produce L-lysine using the set of enzymes analyzed here. Interestingly, both L-lysine biosynthetic superpathways retain groups of duplicated genes for the biosynthesis of L-leucine and L-arginine (Figure 3), which, as detailed above, probably occurred in the LCA. Thus, there is a possibility that L-lysine biosynthesis was incorporated into metabolism from L-leucine and L-arginine biosynthetic routes.

L-methionine

The biosynthesis of L-methionine can be carried out by at least three different superpathways (Figure 1). One involves the degradation of cystathionine via homocysteine using either cystathionine β-synthase (EC:4.2.1.22) or cystathionine β-lyase (EC:4.4.1.8), followed by methionine synthase (EC:2.1.1.13). These three enzymes are widely distributed across the three domains (Figure 4) and, hence, this branch could occur in the LCA. Alternatively, the second superpathway, also called the L-methionine salvage cycle, which begins with EC:4.4.1.14 via S-adenosyl-L-methionine and finishes in L-methionine using EC:2.6.1.5 via 2-oxo-4-methylthiobutanoate (Figure 1), is widely distributed in Eukarya but almost absent in Archaea and Bacteria. An exception to this distribution is the step from L-methionine to S-adenosyl-L-methionine, which can be catalyzed by one of two analog methionine adenosyltransferases (EC:2.5.1.6). These analogs show an almost perfect anti-correlation in their TDs (Figure 4); one is

Figure 2 *(see following page)*
Average taxonomic distribution of amino acid biosynthetic enzymes widely distributed across the three domains of life. The TDs for enzymes catalyzing the amino acid biosynthetic pathways (vertical labels) were computed by searching for their ortholog distribution across diverse taxonomic groups (horizontal labels). The plot shows enzymes with an average normalized distribution ≥ 50% (see Materials and methods). Amino acid three letter codes in red denote amino acids whose biosynthesis probably occurred in the LCA (detailed in the main text). Four types of seeds were used to look for TDs: the canonical *E. coli* enzymes (gray scale); homolog enzymes - paralogs and orthologs - from other species showing a higher distribution than *E. coli* counterparts (yellow scale); analog enzymes - catalyzing the same reaction and coming from a different structural superfamily - (red scale); and alternolog enzymes and branches - converging in the same end compound, but proceeding via different metabolites - in other species (blue scale). In the vertical labels, subunits of multimeric enzymes are denoted with 'S', analog enzyme machinery is denoted with 'A' and isoenzymes are denoted with 'I'. For example, the annotation 'EC:3.5.1.1(Eco_Ans-AnsB)(A:1/2-I:1/2)' indicates that there are two analog EC:3.5.1.1 enzymes and this annotation corresponds to the first type (A:1/2). In turn, this type has two isoenzymes and this annotation corresponds to the first one (I:1/2), formed by AnsA and AnsB proteins in *E. coli*. The average distribution of orthologs for each route is shown in parentheses following amino acid three letter codes. Biosynthetic enzymes for each amino acid were sorted as they appear downstream in the metabolic flux.

