UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Estudio comparativo del metabolismo de las *Gammaproteobacterias* mediante alineamientos de pasos enzimáticos.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Doctor en Ciencias

PRESENTA:
M. en C. Augusto César Poot Hernández

TUTOR PRINCIPAL
Dr. Ernesto Pérez Rueda
Instituto de Biotecnología, UNAM

MIEMBROS DEL COMITÉ TUTOR
Dra. Rosa María Gutiérrez Ríos
Instituto de Biotecnología, UNAM
Dr. Arturo Carlos II Becerra Bracho
Fac. de Ciencias, UNAM

Ciudad de México. Marzo, 2016

A Ruth, sin su amor y apoyo este trabajo no hubiera sido posible.
A Ollin, quien no ha dejado nunca de mostrarme lo brillante que es el mundo.
A Meztli, cuya promesa me recuerda que la flama de la vida nunca se apaga.
ACPH

Agradecimientos

Este trabajo se realizó en el Departamento de ingeniería celular y biocatálisis del Instituto de Biotecnología de la UNAM bajo la dirección del Dr. Ernesto Pérez Rueda. Adicionalmente, el trabajo fue posible gracias al apoyo de la Dra. Katya Rodríquez Vázquez en el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM.

Agradezco a:

- ... la UNAM, el IBT y el programa de Maestría y Doctorado en Ciencias Bioquímicas por la oportunidad de crecer profesionalmente.
- ... al Dr. Ernesto Pérez Rueda por todas sus enseñanzas y por el completo apoyo que me brindó.

- ... la Dra. Katya Rodríguez Vázquez y el IIMAS de la UNAM por permitirme desarrollar mis habilidades computacionales y por toda su ayuda incondicional.
- ... a los miembros de mi comité tutor, Dra. Rosa María Gutiérrez Ríos y Dr. Arturo Carlos II Becerra Bracho, por todos sus comentarios para mejorar y enriquecer este trabajo.
- ... a los miembros de mi jurado de tesis, Dr. Guillermo Gosset Lagarda, Dr. Daniel Genaro Segura Gonzáles, Dra. Katya Rodríguez Vázquez, Dr. Armando Hernández Mendoza y Dr. León Martínez Castilla, por sus valiosas sugerencias para la escritura de este documento.
- ... al Programa de Apoyo a Estudios de Posgrado (PAEP) de la UNAM por el apoyo para la asistencia a congresos especializados.
- ... a mis compañeros de laboratorio, Nancy Rivera, Dagoberto Armenta, Mario Martínez, Jessica Brambila, Edgar Arenas, Paty Ortegón y Germán Pérez, por siempre estar en disposición de ayudar y por todas las pláticas constructivas.
- ... a Rodrigo González del Cueto por todos los *tips* que me brindó para hacer un uso más eficiente de la computadora y por hacerme aún más *geek*.
- ... a mis padres que nunca dudaron de mí.
- ... a la familia Rincón Heredia que me apoyó en los momentos más difíciles.
- ... a toda mi familia y amigos que siempre están presentes en las buenas y las malas.
- ... al proyecto PAPIIT número IN109011 por el apoyo económico brindado.
- ... al Conacyt, No. beca 209805.

Índice general

Índice de figuras

Estudio comparativo del metabolismo de las *Gammaproteobacterias* mediante alineamientos de pasos enzimáticos.

por

M. en C. Augusto César Poot Hernández

Resumen

El metabolismo es considerado uno de los sistemas biológicos mejor estudiados. En general se considera que está ampliamente conservado, sin embargo existe evidencia creciente de que en realidad es muy diverso. Por este motivo, es necesario generar estrategias que nos permitan estudiar esta diversidad para entender mejor su estructura y evolución.

En este trabajo usamos una estrategia general basada en el alineamiento de cadenas lineales de pasos enzimáticos para estudiar comparativa y sistemáticamente el metabolismo de diversas especies de *Gammaproteobacterias*. Para esto, los mapas metabólicos contenidos en la base de datos KEGG fueron convertidos a cadenas lineales de pasos enzimáticos. Posteriormente estas cadenas fueron alineadas usando un algoritmo de alineamiento basado en programación dinámica de forma similar a los alineamientos clásicos de secuencias de nucleótidos o aminoácidos.

De este modo, generamos una base de datos con más de 7900 cadenas de pasos enzimáticos no redundantes de 40 genomas de *Gammaproteobacterias*. Estas cadenas se alinearon todas contra todas para identificar sus similitudes generales. Así se observó una mayor conservación de los mapas metabólicos relacionados con procesos de síntesis en comparación con procesos de degradación. También se observó que los mapas metabólicos que catalizan reacciones que involucran compuestos similares tienden a tener cadenas de pasos enzimáticos similares. Finalmente se compararon dos rutas metabólicas ancestrales, la parte baja de la glucólisis y la ruta del inosin monofosfato con lo que se observaron diferencias en sus patrones de similitud, sugiriendo que la glucólisis pudo haber sido una ruta donadora de pasos enzimáticos durante la evolución del metabolismo. Esto muestra que nuestra estrategia puede ser útil para identificar pasos enzimáticos similares en diferentes mapas metabólicos y puede ayudar a reforzar un modelo de evolución en mosaico en el metabolismo de las *Gammaproteobacterias*.

Estudio comparativo del metabolismo de las *Gammaproteobacterias* mediante alineamientos de pasos enzimáticos.

by

M. en C. Augusto César Poot Hernández

Abstract

Metabolism is considered one of the best-studied biological systems and in general it is considered as widely conserved, however there is growing evidence that it is actually very diverse. For this reason, it is necessary to generate strategies that allow us to study this diversity to better understand its structure and evolution.

Here, we use a general strategy based on the alignment of linear enzymatic step sequences to systematically compare the metabolism of various species of *Gammaproteobacteria*. For this purpose, the metabolic maps in KEGG database were converted into enzymatic step sequences. Subsequently these sequences were aligned using an alignment algorithm based on dynamic programming, similar to the classical alignments of nucleotides or amino acids.

In this way, we generated a database of more than 7900 non-redundant sequences from 40 *Gammaproteobacteria* genomes. These sequences were aligned all against all in order to identify their overall similarities. In general, we observed a greater conservation of maps related to biosynthesis processes compared to degradation processes. Also, maps associated with the metabolism of similar compounds contain a high proportion of similar enzymatic step sequences. Finally two ancestral metabolic pathways, the lower part of glycolysis and the of inosine monophosphate pathway were compared, showing different patterns of conservation and suggesting that glycolysis may have been a donor of enzymatic steps during metabolism evolution. These results show that our strategy can be useful in identifying similar enzymatic steps in different metabolic maps and can help to reinforce a model of mosaic evolution in the metabolism of *Gammaproteobacteria*.

Capítulo 1

Introducción

El metabolismo se puede definir como el conjunto de todas las reacciones bioquímicas que ocurren en un ser vivo y que le permiten intercambiar energía y materia con el ambiente para llevar a cabo sus procesos vitales [1]. La mayoría de estas reacciones son catalizadas por enzimas, las cuales permiten la transformación de una molécula (sustrato) a otra (producto). Los productos generados por una enzima pueden ser usados cómo sustrato por otra. De este modo, las reacciones se agrupan consecutivamente formando rutas metabólicas y en última instancia, formando una red compleja de interacción. Clásicamente, el metabolismo se ha dividido en rutas individuales para su estudio. Estas rutas o vías metabólicas son secuencias de reacciones enzimáticas con una relación funcional específica. Así, el metabolismo se puede dividir en rutas de síntesis de aminoácidos o de lípidos, o en rutas de degradación de carbohidratos como la glucosa.

Recientemente y gracias a la gran cantidad de proyectos genómicos y al desarrollo de la bioinformática, ha sido posible compilar toda la información conocida de las rutas metabólicas en modelos computacionales integrativos que han permitido caracterizar el metabolismo como una red [2–6] y evidenciar diversas propiedades emergentes que moldean su estructura y evolución [7–9].

En general se considera que el metabolismo es uno de los sistemas biológicos más antiguos y mejor conservados [1]. Sin embargo, recientemente se ha puesto de manifiesto que existe una amplia diversidad no prevista en rutas consideradas centrales como la glucólisis y el ciclo del citrato [10, 11]. Esta diversidad es responsable, en parte, de la existencia de microorganismos que son capaces de vivir en una gran variedad de ambientes que pueden ser inhóspitos para otros seres vivos.

La diversidad que puede existir en el metabolismo está bien ilustrada en el grupo de las *Gammaproteobacterias*, el cual es uno de los clados de bacterias mejor estudiados y con más géneros descritos a la fecha [12]. Este grupo incluye organismos con importancia científica, médica y tecnológica, como *Escherichia coli* uno de los organismos modelo por excelencia. Además, incluye organismos con diferentes estilos de vida y por lo tanto, con diferentes capacidades metabólicas. Por ejemplo: bacterias comensales de mamíferos (ej. *E. coli*) o de moluscos (*Ruthia magnifica* [13]), endosimbiontes obligados de insectos (ej. géneros *Baumannia spp* y *Buchnera spp*), fotoautótrofos (ej. bacterias púrpuras del azufre), patógenos de mamíferos (ej. algunas enterobacteriales y bacterias de los géneros *Yersinia spp* y *Vibrio spp*), generalistas que pueden ser encontrados en una gran variedad de ambientes (género *Pseudomonas*) y organismos autótrofos capaces de oxidar el ion arsenito (*Alkalilimnicola sp* [14]) o de degradar alcanos (*Alcanivorax sp* [15]), entre otros.

1.1. Planteamiento del problema y justificación.

En la actualidad, hay disponible una gran cantidad de información metabólica referente a diversos organismos almacenada en bases de datos especializadas cómo *Kyoto Enzyclopedia of Genes and Genomes* (KEGG) [16] o MetaCyc [17], sin embargo existen relativamente pocos estudios que evalúen a nivel global las similitudes y diferencias existentes entre el metabolismo de distintos organismos.

Se han publicado varios estudios que usan aproximaciones comparativas para analizar el metabolismo de distintos organismos usando metodologías puramente genómicas o ayudadas con estrategias filogenéticas y/o modelos de balances de flujos, los cuales han permitido identificar patrones de conservación de enzimas y reconstrucción de modelos ancestrales [18–21]. Por otro lado, también hay trabajos que estudian el metabolismo desde la perspectiva de las funciones enzimáticas, analizando principalmente el conjunto de números enzimáticos o *Enzyme commission numbers* (*EC numbers*)[1]. Estos trabajos han permitido identificar, entre otras cosas, la alta redundancia de funciones enzimáticas en organismos con genomas más complejos [22] o la ausencia de pasos enzimáticos comunes en organismos con genomas reducidos [23].

Sin embargo, aunque estas estrategias pueden estar auxiliadas por modelos metabólicos [20], no comparan directamente las rutas o redes bioquímicas. El alineamiento de rutas o redes metabólicas es una herramienta emergente que permite hacer este tipo comparaciones de forma mas directa. Esta estrategia consiste en el acomodo de dos o más rutas o redes metabólicas de tal modo que se alcance el máximo número de coincidencias entre ellas, de forma análoga a los métodos de alineamiento de secuencias de aminoácidos o nucleótidos.

Por ejemplo, Dandekar y colaboradores [10] publicaron el primer reporte donde se usa el enfoque de alineamientos de rutas metabólicas. En este trabajo, compararon la ruta de la glucólisis de diversos organismos de los tres dominios celulares y descubrieron que la glucólisis en realidad es una ruta diversa que no necesariamente está conservada en su totalidad en todos los organismos. Por otro lado, Tohsato y colaboradores [24] y Chen y Hofestädt [25–27] reportaron de manera independiente, métodos de alineamiento de cadenas lineales de pasos enzimáticos usando como modelo de evaluación la similitud

[1]Los *EC numbers* son un sistema de clasificación jerárquica de las enzimas que describe el tipo de reacción química que catalizan. Este sistema de clasificación es publicado por la comisión de nomenclatura de la Unión Internacional de Bioquímica y Biología Molecular.

entre los *EC numbers*. Complementariamente, Tohsato y Nishimura [28] reportaron un algoritmo de alineamiento que considera la similitud en la estructura química de los sustratos y los productos que intervienen en las reacciones bioquímicas. Además de estos trabajos se han reportado un gran número de métodos para el alineamiento topológico de redes metabólicas y de otras redes biológicas que, aunque en teoría pueden dar una mayor cantidad de información, han mostrado ser aún difíciles de implementar, comparar e interpretar [29].

La mayoría de estos métodos son desarrollados para ser distribuidos como herramientas computacionales tales cómo *MetaPathwayHunter* [30], *PathAligner* [27], *MetNetAligner* [31], entre otros. Sin embargo, uno de los problemas que emergen de esta diversidad es que existe poca consistencia entre los distintos métodos, además, en general, han sido poco adoptados y no se han utilizados para estudiar el metabolismo en un marco conceptual metabólico-evolutivo.

La comparación sistemática del metabolismo intra e inter especies nos permitiría identificar similitudes y diferencias entre distintos organismos, con las cuales se podrían identificar enzimas o series de enzimas propias de un conjunto específico de organismos que pudieran ser importantes como blancos farmacológicos o para la degradación de xenobióticos. Además, esta información podría ayudar a entender cómo ha evolucionado el metabolismo desde una perspectiva global.

En este trabajo se realizó un análisis comparativo del metabolismo de varias bacterias de la división *Gammaproteobacteria* mediante alineamientos de cadenas lineales de pasos enzimáticos o *Enzymatic Step Sequence* (ESS), generadas a partir de los mapas metabólicos almacenados en la base de datos KEGG. Con esta estrategia se identificaron regiones conservadas y variables del metabolismo de las *Gammaproteobacterias* y se sugieren las relaciones funcionales que pueden existir entre los distintos tipos de metabolismos.

1.2. Estrategia general para el estudio comparativo del metabolismo mediante alineamientos de pasos enzimáticos.

Recientemente en nuestro grupo de investigación se desarrolló un método para el alineamiento pareado y múltiple de cadenas lineales de pasos enzimáticos (ESS) basado en un algoritmo genético [32]. Este método compara los pasos enzimáticos representados por sus números enzimáticos (*EC numbers*) usando un criterio de evaluación basado en la entropía de cada columna del alineamiento. El resultado se puede visualizar y analizar de forma similar a los alineamientos de secuencias de aminoácidos o nucleótidos, de modo que se pueden identificar eventos evolutivos discretos como la inserción y/o eliminación de pasos enzimáticos (individuales o en bloques) y la sustitución de enzimas (*mismatches*) (figura 1-1).

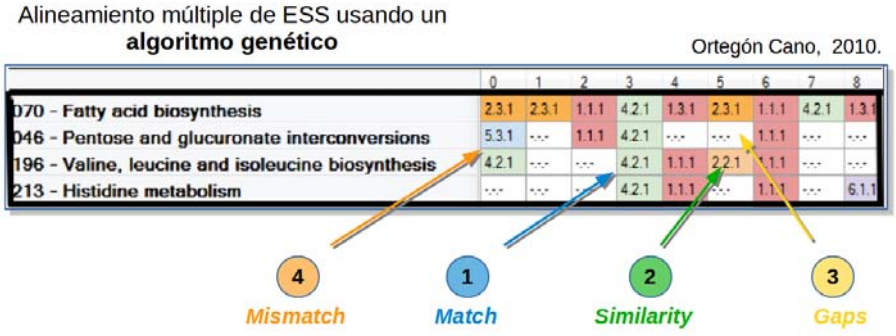

Figura 1-1. *Ejemplo de alineamiento múltiple de ESSs [32].*

El algoritmo de alineamiento se implementó como parte de una estrategia general para el estudio comparativo y sistemático del metabolismo. La estrategia implica 3 pasos principales (figura 1-2). El primer paso es la generación de las ESS a partir de los mapas metabólicos de KEGG usando el algoritmo *Breadth-First Search* (BFS). Posteriormente, las ESS se alinean por pares todas contra todas. Finalmente, se usan métodos de agrupa-

miento para identificar las ESS similares entre si, con las cuales se pueden hacer diversos análisis, por ejemplo, alineamientos múltiples. De este modo, se pueden identificar módulos de pasos enzimáticos conservados entre distintas especies o entre distintas partes del metabolismo, así como rutas metabólicas alternativas o propias de una especie.

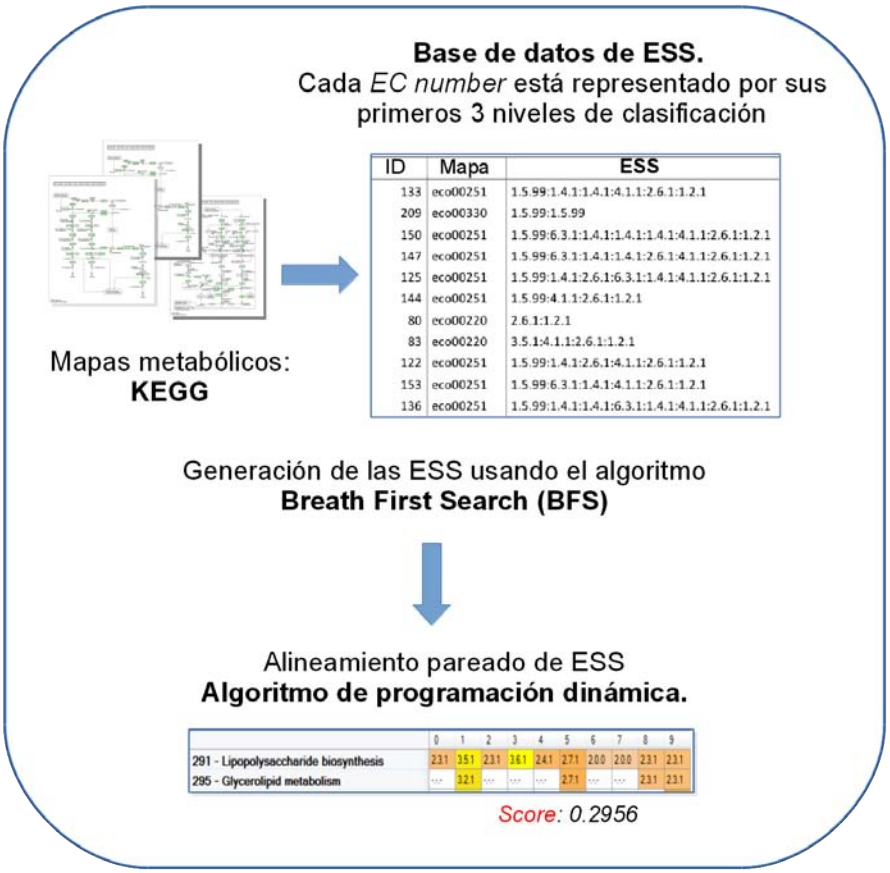

Figura 1-2. *Estrategia general para el estudio comparativo del metabolismo.*

Esta estrategia se usó para analizar el metabolismo de la bacteria *E. coli*, con lo cual se identificaron varios grupos de cadenas de pasos enzimáticos similares. En la figura 1-3 se muestra un resumen del alineamiento de uno de estos grupos (*cluster* 21). El grupo incluyó secuencias relacionadas con el metabolismo de distintos carbohidratos y mostró que la metodología puede identificar pasos enzimáticos similares catalizados por enzimas homólogas, en rutas metabólicas diferentes.

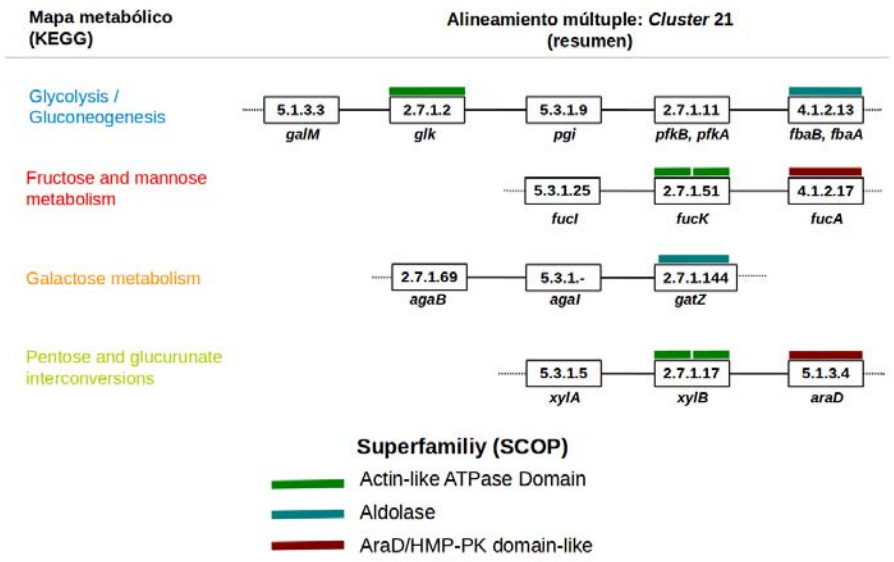

Figura 1-3. *Resumen del alineaminento múltiple de ESSs de distintas rutas de degradación de carbohidratos de E. coli. Los recuadros de colores muestran los dominios proteicos que se repiten en dos o más proteínas dentro del alineamiento. Los dominios proteicos fueron asignados usando la base de datos Superfamily [33].*

Las principales características de este método son:

- Las ESS son derivadas de la información topológica de los mapas metabólicos, por lo que representan en cierta medida la estructura de la red.

- El alineamiento usando cadenas de pasos enzimáticos codificados por sus *EC numbers* permite hacer un análisis del metabolismo con un enfoque funcional sin usar directamente las secuencias de las proteínas involucradas.

- El método de alineamiento utiliza los primeros tres niveles de clasificación de los números enzimáticos, por lo que permite conocer las similitudes catalíticas generales, sin tomar en cuenta los posibles cambios de reconocimiento de sustrato principal. Esto facilita la comparación entre distintas rutas metabólicas y es congruente con observaciones previas cómo la tendencia a la conservación de función general de las enzimas [34, 35] y la falta de pasos enzimáticos comunes en organismos con

genomas reducidos [23]. Adicionalmente, cada paso enzimático de 3 niveles está referenciado a su correspondiente versión de 4 niveles y a la lista de genes asignados para cada organismo, lo que facilita análisis posteriores más detallados.

- El alineamiento también permite la inclusión de *gaps*, lo cual nos permitiría identificar pasos enzimáticos insertados o eliminados de una ruta metabólica determinada.

1.3. Objetivos.

Con base en lo expuesto anteriormente, el objetivo general de este trabajo fue identificar similitudes y diferencias en el metabolismo de diversas *Gammaproteobacterias* mediante alineamientos de pasos enzimáticos.

Para cumplir este objetivo, se plantearon los siguientes objetivos particulares:

1. Generar una base de datos de ESS de distintas especies de *Gammaproteobacterias* a partir de la información contenida en KEGG.

2. Alinear por pares todas las cadenas de pasos enzimáticos.

3. Usar los alineamientos pareados para identificar similitudes entre especies y entre mapas metabólicos.

Capítulo 2

Materiales y métodos.

Para cumplir nuestro objetivo general, se siguió la estrategia ilustrada en la figura 1-2. En breve, se generó una base de datos de ESS a partir de los mapas metabólicos obtenidos de la base de datos KEGG. Posteriormente, las secuencias generadas se alinearon todas contra todas usando un algoritmo basado en programación dinámica, lo cual nos permite medir la similitud entre ESS. Finalmente, esta similitud se usó en conjunto con estrategias de agrupamiento para estudiar la similitud metabólica entre especies y la similitud entre mapas metabólicos. A continuación de describe a profundidad esta estrategia.

2.1. Selección de organismos de estudio.

Las *Gammaproteobacterias* son uno de los grupos de organismos mejor estudiados con una amplia y reconocida diversidad metabólica.

