UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

UNA PERSPECTIVA DE REDES SOBRE LA EVOLUCIÓN DEL METABOLISMO POR DUPLICACIÓN GÉNICA

T E S I S

QUE PARA OBTENER EL GRADO DE DOCTOR EN CIENCIAS BIOQUÍMICAS

P R E S E N T A

JUAN JAVIER DIAZ MEJIA

DIRECTOR DE TESIS
DR. LORENZO PATRICK SEGOVIA FORCELLA

CUERNAVACA, MORELOS **2007**

Ciencias Bioquímicas

F Q I B T I F C

UNAM POSGRADO
Ciencias Bioquímicas

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS BIOQUÍMICAS

OF. PMDCB. IBT.281.2007

ING. LEOPOLDO SILVA GUTIERREZ
Director General de Administración Escolar, UNAM
P R E S E N T E

Por medio del presente me permito informar a usted que en la reunión del Subcomité Académico, del día 14 de mayo de 2007, se acordó nombrar el siguiente Jurado para Examen de Doctor en Ciencias del **BIÓL. JUAN JAVIER DÍAZ MEJÍA** con **número de cuenta 9335037-6**, con la tesis titulada, **"UNA PERSPECTIVA DE REDES SOBRE LA EVOLUCIÓN DEL METABOLISMO POR DUPLICACIÓN GÉNICA"** dirigida por el Dr. Lorenzo Segovia Forcella.

PRESIDENTE:	Dr. Guillermo Gosset Lagarda
SECRETARIO:	Dr. Ricardo Canek Rodríguez de la Vega
VOCAL:	Dr. Maximino Aldama González
SUPLENTE:	Dr. Lorenzo Segovia Forcella
SUPLENTE:	Dr. Gabriel del Río Guerra

Sin otro particular por el momento, aprovecho la ocasión para enviarle un cordial saludo.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPIRITU"
Cuernavaca, Mor., 15 de agosto de 2007
EL COORDINADOR DEL PROGRAMA

DR. RAFAEL VÁZQUEZ PUHALT

C.c.p. Archivo

Agradecimientos

Quisiera externar mi agradecimiento al Consejo Nacional de Ciencia y Tecnología (CONACYT) y a la Dirección General de Estudios de Posgrado de la UNAM por los fondos proporcionados para la realización de este proyecto de doctorado, tanto por la subvención No. 43502 del CONACYT al grupo del Dr. Lorenzo Segovia, como por las becas para estudios de posgrado y los apoyos para asistir a congresos y estancias de investigación que me otorgaron ambas dependencias. De igual manera agradezco a la Fundación Fulbright-Garcia Robles y COMEXUS por la beca que me otorgaron para mi estancia de investigación en la Universidad de Notre Dame, EUA.

También quisiera agradecer encarecidamente a la Unidad de Cómputo del Instituto de Biotecnología (IBT) de la UNAM por las facilidades prestadas para el mantenimiento del 'Cluster Sputnik II', y al Macroproyecto de Tecnologías de la Información y la Computación de la UNAM por el financiamiento otorgado para el mismo. De igual forma, no puedo dejar de agradecer a la Unidad de Docencia y a la Unidad de Biblioteca del IBT por todas las facilidades prestadas durante mis estudios de posgrado.

Agradecimientos 2.0

Tal vez este es uno de los momentos más esperados por un estudiante… cuando ya pasados los agradecimientos obligados, y estando a punto de imprimir la tesis, se puede uno explayar de lo lindo... a fin de cuentas esta parte es la que todo mundo lee… más que el 'abstract'…

Primero que nada quiero agradecer a la Raza que con sus impuestos patrocinan las locuras de los estudihambres

724 Mil gracias a la Familia Linares de South Bend que me dieron nuevos ojos para ver al mundo del otro lado del Río Bravo. Y a Martín Peralta por presentarme con su familia chicana.