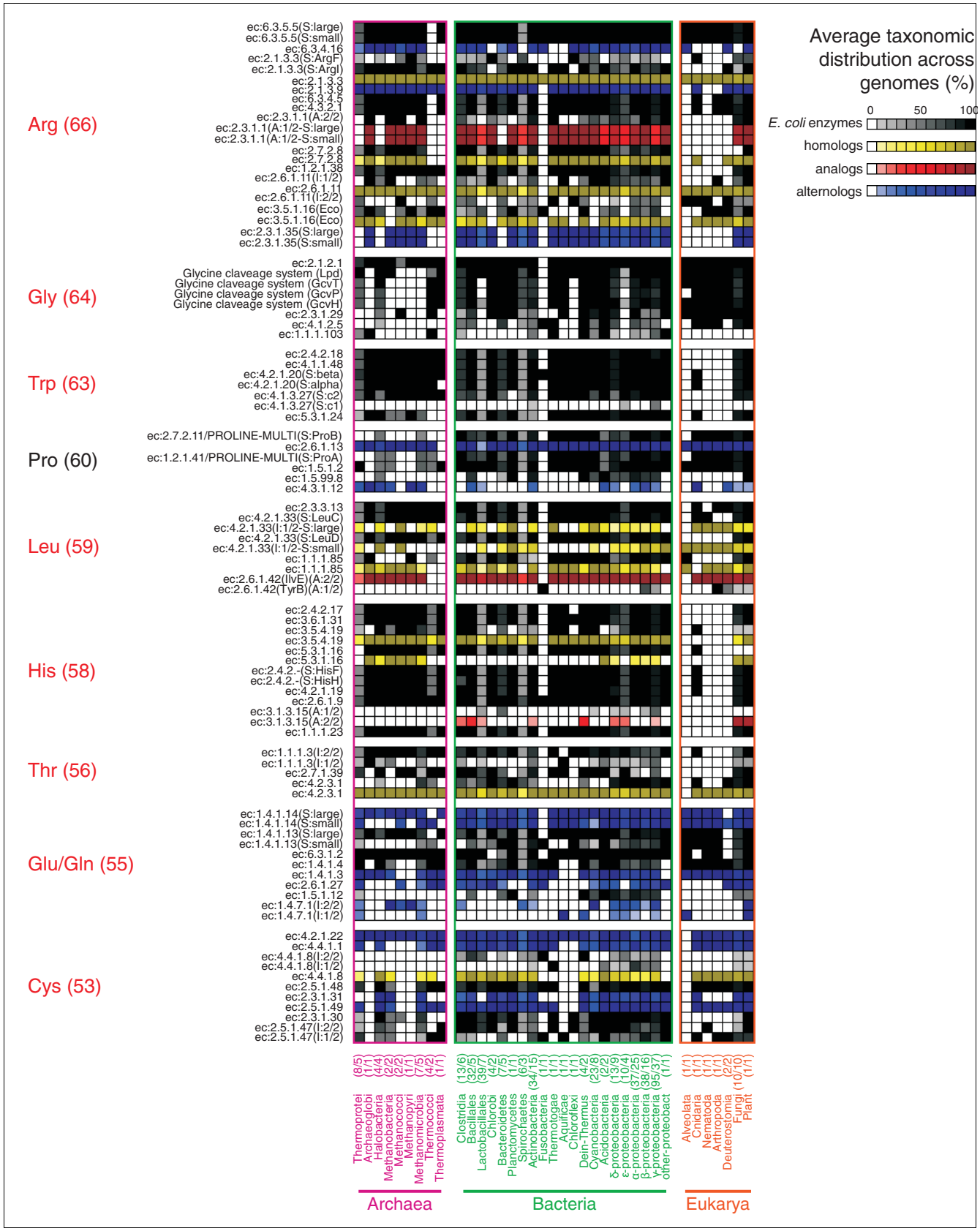

Figure 2 *(see legend on previous page)*

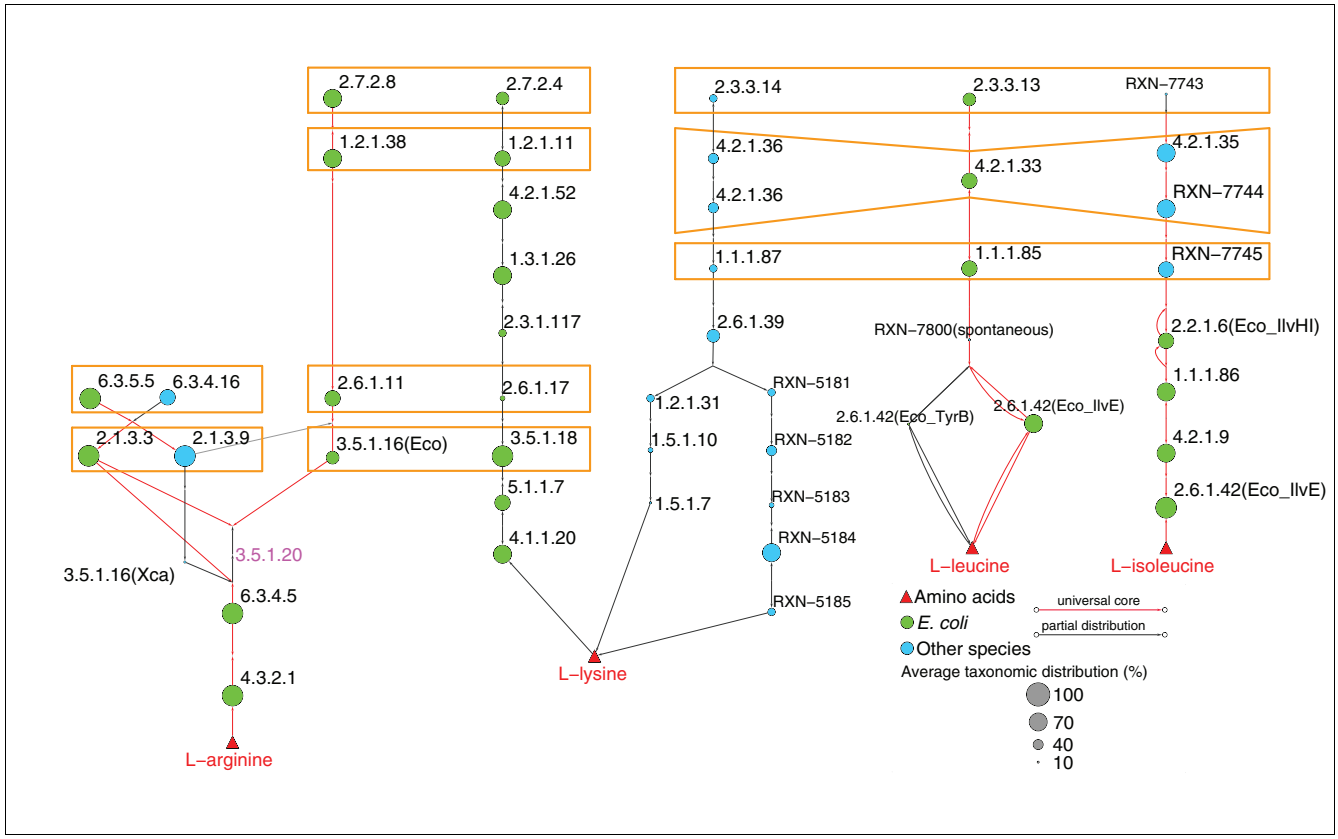

Figure 3
Retention of duplicates as groups instead of as single entities. Orange frames indicate pairs of duplicated genes (paralog enzymes) retained as groups instead of as single entities between the biosynthesis of L-arginine, L-lysine, L-leucine and L-isoleucine.

restricted to Archaea, while the other occurs in Bacteria and Eukarya. Complementarily, a third superpathway, characterized in plants as the so-called S-adenosyl-L-methionine cycle, converts S-adenosyl-L-methionine to L-methionine via S-adenosyl-L-homocysteine (Figure 1). We found that one of this cycle's enzymes, S-adenosylhomocysteine hydrolase (EC:3.3.1.1), is widely distributed across the three domains. In summary, we suggest that the LCA was able to produce L-methionine, degrading cysthationine via homocysteine.