Se seleccionó el metabolismo de 40 especies de *Gammaproteobacterias* de las 275 presentes en la base de datos KEGG hasta Junio de 2011. Estas cepas fueron seleccionadas usando la clasificación de genomas no redundantes reportada en [36] con un umbral de

similitud de 0.7. Dicho umbral está basado en los valores de *bit scores* [1] obtenidos con el alineamiento de las proteínas ortólogas de dos organismos usando el algoritmo *BLAST*. De modo que este valor de umbral es una medida de la similitud entre dos genomas y permite filtrar todos aquellos genomas con una similitud mayor a un valor dado. En este caso, el valor de 0.7 significa que se están agrupando aquellos genomas con una similitud (en *bit score*) mayor al 70 % entre sus proteínas homólogas. Dichos grupos están representados por aquellos genomas de cada grupo que contienen la mayor cantidad de familias proteicas compartidas con otros grupos [36].

De este modo, se seleccionaron a los 40 genomas (o cepas) previamente mencionados debido a que, cualitativamente, representan al clado y a que representan un conjunto de genomas no redundantes, por lo que se puede mitigar el sesgo que pudiera provocar la inclusión de genomas muy parecidos. En la figura 2-1 se muestra la posición taxonómica de las cepas seleccionadas en la filogenia de las *Gammaproteobacerias* reportada en la referencia [12].

En la tabla 2-1 se muestra la lista de genomas incluidos en este trabajo. El identificador de KEGG se usará en el resto de este escrito para identificar cada genoma. También se indica el tamaño del genoma en número de marcos abiertos de lectura (*ORF*s por sus siglas en inglés), el número de *EC numbers* y el número de proteínas asignadas a algún número enzimático. Este conjunto incluye cepas con un tamaño de genoma que va de 182 a 6778 *Open Reading Frames* (*ORF*s).

ID*	Cepa	ORFs**	*EC numbers*+	Enzimas++
aci	*Acinetobacter* ADP1	3308	648	1008
aha	*Aeromonas hydrophila* ATCC7966	4121	751	1218

[1] El *bit score* es un valor generado por el algoritmo *Basic Local Alignment Tool* (*BLAST*) que determina una medida de qué tan bueno es un alineamiento. Este valor depende de la longitud del alineamiento, la proporción de posiciones correctamente alineadas (similitudes e identidades) y la cantidad de *gaps* que contiene.

Continuación ...

ID	Cepa	ORFs	*EC numbers*	Enzimas
aeh	*Alakalilimnicola ehrlichei* MLHE-1	2865	584	850
abo	*Alcanivorax borkumensis* SK2	2755	606	931
bci	*Baummania cicadellinicola* Hc	595	298	346
buc	*Buchnera aphidicola*	574	278	331
bcc	*Buchnera aphidicola* Cc	362	161	195
bfl	*Candidatus Blochmania floridanus*	583	288	351
bpn	*Candidatus Blochmania pennsylvanicus* BPEN	610	301	365
rma	*Candidatus Ruthia magnifica* Cm	976	396	488
crp	*Carsonella ruddii*	182	69	78
csa	*Chromohalobacter salexigens* DSM3043	3298	679	971
cps	*Colwellia psychrerythraea* 34H	4910	684	1075
cbu	*Coxiella burnetii*	1847	414	530
eco	*Escherichia coli* K12	4150	859	1369
ftu	*Francisella tularensis holarctica*	1604	431	581
hdu	*Haemophilus ducreyi* 35000HP	1717	409	546
hch	*Hahella chejuensis* KCTC2396	6778	734	1172
hha	*Halorhodospira halophila* SL1	2407	567	753
ilo	*Idiomarina loihiensis* L2TR	2628	557	752
lpn	*Legionella pneumophila* Paris	2943	550	788
msu	*Mannheimia succiniciproducens* MBEL55E	2369	563	780
maq	*Marinobacter aquaeolei* VT8	4272	640	988
mca	*Methylococcus capsulatus* Bath	2956	581	844
noc	*Nitrosococcus oceani* ATCC19707	3018	599	851
ppr	*Photobacterium profundum* SS9	5489	758	1254
plu	*Photorhabdus luminescens*	4683	732	1086

Continuación ...

ID	Cepa	ORFs	*EC numbers*	Enzimas
pat	*Pseudoalteromonas atlantica* T6c	4281	719	1063
pha	*Pseudoalteromonas haloplanktis* TAC125	3489	672	1002
pfl	*Pseudomonas fluorescens* Pf-5	6139	822	1432
pcr	*Psychrobacter cryohalolentis* K5	2511	600	820
pin	*Psychromonas ingrahamii* 37	3548	677	997
sde	*Saccharophagus degradans* 2-40	4007	604	876
saz	*Shewanella amazonensis* SB2B	3645	690	1022
shm	*Shewanella* ANA-3	4014	678	1025
sdn	*Shewanella denitrificans* OS217	3754	644	937
tcx	*Thiomicrospira crunogena* XCL-2	2196	500	695
vfi	*Vibrio fischeri* ES114	3817	694	1120
vvu	*Vibrio vulnificus* YJ016	4433	717	1096
xca	Xanthomonas campestris vesicatoria 85-10	4469	718	1078

Tabla 2-1: *Lista de Gammaproteobacterias incluidas en este trabajo y algunos datos genómicos.*
* *Clave de identificación de cada genoma en la base de datos KEGG.*
** *Número de ORFs (siglas en inglés de Open Reading Frame) identificados en el genoma según KEGG.*
$^{+}$ *Número de EC numbers diferentes asignados al genoma.*
$^{++}$ *Número de proteínas que tiene asignado al menos un número enzimático en la base de datos KEGG.*

2.2. Construcción de la base de datos de *ESSs*.

El metabolismo de cada organismo en forma de red se transformó a cadenas lineales de paso enzimáticos con el fin de poder comparar estas cadenas usando un algoritmo de alineamiento. En este contexto, se definió una cadena de pasos enzimáticos o ESS cómo una colección lineal de números enzimáticos consecutivos. Estas cadenas fueron extraídas

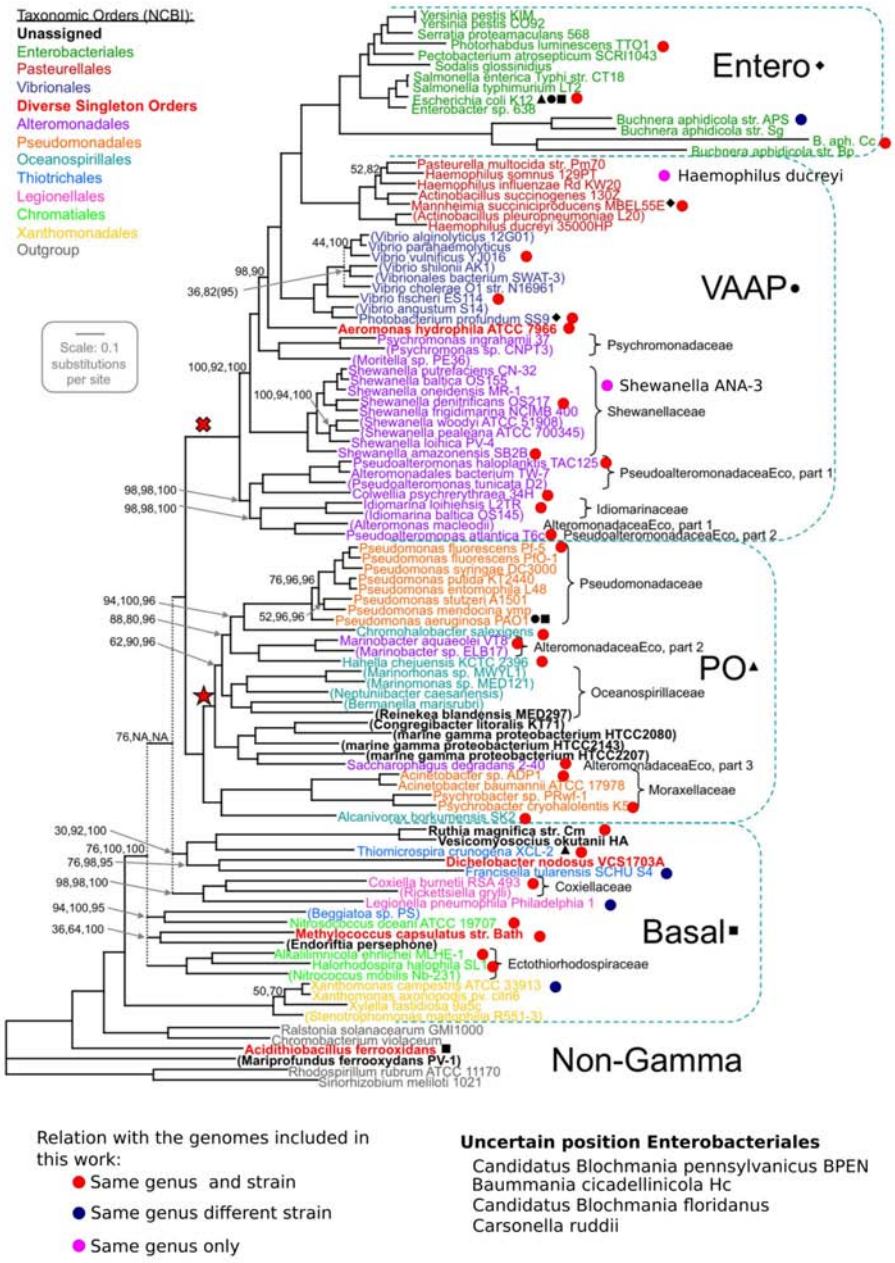

Figura 2-1. *Distribución de los genomas usados en este trabajo en la filogenia de las Gammaproteobacterias. La filogenia está generada mediante la concatenación de 356 alineamientos de proteínas homólogas. Los círculos representan los genomas incluidos en este trabajo. Los círculos rojos son aquellos cuya cepa corresponde exactamente con la cepa usada para construir la filogenia. Los círculos azules corresponden a genomas del mismo género y especie pero diferente cepa a la usada en la filogenia. Finalmente, los círculos rosas corresponden a genomas del mismo género pero diferente especie a los usados en la filogenia. Tomada y modificada de [12]*

de los mapas metabólicos de la base de datos KEGG.

Para cada organismo se descargaron los archivos *kgml* (versión 0.71) de la base de datos KEGG (Junio de 2011). Estos archivos describen los mapas metabólicos tal y como se pueden consultar en la versión web de la base de datos. Usando esta información se creó una representación en forma de grafo dirigido para cada mapa metabólico, donde los nodos representan las enzimas (proteínas), los vértices representan las relaciones sustrato/producto entre ellas y la dirección de los vértices representa la reversibilidad de las reacciones según KEGG. En la figura 2-2 se muestrean dos ejemplos del modelado de dos mapas metabólicos. El modelo del mapa metabólico de la figura 2-2b tiene forma de flor debido a que uno de los nodos representa un complejo protéico que lleva a cabo varias reacciones dentro del mismo mapa metabólico. Complementariamente, las enzimas aisladas no son usadas para la generación de las ESS.

Estos grafos fueron usados para construir un conjunto de árboles usando el algoritmo de búsqueda a lo ancho, *Breadth First Search* (*BFS*), a partir de nodos específicos llamados nodos de inicio [2]. Los nodos de inicio se definieron de acuerdo a dos criterios: primero, todos aquellos nodos cuyo sustrato no es producido por ninguna reacción dentro del mapa metabólico; segundo, todos aquellos nodos que sirven como conexión para otro mapa metabólico y que tienen una conectividad (número de vecinos) menor a 3. [3]. Estos criterios buscan representar las entradas bioquímicas de cada mapa metabólico. Posteriormente, los árboles de *BFS* se usaron como guía para construir las ESS, para lo cual se identificaron todos los nodos terminales de un árbol y se trazaron sus correspondientes caminos hasta el nodo raíz, anotando en el trayecto la información de cada paso enzimático visitado. El modelaje de la red de cada mapa metabólico y el análisis de *BFS*

[2]En el contexto del grafo, cualquier nodo puede ser usado cómo nodo de inicio para BFS, de modo que en el caso descrito aquí, se busca que dichos nodos tengan algún tipo de relevancia biológica.

[3]Se usó una conectividad menor a 3 para limitar el número de *ESSs* generadas considerando aquellas reacciones que a los sumo tengan dos vecinos. Es decir, cuyo sustrato sea producido por una sola enzima y cuyo producto sea usado por una sola enzima (conectividad = 2).

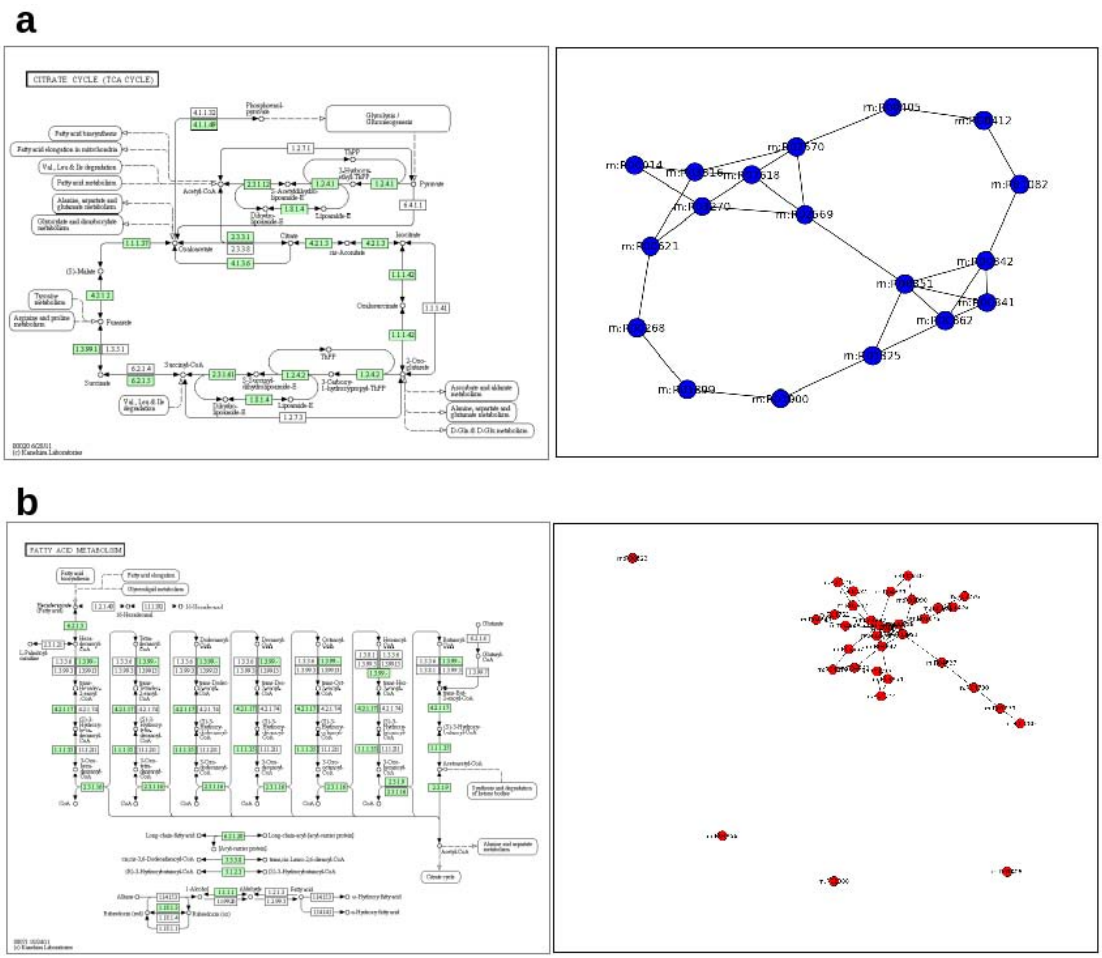

Figura 2-2. *Ejemplos de representación en forma de grafo de dos mapas metabólicos de BFS. (a) Mapa del ciclo del ácido cítrico de E. coli K12 (eco00020). (b) Mapa del metabolismo de ácidos grasos de E. coli K12 (eco00071). Del lado izquierdo se muestran los mapas metabólicos tal y como se obtienen en KEGG. Las enzimas indicadas en verde son aquellas que se encuentran en E. coli K12. Del lado derecho se muestran los grafos construidos a partir de cada mapa metabólico. Las etiquetas son los identificadores de cada reacción.*

se realizaron usando la librería de Python *Networkx* [37] [4]. En la figura 2-3 se muestra un ejemplo de la generación de un árbol de *BFS* a partir de un grafo esquemático y se indican las secuencias de nodos obtenidas. Se observa que el algoritmo *BFS* se encarga naturalmente de los ciclos existentes en un grafo ya que una vez que un nodo es visitado, no vuelve a ser visitado por el algoritmo.

[4] https://networkx.github.io/

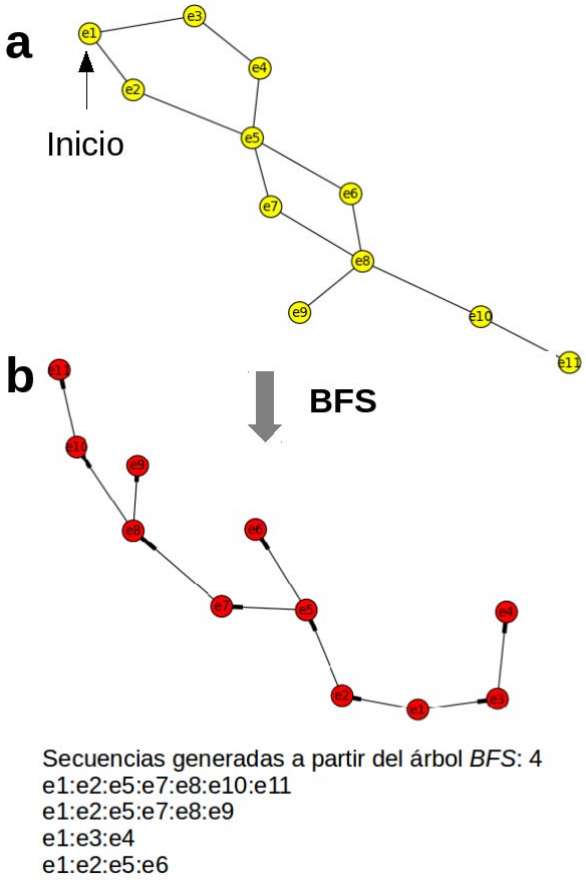

Figura 2-3. *Ejemplo de construcción de cadenas lineales de pasos enzimáticos. En (a) se muestra un grafo esquemático con estructura similar a un mapa metabólico. En (b) se muestra el árbol de BFS generado a partir del grafo en (a) usando el nodo indicado con una flecha como nodo de inicio. También se indican las cadenas de pasos enzimáticos derivadas del árbol en (b).*

Las ESS fueron almacenadas en una base de datos estructurada usando el programa *SQLite 3* [5]. En esta base de datos están representadas las ESS en las siguientes versiones:

1. Secuencia de genes. Cada elemento de la cadena es un gen o grupos de genes que realizan una reacción.

2. Secuencia de *EC numbers*, 4 niveles. Cada elemento de la cadena es un número enzimático usando los 4 niveles de clasificación.

3. Secuencia de *EC numbers*, 3 niveles. Cada elemento de la cadena es un número enzimático usando los primeros 3 niveles de clasificación.

[5] www.sqlite.org

Los primeros tres niveles son suficientes para representar la similitud funcional y general entre dos reacciones bioquímicas [7,8,22,23], de modo que se usó la representación de los primeros 3 niveles de clasificación para comparar las cadenas de paso enzimáticos. Por esta razón el término ESS se refiere a la cadena donde los pasos enzimáticos están representados por los primeros 3 niveles de clasificación de los *EC numbers*.

La base de datos de ESS tiene una redundancia natural, debido a que una ESS puede estar representada de forma idéntica en distintas especies o mapas metabólicos. De modo que se creó una base de datos secundaria no redundante llamada nrESS. En esta base datos solo se representan las ESS únicas. Todos los análisis posteriores se realizaron sobre la base de datos nrESS refiriéndose a los datos originales cuando fue necesario.

2.3. Algoritmo de programación dinámica para el alieamiento de *ESSs*.

Con el fin de identificar la similitud global entre dos ESSs, se implementó un algoritmo de alineamiento de programación dinámica basado en el algoritmo Needelman-Wunsh (NW), tal y cómo está propuesto en la referencia [41]. Este algoritmo es notable por ser uno de los primeros métodos desarrollados para identificar similitudes entre dos secuencias de aminoácidos [38]. En este trabajo se modificó el algoritmo general para identificar las similitudes entre dos ESS. En el anexo A se muestran algunos ejemplos de alineamientos pareados realizados con el algoritmo descrito aquí.

2.3.1. Similitud entre dos *EC numbers*.

Para implementar el algoritmo de alineamiento, primero fue necesario definir la similitud entre dos *EC numbers*. La similitud entre los primeros tres niveles de dos *EC numbers*

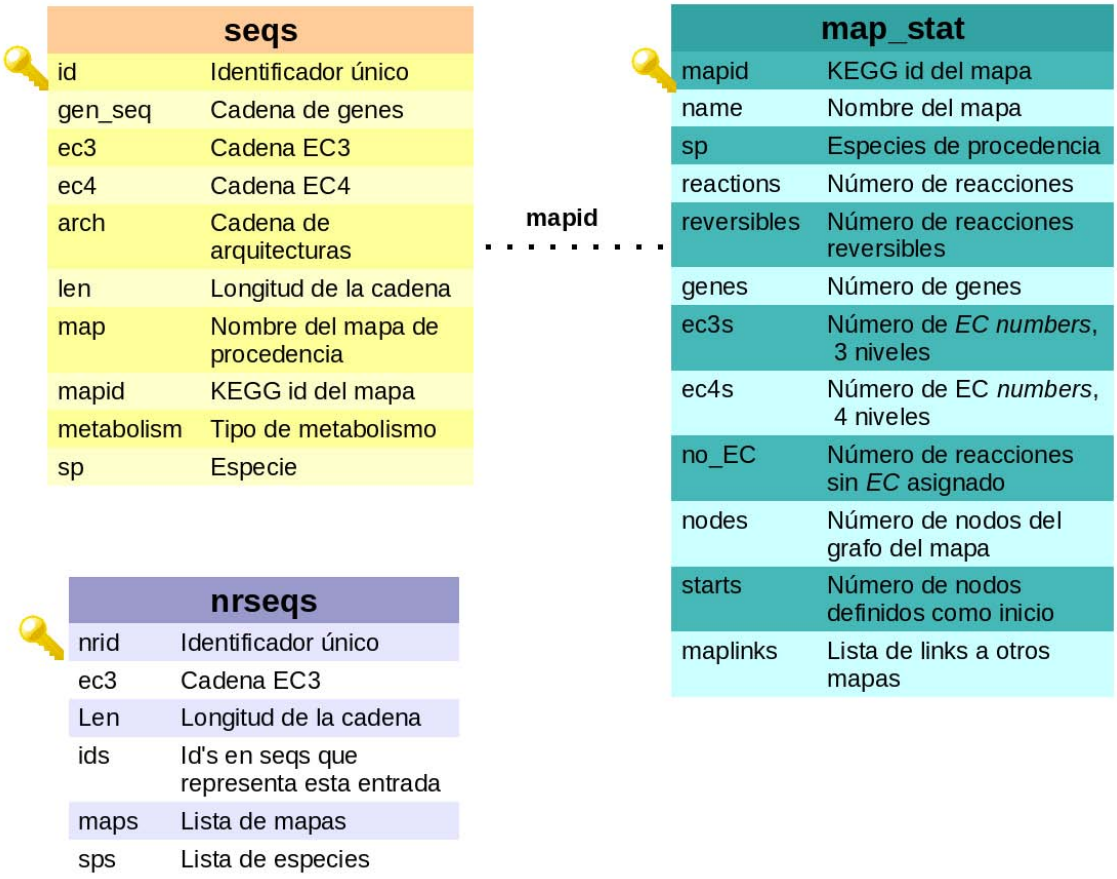

Figura 2-4. *Estructura de la base de datos de ESSs. La base de datos está compuesta por tres tablas. La tabla seqs contiene la información de las ESS originales tal y cómo fueron generadas en el procedimiento descrito en el texto. La tabla nrseqs contiene las nrESS, es decir, las cadenas de pasos enzimáticos no redundantes y su relación con la tabla seqs. La tabla map_stat contiene las estadísticas de la generación de las ESS por mapa metabólico. Contiene información cómo número de genes, reacciones, nodos de inicio, etc.*

$(l1.l2.l3)$, $ECS(EC_1, EC_2)$ fue evaluada de acuerdo a la siguiente ecuación normalizada derivada de trabajos previos de nuestro grupo [32, 39].

$$ECS(EC_1, EC_2) = \frac{w_1 H_1 + w_2 H_2 + w_3 H_3}{w_1 w_2 w_3} \tag{2-1}$$

Donde w_1, w_2, w_3 corresponden a los factores ponderados para cada nivel de jerarquía de los *EC numbers* y H_1, H_2, H_3 corresponde a la entropía medida para cada nivel. Los factores ponderados w_1, w_2, w_3 fueron usados para representar la jerarquía inherente a

la clasificación enzimática. El primer nivel de clasificación de los *EC numbers* describe el tipo general de reacción; el siguiente nivel define un subconjunto del anterior y así sucesivamente. De modo que los niveles superiores tienen un mayor peso para identificar la similitud general de las reacciones catalizadas por dos enzimas. De este modo se asignaron los valores de 15, 10, 1 para los factores w_1, w_2, w_3. Estos valores se seleccionaron empíricamente en función de la calidad de los alineamientos generados [32, 39]. La entropía para cada nivel de clasificación, H_l , fue calculada de acuerdo a la siguiente ecuación basada en la teoría de la información [40]:

$$H_l = \sum_{i=1}^{s} p_i log_2(p_i) \tag{2-2}$$

Donde s representa el número de símbolos o clasificaciones diferentes en el nivel l y p_i es la probabilidad de encontrar la clasificación i en el nivel l. Esta ecuación fue propuesta para ser usada como evaluación de alineamientos múltiples de ESSs [32]. Congruentemente, el cálculo de la entropía puede ser simplificado para el caso de la comparación de dos *EC numbers* de la siguiente manera: si las clasificaciones de ambos *EC numbers* en el nivel l es igual, entonces el valor $H_l = 0$ y si son diferentes, entonces $H_l = 1$. De este modo, cuando el valor de entropía de dos *EC numbers* es alto (cercano a 1), entonces la similitud (información) es pequeña; mientras que cuando el valor de entropía es bajo (cercano a 0), entonces la similitud es alta. Adicionalmente, la evaluación de la entropía toma en cuenta la jerarquía de los *EC numbers*. Es decir, si el nivel de clasificación l de dos *EC numbers* es diferente, entonces la entropía del nivel $l + 1$ es igual a 1 de forma automática. De este modo se mantiene la jerarquía del sistema de clasificación y se evita que los *EC numbers* que difieran en los niveles de clasificación altos tengan valores de entropía bajos (menores a 1). Finalmente, se usa esta simplificación de la ecuación (2-2) y los factores de peso w_1, w_2, w_3 en la ecuación (2-1) para calcular la similitud entre dos

EC numbers. En la figura 2-5 se muestra la similitud correspondiente a la comparación de cada par de *EC numbers* presentes en la base de datos de ESSs creada en este trabajo. Los valores más cercanos a 0 representan *EC numbers* más similares y aquellos cercanos a 1 representan *EC numbers* más diferentes.