También quiero agradecer a mi comité tutoral, Lorenzo, Neto y Sergio que siempre fueron una guía para contestar mis necedades.

A los Lorenzos, léase todos: los que se fueron antes que yo y los que se quedan un ratón más, por aguantar presión cuando me puse más loco de lo normal y por los Ramones, los Pixies, Pink Floyd, Moby y los 30 GB de música de años recientes.

A la comunidad de los miercoles@ibt.unam.mx por tantos momentos chidos y a todos los que no quedaron ciegos después unas burbujitas 1 molar…

A mis Padres y Hermanos

Índice

1. Resumen

Antecedentes

Los seres vivos son sistemas complejos cuyo estudio se ha visto favorecido con el '*boom'* de la Teoría de Redes. La duplicación génica es una las principales fuentes de versatilidad metabólica. Algunos modelos sobre la evolución de las redes metabólicas por duplicación génica recrean las propiedades topológicas globales del metabolismo, pero la forma en que sus preceptos han sido atendidos es relativamente simplista. En este trabajo usamos una perspectiva de redes para determinar la influencia de algunas restricciones funcionales del metabolismo sobre la retención de genes duplicados.

Resultados

Hemos comparado el contenido de dominios de las enzimas que conforman las redes metabólicas de varias especies para detectar genes duplicados y descubrimos que existe una alta retención de duplicados entre enzimas que catalizan reacciones consecutivas, como en el caso de las ligasas de la biosíntesis del peptidoglicano. Derivado de esto, las redes metabólicas tienen una alta retención de duplicados dentro de sus módulos funcionales, y encontramos que un acoplamiento bioquímico preferente entre los tipos de reacciones en el metabolismo explica parcialmente este sesgo. Una situación similar ocurre en las redes de interacciones enzima-enzima, pero no en las de interacciones proteína-proteína no-enzimáticas, ni en las de regulación de la transcripción génica, lo cual sugiere que la retención de duplicados es producto de leyes bioquímicas que gobiernan las relaciones sustrato-enzima-producto. Adicionalmente, confirmamos que existe una alta retención de duplicados entre enzimas que catalizan reacciones químicamente similares, como en el metabolismo de ácidos grasos. Sin embargo, también entre reacciones químicamente diferentes hay una retención de duplicados significativamente alta. Finalmente, detectamos una significativa retención de genes duplicados en grupo que son capaces de general incluso rutas metabólicas completas.

Conclusiones

Nuestros resultados indican que el siguiente paso en el modelado '*in silico'* del origen y la evolución de distintos tipos de redes biológicas debe tratar de capturar tanto las características generales que éstos comparten como las que los diferencian; por ejemplo, el acoplamiento bioquímico preferente entre los tipos de reacciones de las redes metabólicas. Sugerimos que algunos modelos tradicionales de evolución metabólica como el *paso-a-paso* y el *de-mosaico* considerados antagónicos en realidad son complementarios.

1. Abstract

Background

Living organisms are complex systems whose study has been favored with the 'boom' of Networks Theory. Gene duplication is one of the main sources of metabolic versatility. Some models on the evolution of metabolic networks by gene duplication recreate the general topological properties of metabolism, but their assumptions are relatively simplistic. We used a network-based approach to determine the influence of some metabolic functional constraints on the retention of duplicated genes.

Results

We detected duplicated genes by looking for enzymes sharing homologous domains in the metabolic networks from various species, and uncovered an increased retention of duplicates for enzymes catalyzing consecutive reactions, as illustrated by the ligases acting in the biosynthesis of peptidoglycan. As a consequence, metabolic networks show a high retention of duplicates within functional modules, and we found a preferential biochemical coupling of reactions that partially explains this bias. A similar situation was found in enzyme-enzyme interaction networks, but not in neither non-enzymatic protein interaction networks nor in gene transcriptional regulatory networks, suggesting that the retention of duplicates results from the biochemical rules governing substrate-enzyme-product relationships. Additionally, we confirmed a high retention of duplicates between chemically similar reactions, as illustrated by fatty-acid metabolism. The retention of duplicates between chemically dissimilar reactions is, however, also greater than expected by chance. Finally, we detected a significant retention of duplicates as groups, instead of single pairs, capable to generate full metabolic routes.