L-valine and L-isoleucine

The terminal four steps in the biosynthesis of L-valine and L-isoleucine employ a common set of widely distributed enzymes, from EC:2.2.1.6 to branched-chain amino-acid aminotransferase (EC:2.6.1.42) (Figure 4). This set was not found, however, in animals, again with the exception of *N. vectensis*. Complementarily, five alternolog branches can catalyze the initial steps of L-isoleucine biosynthesis, converging in 2-oxobutanoate, which is, in turn, a substrate of acetolactate synthase (EC:2.2.1.6) (Figure 1). We found that the canonical *E. coli* branch carrying out these steps via propionate uses EC:2.7.2.15 and EC:2.3.1.8 and is sparingly distributed among bacterial genomes. In contrast, the alternolog branch characterized in spirochaetes, proceeding via (R)-citramalate (Figure 1), uses isopropylmalate isomerase (EC:4.2.1.35) and β-isopropylmalate dehydrogenase (no EC number assigned), and both enzymes are widely distributed across the three domains (Figure 4). These results clearly exemplify that the *E. coli* canonical pathways are not necessarily the most widely distributed ones and, thus, alternolog pathways must be included in evolutionary analysis. Additionally, this branch participates in the retention of a group of duplicated genes catalyzing consecutive reactions in the biosynthesis of L-lysine, L-leucine and L-isoleucine (Figure 3). Taking together the wide distribution of the spirochaetes-like branch and the enzymes shared between L-valine and L-iso-

Figure 4 *(see following page)*
Average taxonomic distribution of amino acid biosynthetic enzymes partially distributed across the three domains of life. TDs for enzymes with an average normalized distribution <50% (see Materials and methods). Labels and colors are as in Figure 2.

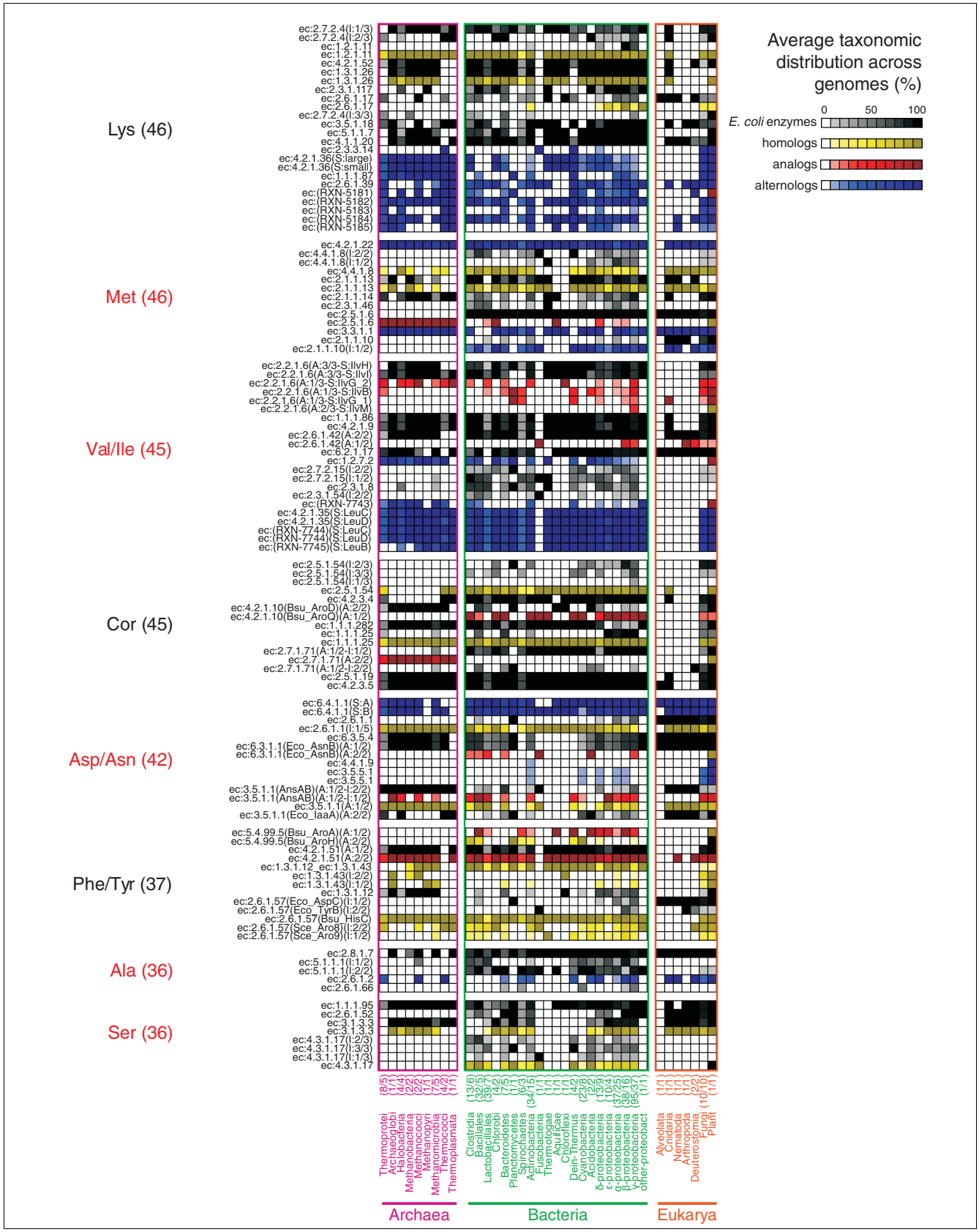

Figure 4 *(see legend on previous page)*

leucine biosynthesis, we suggest that the LCA and even contemporary species could combine these branches to synthesize both amino acids.