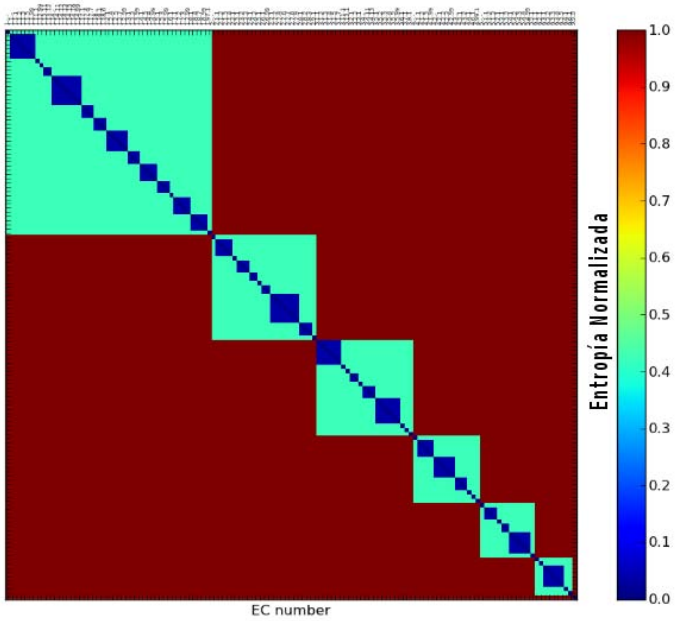

Figura 2-5. *Matriz de similitud entre EC numbers. La entropía normalizada de cada comparación se realizó usando la ecuación (2-1). Los números enzimáticos están ordenados de a cuerdo a su clasificación, de modo que se observa la jerarquía de este sistema de clasificación.*

La matriz de similitud S, ilustrada en la figura 2-5, es necesaria para la implementación del algoritmo de alineamiento NW y es análoga a las ya conocidas matrices de sustitución de aminoácidos como las PAM o las BLOSUM [6]. En total describe la similitud de 135 *EC numbers* diferentes en sus primeros tres niveles de clasificación y adiciona el número 9.9.9 que fue usado en este trabajo para representar aquellas enzimas que no tienen asignado ningún *EC number* en KEGG. Este número solo es similar a sí mismo.

Este esquema de evaluación de similitud entre *EC numbers* fue creado con la intención

[6]El método mostrado aquí busca evaluar la similitud general entre las reacciones químicas de dos enzimas usando sus *EC numbers*. Esto es fundamentalmente diferente a la forma en que se construyen las matrices de sustitución PAM o BLOSUM. Estas matrices se construyen usando la frecuencia observada de sustitución de un aminoácido por otro en un alineamiento múltiple.

de ser útil en un contexto de alineamientos múltiples y Algoritmo genéticos (AGs), donde mostró ser cualitativamente más eficiente que la aproximación clásica de suma de pares [32].

2.3.2. Algoritmo de alineamiento.

El algoritmo NW es un algoritmo exhaustivo que encuentra el alineamiento óptimo entre dos secuencias en función de un esquema de evaluación o *scoring* determinado. Esto lo hace encontrando el alineamiento global con máximo *score* en el espacio de posibles alineamientos entre dos secuencias. Este algoritmo se implementó cómo está descrito en la referencia [41] y se adaptó para funcionar con secuencias lineales de pasos enzimáticos, ESS. La principal diferencia en el algoritmo presentado aquí con respecto al algoritmo original, es que se busca el alineamiento con un *score* mínimo en lugar del máximo. Esto se debe a que, cómo se mencionó en la sección anterior, la similitud entre *EC numbers* está media en términos de entropía. De modo que un *score* mayor significa mayor entropía y por lo tanto, menor similitud. Mientras que un *score* pequeño significa menor entropía y por lo tanto, mayor similitud.

Para llevar a cabo el alineamiento de dos ESSs, ESS_1 y ESS_2 mediante el algoritmo NW se crea una matriz M con forma $(n+1)\times(m+1)$. Donde n es la longitud de la cadena ESS_1 y m es la longitud de ESS_2, ambas medidas en número de pasos enzimáticos. De este modo la matriz representa las comparaciones de todos los *EC numbers* en la secuencia ESS_1 contra todos los *EC numbers* en la secuencia ESS_2. La primer fila y columna se llenan con valores correspondientes a la penalización del *gap*. En este caso, la penalización por omisión de un *gap* es igual a 1. Un *gap* representa un espacio en el alineamiento, el cual se genera debido a que un paso enzimático en una ESS no puede alinearse con algún paso enzimático en la otra. Biológicamente hablando, se puede tratar de una enzima

ausente o agregada a una secuencia de pasos enzimáticos determinada.

Posteriormente, la matriz M se llena en su totalidad de acuerdo a la siguiente regla:

$$M_{i,j} = min \begin{cases} M_{i-1,j-1} + S_{EC_i,EC_j} \\ M_{i,j-1} + gap \\ M_{i-1,j} + gap \end{cases} \tag{2-3}$$

Donde S_{EC_i,EC_j} es el valor de similitud entre los *EC numbers* EC_i y EC_j en las ESSs 1 y 2 respectivamente. Esta ecuación significa que la celda $M_{i,j}$ obtiene el valor mínimo de tres posibles casos. El primer de ellos representa la adición de la celda anterior en diagonal con el valor de similitud de los *EC numbers* correspondientes proveniente de la matriz de similitud S. Por lo tanto, este primer caso representa el alineamiento de dos pasos enzimáticos consecutivos en ambas ESSs. Los otros dos casos representan la adición de una de las celdas anteriores en vertical o en horizontal con el valor de penalización del *gap*. De este modo, ambos casos representan la adición de un espacio en alguna de las ESSs del alineamiento.

Una vez que todas las celdas de la matriz M han sido llenadas, se traza el recorrido que se siguió desde la última celda hasta el origen. De este modo se obtiene el alineamiento óptimo que minimiza el *score* de la comparación de dos ESSs. En términos prácticos, un alineamiento, es la adición de *gaps* en las posiciones adecuadas de dos secuencias para maximizar la similitud entre ellas. De modo que para finalizar, se presentan las ESSs con los *gaps* correspondientes. El símbolo usado para representar los *gaps* en un alineamiento fue "-.-.-".

2.3.3. Función de evaluación del alineamiento

Finalmente, después de la aplicación del algoritmo NW, el alineamiento es evaluado usando una función normalizada que mide los mismos parámetros previamente descritos. La principal razón de usar esta evaluación es tener una medida de similitud normalizada que ocurra en el rango de 0 y 1. Adicionalmente, estos datos son comparables con los resultados previos de nuestro grupo donde se usó un AG [32,39] para realizar alineamientos múltiples. La ecuación (2-4) asigna un valor de *score* al alineamiento:

$$score = 0.95H + 0.05GP \tag{2-4}$$

Donde H representa la homogeneidad (entropía) del alineamiento y GP representa la penalización para los *gaps*. El peso ponderado para cada término de la ecuación fue asignado de forma empírica, dando más valor a la similitud u homogeneidad en el alineamiento que a la presencia de *gaps*. La homogeneidad es medida por la siguiente ecuación:

$$H = \frac{\sum_{i=1}^{n} S_{EC_{1i},EC_{2i}}}{n} \tag{2-5}$$

Donde n representa el número de pares de *EC numbers* alineados y $S_{EC_{1i},EC_{2i}}$ representa la similitud entre un par de *EC numbers* en la posición i del alineamiento. Esta similitud es obtenida de la matriz de similitud S. En otras palabras, esta ecuación calcula la similitud promedio de los *EC numbers* alineados.

Por otro lado, la penalización de los *gaps* está dada por la ecuación (2-6):

$$GP = \frac{\sum_{i=1}^{ns} \frac{GB_i}{TG_i}}{ns} \tag{2-6}$$

Donde GB_i es el número de bloques de *gaps* en la ESS i, TG_i es el número total

de gaps en la ESS i y ns representa el número de ESSs en el alineamiento ($ns = 2$ para el caso de los alineamientos pareados). En términos prácticos, esta ecuación mide la concentración de los *gaps*, es decir, que tan dispersos están en el alineamiento. El numerador se tiende a ser más pequeño conforme existan menos bloques de *gaps* en el alineamiento y por lo tanto la penalización es menor. Esta ecuación tiene el supuesto de que, al igual que en el caso de los nucleótidos o los aminoácidos, es más parsimonioso encontrar un solo *gap* grande que varios *gaps* pequeños. Los *gaps* presentes al principio y al final del alineamiento no son contados para esta evaluación. Aunque la penalización de los *gaps* trata de estar basada en un principio metodológico y filosófico ampliamente usado (parsimonia), es importante reconocer que los procesos evolutivos de inserción o eliminación de enzimas en un contexto metabólico pueden ser mucho más complejos.

Al igual que en el caso de la similitud entre *EC numbers*, el *score* obtenido con la ecuación (2-4) es un valor normalizado que va de 0 a 1. Como está basado en la medición de la entropía, el *score* del alineamiento entre dos ESSs es inversamente proporcional a la similitud de las mismas. Esto es, mientras más cercano a 0 sea el valor, más similares son las ESSs y viceversa.

Toda la metodología descrita hasta este punto está escrita en módulos de Python donde la mayoría de los parámetros (como la penalización por *gaps*) pueden ser modificados para cambiar el comportamiento de los algoritmos.

2.4. Evaluación estadística de los alineamientos pareados de *ESSs*.

Con el fin de poder evaluar la relevancia estadística de los alineamientos entre nrESS, se construyeron 10 bases de datos aleatorias. Cada base de datos se construyó reorde-

nando al azar todos los *EC numbers* de la base de datos nrESS original, manteniendo la longitud de las ESS y la proporción relativa de cada *EC number*. Estas bases de datos se compararon con la base de datos real para establecer un *score* umbral que permitió diferenciar entre los alineamientos con información y lo que pueden deberse únicamente al azar. El *score* umbral se seleccionó siguiendo los siguientes criterios:

- Aquel *score* en los datos reales con una mayor dispersión con respecto a los *scores* obtenidos de las bases de datos aleatorias y
- Aquel *score* en el cual se perdieran menos del 1 % de los nrESSs por el proceso de tamizado.

El valor que cumplió ambos criterios fue $score = 0.3$ por lo que resto de los análisis de este trabajo se realizaron filtrando los alineamientos y usando únicamente aquellos con un $score \leq 0.3$.

2.5. Identificación de pasos enzimáticos conservados en *Gammaproteobacteria*.

La información de similitud que proveen los alineamientos pareados de la base de datos nrESS se usó para identificar las regiones funcionalmente conservadas en el metabolismo de las *Gammaproteobacterias*. Para lograr esto, se seleccionó el siguiente criterio de conservación: dos ESS son consideradas conservadas si su alineamiento tiene un *score* menor o igual al umbral de similitud (0.3) y si en conjunto se encuentran en más del 75 % de los genomas de estudio. Este criterio considera que una ESS conservada debe estar al menos en aquellos genomas que no son considerados pequeños, i.e. con más de 2000 marcos de lectura abiertos (30 de los 40 genomas incluidos en este trabajo, ver tabla 2-1). De las nrESS que cumplieron este criterio se seleccionaron aquellas que corresponden

a alineamientos entre ESS del mismo mapa metabólico. Estas se usaron para mapear aquellos pasos enzimáticos conservados en cada mapa metabólico.

2.6. Agrupamiento de mapas metabólicos.

De forma similar se usó la información generada de los alineamientos pareados para agrupar los mapas metabólicos en función de su similitud funcional. Para hacer esto, se seleccionaron aquellos alineamientos con un *score* menor o igual al umbral de similitud (0.3) y se contó el número de alineamientos compartidos por cada par de mapas metabólicos. Los conteos se usaron para construir una matriz de similitud normalizada que fue utilizada para realizar un análisis de agrupamiento jerárquico usando el programa MeV4 [7]. La similitud entre los mapas fue calculada usando la correlación de Spearman y el agrupamiento se realizó con el método de unión promedio. Se usó un umbral de corte de 0.46 de la longitud del dendograma para definir grupos de mapas metabólicos similares. El agrupamiento se repitió usando la correlación de Kendall y distintos métodos de agrupamiento obteniendo en general resultados similares. El dendograma se visualizó usando el módulo de Python E.T.E. 2 [42] [8].

[7] http://www.tm4.org/mev.html
[8] http://etetoolkit.org/

Capítulo 3

Resultados y discusión.

3.1. Organismos de estudio.

Una lista completa de los genomas incluidos en este trabajo se muestran en la tabla 2-1 y su posición filogenética se muestra la figura 2-1. El conjunto de genomas seleccionadas incluyen organismos con un tamaño de genoma variado que van de 182 a 6778 *ORF*s y un número de *EC number*s diferentes que van de 69 a 859 cómo se muestra en la parte superior de la figura 3-1.

Para tener una idea de la importancia de las enzimas en el genoma de cada organismo, se graficó la proporción de proteínas asignadas a algún número enzimático (enzimas) en función del tamaño del genoma (figura 3-1 panel inferior). Se aprecia que los organismos con genomas pequeños tienden a tener una mayor proporción de enzimas, lo cual sugiere que existe alguna restricción en el número (y posiblemente en la naturaleza) de las enzimas que puede tener un organismo. La naturaleza de estas restricciones aún no ha sido aclarada y parece ser que la reducción y conservación de enzimas en los organismos con genomas pequeños depende de la historia evolutiva de cada linaje, dado que se ha demostrado que no existe un conjunto conservado de *EC number* constante en los organismos

con genomas pequeños [23].

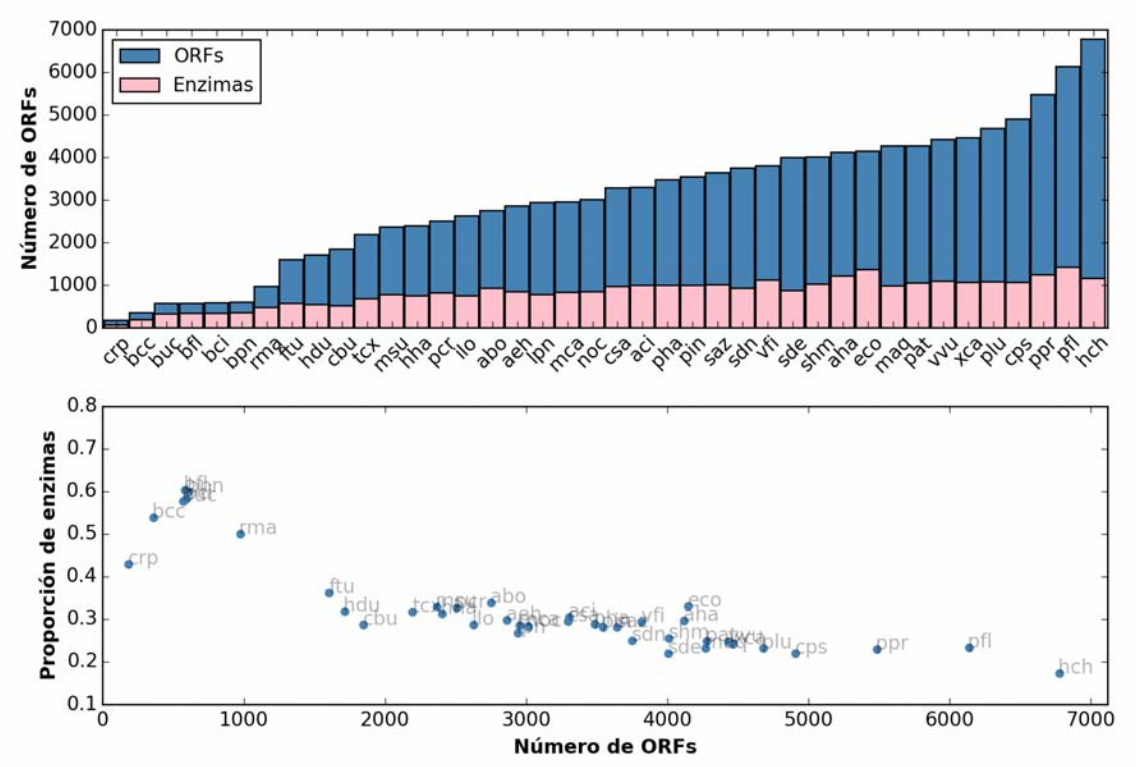

Figura 3-1. *Número de ORFs y proporción de enzimas en los genomas de las Gammaproteobacterias. En el panel superior se muestra el número crudo de ORFs y de enzimas asignadas a cada genoma en KEGG. Los genomas están ordenados de acuerdo al número de ORFs de menor a mayor. En el panel inferior, se muestra la proporción de enzimas del total de proteínas (ORFs) de cada genoma.*

Tomando como punto de partida esta observación realizamos un análisis de las proporciones de enzimas en 794 genomas no redundantes de procariontes y se comparó con el repertorio de factores de transcripción para cada genoma [21]. Se encontró que el comportamiento observado aquí se extiende a todos los procariontes y que el comportamiento es inverso para el caso de los factores de transcripción. Estos resultados sugieren una restricción funcional que limita el tamaño del metabolismo (tanto para genomas pequeños como para genomas grandes) y muestran que los genomas de mayor tamaño tienden a "invertir" más recursos a la regulación genética que a aumentar el tamaño de la red metabólica.

3.2. Creación de la base de datos de *ESSs*.

Para poder estudiar de forma comparativa el metabolismo de las *Gammaproteobacterias* su metabolismo fue fragmentado sistemáticamente usando el algoritmo BFS como fue descrito en la sección de métodos. Este algoritmo ha sido usado previamente para identificar la ruta más corta entre dos metabolitos en una red metabólica [43]. De este modo, el metabolismo se transforma en cadenas lineales de pasos enzimáticos (ESS) que pueden ser comparadas usando un algoritmo de alineamiento de secuencias. En este contexto, se definió una ESS cómo una colección lineal de pasos enzimáticos consecutivos desde un substrato hasta un producto dado, de forma similar a la definición formal de ruta metabólica previamente propuesta [25,27]. De este modo, una ESS representa un conjunto consecutivo de pasos enzimáticos donde cada paso enzimático está representado por los primeros tres niveles de clasificación de su *EC number*. La codificación de los pasos enzimáticos permite estudiar el metabolismo desde una perspectiva funcional la cual posibilita considerar procesos como la convergencia evolutiva. Además el uso de los primeros tres niveles de clasificación de los *EC numbers* nos permite identificar similitudes en los procesos catalíticos generales [7,8,23] lo que en última instancia nos permitiría estudiar similitudes lejanas cómo las que pueden existir entre distintas partes del metabolismo.

Siguiendo esta estrategia, se analizaron en total 2973 mapas metabólicos de la base de datos KEGG provenientes de 40 especies de *Gammaproteobacterias*. De estos, 2284 mapas generaron al menos una ESS. Los 689 mapas restantes no generaron ninguna secuencia debido a que contienen pocas enzimas, describen vías de ramificación o mecanismos de transporte y/o las enzimas no están conectadas entre sí. En total, se generaron 36,621 ESS con longitudes que van de 2 a 17 pasos enzimáticos con una longitud media de 5 y una moda de 3 (figura 3-2a). Adicionalmente, se identificó una correlación positiva entre el tamaño del genoma (figura 3-2b), medido en número de marcos abiertos de lectura

(ORFs por sus siglas en inglés) y el número de ESS generado ($r^2 = 0.78, p = 1.7 * 10^{-14}$). De igual modo, se observó una correlación positiva, aunque menos fuerte, entre el tamaño de cada mapa metabólico (ORFs) y el número de ESS generadas ($r^2 = 0.58, p \approx 0$). Estos resultados sugieren que el número de ESS refleja en cierta medida el incremento de la complejidad de un metabolismo en función del tamaño del genoma.

Una observación natural derivada de la construcción de las ESS es su redundancia, esto es, las ESS idénticas derivadas de distintos organismos. Para reducir esta redundancia y para facilitar los análisis posteriores se identificaron todas las ESS idénticas y se dejó una cadena representativa. De este modo se definió un conjunto de cadenas no redundantes, nrESS. El histograma de longitud del conjunto de datos nrESS tiene una distribución similar a los datos originales con una longitud media de 5.4 y una moda de 4 (figura 3-2b). Los análisis posteriores fueron realizados sobre el conjunto de datos nrESS con referencia a los datos originales cuando sea necesario.

3.3. Alineamiento de *ESSs* y análisis estadístico.

Las ESS del conjunto de datos nrESS fueron comparadas todas contra todas usando un algoritmo de alineamiento basado en el algoritmo Needleman-Wunsh de programación dinámica. Cada alineamiento es evaluado usando una función normalizada basada en entropía que arroja valores en el rango de 0 a 1. Como la función está basada en entropía, una evaluación (*score*) con un valor cercano a 0, significa menor entropía y por lo tanto, mayor homogeneidad en el alineamiento. Por esta razón, los *scores* cercanos a 0 significan alineamientos entre nrESS similares, mientras que los *scores* cercanos a 1 significan alineamientos entre nrESS diferentes. En el anexo A se muestran algunos ejemplos de alineamientos entre nrESS tomadas aleatoriamente de la base de datos.