Conclusion

Our results indicate that the next step in '*in silico'* modeling of the origin and evolution of biological networks must capture as general characteristics shared by these networks as those differentiating them, such as the preferential biochemical coupling of reactions in metabolic networks. We suggest that some traditional models on metabolic evolution such as the *stepwise* and *patchwork* models are not independent of each other but complementary.

2. Introducción

En 1987, un artículo publicado en la revista *Nature* [1] sugirió que el trastorno maníaco depresivo o trastorno bipolar en los humanos podría estar causado por una mutación en el cromosoma 11. Una década más tarde otros grupos describieron diferentes mutaciones en los cromosomas 6, 13, 15, 1 y 5 como las posibles causas de este padecimiento. Aparentemente, no existe un gene único responsable del trastorno bipolar sino diversos genes que al interactuar, son responsables conjuntamente. Lo que es claro a la fecha, es que para entender muchas enfermedades y diversos procesos celulares, como la transmisión de señales y el metabolismo, necesitamos pensar que los seres vivos están formados por diversos componentes celulares que trabajan de forma conjunta y no basta con analizarlos por partes. Hay que descubrir las redes de sus interacciones y analizarlas como un todo, desde los puntos de vista topológico y dinámico. Este tipo de enfoque ha sido impulsado por los estudios de funciones moleculares puntuales y algunas de sus interacciones, desde una perspectiva reduccionista, pero los principios generales que rigen la estructura y función de las redes biológicas están siendo descubiertos gracias a la cooperación entre las Ciencias Genómicas, la Biología Celular, la Bioquímica y la Teoría Evolutiva, echando mano del poder analítico de las llamadas ciencias sintéticas, como la Ingeniería y la Computación [2], dando lugar a la llamada Biología de Sistemas.

"En la Biología de Sistemas… se trata de unir más que separar, de integrar más que reducir. Por lo cual, se requiere que desarrollemos nuevas formas de pensar acerca de esa integración, que sean tan rigurosas como sus contrapartes reduccionistas, pero diferentes. Esto es, cambiar nuestra filosofía en el sentido más amplio" [3].

En los siguientes apartados hacemos un recuento de los trabajos que han hecho que las redes en la biología pasen de ser una noción teórica, a una de las herramientas más importantes para entender a los seres vivos.

2.1 Antecedentes

2.1.1 La noción de redes en el estudio de los sistemas complejos

Se define a una red como el conjunto de nodos y conexiones entre esos nodos que conforman a un sistema. El estudio de las redes tiene un amplia tradición en las matemáticas, remontándose al trabajo de Euler que resolvió el llamado problema de los Puentes de Königsberg en 1736: imaginemos dos islas, en el río Pregel que cruza Königsberg, que se unen entre ellas y con tierra firme mediante siete puentes (Figura 1). El problema se plantea con la siguiente pregunta ¿es posible dar un paseo empezando por una de las cuatro partes de tierra firme, cruzando cada puente una sola vez y volviendo al punto de partida? Euler enfocó el problema representando cada parte de tierra por un nodo y cada puente por una conexión (Figura 1). Entonces, el problema anterior se puede trasladar a la siguiente pregunta: ¿se puede recorrer la red iniciando y terminando en el mismo nodo, sin repetir las conexiones? Euler demostró que esto no es posible puesto que el número de conexiones que inciden en cada nodo no es par, lo cual es una condición necesaria para entrar y salir de cada nodo, y para regresar al nodo de partida, por caminos distintos en todo momento. Aunque actualmente este problema pudiera parecer trivial, representa el punto de partida de lo que en nuestros días llamamos la Teoría de Redes. Algunas de las redes estudiadas hoy en día, al ser sistemas complejos[*], presentan muchos más nodos y conexiones que los puentes de Königsberg, y es gracias a los estudios analíticos[†] heredados de las matemáticas y a los estudios empíricos provenientes de las ciencias sociales que la noción de redes ha penetrado en una amplia gama de disciplinas científicas, convirtiéndose probablemente en la herramienta multidisciplinaria más importante para entender a los sistemas complejos, incluyendo los biológicos.