Chorismate

Chorismate is not an amino acid itself, but it is a key compound in the biosynthesis of aromatic amino acids and we consider the distribution of their catalyzing enzymes particularly interesting. The biosynthesis of chorismate comprises seven steps, the last two being catalyzed by two widely distributed enzymes, 3-phosphoshikimate-1-carboxyvinyltransferase (EC:2.5.1.9) and chorismate synthase (EC:4.2.3.5). Complementarily, the first two steps are catalyzed by enzymes widely distributed in Bacteria and some Eukarya, but absent in Archaea. A recent report suggesting a novel pathway for the biosynthesis of aromatic amino acids and *p*-aminobenzoic acid in the archaeon *Methanococcus maripaludis* helps to understand this distribution [32]. Additionally, three intermediate steps are catalyzed by scarcely distributed analog and alternolog enzymes as follows. First, the transformation of 3-dehydroquinate to 3-dehydro-shikimate can be catalyzed by two analog 3-dehydroquinate dehydratases (EC:4.2.1.10). *B. subtilis* possesses both analogs, while Archaea, some Eukarya and a few Bacteria carry only the type II enzyme (Figure 4) belonging to the aldolase (TIM-barrel) superfamily. In contrast, the majority of Bacteria, including *E. coli*, uses the type I enzyme (Figure 4) belonging to the 3-dehydroquinate dehydratase superfamily. Second, in *E. coli* there are two paralogs catalyzing the conversion of 3-dehydro-shikimate to shikimate. One of them, $NADP^+$-dependent EC:1.1.1.25, is widely distributed, while EC:1.1.1.282 (using either NAD^+ or $NADP^+$, and either quinate or shikimate) is sparsely distributed. In contrast, *B. subtilis* has only the $NADP^+$-dependent shikimate dehydrogenase and, when its sequence is used as a seed for BRHs, we found more orthologs than with the *E. coli* counterparts (Figure 4). This finding is probably caused by cross-matches between the *E. coli* paralogs during the construction of TDs. Third, the transformation of shikimate to shikimate-3-phosphate can be catalyzed by two analog shikimate kinases (EC:2.7.1.71). The archaeal-type belongs to the GHMP kinase superfamily, while the bacterial/eukaryotic-type belongs to the superfamily of P-loop containing nucleoside triphosphate hydrolases. Interestingly, there is an almost perfect anti-correlation between the TDs of these enzymes (Figure 4). Animals, including *N. vectensis*, have lost all enzymes catalyzing intermediate steps in chorismate biosynthesis, supporting the fact that aromatic amino acids (L-histidine, L-trypthopan, L-phenylalanine, and L-tyrosine) are essential for humans. Summarizing, we found that the lower portion of chorismate biosynthesis, converting 3-dehydro-shikimate to chorismate, is widely distributed across the three domains, suggesting that it probably occurred in the LCA. In contrast, the upper and intermediate portions of this route appear to have originated independently in specific lineages during evolution.

L-aspartate and L-asparagine

The biosynthesis and inter-conversion of L-aspartate and L-asparagine are mediated by a diverse set of alternolog enzymes (Figure 1), most of which have been characterized in *E. coli* and are sparsely distributed. Nevertheless, aspartate aminotransferase (EC:2.6.1.1) and pyruvate carboxylase (EC:6.4.1.1) are able to produce L-aspartate from pyruvate, via oxaloacetate, and both enzymes are widely distributed across the three domains (Figure 4). Complementarily, the conversion of L-aspartate to L-asparagine can be carried out by three asparagine synthetases, two of which are glutamine dependent (EC:6.3.5.4) while the other is ammonia dependent (EC:6.3.1.1). Both EC:6.3.1.1 type 1 and EC:6.3.5.4 belong to the adenine nucleotide alpha hydrolases-like superfamily and are widely distributed across the three domains (Figure 4). In contrast, the production of L-aspartate and L-asparagine via 3-cyano-L-alanine, which is mediated by β-cyano-L-alanine-synthase (EC:4.4.1.9) and two paralog nitrilases (EC:3.5.5.1), appears to be restricted to plants, cyanobacteria and α-proteobacteria (Figure 4). This distribution could be the product of horizontal gene transfer among these clades, probably by symbiosis - as some α-proteobacteria are symbionts and parasites of plants - or by endosymbiosis - because cyanobacteria are considered descendants of plastid ancestors in plants. We did not detect any other possible horizontal gene transfer events in these routes using a database of putative horizontally transferred genes in prokaryotic complete genomes [33]. Finally, the two analog asparaginases (EC:3.5.1.1), converting L-asparagine to L-aspartate, show anti-correlated TDs. One of them, from the glutaminase/asparaginase superfamily, was found in Archaea, some Bacteria, Fungi and Animals (Figure 4), while the second one, from the superfamily of amino-terminal nucleophile aminohydrolases shows a distribution similar to that of EC:4.4.1.9 and EC:3.5.5.1. In summary, the LCA probably was not able to produce either L-aspartate or L-asparagine via the modern canonical alternologs (nitrilase and asparaginase), but could via the degradation of oxaloacetate using the branches described above.

L-tyrosine and L-phenylalanine

There are at least five branches diverging from prephenate for the biosynthesis of L-tyrosine and L-phenylalanine. Two of them proceed via phenylpyruvate and use one of the two widely distributed analog prephenate dehydratases (EC:4.2.1.51). Another two branches proceed via L-arogenate and use either arogenate dehydrogenase (EC:1.3.1.43) to synthesize L-tyrosine or arogenate dehydratase (EC:4.2.1.91) to synthesize L-phenylalanine. EC:1.3.1.43 occurs in Bacteria and some Archaea, while EC 4.2.1.91 has no assigned enzyme (nor gene) sequences. The fifth branch uses prephenate dehydrogenase (EC:1.3.1.12) followed by an aromatic-amino acid aminotransferase (EC:2.6.1.57). *E. coli*, *B. subtilis* and *S. cerevisiae* have two EC:2.6.1.57 and all of them can be classified in the PLP-dependent transferase superfamily, with the exception of AroJ in *B. subtilis*, whose sequence is unknown.