De estas comparaciones, se observó que los *scores* siguen una distribución similar a

una distribución de valor extremo de Gumbel con la mayor proporción de alineamientos con *scores* cercanos a 1. Esto significa que la mayoría de los alineamientos ocurren entre secuencias con poca o nula similitud. Para evaluar estadísticamente estos resultados, se crearon 10 bases de datos de ESSs aleatorias (ver sección 2.4) y se analizaron de la misma forma que los datos originales. La distribución promedio de los *scores* de estos alineamientos se comparó con la distribución de los datos reales (figura 3-2c, d). Con base en esta comparación, se observó una sobrerrepresentación de *scores* cercanos a 0 en los alineamientos de los datos reales en comparación con los alineamientos de los datos aleatorios.

Para evaluar mejor esta sobrerrepresentación, se calculó la desviación de los datos reales con respecto a la media ± desviación estándar de las 10 bases de datos aleatorias (figura 3-2d). Se observa que la línea correspondiente a los datos reales se intersecta con los valores de los datos aleatorios en un *score* aproximado a 0.65; sugiriendo que este valor puede ser él límite para identificar similitudes distantes, por lo que cualquier *score* cercano a el puede considerarse completamente aleatorio. Por otro lado, para trabajar con los alineamientos más significativos, se estableció un umbral de similitud de 0.3. Este valor representa la máxima dispersión (más de 200 desviaciones estándar) de los valores reales con respecto a los valores aleatorios en la cual se pierde la menor cantidad de ESS con poca similitud global. Usando este *score* se incluyen a más de 99 % de las cadenas en la base de datos nrESS. Este umbral corresponde al 0.26 % de todos los alineamientos entre las nrESS reales (81,520 de 31,756,465) y al 0.04 % ± 0.001 % de los alineamientos en las bases de datos aleatorias. Estos resultados muestran que nuestro método puede identificar ESSs similares excluyendo la posibilidad de que los alineamientos sean generados únicamente por el azar.

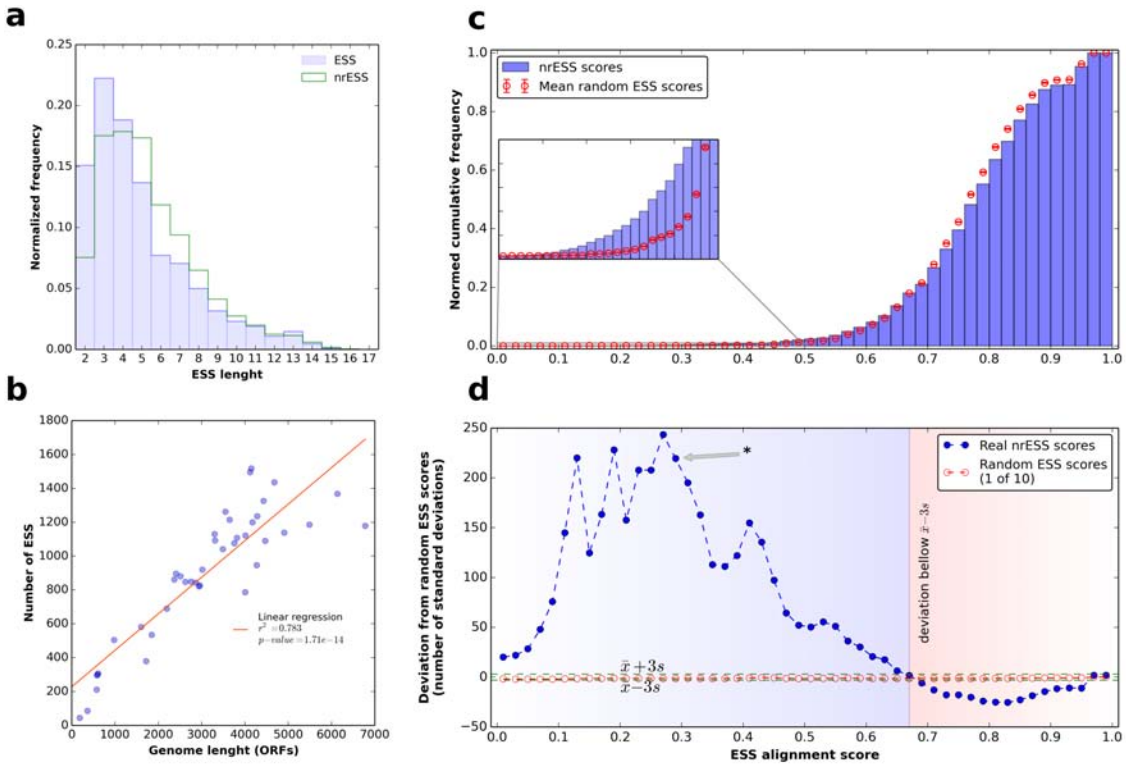

Figura 3-2. *Base de datos de ESSs y validación estadística de los alienamientos entre nrESSs. a) Histograma de longitudes de las bases de datos de ESS y nrESS. b) Número de ESS generadas por organismo en función del tamaño del genoma. El tamaño del genoma está medido en número de marcos abiertos de lectura (ORFs). Cada punto representa a un organismo. c) Histograma acumulativo de los scores de los alineamientos pareados de la base de datos nrESS y de 10 bases de datos aleatorias. Las barras azules representan los scores de los datos reales y los puntos rojos representan las medias ± desviación estándar de los scores de las 10 bases de datos aleatorias. Las desviaciones estándar son tan pequeñas que no se aprecian en la figura. El recuadro muestra una amplificación de la parte de la gráfica en el rango de scores de 0 a 0.5. d) Desviación de los scores reales con respecto a la media ± desviación estándar de scores de las bases de datos aleatorias. Los puntos azules corresponden a la desviación de los datos reales y los puntos rojos corresponden a la desviación de una de las 10 bases de datos aleatorias. Las lineas punteadas grises representan 3 desviaciones estándar de la media. El (*) denota el valor de desviación en un score = 0.3. Este score fue usado como umbral de similitud tal y como se describe en el texto.*

3.4. Identificación de pasos enzimáticos conservados en *Gammaproteobacterias.*

Con el fin de identificar las similitudes en los distintos mapas metabólicos de las *Gammaproteobacterias* y saber si existe algún grupo de rutas metabólicas comunes, se

seleccionó un conjunto de nrESS *funcionalmente conservados*. En este contexto, el concepto *funcionalmente conservado* se refiere a la conservación en la función de las enzimas (*EC numbers*), por lo que una función conservada puede deberse a diversos fenómenos evolutivos cómo la herencia vertical, herencia horizontal o convergencia funcional. En este contexto, se definió que dos nrESS se consideran funcionalmente conservadas si su alineamiento tiene un *score* menor o igual al umbral (≤ 0.3) y si, en conjunto, están presentes en más del 75 % de las especies analizadas. Basados en esta definición, el conjunto obtenido incluye 1484 nrESS provenientes de 74 mapas metabólicos diferentes. El 68.55 % de los alineamientos dentro de este conjunto corresponde a alineamientos entre secuencias del mismo mapa metabólico y el 31.43 % corresponde a alineamientos entre ESSs de distintos mapas (figura 3-3a).

Para conocer la conservación funcional de cada mapa metabólico seleccionamos únicamente aquellos alineamientos que ocurren entre nrESS provenientes del mismo mapa metabólico (aristas verdes en la figura 3-3a). La proporción de ESSs de cada mapa metabólico incluidas en este conjunto se muestran en la figura 3-3b. La mayoría de las nrESS corresponden al metabolismo de nucleótidos, seguido del metabolismo de carbohidratos, cofactores y vitaminas, amino ácidos y lípidos. En contraste, los mapas metabólicos relacionados con metabolismo de xenobióticos y degradación, biosíntesis de otros aminoácidos y metabolismo de terpenoides y policétidos, entre otros, representan menos del 5 % del total de nrESSs clasificados cómo funcionalmente conservados.

Usando estos alineamientos se mapeó la posición de los pasos enzimáticos *funcionalmente conservados* en cada mapa metabólico. Por cada alineamiento se seleccionaron aquellas columnas con *EC numbers* idénticos alineados y se mapeó su posición en cada mapa metabólico. Posteriormente se calculó la proporción de pasos enzimáticos(*EC numbers*) clasificados cómo funcionalmente conservados en relación al total de pasos enzimáticos presentes en las *Gammaproteobacterias* para cada mapa metabólico (figu-

ra 3-3c). Usando esta información, se clasificaron los mapas metabólicos en 4 categorías. 1) Mapas altamente (funcionalmente) conservados, aquellos mapas donde mas del 70 % de sus *EC numbers* fueron clasificados como funcionalmente conservados; 2) mapas moderadamente conservados, aquellos con una proporción entre el 30 % y el 70 %; 3) mapas poco conservados, aquellos con una proporción entre el 1 % y el 30 %; y 4) mapas variables, aquellos con ningún paso enzimático considerado funcionalmente conservado. Con base en lo anterior, menos de un tercio de los mapas analizados (24 de 86) caen dentro de las categorías de altamente o moderadamente conservados, mientras que más de dos tercios son considerados poco conservados o variables. Además, más de la mitad de los mapas metabólicos no contiene ESS consideradas funcionalmente conservadas en las *Gammaproteobacterias*. Esto sugiere que existe una alta variabilidad en el metabolismo de este clado y que en realidad son relativamente pocos los fragmentos del metabolismo que pueden considerarse comunes.

En detalle, los mapas metabólicos clasificados como altamente conservados incluyen mapas metabólicos con procesos importantes como la biosíntesis de ácidos grasos (map00061), el metabolismo de algunos aminoácidos (00290 biosíntesis de valina, leucina e isoleucina; 00300, biosíntesis de lisina), componentes de la pared celular (00540 biosíntesis de lipopolisacáridos; 00550 biosíntesis de peptidoglicanos), metabolismo de algunos cofactores (00770 biosintesis de CoA y pantotenato; 00780 metabolismo de la biotina; 00785 metabolismo del ácido lipoico) y la biosíntesis de la novobiocina (00401)[1]. El caso de los amino ácidos es congruente con el hecho de que las rutas de síntesis de aminoácidos como la valina, leucina, isoleucina y lisina son consideradas ancestrales a los

[1]Aunque la biosíntesis de la novobiocina aparece conservada funcionalmente, esto es el efecto de pasos enzimáticos que se repiten en dos mapas metabólicos diferentes. En la base de datos de *nrESSs* solo aparecen dos cadenas de pasos enzimáticos para la biosíntesis de la novobiocina. Ambas cadenas corresponden a la siguiente reacción: prefenato $\longleftrightarrow$ 4-hidroxi fenilpiruvato $\longleftrightarrow$ L-tirosina. Esta reacción química también está presente en el mapa metabólico que describe la biosíntesis de fenilalanina, tirosina y triptófano (mapa 00400). Por esta razón y en sentido estricto, la síntesis de novobiocina no está conservada en las *Gammaproteobacterias*

3 dominios celulares [44, 45].

Por otro lado, en los mapas definidos cómo moderadamente funcionalmente conservados se encuentran el metabolismo de las purinas (map00230) y pirimidinas (map00240), la glucólisis/gluconeogenesis (map00010), el ciclo del citrato (map00020), metabolismo de glicerofosfolípidos (map00564), terpenoides (00900) y algunos cofactores cómo riboflavina (map00740), nicotinamida (map00760), folato (map00790) y tiamina (map00730). En particular es interesante notar que la parte central del mapa de la glucólisis (ruta Embden-Meyerhof) está parcialmente conservada, es decir, la sección "alta" que corresponde a la transformación de las hexosas es considerada como variable, mientras que la sección "baja" correspondiente a las conversiones de las triosas hasta la oxidación de piruvato en acetil-CoA es considerada conservada (figura 3-3d). En la parte "alta", los pasos enzimáticos catalizados por la 6-fosfofructoquinasa (2.7.1.11) y la fructosa bifosfato aldolasa (4.1.2.13), al igual que las enzimas que permiten la entrada de las hexosas a la glucólisis son consideradas variables. De manera similar, las enzimas relacionadas con la fermentación láctica y etanólica y la enzima de la gluconeogénesis, fructosa bifosfatasa (3.1.3.11), son clasificadas como funcionalmente variables. Adicionalmente, la ruta alternativa Entner-Doudoroff, incluida en el metabolismo de las pentosas fosfato (map00030), está clasificada como variable. Estos resultados son consistentes con el trabajo previamente descrito (sección 1.1) donde se observó que la glucólisis es una ruta metabólica más plástica de lo que se pensaba y que la parte "baja"es la única que esta completamente conservada en los tres dominios celulares [10].

Otro ejemplo, es el caso del metabolismo de las purinas (00230) y las pirimidinas (00240). En general ambos mapas muestran conservación funcional en las reacciones que implican la síntesis de ribo y desoxiribonucleótidos mono, di y trifosfatados. En el caso del metabolismo de las purinas, la ruta de la inosina monofosfato, IMP (módulo de KEGG M00048) está clasificada como funcionalmente conservada. Esta es la única ruta conocida

para la síntesis *de novo* de nucleótidos purínicos. En congruencia, la ruta clásica para la síntesis de GTP está completamente conservada (M00050) y la ruta clásica para la síntesis de ATP (M00048) está parcialmente conservada. Por otro lado, las rutas de salvamento, la ruta de degradación de nucleótidos vía xantina y la ruta de utilización de alantoato son considerados variables. De modo que, en general, se observa conservación funcional en vías relacionadas con síntesis y variación en vías relacionadas con degradación. Estos resultados refuerzan la noción de que la biosíntesis de purinas es una ruta metabólica antigua [23, 46, 47].

Un patrón similar se observa en otros mapas metabólicos catalogados como moderadamente funcionalmente conservados. Por ejemplo, en el metabolismo de glicerofosfolípidos (000564) se observa que las rutas que van de CDP-diacilglicerol a fosfatidil glicerol, fosfatidil serina y fosfatidil etanolamina están funcionalmente conservadas, mientras que la ruta biosintética a fosfatidil colina y las rutas de degradación son consideradas variables. Análogamente, se observa conservación en las vías de síntesis de cofactores cómo la tiaminadifosfato (00730), NAD+ y NADP+ (00760) y tetrahidrofolato (00790). En conjunto, es posible deducir un patrón de conservación funcional en el metabolismo de las *Gammaproteobacteria*, en el cual algunos mapas metabólicos contienen un *core* de pasos enzimáticos conservados relacionados con procesos biosintéticos que están acompañados con un conjunto de pasos enzimáticos variables que incluyen principalmente procesos de degradación. Estos pasos variables pueden representar posibles rutas alternativas en diferentes organismos y/o diferentes nichos ecológicos como ha sido propuesto previamente [44, 45].

El grupo de mapas metabólicos clasificados como poco funcionalmente conservados incluye varios procesos importantes tales como el metabolismo de aminoácidos, metabolismo de ácidos grasos (beta oxidación) y el metabolismo de glicerofosfolípidos. En particular, es posible identificar varias reacciones alternativas en el mapa que describe el

metabolismo de la alanina, el aspartato y el glutamato (00250) sugiriendo la existencia de rutas alternativas para producir dichos compuestos. Por ejemplo hay tres posibles enzimas que catalizan la conversión de L-glutamato a L-glutamina: una mediante una reacción de ligación (glutamina sintetasa, 6.3.1.2) y dos mediante hidrólisis reversible (glutaminasa, 3.5.1.2 y L-glutamina [L-aspargina] amidohidrolasa, 3.5.1.38). En particular, la L-glutamina [L-aspargina] amidohidrolasa también cataliza la deaminación de la aspargina a aspartato. Este resultado es congruente con la observación previa de pasos enzimáticos alternativos en el metabolismo de ciertos aminoácidos [44] y refuerza el argumento de que las redes de síntesis de ciertos aminoácidos son más flexibles.

Una observación similar se aprecia en el metabolismo de la cisteina y la metionina (00270). En este mapa metabólico, la ruta canónica para la síntesis de metionina a partir de aspartato (M00017) está conservada de forma fragmentada en las *Gammaproteobacterias*. Sin embargo, se identificaron algunos pasos enzimáticos que pueden servir como rutas alternativas para la síntesis de metionina. Además, varios de estos paso enzimáticos están clasificados como conservados lo cual sugiere la presencia de otras rutas importantes para la síntesis de metionina diferentes de la ruta canónica.

Finalmente, los mapas clasificados como variables pertenecen a distintos tipos de metabolismo. Algunos de ellos contienen pocos pasos enzimáticos o pasos enzimáticos que no están conectados por alguna relación sustrato/producto, lo cual sugiere la ausencia de estos metabolismos en *Gammaproteobacteria*. Sin embargo, también entran en esta clasificación algunos mapas densamente poblados de pasos enzimáticos, como son el metabolismo de compuestos de selenio, el metabolismo de la galactosa, el metabolismo de las pentosas fosfato y las interconversiones del glucuronato. En general en esta categoría se describen procesos de degradación de compuestos poco comunes (como los xenobióticos) o procesos para la utilización de fuentes alternativas de carbono (metabolismo de carbohidratos). En conjunto, estos resultados apoyan la idea propuesta arriba relacionada

con la alta variación de los procesos de degradación, lo que pone en evidencia la posibilidad de patrones diferenciales de conservación y reclutamiento enzimáticos en el clado. Adicionalmente, estos resultados respaldan la preponderancia del metabolismo central del carbono y de las rutas anabólicas en la evolución del metabolismo [19, 23, 48].

En resumen, nuestro análisis permite la identificación de un *core* de pasos enzimáticos funcionalmente conservados en *Gammaproteobacteria*. Este *core* incluye principalmente el metabolismo central del carbono (parte "baja" de la glucólisis, ciclo del citrato) y las rutas biosintéticas para nucleótidos, cofactores y algunos aminoácidos. El *core* está complementado con un conjunto de pasos enzimáticos variables, los cuales incluyen principalmente rutas de degradación de carbohidratos, aminoácidos y xenobióticos que pueden ser esenciales para cada estilo de vida. Además se pueden identificar la presencia de rutas alternativas para la síntesis de varios aminoácidos.

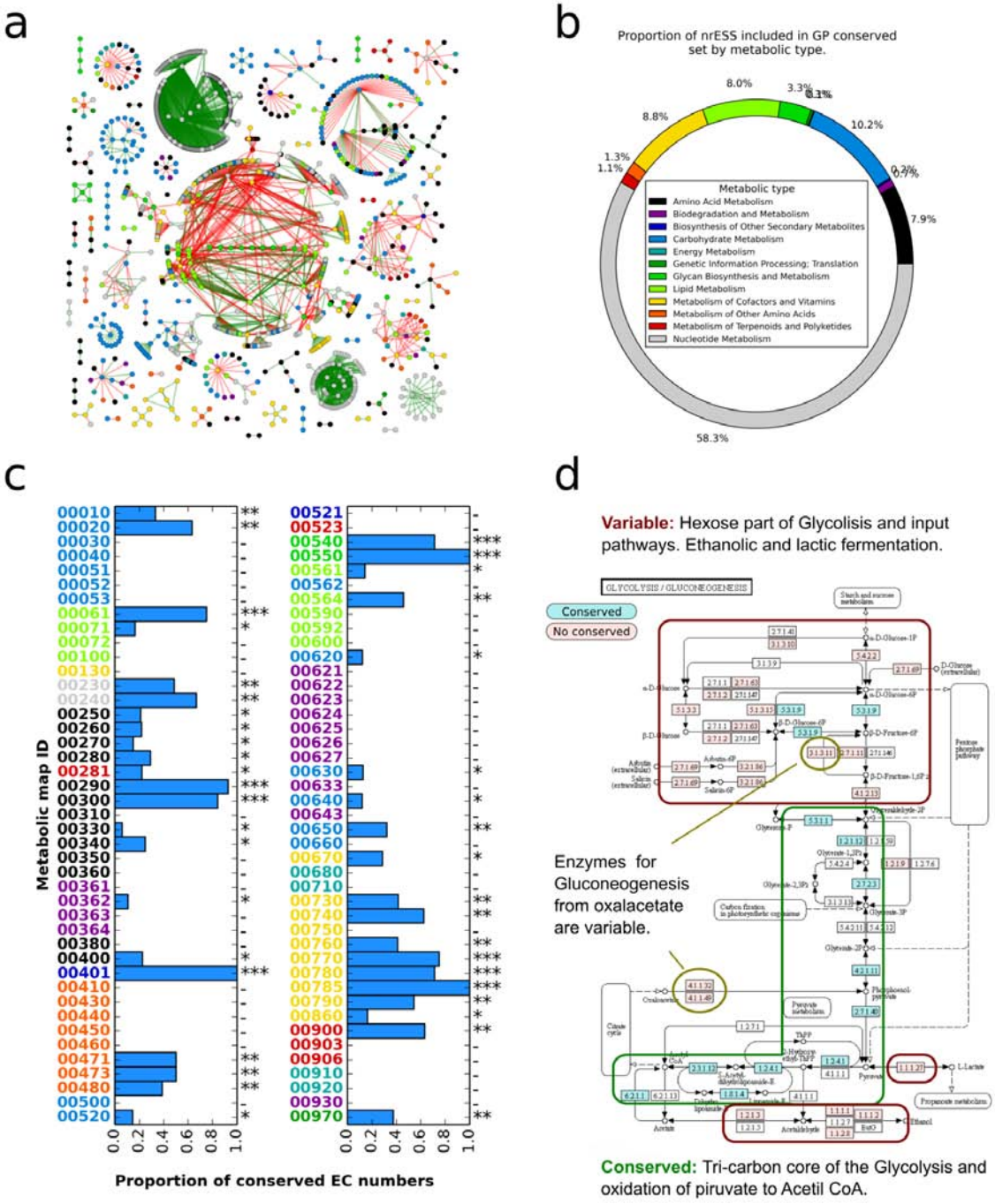

Figura 3-3. *Pasos enzimáticos funcionalmente conservados en las Gammaproteobacterias.*

Figura 3-3 (*continuación*). *a) Representación de red de las relaciones entre nrESS del conjunto funcionalmente conservado. Cada Nodo representa una nrESS. Los vértices verdes representan alineamientos entres ESS del mismo mapa metabólico y los vértices rojos representan alineamientos entre ESS de distintos mapas metabólicos. El color del nodo representa el tipo de metabolismo de donde se originó cada ESS. b) Proporción de nrESS de cada tipo metabólico incluidos en el conjunto funcionalmente conservado. c) Proporción de EC numbers clasificados como funcionalmente conservados de cada mapa metabólico. Cada mapa metabólico está codificado usando la clave de 5 números de KEGG. El color de la fuente indica el tipo metabólico. Los mapas metabólicos fueron clasificados en función del porcentaje de pasos enzimáticos funcionalmente conservados (ver texto): *** funcionalmente conservados; ** parcialmente conservados; * poco conservados; - variables. d) Pasos enzimáticos funcionalmente conservados en el mapa de la Glucólisis/Gluconeogénesis (map00010). Los pasos clasificados cómo conservados se muestran en color cian y los pasos clasificados cómo variables se muestran en color rosa. Las cajas en color blanco indican pasos enzimáticos que no están presentes en las Gammaproteobacterias. Se indican las regiones del mapa metabólico consideradas conservadas y variables. Detalles en el texto.*

3.5. Agrupación de mapas metabólicos en función de la similitud de sus *ESSs*.

Con el fin de conocer las posibles relaciones funcionales que pueden existir entre los distintos mapas metabólicos analizamos la similitud que existe entre ellos por medio de las comparaciones de las nrESS. Para esto, se cuantificó el número de alineamientos entre cada par de mapas metabólicos con un *score* ≤ 0.3. La matriz resultante se normalizó por fila con respecto al total de alineamientos de cada mapa metabólico. Las filas de la matriz normalizada se usaron como vectores de similitud para el análisis de agrupamiento (*clustering*) jerárquico usando cómo medida de similitud la correlación de Spearman (figura 3-4).