El interés científico por la estructura de las redes inició en la década de 1940 con los trabajos de Solomonoff y Rapoport quienes sugirieron la necesidad de analizar las propiedades estadísticas generales de las redes, más que las características individuales de sus nodos. En

[*] Las definiciones de **sistemas complejos** son muchas, pero todas conllevan al hecho de que sus características emergen por la interacción de sus partes, por lo que pueden ser descritos solamente al analizarlos como un todo. De igual manera, se les puede estudiar como una entidad que interactúa con su ambiente.

[†] Los **estudios empíricos** permiten determinar las características estructurales y/o funcionales de un sistema a través de experimentación y observación fenomenológica, contemplando tanto pruebas acertadas como errores. También permiten determinar las relaciones que las propiedades del sistema guardan entre sí y con el ambiente. Complementariamente, los **estudios analíticos**, tratan de generar modelos sustentados matemáticamente que describen al sistema, y se retroalimentan con los estudios empíricos.

particular, encontraron que durante la formación de las redes aleatorias*, es posible llegar a un punto en que exista un 'componente gigante' de la red en el cual todos sus nodos están conectados, y que la formación de ese componente depende del grado de conectividad promedio de la red† [4]. Definiendo en 1951 lo que hoy conocemos como la 'transición de fase', pero sus contribuciones permanecieron en la oscuridad. Diez años mas tarde, Erdős y Rényi, considerados los padres de la teoría moderna de grafos aleatorios, llegaron a la misma conclusión generando un modelo que muestra que la probabilidad de que una red posea un propiedad Q, se aproxima a 1 cuando el tamaño de la red N → ∞. O dicho de otra forma, casi cualquier red aleatoria con N nodos posee la propiedad Q [5,6]. En conjunto, estos estudios analíticos manifiestan la importancia de estudiar a los sistemas complejos como un todo, más que por partes.

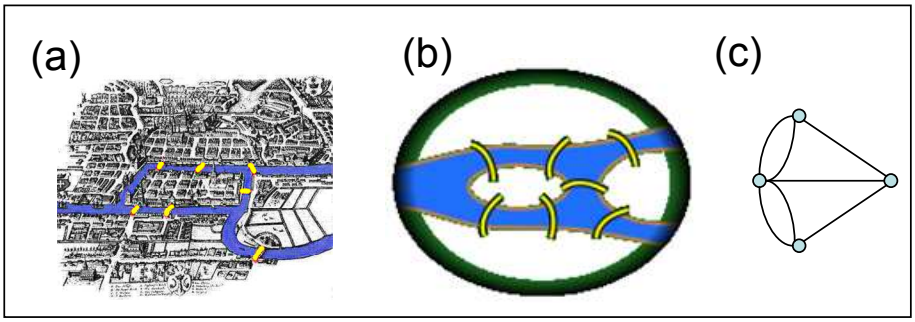


Figura 1. El problema de los puentes de Königsberg. (a) La ciudad de Königberg (actualmente Kalinin-grado, Rusia) en la época de Euler con sus siete puentes en ama-rillo. **(b)** una abstracción de la misma ciudad. **(c)** La red con que Euler representó el problema de los puentes. Cada nodo corresponde a una porción de tierra y cada conexión a un puente. La idea es recorrer todos los nodos y todas las conexiones, empezando y terminando en el mismo nodo y pasando una sola vez por cada conexión.