However, it is difficult to establish orthology relationships between these enzymes because none of them are BRHs and, thus, are putatively paralogs. Apparently, this high diversity is maintained by differential expression and multifunctional properties of these enzymes. For instance, TyrB in *E. coli* is approximately 1,000-fold more active on aromatic substrates than its paralog AspC, which is more specific for aspartate. Similarly, in *B. subtilis*, HisC is more active than AroJ on phenylalanine and tyrosine, in spite of its primary activity on histidinol-phosphate, and *S. cerevisiae* uses Aro8 preferentially in anabolism and Aro9 in catabolism [34,35]. HisC could represent one of the most ancestral lineages in this family because it is the only member widely distributed across the three domains. This agrees with the fact that biosynthesis of L-histidine, the pathway in which HisC preferentially participates, is proposed to be ancestral (see above). The wide distribution of HisC and two analog EC:4.2.1.51 enzymes suggests that the biosynthesis of phenylalanine and tyrosine is also ancient. Nevertheless, the step preceding these enzymes, from chorismate to prephenate, can be catalyzed by one of the two analog chorismate mutases (EC:5.4.99.5); these show a sparse distribution, with some representatives in firmicutes, proteobacteria and plants. For instance, *E. coli* possesses two paralog EC:5.4.99.5 enzymes belonging to the chorismate mutase II superfamily and they are fused to domains catalyzing EC:1.3.1.12 and EC:4.2.1.51 activities, while *B. subtilis* also has two EC:5.4.99.5 enzymes, one from the chorismate mutase II superfamily and the other from the YjgF-like superfamily. Neither of these EC:5.4.99.5 domains is widely distributed; thus, it is difficult to establish whether the LCA was able to synthesize L-phenylalanine and L-tyrosine.

L-alanine

There are four alternolog single steps to synthesize L-alanine (Figure 1). One of them uses cysteine desulfurase (EC:2.8.1.7) to degrade L-cysteine and is widely distributed across the three domains. Given that L-cysteine biosynthesis probably occurred in the LCA (see above), we suggest that biosynthesis of L-alanine could occur in the LCA via this step. In contrast, alanine racemase (EC:5.1.1.1) isomerizes D-alanine to L-alanine and is constrained to Bacteria. Alanine aminotransferase (EC:2.6.1.2) converts L-glutamate and pyruvate to L-alanine and is widely distributed in Eukarya but poorly represented in Bacteria and Archaea, whereas valine-pyruvate aminotransferase (EC:2.6.1.66) degrades L-valine to L-alanine and was detected only in few Bacteria (Figure 4).

L-serine

Finally, there are four proficient branches to synthesize L-serine. The first, proceeding via 3-phospho-hydroxypyruvate, is catalyzed by three enzymes, two of them, 3-phosphoglycerate dehydrogenase (EC:1.1.1.95) and 3-phosphoserine phosphatase (EC:3.1.3.3) are widely distributed, but the third, phosphoserine aminotransferase (EC:2.6.1.52) is restricted to Eukarya and some Bacteria. The second branch is a single step converting ammonia and pyruvate to L-serine by L-serine ammonia-lyase (EC:4.3.1.17), and is restricted to some Bacteria. The third and fourth branches are closely related to the biosynthesis of L-cysteine and L-methionine, inter-converting cystathionine and homocysteine by either EC:4.2.1.22 or EC:4.4.1.8, which are widely distributed across the three domains (see above). Given that L-cysteine and L-methionine could exist in the LCA, the biosynthesis of L-serine could also exist via these enzymes. In fact, this chain of widely distributed enzymes can be extended to the biosynthesis of L-alanine (Figure 1), and all of them together constitute the larger succession of reactions that probably existed in the LCA.

In summary, our results have uncovered a set of 64 enzyme domains participating in the biosynthesis of at least 16 out of the 20 proteinogenic amino acids that tentatively occurred in the LCA. Figure 5 shows a marked bias in the taxonomic distribution of this set of domains with respect to the general trend for the whole metabolism and other less conserved parts of amino acid biosynthesis, suggesting that branches in other metabolic processes could also posses a hidden universality.

Conclusion

We have carried out a comprehensive analysis of the origin and evolution of amino acid biosynthesis. Our strategy combines genomic tools with a network perspective to identify a core of widely distributed enzymes that probably occurred in the LCA, synthesizing at least 16 out of the 20 standard amino acids. This proposal does not imply, however, that the full biosynthetic routes for these 16 amino acids appeared early, but only some of their branches that could satisfy the minimal biochemical and structural requirements. It is important to note that some species such as parasites and free living animals, including mammals, can lack significant portions of this 'universal' set because they can import amino acids from their hosts or include it in their diet. In parasites, these absences have been attributed to secondary loss. Our results show that most basal animal lineages and other Eukarya posses these universal branches and, thus, their absence in the animal kingdom apparently is also due to secondary losses. Further studies on the possible occurrence of the remaining four standard amino acid biosynthetic routes - for L-proline, L-lysine, L-phenylalanine and L-tyrosine - are necessary as some portions of these pathways are also widely distributed and some 'lost' reactions could fill the gaps.

One of the major biological roles of amino acids is that they are protein constituents; thus, an emerging question from our results is whether this core of amino acids could be sufficient for the LCA protein repertoire? Recently, Atchley *et al.* [36], grouped amino acids according to almost 500 attributes, ranging from structural to biochemical and biophysical properties, producing a multidimensional representation of amino acid variability. Mapping the putative core of 16