Usando este agrupamiento, definimos 24 grupos de mapas metabólicos. De estos, 5 contienen más de tres mapas metabólicos, 5 contienen entre 2 y 3 mapas y catorce son grupos simples (1 mapa). Los grupos con más de 3 mapas fueron considerados grupos mayores. El primero de estos grupos, incluye mapas relacionados con la degradación de compuestos aromáticos como el amino benzoato, bisfenol, tolueno y naftaleno, entre otros. El segundo grupo incluye mapas relacionados con el metabolismo de amino ácidos

hidrofílicos cómo el triptófano, fenilalanina y tirosina y la degradación de otros compuestos aromáticos cómo el xileno y las dioxinas. El tercer grupo contiene el metabolismo de ácidos caboxílicos (ácidos grasos y butanoato), lípidos de cadena alifática larga (limoneno y geraniol), degradación de valina, leucina e isoleucina y el metabolismo del piruvato y el ciclo del citrato. Estos resultados sugieren una posible relación funcional y evolutiva entre el metabolismo del piruvato y el ciclo del citrato con el metabolismo de compuestos con cadenas alifáticas hidrófobas.

El cuarto grupo incluye mapas relacionados con el metabolismo de carbohidratos, fijación de carbono, metabolismo de algunos cofactores (CoA y folato), terpenoides, glicerolípidos y compuestos con azufre (incluyendo la metionina y cisteina) y selenio. Finalmente, el quinto grupo contiene el metabolismo de nucleótidos, peptidoglicanos, metabolismo del nitrógeno y de otros cofactores que contienen nitrógeno cómo la tiamina y la nicotinamida.

En resumen, se identificó una posible tendencia, donde los mapas metabólicos que describen la transformación de moléculas químicamente similares también contienen cadenas de pasos enzimáticos similares. Adicionalmente, es posible que estos resultados refuercen la noción de reclutamiento enzimáticos entre rutas cómo ha sido propuesto previamente (ej, [34, 49]). Hasta donde sabemos, este es el primer intento de establecer las similitudes entre mapas metabólicos usando de forma sistemática la información funcional de los *EC numbers* y la comparación de rutas metabólicas. Es importante mencionar que estas conclusiones tienen un alcance limitado al metabolismo cubierto por las *Gammaproteobacterias*, por lo que puede ser bastante constructivo extender este análisis usando todo el metabolismo conocido.

Con base en estas observaciones se buscó estudiar a mas detalle cómo pueden ser los patrones de similitud entre distintos mapas metabólicos y si estos patrones pueden dar pistas de posibles eventos de reclutamiento enzimático en el metabolismo de las *Gamma-*

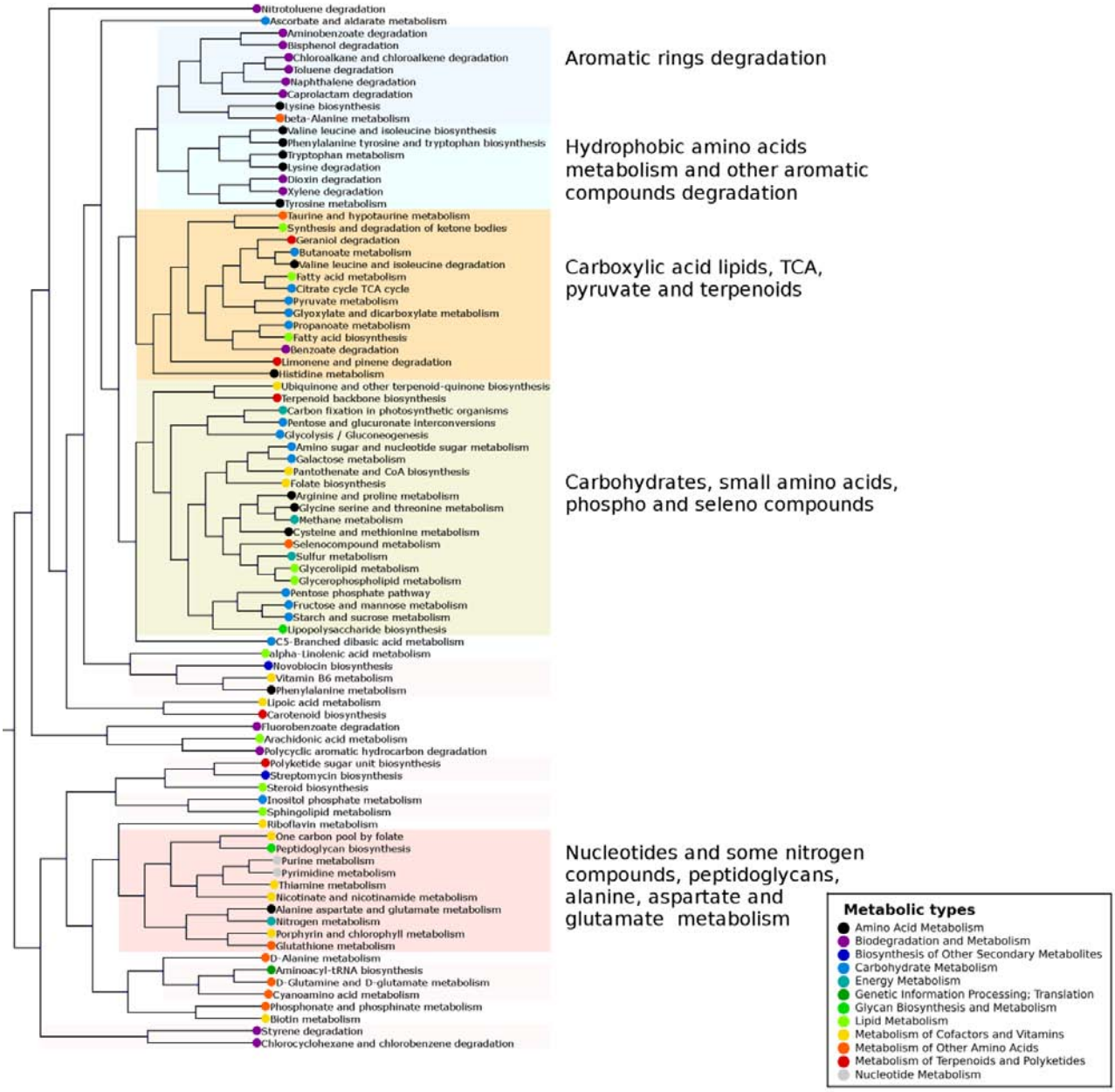

Figura 3-4. *Los mapas metabólicos que incluyen compuestos similares tienden a tener ESSs similares. Se indican los grupos identificados y el tipo de metabolismo de cada mapa metabólico.*

proteobacaterias. Para explorar esta posibilidad se seleccionaron las ESS que describen la parte baja de la glucólisis y la vía de IMP. Ambas vías fueron consideradas funcionalmente conservadas en este trabajo y evolutivamente conservadas en trabajos previos [10, 23, 50].

Estas ESS se compararon contra toda la base de datos nrESS usando el algoritmo de alineamiento con un sistema de evaluación optimizado para identificar las regiones de máxima similitud. Se usó el umbral de similitud previamente descrito ($score \leq 0.3$) para considerar un alineamiento cómo positivo (*hit*).

De estas comparaciones se observó que el número de *hits* significativos identificados para ambas ESS es mayor a lo que se esperaría encontrar en las bases de datos aleatorias (figura 3-5). Sin embargo, la diferencia relativa es mayor en el caso de la vía de IMP que en el caso de la parte baja de la glucólisis. La diferencia relativamente pequeña en el número de alineamientos significativos para el caso de la glucólisis comparada con lo obtenido con las bases de datos aleatorias puede ser explicada en parte por el número de ocurrencias de sus *EC numbers* en las bases de datos nrESS y KEGG (figura 3-5c). Los números enzimáticos 1.2.1 (oxidorreductasas que actúan en un grupo oxo o aldehído con NAD+ o NADP+ como aceptor), 2.7.1 (fosfotransferasas con grupo alcohol como aceptor) y 4.2.1 (hidrolasas de carbón-oxígeno) están entre los primeros 10 en abundancia en la base de datos nrESS.

Adicionalmente, se observó que aunque el número crudo de alineamientos significativos fue mayor para el caso de las vía del IMP (381) en comparación de la glucólisis (148), el número de alineamientos con algún otro mapa metabólico es mayor en el caso de la glucólisis (37 contra 19). Este resultado muestra que la parte baja de la glucólisis tiene una amplia similitud con otros procesos metabólicos y sugiere procesos catalíticos similares usados para transformar distintos compuestos en otras partes del metabolismo. Además, esta observación sugiere posibles eventos de reclutamiento a partir de la glucólisis. En contraste, la mayoría de los *hits* obtenidos con la vía del IMP corresponden alineamientos dentro del mapa del metabolismo de las Purinas. Aunque ambas rutas metabólicas pueden ser consideradas antiguas y están clasificadas como funcionalmente conservadas, lo patrones de similitud funcional son diferentes y posiblemente reflejan las

restricciones funcionales del reclutamiento enzimático que pueden tener relación con la ubicuidad de ciertos tipos de compuestos o grupos funcionales.

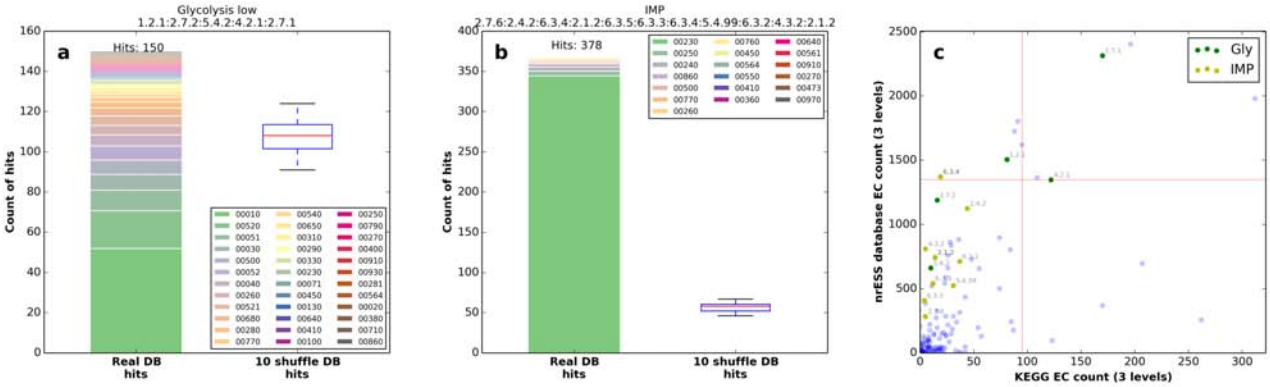

Figura 3-5. *Similitud de la parte baja de la glucólisis y la ruta del IMP con el resto del metabolismo. Comparación de la parte baja de la glucólisis (a) y de la vía del IMP (b) contra la base de datos nrESS (barra) y contra las 10 bases de datos aleatorias (gráficas de caja). El código de color representa la fracción de alineamientos correspondiente a cada mapa metabólico. c) Relación entre el número de EC numbers que comparten los primeros tres niveles de clasificación en la base de datos KEGG y el número de veces que aparece cada número enzimático en la base de datos nrESS. Se indican los EC numbers que constituyen la parte baja de la glucólisis y la vía del IMP. Las lineas rojas indican los 10 EC numbers más abundantes en cada base de datos.*

Capítulo 4

Resumen y conclusiones.

En este trabajo se ha mostrado el uso de una estrategia sistemática para el estudio comparativo del metabolismo. Esta estrategia se aplicó para buscar las similitudes y diferencias en el metabolismo de un clado diverso y bien estudiado, *Gammaproteobacteria*. El método se basa en alineamientos de secuencias de pasos enzimáticos, llamados ESS. En una ESS cada paso enzimático está codificado por su *EC number*. De modo que este método de comparación permite conocer las similitudes funcionales del metabolismo sin tomar en cuenta de forma directa la información de las secuencias biológicas, por lo que, en principio, el método permitiría estudiar procesos cómo la convergencia funcional.

Con este análisis se identificó un conjunto de pasos enzimáticos funcionalmente conservados en las *Gammaproteobacterias* donde sobresalen principalmente rutas metabólicas anabólicas y relacionadas con el metabolismo central del carbono y de cofactores. Por otro lado, las partes variables del metabolismo están involucradas en general con procesos catabólicos que pueden estar relacionados con nichos ecológicos específicos. Estos resultados concuerdan con estudios previos y sugieren procesos diferenciales de pérdida y ganancia de pasos enzimáticos dentro del metabolismo.

Adicionalmente, se creó un agrupamiento de mapas metabólicos en función de la

similitud de sus ESS. De este modo, se observó que en general los mapas metabólicos que catalizan la conversión de moléculas químicamente similares presentan ESS similares. Complementariamente se analizaron los patrones de similitud de dos rutas metabólicas consideradas conservadas y se observó que ambas tienen patrones de similitud diferentes, reforzando la idea de procesos diferenciales de reclutamiento. Además, aparentemente la parte baja de la glucólisis pudo haber sido una importante vía donadora de pasos enzimáticos en el resto del metabolismo. Esta idea debe ser abordada más a fondo.

Actualmente, estamos trabajando en crear una base de datos que incluya una mayor cantidad de organismos y en extender nuestras observaciones con respecto a la presencia de enzimas homólogas y análogas en los alineamientos de ESS.

Apéndice A

Ejemplos de alineamientos.

En esta sección se muestran algunos ejemplos de alineamientos de ESSs usando en algoritmo propuesto en este trabajo. Los alineamientos corresponden a cadenas de pasos enzimáticos tomados aleatoriamente de la base de datos de cadenas no redundantes, *nrESS*.

nrid	1	2	3	4	5	6	7	8	9	10	11	12	13
7454	-.-.-	6.3.4	2.1.2	6.3.5	6.3.3	6.3.4	5.4.99	6.3.2	4.3.2	2.4.2	2.7.1	-.-.-	-.-.-
2528	2.4.2	6.3.4	2.1.2	6.3.5	6.3.3	6.3.4	5.4.99	6.3.2	4.3.2	2.7.4	2.7.1	2.7.6	3.6.1

score = **0.25015**

nrid	1	2	3	4
7883	6.4.1	2.3.1	2.7.2	1.2.1
7882	6.4.1	2.3.1	2.7.2	-.-.-

score = **0.2375**

nrid	1	2	3	4	5	6	7	8	9	10
1388	1.5.99	1.2.1	2.6.1	4.1.1	3.5.1	2.1.3	2.7.2	1.4.1	3.5.1	1.2.1
1383	1.5.99	1.2.1	2.6.1	4.1.1	3.5.1	2.1.3	2.7.2	1.4.1	1.5.1	1.5.1

score = **0.13519**

nrid	1	2	3	4	5	6	7
3614	2.7.2	1.2.1	1.1.1	2.3.1	2.5.1	2.3.1	2.5.1
1748	2.1.3	4.3.1	1.5.99	-.-.-	1.5.1	2.3.1	-.-.-

score = **0.68269**

nrid	1	2	3
1653	-.-.-	2.1.3	3.5.1
1542	2.1.1	2.1.1	1.3.1

score = **0.64551**

nrid	1	2	3	4	5
6600	4.3.2	3.5.3	4.3.1	1.5.1	2.7.2
4214	3.1.1	3.1.4	4.3.1	-.-.-	-.-.-

score = **0.65038**

nrid	1	2	3	4	5	6
3857	-.-.-	-.-.-	-.-.-	2.7.4	3.6.1	-.-.-
1288	1.5.1	3.5.1	1.2.1	2.6.1	3.5.3	2.3.1

score = **0.76731**

nrid	1	2	3
6781	4.4.1	4.3.1	-.-.-
185	1.1.1	2.7.1	1.1.1

score = **0.95**

nrid	1	2	3	4	5	6	7	8
3739	2.7.2	1.4.1	2.3.1	2.7.2	1.2.1	2.6.1	3.5.1	2.1.3
2760	2.5.1	-.-.-	-.-.-	4.2.1	1.3.1	-.-.-	-.-.-	-.-.-

score = **0.82548**

Apéndice B

Artículos publicados.

Los artículos que fueron publicados directamente como productos de este trabajo o como productos derivados se listan a continuación. El primero de ellos se anexa en este documento.

- [1]**Poot-Hernandez AC**, Rodriguez-Vazquez K, Perez-Rueda E. 2015. The alignment of enzymatic steps reveals similar metabolic pathways and probable recruitment events in Gammaproteobacteria. *BMC Genomics*. 16:957.

- Ortegon P, **Poot-Hernández AC**, Perez-Rueda E, Rodriguez-Vazquez K. 2015. Comparison of Metabolic Pathways in Escherichia coli by Using Genetic Algorithms. *Comput Struct Biotechnol J*. 2015 Apr 9;13:277-85.

- Martínez-Núñez MA, **Poot-Hernandez AC**, Rodríguez-Vázquez K, Perez-Rueda E. 2013. Increments and duplication events of enzymes and transcription factors influence metabolic and regulatory diversity in prokaryotes. *PLoS One*. 2013 Jul 29;8(7):e69707.

[1]Este artículo fue presentado como requisito para el trámite de titulación de Doctorado.

Poot-Hernandez et al. BMC Genomics (2015) 16:957
DOI 10.1186/s12864-015-2113-0

BMC Genomics

RESEARCH ARTICLE Open Access

The alignment of enzymatic steps reveals similar metabolic pathways and probable recruitment events in *Gammaproteobacteria*

Augusto Cesar Poot-Hernandez[1,2*], Katya Rodriguez-Vazquez[2] and Ernesto Perez-Rueda[1*]

Abstract

Background: It is generally accepted that gene duplication followed by functional divergence is one of the main sources of metabolic diversity. In this regard, there is an increasing interest in the development of methods that allow the systematic identification of these evolutionary events in metabolism. Here, we used a method not based on biomolecular sequence analysis to compare and identify common and variable routes in the metabolism of 40 *Gammaproteobacteria* species.

Method: The metabolic maps deposited in the KEGG database were transformed into linear Enzymatic Step Sequences (ESS) by using the breadth-first search algorithm. These ESS represent subsequent enzymes linked to each other, where their catalytic activities are encoded in the Enzyme Commission numbers. The ESS were compared in an all-against-all (pairwise comparisons) approach by using a dynamic programming algorithm, leaving only a set of significant pairs.

Results and conclusion: From these comparisons, we identified a set of functionally conserved enzymatic steps in different metabolic maps, in which cell wall components and fatty acid and lysine biosynthesis were included. In addition, we found that pathways associated with biosynthesis share a higher proportion of similar ESS than degradation pathways and secondary metabolism pathways. Also, maps associated with the metabolism of similar compounds contain a high proportion of similar ESS, such as those maps from nucleotide metabolism pathways, in particular the inosine monophosphate pathway. Furthermore, diverse ESS associated with the low part of the glycolysis pathway were identified as functionally similar to multiple metabolic pathways. In summary, our comparisons may help to identify similar reactions in different metabolic pathways and could reinforce the *patchwork model* in the evolution of metabolism in *Gammaproteobacteria*.

Keywords: Metabolism, Pathway alignment, Gammaproteobacteria, Enzyme commission number

Background

The study of the evolution of metabolism is central to understanding the adaptive processes of cellular life, the emergence of high levels of organization (multicellularity), and the diversity and complexity of the living world [1, 2]. At present, the large-scale information derived from genomic and proteomic studies has allowed the development of databases devoted to organizing the metabolic processes, such as the KEGG [3] and MetaCyc [4]. The information contained in these databases can be used to generate an integrative perspective of cellular functioning.

Metabolism can be considered one of the most ancient biological networks, where the nodes represent substrates and/or enzymes and the edges represent the relationships among them. From this perspective, the study of metabolic networks has focused on describing topological properties and has showed the existence of a structured network architecture [5–7]. Another relevant feature of metabolic networks is their modularity [8, 9], where each module is a discrete entity of elementary components (enzymes and substrates) that performs a certain task, separable from the functions of other modules. The elements of each module are related to each

* Correspondence: apoot@ibt.unam.mx; erueda@ibt.unam.mx
[1]Departamento de Microbiología Molecular, Instituto de Biotecnología, UNAM, Av. Universidad 2001, Cuernavaca, Morelos CP 62210, México
Full list of author information is available at the end of the article

other and may be subjected to the same evolutionary process, such as amino acid biosynthesis, where a high rate of duplication events has been identified [10]. In this regard, metabolic pathways exhibit high retention of duplicates within functional modules and a preferential biochemical coupling of reactions. This retention of duplicates may result from the biochemical rules governing substrate-enzyme-product relationships [11–13].

In this work, we ask whether there are groups of similar reactions in different or in the same metabolic pathways, which might suggest a transfer of enzymatic activities, and whether these groups can be used to define common and variable metabolic pathways in 40 organisms belonging to the *Gammaproteobacteria* division. *Gammaproteobacteria* are excellent models to consider because they contain a large diversity of species [14], such as the bacterium *Escherichia coli*K-12, for which a large number of molecular and functional mechanisms have been elucidated. In addition, *Gammaproteobacteria* include organisms widely distributed throughout diverse environments, such as the endocommensal bacterium *Ruthia magnifica* [15], obligate endosymbionts *Baumannia* sp. and *Buchnera* sp., photoautotrophs such as *Halorhodospira halophile* [16], and mammal pathogens, such as *Yersinia* spp. and *Vibrio* spp., among others [17, 18].

To this end, we implemented a general strategy that considers the transformation of the metabolic maps deposited in the KEGG database into linear Enzymatic Step Sequences (ESS) and their posterior comparison with a dynamic programming sequence alignment algorithm. From these comparisons, we show that maps associated with the metabolism of similar compounds also contain a high proportion of similar ESS. In addition, we evaluate the possible contribution of two ancient pathways, glycolysis and IMP, to the metabolic pathways growth. Finally, we consider that our comparisons may provide clues reinforcing the *patchwork model* in the evolution of metabolism in *Gammaproteobacteria*.

Results

Construction and comparison of ESS

In order to evaluate the commonalities and differences in the metabolism of organisms belonging to the *Gammaproteobacteria* division, a collection of ESS was generated from their corresponding metabolic maps. In this regard, an ESS was defined as a linear collection of consecutive enzymatic reactions from a given substrate to a given product, in a similar way as a previously proposed definition of metabolic pathways [19, 20]. To do this, the breadth-first search (BFS) algorithm was used, as we describe in Material and Methods. This algorithm allows the systematic fragmentation of metabolic pathways for the alignment analysis, and it has been used to identify the shortest pathway between compounds in metabolic networks [21]. Therefore, each ESS was reconstructed following subsequent reactions in each metabolic map. The enzymes related to each reaction were represented by using the first three levels of the Enzyme Commission (EC) number classification to describe their general type of chemical reaction, as it was previously suggested [22]. In total, 2973 KEGG maps from 40 species were analyzed, of which 2284 generate at least one sequence. The remaining 689 maps did not generate any sequence because they contain few enzymes, contain ramification pathways or describe transport mechanisms, or the enzymes do not have connections among them (Additional file 1: Table S1). Therefore, the length distribution of the total 36,621 constructed ESS ranged from 2 to 17 enzymatic steps, with a mean length of 5 and a mode equal to 3 (Fig. 1a). In addition, we found a correlation between the genome size (in open reading frames, or ORFs) and the number of ESS, as large genomes generated more ESS ($r^2 = 0.78$, $p = 1.7\text{x}10^{-14}$) than small genomes (Fig. 1b). In a similar way, the number of ESS generated per metabolic map also depended on the number of ORFs associated with each map ($r^2 = 0.581$, $p \approx 0$) (Additional file 2: Figure S1). These results suggest that the number of ESS reflects to some extent the increased complexity in metabolism as a function of the number of proteins contained in an organism.

A natural observation that emerged from these sequences concerns their redundancy, i.e., identical ESS derived from different organisms. To reduce this redundancy and to facilitate the subsequent analysis, identical sequences were identified and excluded, leaving a representative of them and defining the non redundant ESS (nrESS) dataset. From this filtering, 7970 different nrESS were considered for posterior analyses. The nrESS length histogram was similar to that for the complete set of ESS, with a mean length of 5.4 and a mode equal to 4 (Fig. 1a). In this report, we refer only to the nrESS.

In a second step, the nrESS were compared by using the dynamic programming Needleman and Wunsh (NW) algorithm in an all-against-all strategy. The alignment generated by this algorithm was evaluated by using an entropy based normalized function that yields values in the interval from 0 to 1. Hence, values close to 0 mean less entropy and more homogeneous columns in the alignment, reflecting more similar nrESS. Conversely, values close to 1 reflect dissimilar nrESS.

From these comparisons, we found that the distribution of the scores resembled an extreme value Gumbel distribution (Additional file 2: Figure S2), with the highest proportion of the scores close to 1, i.e., the major proportion of alignments occurs between dissimilar nrESS. To evaluate the statistical significance of all comparisons, 10 random databases were generated by

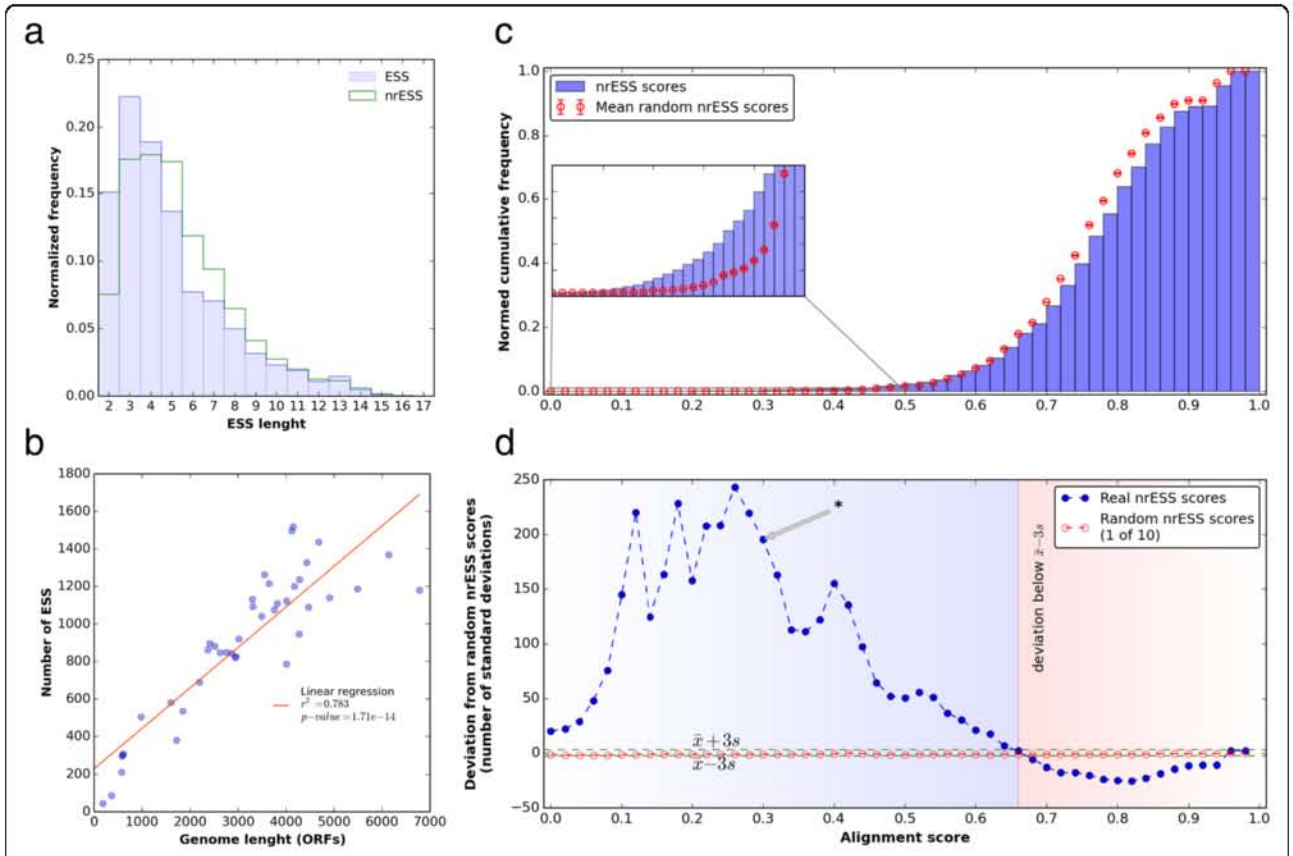

Fig. 1 Construction and alignment of Enzymatic Step Sequences (ESS). **a** Histogram of the distribution of ESS and nrESS according to their lengths. **b** Number of total ESS constructed per organism. The *x*-axis represents the genome size according to the number of ORFs, and the *y*-axis corresponds to the total number of ESS generated using the strategy presented in the text. Each point corresponds to one organism. **c** Cumulative histogram of the nrESS pairwise alignment scores. Blue bars represent the real nrESS scores; red points represent the mean distribution ± SD from random ESS. The SD were so small that they are not visible in the plot. The inset is an amplification of the lower portion of the plot and includes the scores from 0 to 0.5. **d** Deviation of the real nrESS score histogram relative to the mean ± SD of the random datasets. The blue dots correspond to the real data dispersion, and red dots correspond to single random database dispersions. Dotted green lines represent 3 SD from the mean of random histograms. The asterisk denotes the dispersion at 0.3. This score was used as the similarity threshold

shuffling the real nrESS, maintaining the EC composition and length sizes. The random databases were analyzed in the same all-against-all fashion, and the resulting scores were compared against real alignment scores. In Fig. 1c we show the cumulative histogram of the alignment scores of the real and random datasets. Based on this analysis, scores close to 0 are overrepresented in real data compared to random nrESS. To evaluate this overrepresentation, the deviation of the real dataset relative to the mean ± standard deviation of the 10 random datasets was calculated (Fig. 1d). According to these data, the real and random scores intersect at 0.65, suggesting that this value is the limit to identify distant similarities; therefore, an alignment with a score of ≈ 0.65 may be considered clearly random. Based on this information, a significant alignment threshold was established to analyze the most of the nrESS, with not compromising the statistical relevance. Therefore, a score of ≤0.3 was established as threshold. This value represents the higher dispersion (>195 SD) of the random data (Fig. 1d, asterisk) with the lowest loss of nrESS, i.e., more than 99 % of the real nrESS were included (Additional file 2: Figure S3). This threshold also corresponds to 0.26 % of all nrESS alignments (81,520 of 31,756,465) and includes 7907 out of 7970 nrESS. In contrast, from the alignments associated with the 10 random databases (31,756,465 for each dataset), only 0.04 ± 0.001 % (13,827 ± 308) of the total alignments exhibited a threshold of ≤0.3. These results show that our method can be used to identify similar nrESS with significant scores, excluding the possibility of finding such similar nrESS by random chance. Here, we report information concerning our comparisons of these nrESS related to metabolism in diverse bacterial organisms.

Pairwise alignments of nrESS identify a core of common metabolic pathways in *Gammaproteobacteria*

In order to evaluate the similarity of the metabolic maps in *Gammaproteobacteria* and whether there is a group

of *functionally conserved* pathways in these organisms, a set of similar nrESS was defined. In this context, the term *functionally conserved* refers to the identification of similar nrESS that may be common to *Gammaproteobacteria*. Two nrESS were considered as *functionally conserved* if their alignment had a score below the threshold (≤0.3) and, in conjunction, they were present in more than 75 % of the species analyzed. Based on this definition, the set included 1484 sequences from 74 different metabolic maps, with 69 % of the total alignments corresponding to alignments between the same metabolic maps (1805 of 2633), whereas 31 % corresponded to alignments between different metabolic maps (Fig. 2a).

To assess the nrESS similarity of each metabolic map as an indicator of functional conservation, we used the alignments that occurred within them (green edges in Fig. 2a), and we named this dataset the Metabolic Map Functional Conserved Dataset (MMFCD). The proportion of each metabolic type represented in this dataset is shown in Fig. 2b, and corresponds primarily to nrESS of the metabolism of nucleotides, followed by the metabolism of carbohydrates, cofactors and vitamins, amino acids, and lipids. In contrast, the pathways for xenobiotic biodegradation and metabolism, biosynthesis and other secondary metabolism, metabolism of other amino acids, and metabolism of terpenoids and poliketides, among others, represent less than 5 % of the total nrESS included in the dataset. From these alignments, we mapped the position of the highly similar nrESS in the corresponding metabolic map to determine the proportion of the *functionally conserved* EC numbers in relation to the total EC numbers present in *Gammaproteobacteria* (Fig. 2c). Using this information, we classified the metabolic maps into four groups: 1) maps with more than 70 % of the EC numbers identical, i.e. highly *functionally conserved*; 2) moderately *functionally conserved* maps, with percentages between 30 % and 69 %; 3) barely *functionally conserved*, i.e., those maps with percentages between 1 % and 29 %; finally, 4) variable maps, i.e., with 0 % EC classified as *functional conserved*. From these data, less than one-third of the analyzed maps (24 of 86) were classified as highly or moderately *functionally conserved*, while more than two-thirds were considered as barely *functionally conserved* or variable. All these data showed that more than half of the metabolic maps analyzed did not exhibit common nrESS in *Gammaproteobacteria* and, by consequence, may be considered variable, suggesting a high variability in the metabolism of this bacterial division.

In detail, maps classified as highly *functionally conserved* are related to important processes, like the pathways for fatty acid biosynthesis (map00061), metabolism of some amino acids (00290 for valine, leucine, and isoleucine biosynthesis; 00300 for lysine biosynthesis), components of the cell wall (00540 for lipopolysaccharide biosynthesis; 00550 for peptidoglycan biosynthesis), metabolism of some cofactors (00770 for pantothenate and CoA biosynthesis; 00780 for biotin metabolism; 00785 for lipoic acid metabolism), and novobiocin biosynthesis (00401). These functional similarity also correlate with the fact that amino acid metabolism pathways for valine, leucine, isoleucine, and lysine have been identified as pathways with diverse duplicated genes in the three cellular domains of life [10, 23].

The second group includes those maps defined as moderately *functionally conserved*. In this category were included the pathways for metabolism of purines (00230) and pyrimidines (00240), glycolysis/gluconeogenesis (00010), the citrate cycle (00020), metabolism of glycerophospholipids (00564), terpenoids backbone (00900), and some cofactors, like riboflavin (00740), nicotinamide (00760), folate (00790), and thiamine (00730). It is interesting that the central part of glycolysis (00010), the Embden-Meyerhof pathway, is partially conserved among *Gammaproteobacteria* (Fig. 2d), whereas the core pathway that comprises the tricarbon compounds is widely *functionally conserved* among the analyzed organisms, including the oxidation of pyruvate to acetyl CoA. In the hexose section, the enzymatic steps catalyzed by 6-phosphofructokinase (EC2.7.1.11) and fructose biphosphate aldolase (EC4.1.2.13) are considered variable. A similar result was observed with the glycolysis input, where the mechanisms by which the hexoses enter the pathway are variable. In addition, the enzymatic steps to transform pyruvate to lactate and the ethanolic fermentation from acetate are also variable. In a similar way, gluconeogenesis from oxaloacetate is partially *functionally conserved* in *Gammaproteobacteria*, where the enzymes allowing the input from the oxaloacetate (phosphoenol pyruvate carboxykinase, EC4.1.1.49 and 4.1.1.32) and the enzyme that dephosphorylates fructose 1,6-bisphosphate to fructose 6-phosphate (fructose biphosphatase, EC3.1.3.11) are considered variable. These results are congruent with those from a previous study, where it was concluded that glycolysis is a plastic pathway and that the lower part of the glycolysis pathway is the more conserved section among the three cellular domains [24].

Another example is the case of purine (00230) and pyrimidine (00240) metabolism. In general, both metabolic maps show *functionally conserved* reactions that converge the synthesis of mono-, di-, and triphosphate ribonucleotides and deoxyribonucleotides. In the case of purine metabolism (Fig. 3), the biosynthetic pathway for the main precursor to the synthesis of purine nucleotides [25], inosine monophosphate (IMP) is completely conserved in *Gammaproteobacteria* (KEGG module

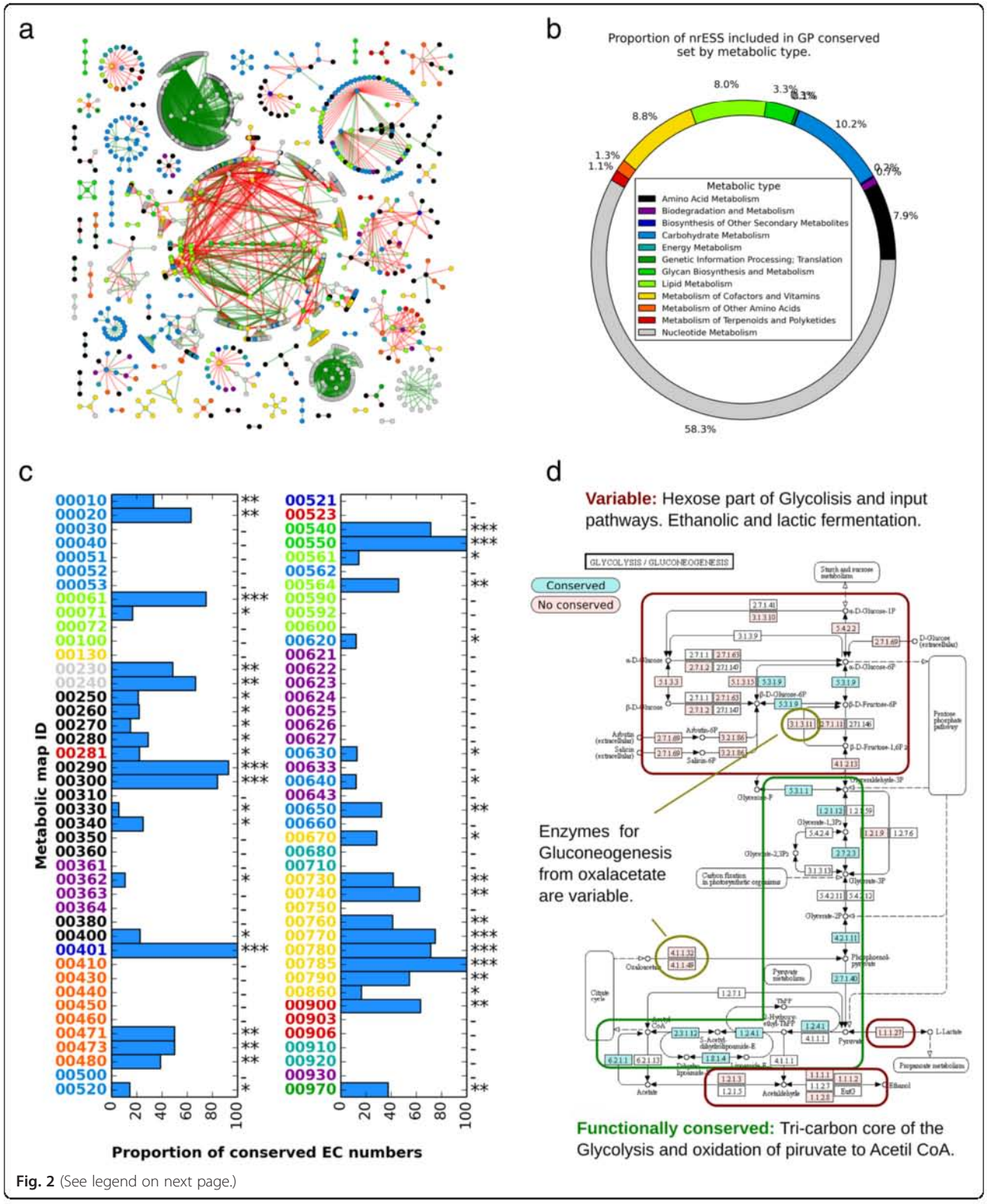

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Functional conserved and variable nrESS in *Gammaproteobacteria*. **a** Graph representation of the relationships between the set of conserved ESS in *Gammaproteobacteria*. The nodes represent the nrESS and the edges show the alignments among them. Green edges are alignments between the same metabolic map, and the red ones represent alignments between different metabolic maps. The nodes are colored according to the metabolism type, as indicated. The alignments between ESS from the same map were selected as the Metabolic Map Functional Conserved Dataset (MMFCD). **b** Proportion of ESS from each metabolic type included in the MMFCD. **c** Proportion of EC numbers conserved in the alignments of the MMFCD for each metabolic map. The metabolic maps are represented by the KEGG map ID, and the colors indicate the metabolic type. The metabolic maps were classified according to the proportion of functional conserved EC numbers, as follows: conserved (***), more than 70 % of EC numbers conserved; moderately conserved (**), between 30 and 69 % of numbers conserved; barely conserved (*), between 1 and 30 % of numbers conserved; non conserved (-), 0 % of numbers conserved. **d** Functional conservation of the glycolysis/gluconeogenesis metabolic map in *Gammaproteobacteria*. The map represents the functional mapping of conserved nrESS (3 levels of EC classification) in the alignments of the MMFCD (cyan). In addition, the nrESS present in at least one of the species analyzed but not in the conserved set are shown (pink). The white steps represent enzymes not present in any of the species analyzed. In boxes and circles are represented the conserved and variable regions in the metabolic map. See text for details

M00048). Congruently, the classical synthesis of ATP (module M00049) and GTP (module M00050) from IMP is also *functionally* conserved, although there are many variable enzymatic steps that catalyze the production of nucleosides and nitrogenous bases. Finally, the pathways for degradation of purines via xanthine and allantoate utilization are variable in *Gammaproteobacteria*. Therefore, in purine metabolism we observed a general *functional conservation* of synthetic pathways and a general *non-functional conservation* of degradation pathways. These results, in conjunction with recent data suggesting that nucleotide metabolism is highly conserved across all the organisms [26], reinforce the notion that purine biosynthesis is one of the more ancient metabolic pathways [1, 27].

A similar conservation pattern is observed in other metabolic maps classified as moderately *functionally conserved*, such as the pathway for glycerophospholipid metabolism (00564). We found that the biosynthetic pathways to CDP-diacylglycerol and then to phosphatidyl glycerol, phosphatidyl serine, and phosphatidyl ethanolamine are conserved, while the biosynthetic pathway to phosphatidyl choline and the degradation pathways are variable. A similar result arises for the biosynthesis

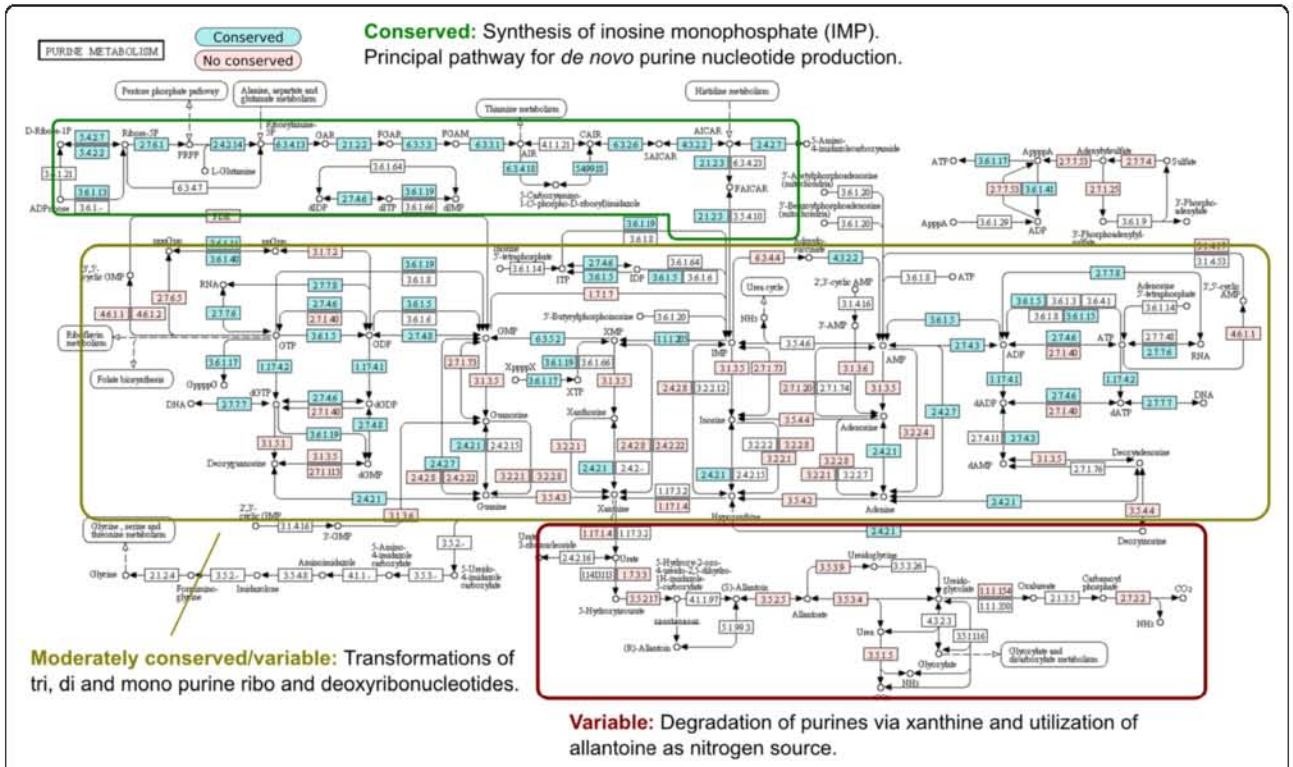

Fig. 3 Functional conservation of purine metabolism in *Gammaproteobacteria*. The map shows the mapping of functional conserved enzymatic steps (three levels of EC classification) in the alignments of the metabolic map conserved dataset (cyan). The enzymatic steps present in at least one of the species analyzed but not in the conserved set are shown (pink). The white steps represent enzymes do not present in *Gammaproteobacteria*. In boxes are represented the conserved and variable regions in the metabolic map

of cofactors like thiamine-diphosphate (map 00730), riboflavin (map 00740), NAD^+ and $NADP^+$ (map 00760), and tetrahydrofolate (map 00790). In conjunction, it is possible to deduce a *functional conservation* pattern for *Gammaproteobacteria*, where some metabolic maps contain a biosynthesis-related core of similar enzymatic steps, and some variable steps that include the degradation of various compounds. These variable or dispensable steps may represent possible alternative pathways in different organisms and/or in different ecological niches, as has been previously suggested [10, 28].

The group of metabolic maps classified as barely *functionally conserved* includes important processes, such as amino acid metabolism, fatty acid degradation (beta-oxidation), and glycerolipid metabolism. In this context, we identified many variable reactions in the map that describes alanine, aspartate, and glutamate metabolism (map 00250), suggesting the existence of alternative pathways to produce these compounds. In this regard, there are three possible enzymes that catalyze the conversion of L-glutamate to L-glutamine: one of them by a ligation reaction (glutamine synthetase, EC6.3.1.2) and two by reversible hydrolysis (glutaminase, EC3.5.1.2, and L-glutamine (L-asparagine) amidohydrolase, EC3.5.1.38). In particular, the L-glutamine (L-asparagine) amidohydrolase also catalyzes the deamination of asparagine to aspartate. This finding suggests more flexible networks for the production of amino acids and reinforces the notion of various alternative enzymes for the production of amino acids [10]. A similar observation arises for cysteine and methionine metabolism (map 00270), for which alternative pathways were also identified. For example, the pathway to produce methionine from aspartate (module M00017) is not completely conserved in *Gammaproteobacteria*; nevertheless, there are some alternative enzymes that may work as alternative paths for the synthesis of methionine. Interestingly, some of these alternative enzymes were identified as functionally conserved in this work, suggesting not only the absence of a conserved canonical route but also important alternative enzymatic steps.