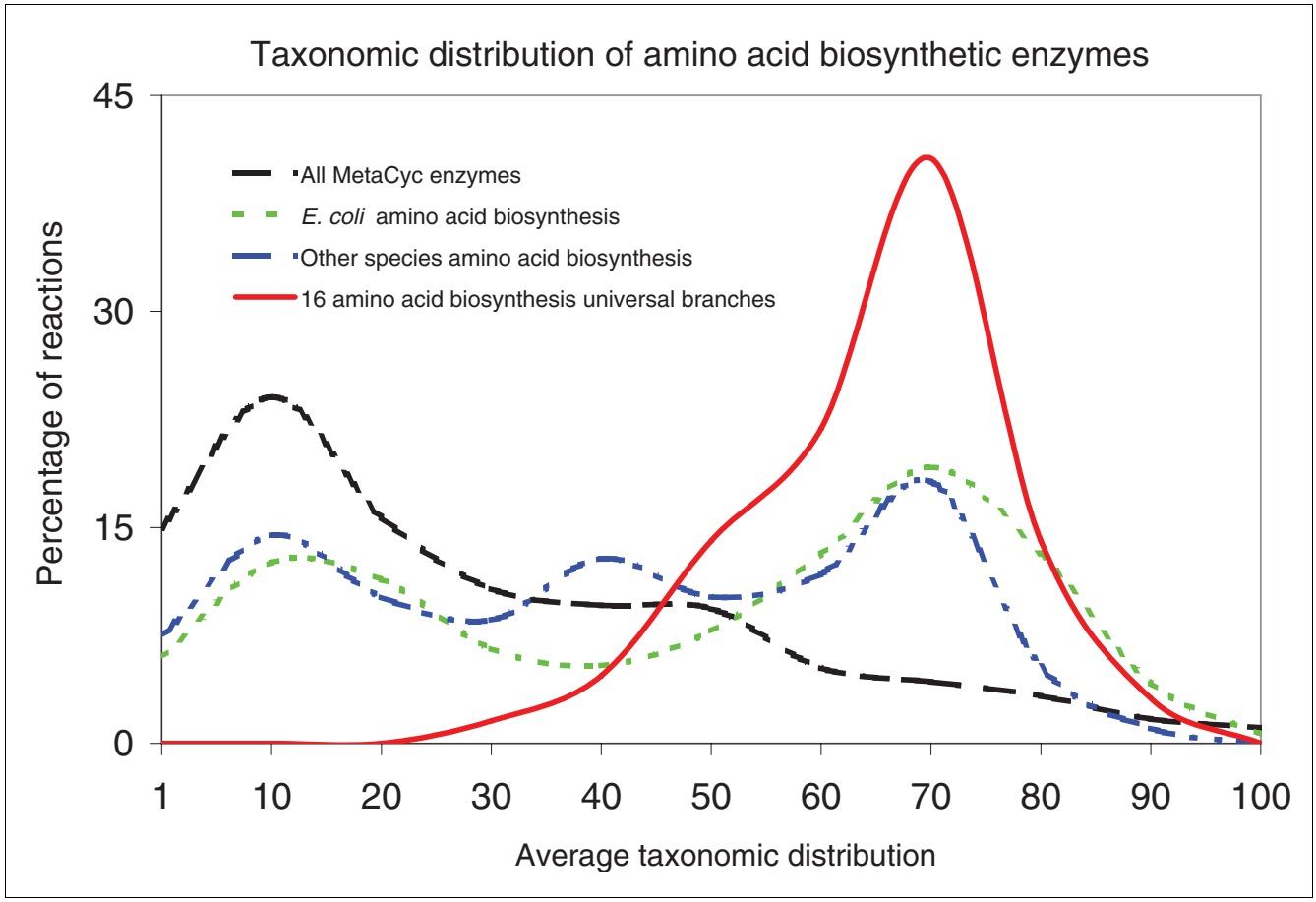

Figure 5
Taxonomic distribution of amino acid biosynthesis. Conservation of amino acid biosynthesis from the *E. coli*-centric and multi-organismal seed perspectives. The general trend in the whole of metabolism (MetaCyc), using a manually depurated set of enzyme domains, is also shown. The 16 amino acid biosynthesis universal branches show a maximum of around 45% of reactions (y-axis) in 70% of sampled genomes (x-axis) when all MetaCyc enzymes have a maximum of 24% of reactions in only 10% of genomes.

ancient amino acid biosynthetic branches onto the Atchley *et al*. plot (Figure 1 in Ref. [36]) suggests that the LCA was able to populate all the regions of amino acid variability space, so this core could be sufficient for protein functions in early biosystems. Most of the universal branches found in this work are connected to each other (Figure 1), allowing the possibility that they feedback and complete a minimal set of enzyme-driven reactions for the biosynthesis of amino acids. Interplay between the variability and selective pressures resulting from increasing genome sizes and protein structure complexity could promote the incorporation of novel amino acids to this core.

Additionally, we identified alternative branches and routes (paralogs, analogs and alternologs) reflecting the adoption of specific amino acid biosynthetic strategies by taxa, probably due to differences in their life-styles. Eleven out of the twenty amino acid biosynthetic routes revealed an important contribution of paralogy to the generation of diversity. In particular, we corroborated that the retention of gene duplicates as groups, instead of as single entities, is an important factor in the evolution of metabolism. Furthermore, analog enzymes contribute in eight out of the twenty standard amino acid biosynthetic routes, while alternolog routes participate in nine. This implies that analog enzymes and alternolog branches contribute almost as much as gene duplication to genetic buffering in the biosynthesis of amino acids. Further studies are necessary to determine the generality of these observations and to complement them with observations from alternative reactions modeling fluxes in metabolism [37]. In conclusion, we suggest that despite a core of amino acid biosynthetic branches being inherited from ancient systems, the whole contemporary repertoire has been originated independently by lineages according to their environmental resources as reflected by the high diversity of anabolic branches.

In this sense, we consider that one of the goals of the two step strategy presented here (*E. coli* and multi-organismal TD seeds) is that it uses not only a traditional model organism for

genomic analyses, but also as many species as available in current databases. This is important because 8 out of the 20 amino acid biosynthetic routes (L-cysteine, L-serine, L-alanine, L-isoleucine, L-arginine, L-aspartate, L-proline and L-methionine) were quite sparingly distributed from an *E. coli*-centric perspective, but widely distributed when adding the orthologs, paralogs, analogs and alternologs from other species, revealing the universal nature of some of these routes. Further studies are necessary to determine the generality of these findings, not only in metabolic networks but also in other biological processes.