Finally, the *variable* maps include a high diversity of metabolisms types. Some of them contain few or fragmented enzymatic steps present in at least one *Gammaproteobacteria* species, suggesting the absence of those metabolic maps in this clade. However, other maps contain many enzymes present in *Gammaproteobacteria*; such as those for seleno compound metabolism, galactose metabolism, pentose phosphate and pentose metabolism, and glucuronate metabolism, among others. In general, the maps classified in this category represent pathways for degradation of uncommon compounds (xenobiotics) or for alternative carbon sources (carbohydrate metabolism). Altogether, these observations in addition to supporting the previously proposed idea concerning the reduced conservation of degradation related pathways; reinforce the notion of differential enzyme recruitment across the clade. Also, our results support the proposed preponderance of central carbon and anabolic pathways in the evolution of metabolism [2, 8, 27, 29].

In summary, all these data allow the identification of a *core* of similar enzymatic steps in *Gammaproteobacteria*. This *core* includes primarily reactions of the central carbon metabolism (low part of glycolysis and tricarboxylic acid cycle), and the biosynthetic pathways for nucleotides, cofactors and some amino acids. In addition this *core* is complemented with a set of variable pathways that primarily includes degradation pathways for carbohydrates, amino acids and xenobiotics that may be essential to the particular life style of each organism.

The complete set of functional conservation of metabolic maps in *Gammaproteobacteria* is available as KEGG weblinks in Additional file 3: File S2.

Metabolic maps that convert similar compounds also share similar nrESS

In this section, we asked whether the similarities between nrESS might help to identify those metabolic maps that convert similar compounds and uncover explicitly the functional relations between metabolic maps. In this regard, we explored the general similarities between metabolic maps identified by nrESS comparisons. To do so, the total number of shared alignments between each pair of metabolic maps was calculated, considering only those alignments with scores of ≤0.3. The counts of the shared alignments were normalized to the total number of alignments for each map and used as similarity vectors in a hierarchical clustering analysis with the Spearman rank correlation as similarity measure (Fig. 4). Considering a cutoff of 0.46 of the total length of the clustering tree, we defined a total of 24 different metabolic map clusters. Similar results were obtained using Kendal rank correlation and self-organizing maps. This analysis showed 5 clusters that included more than 3 metabolic maps, 5 clusters that included 2 or 3 maps, and 14 maps that were considered singletons. The first of the major clusters included metabolic maps related to the degradation of aromatic compounds, such as amino benzoate, bisphenol, toluene, and naphthalene, among others. The second major cluster included hydrophobic amino acid metabolism (e.g., for tryptophan, phenylalanine, and tyrosine) and aromatic compound degradation (e.g., of xylene and dioxins). The third cluster contained carboxylic acid (fatty acids and butanoate), long aliphatic chain lipids (e.g., limonene and geraniol), valine, leucine, and isoleucine degradation, and pyruvate and the TCA cycle maps. This result suggests a functional similarity between pyruvate

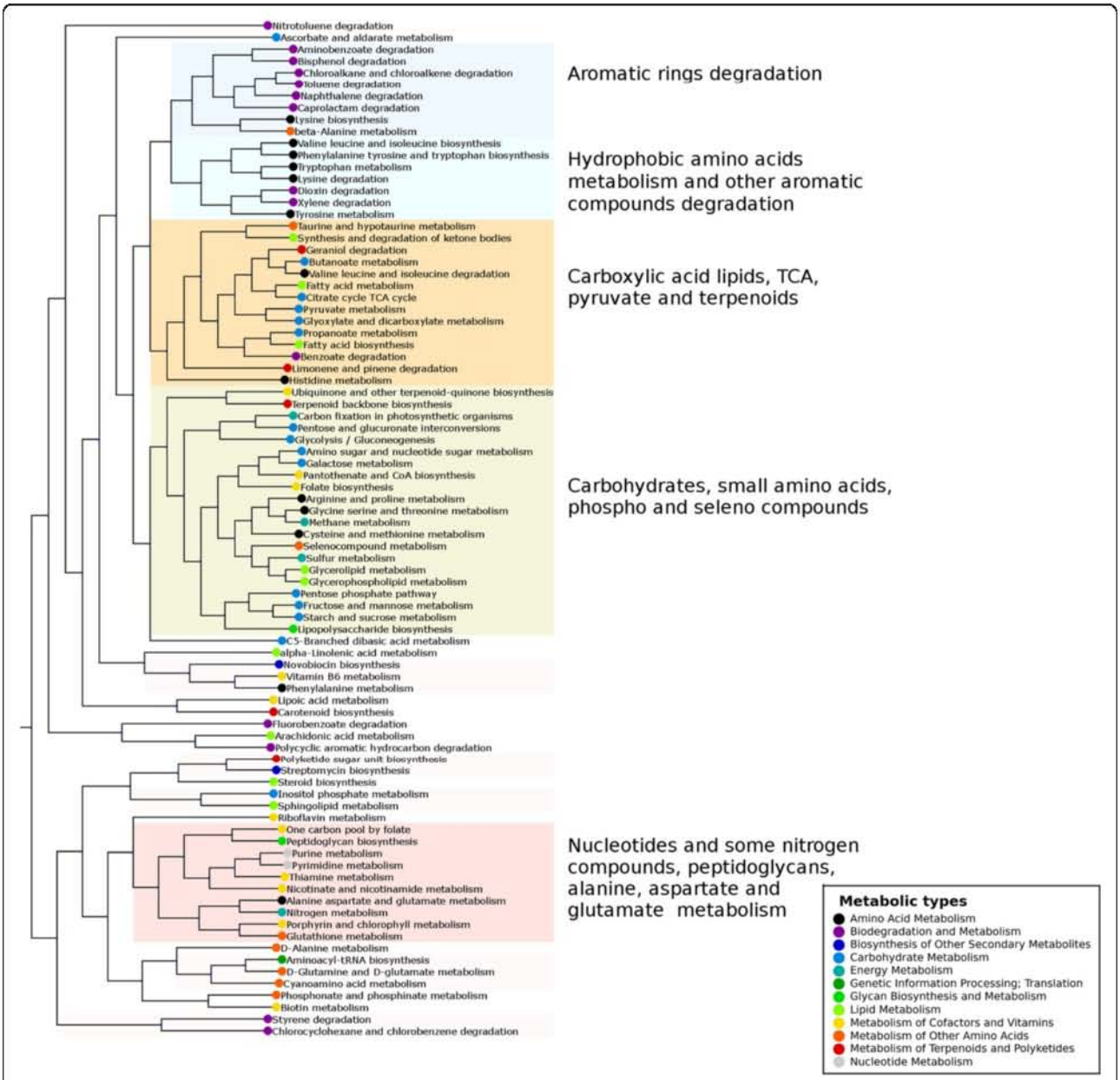

Fig. 4 The metabolic map that includes similar compounds exhibits similar ESS. The hierarchical clustering was constructed by comparing the normalized counts of shared pairwise nrESS alignments, using the Spearman rank correlation and average linkage method. A cutoff of 0.46 of the total length of the clustering was used to define the number of clusters. The major and minor (pale pink) clusters are shown, and their corresponding KEGG metabolic types are indicated

and the TCA cycle and the synthesis of aliphatic chain hydrophobic carboxylic acids. The fourth cluster includes the metabolism of carbohydrates, carbon fixation, metabolism of some cofactors (CoA and folate), terpenoids, glycerolipids, and sulfur (including methionine and cysteine) and seleno compounds. Finally, the fifth cluster contains the metabolism of nucleotides, peptidoglycans, nitrogen metabolism, and other nitrogen-containing cofactors, like hiamine and nicotinamide. In summary, we identified a trend where metabolic maps describing the transformations of chemically similar molecules also contained similar nrESS, probably as consequence of enzymatic recruitment.

Similar nrESS suggest that enzyme recruitment is a frequent event in metabolism of *Gammaproteobacteria*

Based on the previous sections, we ask if *functionally conserved* pathways can be used to identify the possible recruitment patterns in the metabolism of *Gammaproteobacteria*. In this context, the corresponding nrESS of the lower part of the glycolysis pathway and the IMP

pathway for *de novo* synthesis of purines were used to scan the complete nrESS dataset. Both pathways are considered ancient [24, 30]. We used the NW algorithm with a score threshold of 0.3. To determine the significance of the alignments we also scanned the random nrESS. From these comparisons, we found that the numbers of significant matches for the lower part of the glycolysis pathway (Fig. 5a) and the IMP pathway (Fig. 5b) were greater than those expected by chance; however, the difference relative to the random databases was greater for IMP than for glycolysis. In addition, we determined that although the raw number of hits was greater for IMP (381) than for glycolysis (148), the number of alignments with other maps was greater for glycolysis than for IMP (37 versus 19 different metabolic maps). The relatively small difference in the number of significant matches obtained for glycolysis compared to the random databases may be explained in part by the number of occurrences of its constituent EC numbers in the nrESS database and in the KEGG database (Fig. 5c). The EC numbers 1.2.1 (oxidoreductases acting on the aldehyde or oxo group of donors and with NAD^+ or $NADP^+$ as acceptor), 2.7.1 (phosphotransferases with an alcohol group as acceptor), and 4.2.1 (carbon-oxygen hydrolyases) are within the top 10 in nrESS database abundance. This result shows a broad similarity of the lower part of the glycolysis pathway with many other metabolic processes and suggests similar catalytic processes are used to transform some compounds in different metabolic maps. In turn, this observation suggests an outstanding proportion of enzyme recruitment events from glycolysis to other metabolic pathways and may reflect the utilization of similar products generated for similar reactions in different metabolic maps. On the other hand, the major number of hits for the IMP pathway corresponds to alignments within the same map, suggesting that this pathway has increased its size by duplication and recruitment of its own enzymes. Although both metabolic pathways may be considered ancient and were classified as functional conserved in this study, the patterns of functional similarities are different and may reflect the constraints of enzyme recruitment and the ubiquity of some types of compounds.

Discussion and conclusions

In this work we used a simple workflow for the comparative study of metabolism through the alignment of linear sequences of ESS. The metabolic maps stored in KEGG were transformed into linear ESS by using an exhaustive and well-defined graph search algorithm. Then, the ESS were compared to identify the commonalities and differences between them. This approach allows the identification of similarities at the Enzymatic Step Sequences (ESS) level in a set of metabolic pathways. In this regard, the use of the functional information of the enzyme activity rather than the (protein and DNA) sequence information suggest that metabolism comprises a complex and dynamic network that may have different proteins to achieve the same or similar function.

Diverse methods for the alignment of biological networks have been suggested, such as protein-protein interaction networks (for some examples see references [31–33]) and metabolic networks (for some examples see references [34–38]), mainly based on protein homology and/or network topology. However they consider a small number of organisms or general metabolic maps

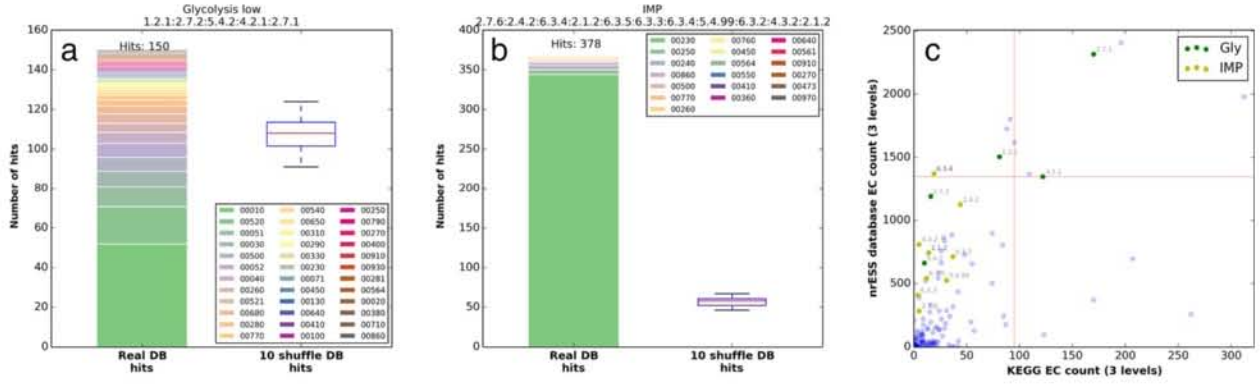

Fig. 5 The similarities between nrESS suggest enzyme recruitment. Two nrESS representing the lower part of the glycolysis pathway (**a**) and the Inosine Monophosphate (IMP) pathway (**b**) were compared against the nrESS database (bar) and 10 random databases (boxplot). The number of significant hits (score of ≤0.3) was plotted. The color code indicates the proportion of hits corresponding to each metabolic map. **c** The relation between the number of EC numbers sharing the first 3 levels of classification in the KEGG database versus the number of occurrences of each EC number in the nrESS database

of KEGG database. Also, many of these methods are in general difficult to compare with each other, as has been recently shown by Clack et al. [39].

In this work we used the alignment of linear enzymatic step sequences, similar to the previously described approaches [20, 40], where a general strategy for the systematic analysis of the metabolism in a multigenome scale was additionally implemented. The linear enzymatic alignment approach described here allows gaps using the NW algorithm, uses a random data comparison, and allows the identification of distant similarities like those observed between metabolic maps. To our knowledge, this is first time that these methods are used to compare systematically the metabolism of a well-studied and metabolic diverse clade. Therefore, our approach is able to capture directly the information contained in the individual metabolic networks of each organism.

Based on these comparisons, we detected a *core* of metabolic pathways associated with central carbohydrate metabolism, lipid, cell wall, and cofactors, and biosynthetic pathways. In contrast, variable maps are those associated with degradation pathways, except the glucose-related pathways and the TCA cycle. In addition, amino acid metabolism is an example of a pathway with multiple routes to yield similar compounds from different routes, characterized for alternative pathways.

In addition, two scenarios can be suggested to exemplify the growth of the metabolism. The first one, associated to the glycolysis, where a significant proportion of functional similarities from this pathway were observed in other metabolic pathways; suggesting the utilization of similar substrates/products processed by similar reactions in different metabolic maps. The second scenario is associated to the high number of hits for the IMP pathway associated to alignments within the same map, suggesting that this pathway possibly has increased its size by duplication and recruitment of its own enzymes and arising the possibility of major biochemical coupling restrictions for the recruitment of the enzymes in the IMP pathway. Therefore, the different patterns of ESS similarities of two ancient pathways suggest that the recruitment of catalytic activities in the metabolism is restricted by the metabolic context, being not a random phenomenon. Albeit our data suggest functional and, probably, evolutionary conservation of diverse catalytic steps, additional information must be considered to have a better approximation of metabolism evolution, such as gene transfer and gene loss, among other processes. For this reason, we do not exclude the possibility of diverse genetic phenomena, such as the continuum interchange of genetic material that diminishing the border between bacterial species, as it has been recently described in *E. coli* bacterial strains, where a small proportion of universal protein families [41] and a large proportion of specific families [42] have been found. In this regard, the functional conservation of metabolic steps was evaluated in a representative group of species selected with a genome similarity score of 0.7, as described by [43], capturing the general diversity of the *Gammaproteobacteria* metabolism.

Therefore, the method described here is able to identify alternative enzymes involved in similar metabolic processes, and although the conclusions can be restricted to the metabolism covered by *Gammaproteobacteria*, the method can be extended to any organism or clade for which there is metabolic information. Finally, we understand that the approach described here does not consider the effect of promiscuous enzymes, defined as those enzymes with more than two different E.C. numbers. However, previous analysis have described that around 10 % of the total enzymatic repertoire in bacterial and archaeal organisms corresponds to promiscuous enzymes [28, 44], suggesting that our results and conclusions are enough robust and can be little influenced by the multifunctional enzymes.

Methods

Selection of proteobacterial species

In this study we included the small-molecule metabolism of 40 different *Gammaproteobacteria* species. These organisms were selected from the 275 *Gammaproteobacteria* genomes included in the KEGG database as of June 2011 [3]. We chose non redundant genomes using the criteria described in reference [43], with a genome similarity score of 0.7, resulting in a set of 40 non redundant *Gammaproteobacteria* species. These organisms are representative of the division as it is shown in Additional file 2: Figure S4. Additional file 4: Table S2 contains the list of the organisms included in the analysis.

Construction of ESS

We downloaded the KGML files (version 0.71) that describe the metabolic maps (pathways) of 40 *Gammaproteobacteria* in June 2011 from the KEGG database. Based on these metabolic maps, the ESS were constructed by using the BFS algorithm. In brief, a directed graphical representation of each metabolic map was created in which the nodes represented enzymes and the edges represented a shared substrate/product between two enzymes. This representation takes into account the reversibility of the reactions. In a posterior step, a group of BFS trees was generated for each metabolic map from a set of initialization nodes, used as roots. An initialization node was defined by two criteria: a node which substrate is not catalyzed by another enzyme in the metabolic map, and a node which substrate comes from another

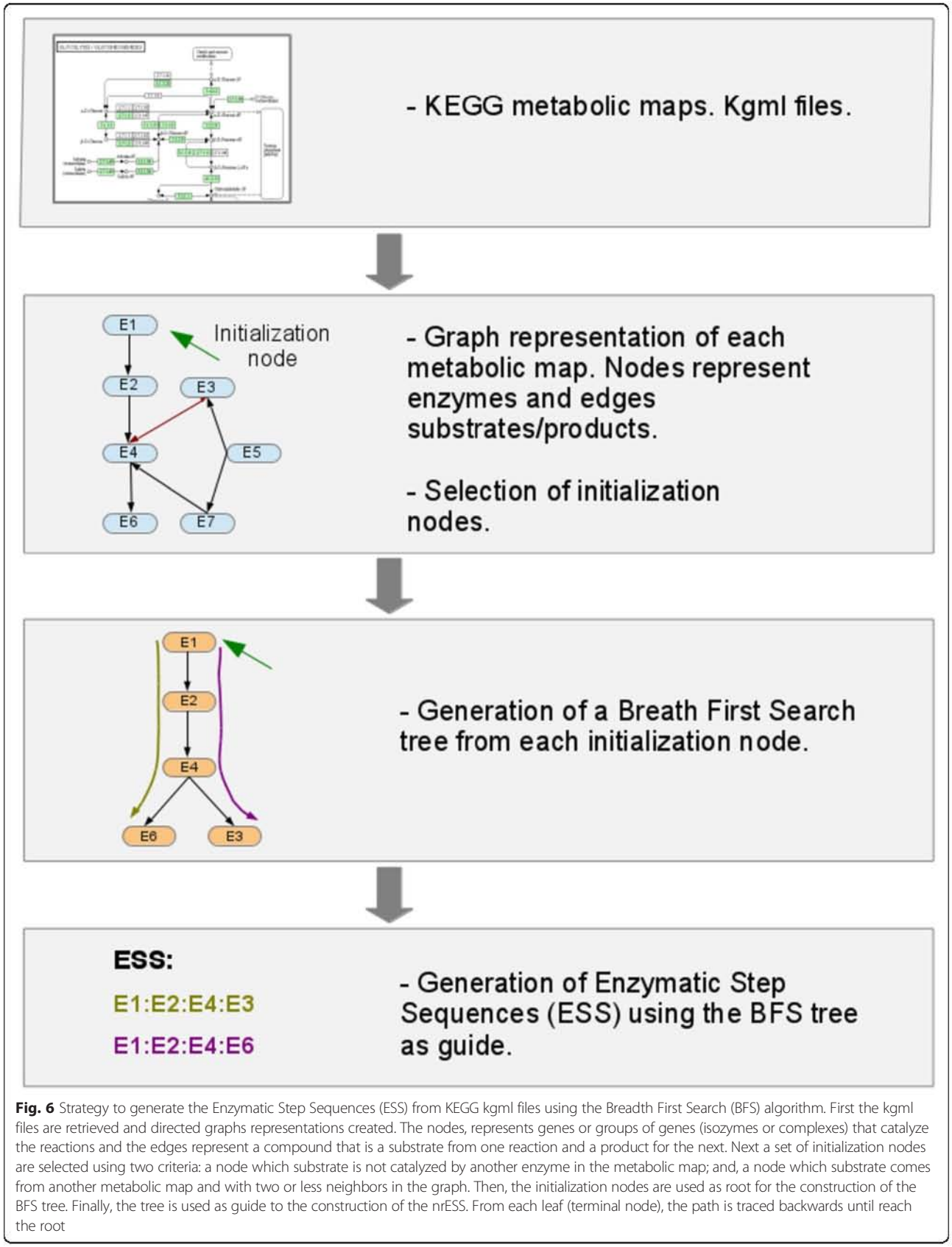

Fig. 6 Strategy to generate the Enzymatic Step Sequences (ESS) from KEGG kgml files using the Breadth First Search (BFS) algorithm. First the kgml files are retrieved and directed graphs representations created. The nodes, represents genes or groups of genes (isozymes or complexes) that catalyze the reactions and the edges represent a compound that is a substrate from one reaction and a product for the next. Next a set of initialization nodes are selected using two criteria: a node which substrate is not catalyzed by another enzyme in the metabolic map; and, a node which substrate comes from another metabolic map and with two or less neighbors in the graph. Then, the initialization nodes are used as root for the construction of the BFS tree. Finally, the tree is used as guide to the construction of the nrESS. From each leaf (terminal node), the path is traced backwards until reach the root

metabolic map and with two or fewer neighbors in the graph. These criteria represent the metabolic input for each pathway; the first criterion considers the substrates not created in the same pathway, and the second one considers the connections with other pathways. Each initialization node was used as a root for the construction of a BFS tree. Finally, each tree was used as a guide for the construction of the ESS. Thus, a BFS tree creates as many ESS as the number of branches it has. The graph representation of the metabolic maps and the BFS trees were generated using the Networkx Python module [45] (Fig. 6).

In a posterior stage, ESS were organized in a relational database. In this database, each EC number contained in a sequence was related to its corresponding protein(s), species, metabolic map. This database has a high degree of redundancy, because an ESS may be the same in different species. Thus, a nrESS dataset was constructed by filtering identical ESS and leaving only one representative. Each ESS in the nrESS dataset is linked to the original ESS. All analyses were conducted using the nrESS dataset and referring to the original data when necessary. The ESS and nrESS data are provided as supplementary material (Additional file 5: File S1).

Comparison of nrESS by pairwise alignments

In order to identify the similarity of the nrESS, we implemented a pairwise alignment algorithm based on the Dynamic Programing Needleman and Wunsch (NW) algorithm as described in reference [46]. This algorithm works in a similar way as the classic tools to align nucleotide or amino acid sequences (Additional file 6: Text S1). We used an EC number weight matrix derived from an entropy-based evaluation function that evaluated the similarities between EC numbers. The weight matrix describes the similarity between the 136 different three levels EC numbers. The number 9.9.9 was used to describe an enzyme with no EC assigned and that was similar only to itself. The similarity between two EC numbers ranged from 0 to 1. Values close to 0 indicate similar EC numbers, and values close to 1 indicate different EC numbers. This matrix takes into account the hierarchy of the EC numbers, giving a value of 1 to all the EC pairs that are different in the first level of classification regardless of whether the second or third numbers are identical. Therefore, the NW algorithm uses the matrix to construct an alignment that minimizes the global score. Finally, the alignment is evaluated by using the normalized entropy-based function. The *score* obtained with such an evaluation function also ranges from 0 to 1, where 0 indicates similar nrESS and 1 dissimilar nrESS. To analyze in more detail the similarities of the low part of the glycolysis and the IMP pathways, their ESS were compared against the nrESS. Examples of nrESS alignments are shown in Additional file 2: Figure S5.

Statistically significant ESS alignments

To determine the statistical significance of the nrESS alignments, we compared the alignment scores of the real database against the scores from10 different random databases. These random sequences were constructed by shuffling the EC number content of the entire database, maintaining the nrESS length and EC composition of the original sequences. Each random database was submitted to the same all-versus-all alignment approach used for the real data, and the distribution of alignment scores considered the mean ± SD. The threshold considered statistically significant corresponded to a score of ≤0.3, i.e., that point with higher dispersion of the real data relative to the mean random databases scores and where the loss of nrESS due to extreme dissimilarity was less than 1 %, i.e. this threshold includes the 99 % of the nrESS.