Materials and methods

Network reconstruction

In bipartite metabolic networks, there are two sets of nodes - enzymes and compounds (substrates, products and cofactors) - and edges relating enzymes with compounds occurring in the same reaction. For instance, if reaction R1 consumes compound C1 and produces C2 and C3, and it is catalyzed by enzyme E1, the following edges are established: C1 → E1, E1 → C2, and E1 → C3. In reversible reactions a second group of links from products to enzymes and, in turn, from enzymes to substrates is added. In this work we reconstructed the bipartite networks derived from three metabolic databases: EcoCyc v8.0 [8] for *E. coli*, MjCyc [10] for *M. jannaschii* and MetaCyc v8.0 [9] for multi-organismal assignments. To obtain information concerning the nodes and edges for each reaction, we used the following files from EcoCyc and MetaCyc: reactions.dat (substrate/product), enzrxns.dat (reversibility) and reaction-links.dat (EC numbers). From MjCyc the corresponding information was retrieved manually from the database's web page. Networks derived from these databases were merged and prepared for presentation with Cytoscape v2.5.2 [38]. Amino acids were highlighted (red triangles in Figure 1) to denote terminal points of pathways and branches into this network. For clarity in presentation, the most highly connected compounds (mainly cofactors) and the terminal non-amino acid metabolites were removed from the network. Additional data file 2 lists these compounds and contains the pairs of nodes used to construct Figure 1. Multifunctional enzyme sequences were split manually according to their functional domain assignments from Swiss-Prot [39]. Thus, in Figure 1 each node represents one reaction catalyzed by a functional domain (or domains in multimeric enzymes). Analogue enzymes - those catalyzing the same reaction but possessing different folds - were detected by comparing the structural domain content among proteins according to the Superfamily database v1. 69 [40] using HMMer [41]. Additional data file 2 contains details for the final set of 537 enzyme functional domains analyzed in this work as well as 32 reactions without known gene/enzymes. Alternologs were detected by manual inspection of the network in Figure 1, looking for branches that, proceeding via different metabolites, converge in a given compound, generally in an amino acid.

Taxonomic distribution

Amino acid sequences from 537 enzymes (functional domains) were tracked in completely sequenced genomes using BLASTP (cutoff E-value = 10^{-20}, and identity percentage >95), this was carried out to obtain the corresponding genomic sequences used as seeds for ortholog detection. Sixty-nine enzymes were excluded because they do not have assigned enzyme (gene) sequences (Additional data file 2). Each genomic sequence seed was used for ortholog detection across 410 non-obligate parasitic genomes (30 Archaea, 363 Bacteria and 17 Eukarya) representing 297 species from 192 genera (Additional data file 2), following the BRH criterion using BLASTP with a cut E-value of 10^{-5}, and a minimum alignment coverage for query and/or subject sequence ≥ 50%. Genomes with less than 1,500 predicted open reading frames (mainly from obligate parasitic genomes) were eliminated from this analysis because they have experienced extensive secondary loses of anabolic enzymes in their genomes, which could introduce noise in TDs. Sequences in the same genome with >95% identity estimated with CD-HIT [42] were grouped into clusters. As reported [43], this procedure reduces the frequency of false-negative results caused by cross-matches between highly similar sequences within a genome. Given the redundancy of sequenced strains for some bacterial species, we systematically depurated the original set of genomes, attempting to obtain a normalized measure of ortholog distribution according to the following steps. In step 1, a TD of enzymes versus species was constructed assigning a value of1 to enzymes with orthologs (BRHs) in ≥ 50% of strains from each species. Otherwise a value of 0 was assigned to the corresponding bit. In step 2, a TD of enzymes versus genera was constructed. In this TD, each vector represents the percentage of species having a value of 1 (assigned in step 1) for each genus. In step 3, a TD of enzymes versus clades was constructed. In this TD each vector represents an average of the genera percentages obtained in step 2, for each clade. Clades correspond to the taxonomic categories from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [7]. This procedure provides a 'normalized average distribution' of enzymes across genomes. The final set of strains and genera for each clade is shown in Additional data file 2.

Abbreviations

BRH, best reciprocal hits; LCA, last common ancestor; TD, taxonomic distribution.

Authors' contributions

GHM and JJDM carried out the analyses under the supervision of EPR and LS. All the authors planned the project and wrote the manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is a graph showing a detailed view of the bipartite network analyzed in this work. Additional data file 2 provides details of enzymes analyzed in this work. Additional data file 3 is a detailed view of pathways and branches analyzed in this work. Additional data file 4 is a bipartite graph of the final metabolic network analyzed in this work (after hub and end compounds removal).

Acknowledgements

We would like to thank Alejandra Covarrubias for helpful discussions and comments during the development of this project. We also want to thank Areli Morán for general help throughout this work. This work was partially supported by grant 43502 from the Mexican Science and Technology Research Council (CONACYT). GHM was supported by doctoral scholarship from CONACYT. JJDM has been partially supported by doctoral and post-doctoral fellowships from CONACYT, DGEP-UNAM and Fulbright-García Robles. EPR was partially supported by a grant (ASTF 224-2005) from EMBO. A substantial part of this work was conducted on the Cluster "Sputnik II" from Instituto de Biotecnología-UNAM sponsored by "Macroproyecto de Tecnologías de la Información y la Computación de la UNAM".