Functional conservation of enzymatic steps in metabolic maps

We used the information provided by the nrESS pairwise alignments to identify the *functionally conserved* enzymatic steps in *Gammaproteobacteria* for each metabolic map. Two nrESS were considered conserved if their alignment scores were below or equal to 0.3 and if, in conjunction, they were present in more than 75 % of the organisms. This criterion was employed because we assumed that a pair of conserved ESS would be shared by at least all of the species with genomes greater than 2000 ORFs, i.e., 30 of the 40 *Gammaproteobacteria* organisms. From the ESS that fulfilled this criterion, we selected those that corresponded to the same metabolic map. This subset of sequences was named the Metabolic Map Functional Conserved Dataset (MMFCD). To identify the conserved ESS, the aligned identical EC numbers from each alignment were mapped in the corresponding position in KEGG metabolic maps.

Clustering of similar metabolic maps

In order to identify the functional similarities among metabolic maps, we selected a subset of nrESS pairwise alignments with score values of ≤0.3. These alignments were used to construct a similarity matrix where each cell corresponded to the count of the alignments shared by each pair of metabolic maps. The rows representing the metabolic maps were normalized by the total alignments in each row. The matrix was used as input to a hierarchical clustering analysis with the program MeV4 (http://www.tm4.org/mev.html). The similarity between maps was calculated with the Spearman's rank correlation, and elements were clustered with the average

method. A cutoff of 0.46 of the total length of the dendogram was used to classify the metabolic maps into groups and is displayed with the E.T.E. 2 Python toolkit [47].

Additional files

Additional file 1: Table S1. Statistics of construction of ESS per metabolic map. Number of reactions, genes, and ESS produced for each metabolic map per organism. (TAB 230 kb)

Additional file 2: Figure S1. Number of ESS generated by metabolic map. In X-axis, denotes the number of enzymes per metabolic map; Y-axis corresponds to the number of ESS. Each point represents one metabolic map. The data were adjusted to a linear model. **Figure S2.** Alignment scores of the nrESS database comparisons. The distribution resembles an extreme value *Gumbel* distribution skewed to the right. Scores are close to 1 and represent alignments of dissimilar sequences. In counterpart, scores close to 0 correspond to alignments between similar sequences. **Figure S3.** Proportion of nrESS included in at least one pairwise alignment as function of the alignment score. X-axis represents the threshold at different values, whereas the Y-axis shows the proportion of nrESS included. The number of excluded nrESS decreases as the alignment score increases. The proportion of included nrESS of the real data is compared with the same proportion of 10 random databases. In the last case, the proportion of included ESS decreases abruptly at scores close to 0. **Figure S4.** *Gammaproteobacteria* coverage of the organisms used in this study. The organisms used in this study are marked as circles. The organisms listed in the upper right corner are not represented in the phylogenetic tree. The tree was taken and modified from [14]. **Figure S5.** Pairwise ESS alignments. A NW algorithm was used in these examples. The aligned pairs of ESS and their corresponding scores are indicated. Gaps in the alignment are indicated by dashes (-.-.-). Significant scores are those with scores ≤ 0.3. (PDF 4362 kb)

Additional file 3: File S2. KEGG weblinks with the functional conservation of metabolic maps in *Gammaproteobacteria. (TXT 56 kb)*

Additional file 4: Table S2. Organisms considered in this study. Genomic and metabolic information associated to each organism analyzed in this work. (DOCX 31 kb)

Additional file 5: File S1. Enzymatic Step Sequence database. (ZIP 1300 kb)

Additional file 6: Text S1. Description of the NW ESS alignment functions. (DOC 72 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
CP-H retrieved all metabolic maps, performed the all sequence analysis and interpreted the results. KR-V interpreted the results; and EP-R interpreted the results. All authors read and approved the final manuscript.

Acknowledgements
We thank to Georgina Hernandez-Montes and Anny Rodriguez for their critical reading of the manuscript and to Rosa María Gutierrez-Rios for her constructive opinions and comentaries. CAP-H acknowledges the support by a PhD fellowship (209805) from CONACyT-México. We also thank the anonymous reviewers for their comments, which help us to improve the manuscript. Support from DGAPA-UNAM (IN-209511), PAPIIT-UNAM (IN-107214) and CONACYT (155116) is gratefully acknowledged.

Author details
[1]Departamento de Microbiología Molecular, Instituto de Biotecnología, UNAM, Av. Universidad 2001, Cuernavaca, Morelos CP 62210, México. [2]Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM, Ciudad Universitaria CP 04510México D.F., México.

Received: 3 February 2015 Accepted: 19 October 2015
Published online: 17 November 2015

References
1. Caetano-Anollés G, Kim HS, Mittenthal JE. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. Proc Natl Acad Sci U S A. 2007;104:9358–63.
2. Braakman R, Smith E. The emergence and early evolution of biological carbon-fixation. PLoS Comput Biol. 2012;8:e1002455.
3. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(Database issue):D355–60.
4. Caspi R, Foerster H, Fulcher C a, Hopkinson R, Ingraham J, Kaipa P, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res. 2006;34(Database issue):D511–6.
5. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási a L. The large-scale organization of metabolic networks. Nature. 2000;407:651–4.
6. Ravasz E, Somera a L, Mongru D a, Oltvai ZN, Barabási a L. Hierarchical organization of modularity in metabolic networks. Science. 2002;297:1551–5.
7. Arita M. The metabolic world of Escherichia coli is not small. Proc Natl Acad Sci U S A. 2004;101:1543–7.
8. Von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, et al. Genome evolution reveals biochemical networks and functional modules. Proc Natl Acad Sci U S A. 2003;100:15428–33.
9. Spirin V, Gelfand MS, Mironov A a, Mirny L a. A metabolic network in the evolutionary context: multiscale structure and modularity. Proc Natl Acad Sci U S A. 2006;103:8774–9.
10. Hernández-Montes G, Díaz-Mejía JJ, Pérez-Rueda E, Segovia L. The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. Genome Biol. 2008;9:R95.
11. Díaz-Mejía JJ, Pérez-Rueda E, Segovia L. A network perspective on the evolution of metabolism by gene duplication. Genome Biol. 2007;8:R26.
12. Light S, Kraulis P. Network analysis of metabolic enzyme evolution in Escherichia coli. BMC Bioinformatics. 2004;5:15.
13. Rison SCG, Thornton JM. Pathway evolution, structurally speaking. Curr Opin Struct Biol. 2002;12:374–82.
14. Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, et al. Phylogeny of gammaproteobacteria. J Bacteriol. 2010;192:2305–14.
15. Newton ILG, Woyke T, Auchtung TA, Dilly GF, Dutton RJ, Fisher MC, et al. The Calyptogena magnifica chemoautotrophic symbiont genome. Science. 2007;315:998–1000.
16. Hirschler-Rea A. Isolation and characterization of spirilloid purple phototrophic bacteria forming red layers in microbial mats of Mediterranean salterns: description of Halorhodospira neutriphila sp. nov. and emendation of the genus Halorhodospira. Int J Syst Evol Microbiol. 2003;53:153–63.
17. Hoeft SE, Blum JS, Stolz JF, Tabita FR, Witte B, King GM, et al. Alkalilimnicola ehrlichii sp. nov., a novel, arsenite-oxidizing haloalkaliphilic gammaproteobacterium capable of chemoautotrophic or heterotrophic growth with nitrate or oxygen as the electron acceptor. Int J Syst Evol Microbiol. 2007;57(Pt 3):504–12.
18. Hara A, Syutsubo K, Harayama S. Alcanivorax which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation. Environ Microbiol. 2003;5:746–53.
19. Chen M, Hofestaedt R. An algorithm for linear metabolic pathway alignment. In Silico Biol. 2005;5:111–28.
20. Chen M, Hofestädt R. PathAligner: metabolic pathway retrieval and alignment. Appl Bioinformatics. 2004;3:241–52.
21. Chou C-H, Chang W-C, Chiu C-M, Huang C-C, Huang H-D. FMM: a web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res. 2009;37(Web Server issue):W129–34.
22. Klein CC, Cottret L, Kielbassa J, Charles H, Gautier C, Ribeiro de Vasconcelos AT, et al. Exploration of the core metabolism of symbiotic bacteria. BMC Genomics. 2012;13:438.
23. Cunchillos C, Lecointre G. Integrating the universal metabolism into a phylogenetic analysis. Mol Biol Evol. 2005;22:1–11.
24. Dandekar T, Schuster S, Snel B, Huynen M, Bork P. Pathway alignment: application to the comparative analysis of glycolytic enzymes. Biochem J. 1999;343(Pt 1):115.
25. Zhang Y, Morar M, Ealick SE. Structural biology of the purine biosynthetic pathway. Cell Mol Life Sci. 2008;65:3699–724.

26. Armenta-Medina D, Segovia L, Perez-Rueda E. Comparative genomics of nucleotide metabolism: a tour to the past of the three cellular domains of life. BMC Genomics. 2014;15:800.
27. Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE. The origin and evolution of modern metabolism. Int J Biochem Cell Biol. 2009;41:285–97.
28. Martínez-Núñez MA, Poot-Hernandez AC, Rodríguez-Vázquez K, Perez-Rueda E. Increments and duplication events of enzymes and transcription factors influence metabolic and regulatory diversity in prokaryotes. PLoS One. 2013;8:e69707.
29. Braakman R, Smith E. The compositional and evolutionary logic of metabolism. Phys Biol. 2013;10:011001.
30. Becerra A, Lazcano A. The role of gene duplication in the evolution of purine nucleotide salvage pathways. Orig Life Evol Biosph. 1998;28:539–53.
31. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A. 2003;100:11394–9.
32. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res. 2004;32(Web Server issue):W83–8.
33. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, et al. Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci U S A. 2005;102:1974–9.
34. Ogata H, Fujibuchi W, Goto S, Kanehisa M. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res. 2000;28:4021–8.
35. Pinter RY, Rokhlenko O, Yeger-Lotem E, Ziv-Ukelson M. Alignment of metabolic pathways. Bioinformatics. 2005;21:3401–8.
36. Alberich R, Llabrés M, Sánchez D, Simeoni M, Tuduri M. MP-Align: alignment of metabolic pathways. BMC Syst Biol. 2014;8:58.
37. Ay F, Kahveci T, DE Crécy-Lagard V. A fast and accurate algorithm for comparative analysis of metabolic pathways. J Bioinform Comput Biol. 2009;7:389–428.
38. Wernicke S, Rasche F. Simple and fast alignment of metabolic pathways by exploiting local diversity. Bioinformatics. 2007;23:1978–85.
39. Clark C, Kalita J. A comparison of algorithms for the pairwise alignment of biological networks. Bioinformatics. 2014;30:2351–9.
40. Tohsato Y, Matsuda H, Hashimoto A. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. Proc Int Conf Intell Syst Mol Biol. 2000;8:376–83.
41. Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proc Natl Acad Sci. 2015;112:10139-46.
42. Lukjancenko O, Wassenaar TM, Ussery DW. Comparison of 61 sequenced escherichia coli genomes. Microb Ecol. 2010;60:708–20.
43. Moreno-Hagelsieb G, Janga SC. Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins Struct Funct Bioinforma. 2008;70:344–52.
44. Martínez-Núñez MA, Rodríguez-Vázquez K, Pérez-Rueda E. The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. Proteins Struct Funct Bioinforma. 2015:n/a–n/a.
45. Hagberg A, Swart P, Chult D. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. p. 11-15.
46. Kinser J. Python for Bioinformatics. USA: Jones & Bartlett Publishers; 2008.
47. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010;11:24.

Referencias

[1] D. Nelson and M. Cox, *Lehninger princípios de bioquímica.* Barcelona: Omega, 4ta ed., 2005.

[2] M. a. Ragan, J. O. McInerney, and J. a. Lake, "The network of life: genome beginnings and evolution. Introduction.," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 364, pp. 2169–75, Aug. 2009.

[3] J. M. Peregrín-Alvarez, C. Sanford, and J. Parkinson, "The conservation and evolutionary modularity of metabolism.," *Genome biology*, vol. 10, p. R63, Jan. 2009.

[4] R. Guimera and L. Amaral, "Functional cartography of complex metabolic networks," *Nature*, vol. 433, no. 7028, pp. 895–900, 2005.

[5] M. Arita, "The metabolic world of Escherichia coli is not small.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 1543–7, Feb. 2004.

[6] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and a. L. Barabási, "The large-scale organization of metabolic networks.," *Nature*, vol. 407, pp. 651–4, Oct. 2000.

[7] S. Light and P. Kraulis, "Network analysis of metabolic enzyme evolution in Escherichia coli.," *BMC bioinformatics*, vol. 5, p. 15, Feb. 2004.

[8] J. J. Díaz-Mejía, E. Pérez-Rueda, and L. Segovia, "A network perspective on the evolution of metabolism by gene duplication.," *Genome biology*, vol. 8, p. R26, Jan. 2007.

[9] S. C. G. Rison and J. M. Thornton, "Pathway evolution, structurally speaking.," *Current opinion in structural biology*, vol. 12, pp. 374–82, June 2002.

[10] T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork, "Pathway alignment: application to the comparative analysis of glycolytic enzymes.," *Biochemical Journal*, vol. 343, no. Pt 1, p. 115, 1999.

[11] M. a. Huynen, T. Dandekar, and P. Bork, "Variation and evolution of the citric-acid cycle: a genomic perspective.," *Trends in microbiology*, vol. 7, pp. 281–91, July 1999.

[12] K. P. Williams, J. J. Gillespie, B. W. S. Sobral, E. K. Nordberg, E. E. Snyder, J. M. Shallom, and A. W. Dickerman, "Phylogeny of gammaproteobacteria.," *Journal of bacteriology*, vol. 192, pp. 2305–14, May 2010.

[13] I. L. G. Newton, T. Woyke, T. A. Auchtung, G. F. Dilly, R. J. Dutton, M. C. Fisher, K. M. Fontanez, E. Lau, F. J. Stewart, P. M. Richardson, K. W. Barry, E. Saunders, J. C. Detter, D. Wu, J. A. Eisen, and C. M. Cavanaugh, "The Calyptogena magnifica chemoautotrophic symbiont genome.," *Science (New York, N.Y.)*, vol. 315, pp. 998–1000, Feb. 2007.

[14] S. E. Hoeft, J. S. Blum, J. F. Stolz, F. R. Tabita, B. Witte, G. M. King, J. M. Santini, and R. S. Oremland, "Alkalilimnicola ehrlichii sp. nov., a novel, arsenite-oxidizing haloalkaliphilic gammaproteobacterium capable of chemoautotrophic or heterotrophic growth with nitrate or oxygen as the electron acceptor.," *International journal of systematic and evolutionary microbiology*, vol. 57, pp. 504–12, Mar. 2007.

[15] A. Hara, K. Syutsubo, and S. Harayama, "Alcanivorax which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation.," *Environmental microbiology*, vol. 5, pp. 746–53, Sept. 2003.

[16] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes.," *Nucleic acids research*, vol. 28, pp. 27–30, Jan. 2000.

[17] R. Caspi, H. Foerster, C. a. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa, M. Krummenacker, S. Paley, J. Pick, S. Y. Rhee, C. Tissier, P. Zhang, and P. D. Karp, "MetaCyc: a multiorganism database of metabolic pathways and enzymes.," *Nucleic acids research*, vol. 34, pp. D511–6, Jan. 2006.

[18] R. Braakman and E. Smith, "The compositional and evolutionary logic of metabolism.," *Physical biology*, vol. 10, p. 011001, Feb. 2013.

[19] R. Braakman and E. Smith, "The emergence and early evolution of biological carbon-fixation.," *PLoS computational biology*, vol. 8, p. e1002455, Jan. 2012.

[20] D. J. Baumler, B. Ma, J. L. Reed, and N. T. Perna, "Inferring ancient metabolism using ancestral core metabolic models of enterobacteria," *BMC Systems Biology*, vol. 7, no. 1, p. 46, 2013.

[21] M. A. Martínez-Núñez, A. C. Poot-Hernandez, K. Rodríguez-Vázquez, and E. Perez-Rueda, "Increments and Duplication Events of Enzymes and Transcription Factors Influence Metabolic and Regulatory Diversity in Prokaryotes," *PLoS ONE*, vol. 8, p. e69707, July 2013.

[22] S. Freilich, L. Goldovsky, C. a. Ouzounis, and J. M. Thornton, "Metabolic innovations towards the human lineage.," *BMC evolutionary biology*, vol. 8, p. 247, Jan. 2008.

[23] C. C. Klein, L. Cottret, J. Kielbassa, H. Charles, C. Gautier, A. T. Ribeiro de Vasconcelos, V. Lacroix, and M.-F. Sagot, "Exploration of the core metabolism of symbiotic bacteria.," *BMC genomics*, vol. 13, p. 438, Jan. 2012.

[24] Y. Tohsato, H. Matsuda, and A. Hashimoto, "A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy.," *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 376–83, Jan. 2000.

[25] M. Chen and R. Hofestädt, "PathAligner: metabolic pathway retrieval and alignment.," *Applied bioinformatics*, vol. 3, pp. 241–252, Jan. 2004.

[26] M. Chen and R. Hofestädt, "Web-based information retrieval system for the prediction of metabolic pathways.," *IEEE transactions on nanobioscience*, vol. 3, pp. 192–9, Sept. 2004.

[27] M. Chen and R. Hofestaedt, "An algorithm for linear metabolic pathway alignment.," *In silico biology*, vol. 5, pp. 111–28, Jan. 2005.

[28] Y. Tohsato and Y. Nishimura, "Metabolic Pathway Alignment Based on Similarity between Chemical Structures," *IPSJ Digital Courier*, vol. 3, pp. 736–745, 2007.

[29] C. Clark and J. Kalita, "A comparison of algorithms for the pairwise alignment of biological networks.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 2351–2359, May 2014.

[30] R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson, "Alignment of metabolic pathways.," *Bioinformatics (Oxford, England)*, vol. 21, pp. 3401–8, Aug. 2005.

[31] Q. Cheng, R. Harrison, and A. Zelikovsky, "MetNetAligner: a web service tool for metabolic network alignments.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 1989–90, Aug. 2009.

[32] P. G. Ortegón Cano, *Alineamiento múltiple de vías metabólicas usando cómputo evolutivo*. Tesis para obtener el grado de maestría, Universidad Nacional Autonoma de Mexico, 2010.

[33] J. Gough, K. Karplus, R. Hughey, and C. Chothia, "Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.," *Journal of molecular biology*, vol. 313, pp. 903–19, Nov. 2001.

[34] S. A. Teichmann, S. C. Rison, J. M. Thornton, M. Riley, J. Gough, and C. Chothia, "The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli.," *Journal of molecular biology*, vol. 311, pp. 693–708, Aug. 2001.

[35] S. A. Teichmann, S. C. Rison, J. M. Thornton, M. Riley, J. Gough, and C. Chothia, "Small-molecule metabolism: an enzyme mosaic," *Trends in Biotechnology*, vol. 19, pp. 482–486, Dec. 2001.

[36] G. Moreno-Hagelsieb and S. Janga, "Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles," *Proteins: Structure, Function, and Bioinformatics*, vol. 70, no. 2, pp. 344–352, 2008.

[37] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, no. SciPy, pp. 11–15, 2008.

[38] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *Journal of molecular biology*, vol. 48, pp. 443–53, Mar. 1970.

[39] P. Ortegon, A. C. Poot-Hernández, E. Perez-Rueda, and K. Rodriguez-Vazquez, "Comparison of Metabolic Pathways in Escherichia coli by Using Genetic Algorithms.," *Computational and structural biotechnology journal*, vol. 13, pp. 277–85, Jan. 2015.

[40] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. July, pp. 379–423, 1948.

[41] J. Kinser, *Python For Bioinformatics*. Jones & Bartlett Publishers, 2008.

[42] J. Huerta-Cepas, J. Dopazo, and T. Gabaldón, "ETE: a python Environment for Tree Exploration.," *BMC bioinformatics*, vol. 11, p. 24, Jan. 2010.

[43] C.-H. Chou, W.-C. Chang, C.-M. Chiu, C.-C. Huang, and H.-D. Huang, "FMM: a web server for metabolic pathway reconstruction and comparative analysis.," *Nucleic acids research*, vol. 37, pp. W129–34, July 2009.

[44] G. Hernández-Montes, J. J. Díaz-Mejía, E. Pérez-Rueda, and L. Segovia, "The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution.," *Genome biology*, vol. 9, p. R95, Jan. 2008.

[45] C. Cunchillos and G. Lecointre, "Integrating the universal metabolism into a phylogenetic analysis.," *Molecular biology and evolution*, vol. 22, pp. 1–11, Jan. 2005.

[46] G. Caetano-Anollés, H. S. Kim, and J. E. Mittenthal, "The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture.," *Procee-*

dings of the National Academy of Sciences of the United States of America, vol. 104, pp. 9358–63, May 2007.

[47] A. Becerra and A. Lazcano, "The role of gene duplication in the evolution of purine nucleotide salvage pathways," *Origins of Life and Evolution of the Biosphere*, vol. 28, pp. 539–553, 1998.

[48] C. von Mering, E. M. Zdobnov, S. Tsoka, F. D. Ciccarelli, J. B. Pereira-Leal, C. A. Ouzounis, and P. Bork, "Genome evolution reveals biochemical networks and functional modules," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 15428–15433, Dec. 2003.

[49] R. A. Jensen, "Enzyme recruitment in evolution of new function.," *Annual review of microbiology*, vol. 30, pp. 409–25, Jan. 1976.

[50] K. Caetano-Anollés and G. Caetano-Anollés, "Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism.," *PloS one*, vol. 8, p. e59300, Jan. 2013.

Abreviaturas

BFS *Breadth First Search*. 15, 39, 40, *Glossary:* *Breadth First Search*

BLAST *Basic Local Alignment Tool*. 10, 39, 40, *Glossary:* *Basic Local Alignment Tool*

EC number *Enzyme commission number*. 15, 16, 39, *Glossary:* *Enzyme commission number*

ORF *Open Reading Frame*. 10, 12, 21, 39, 40, *Glossary:* *Open Reading Frame*

ESS *Enzymatic Step Sequence*. 4–9, 12, 14–16, 18, 19, 22–27, 33, 35–40, *Glossary:* *Enzymatic Step Sequence*

KEGG *Kyoto Enzyclopedia of Genes and Genomes*. 2, 4, 5, 8, 10, 12, 14, 39

Glosario

Basic Local Alignment Tool Algoritmo de alienamiento local altamente eficiente usado para hacer búsquedas de secuencias similares en bases de datos de secuencias biológicas. 10, 39, 40

Breadth First Search Búsqueda a lo ancho. En teoría de grafos, es un algoritmo usado para recorrer o hacer búsquedas en un grafo con prioridad "a lo ancho". Es decir, que recorre el grafo visitando todos los vecinos de un nodo antes de comenzar a visitar los vecinos de los vecinos. Un método complementario podría ser el método de búsqueda a lo profundo o *Deep First Search* (*DFS*). 15, 39, 40

Enzymatic Step Sequence Conjunto de pasos enzimáticos consecutivos representados por los primeros tres niveles de clasificación de la *enzyme comission.*. 39, 40

Enzyme commission number Número enzimático. Sistema de clasificación de las enzimas basado en las reacciones químicas que catalizan. Consiste en cuatro números o niveles ordenados jerárquicamente que describen el tipo, co-factores y sustratos de las reacciones químicas. 15, 39

Open Reading Frame Marco abierto de lectura. Secuencia identificada en un genoma correspondiente a la región codificante de un gen. Comienza con un codón de inicio y termina con un codón de paro. 10, 12, 39, 40

Python Lenguaje de programación libre, interpretado, multiparadigma, multiplataforma y de propósito general cuya filosofía apoya una sintaxis limpia y de fácil lectura. Por ésta última característica, es uno de los lenguajes de programación más usados en la actualidad y ha ido ganando terreno en en las áreas científicas para el análisis de datos. Existen una gran variedad de módulos o bibliotecas para hacer análisis científicos de distintos tipos. . 15, 39