References

1. Horowitz NH: **On the evolution of biochemical synthesis.** *Proc Natl Acad Sci USA* 1945, **31:**153-157.
2. Ohno S: *Evolution by Gene Duplication* New York: Springer; 1970.
3. Jensen RA: **Enzyme recruitment in the evolution of new function.** *Annu Rev Microbiol* 1976, **30:**409-425.
4. Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS: **The 'evolvability' of promiscuous protein functions.** *Nat Genet* 2005, **37:**73-76.
5. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654.
6. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268:**1803-1810.
7. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
8. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: **The EcoCyc Database.** *Nucleic Acids Res* 2002, **30:**56-58.
9. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucleic Acids Res* 2004, **32:**D438-442.
10. Tsoka S, Simon D, Ouzounis CA: **Automated metabolic reconstruction for *Methanococcus jannaschii*.** *Archaea* 2004, **1:**223-229.
11. Huynen MA, Dandekar T, Bork P: **Variation and evolution of the citric-acid cycle: a genomic perspective.** *Trends Microbiol* 1999, **7:**281-291.
12. Nishida H: **Evolution of amino acid biosynthesis and enzymes with broad substrate specificity.** *Bioinformatics* 2001, **17:**1224-1225.
13. Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C: **The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*.** *J Mol Biol* 2001, **311:**693-708.
14. Alves R, Chaleil RA, Sternberg MJ: **Evolution of enzymes in metabolism: a network perspective.** *J Mol Biol* 2002, **320:**751-770.
15. Light S, Kraulis P: **Network analysis of metabolic enzyme evolution in *Escherichia coli*.** *BMC Bioinformatics* 2004, **5:**15.
16. Diaz-Mejia JJ, Perez-Rueda E, Segovia L: **A network perspective on the evolution of metabolism by gene duplication.** *Genome Biol* 2007, **8:**R26.
17. Galperin MY, Walker DR, Koonin EV: **Analogous enzymes: independent inventions in enzyme evolution.** *Genome Res* 1998, **8:**779-790.
18. Kitami T, Nadeau JH: **Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication.** *Nat Genet* 2002, **32:**191-194. (corrigendum in December 2002).
19. Papp B, Pal C, Hurst LD: **Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast.** *Nature* 2004, **429:**661-664.
20. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297:**1551-1555.
21. Hudson AO, Bless C, Macedo P, Chatterjee SP, Singh BK, Gilvarg C, Leustek T: **Biosynthesis of lysine in plants: evidence for a variant of the known bacterial pathways.** *Biochim Biophys Acta* 2005, **1721:**27-36.
22. Miyazaki J, Kobashi N, Nishiyama M, Yamane H: **Functional and evolutionary relationship between arginine biosynthesis and prokaryotic lysine biosynthesis through alpha-aminoadipate.** *J Bacteriol* 2001, **183:**5067-5073.
23. Nishida H, Nishiyama M, Kobashi N, Kosuge T, Hoshino T, Yamane H: **A prokaryotic gene cluster involved in synthesis of lysine through the amino adipate pathway: a key to the evolution of amino acid biosynthesis.** *Genome Res* 1999, **9:**1175-1183.
24. Irvin SD, Bhattacharjee JK: **A unique fungal lysine biosynthesis enzyme shares a common ancestor with tricarboxylic acid cycle and leucine biosynthetic enzymes found in diverse organisms.** *J Mol Evol* 1998, **46:**401-408.
25. Benner SA, Ellington AD, Tauer A: **Modern metabolism as a palimpsest of the RNA world.** *Proc Natl Acad Sci USA* 1989, **86:**7054-7058.
26. Cunchillos C, Lecointre G: **Integrating the universal metabolism into a phylogenetic analysis.** *Mol Biol Evol* 2005, **22:**1-11.
27. Xie G, Keyhani NO, Bonner CA, Jensen RA: **Ancient origin of the tryptophan operon and the dynamics of evolutionary change.** *Microbiol Mol Biol Rev* 2003, **67:**303-342.
28. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization.** *Science* 2007, **317:**86-94.
29. Velasco AM, Leguina JI, Lazcano A: **Molecular evolution of the lysine biosynthetic pathways.** *J Mol Evol* 2002, **55:**445-459.
30. Fani R, Lio P, Lazcano A: **Molecular evolution of the histidine biosynthetic pathway.** *J Mol Evol* 1995, **41:**760-774.
31. Rutter MT, Zufall RA: **Pathway length and evolutionary constraint in amino acid biosynthesis.** *J Mol Evol* 2004, **58:**218-224.
32. Porat I, Sieprawska-Lupa M, Teng Q, Bohanon FJ, White RH, Whitman WB: **Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p-aminobenzoic acid in the archaeon *Methanococcus maripaludis*.** *Mol Microbiol* 2006, **62:**1117-1131.
33. Garcia-Vallve S, Guzman E, Montero MA, Romeu A: **HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes.** *Nucleic Acids Res* 2003, **31:**187-189.
34. Nester EW, Montoya AL: **An enzyme common to histidine and aromatic amino acid biosynthesis in *Bacillus subtilis*.** *J Bacteriol* 1976, **126:**699-705.
35. Weigent DA, Nester EW: **Purification and properties of two aromatic aminotransferases in *Bacillus subtilis*.** *J Biol Chem* 1976, **251:**6974-6980.
36. Atchley WR, Zhao J, Fernandes AD, Druke T: **Solving the protein sequence metric problem.** *Proc Natl Acad Sci USA* 2005, **102:**6395-6400.
37. Segre D, Deluna A, Church GM, Kishony R: **Modular epistasis in yeast metabolism.** *Nat Genet* 2005, **37:**77-83.
38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
39. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
40. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov**

models that represent all proteins of known structure. *J Mol Biol* 2001, **313:**903-919.
41. Eddy SR: **Hidden Markov models.** *Curr Opin Struct Biol* 1996, **6:**361-365.
42. Li W, Jaroszewski L, Godzik A: **Tolerating some redundancy significantly speeds up clustering of large protein databases.** *Bioinformatics* 2002, **18:**77-82.
43. Lozada-Chavez I, Janga SC, Collado-Vides J: **Bacterial regulatory networks are extremely flexible in evolution.** *Nucleic Acids Res* 2006, **34:**3434-3445.