Del lado de los estudios empíricos, Pool y Kochen, inspirados por los trabajos de Rapoport, desarrollaron un análisis en el cual observaron que las personas son más cercanas, en términos de conocidos mutuos, de lo esperaríamos por mera casualidad, y comenzaron a circular sus resultados en 1958 como un manuscrito que fue publicado veinte años después en el número inaugural de la revista Social Networks [7]. Esto inició una serie de reportes que incitaron a Milgram a desarrollar sus famosos experimentos de redes sociales de 'mundo pequeño'‡ con los

* Las **redes aleatorias** se generan partiendo de un conjunto fijo de nodos a los cuales se añaden conexiones al azar, por lo que se pueden formar redes con distinto grado de conectividad según el modelo con el que se añaden conexiones.

† En una red, cada nodo tiene un número definido de conexiones con otros nodos, a lo cual se llama **grado de conectividad** (k). El grado de conectividad promedio de la red (K) es el promedio de k de todos sus nodos. Solomonoff y Rapoport demostraron que en las redes aleatorias, cuando $K < 1$ la red se fracciona en muchos componentes pequeños; en cambio, cuando $K \geq 1$ se forma un componente gigante que contiene a todos los nodos de la red.

‡ En una red aleatoria la distancia promedio (número de pasos) entre sus nodos crece logarítmicamente con el tamaño de la red. En cambio, en las **redes de mundo pequeño** las distancias son más cortas. Además,

cuales en la década de 1960 dedujo que, en promedio, una persona puede contactar a otra por medio de una cadena de tan solo cinco conocidos en común [8]. Esto desde luego antes de las relaciones basadas en el correo electrónico. El estudio de las redes de mundo pequeño tuvo un gran avance con el modelo de Watts y Strogatz [9] que lograron generar redes con nodos a distancias más pequeñas de las esperadas al azar y con un alto grado de agrupamiento (Figura 2).

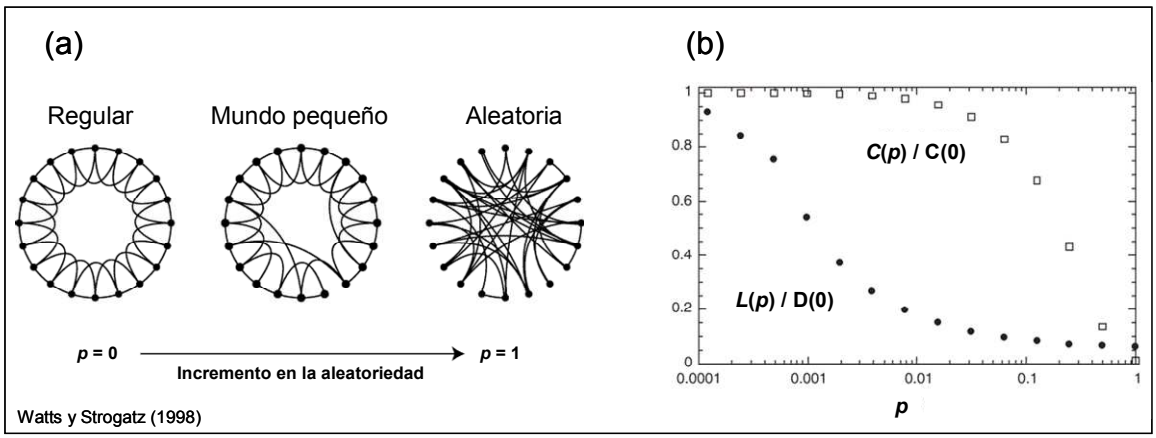


Figura 2. Redes de mundo pequeño. (a) tres tipos de redes que se pueden obtener aumentando la reconexión de sus nodos aleatoriamente. **(b)** Valores de la distancia entre los nodos (*L*) y su agrupamiento (*C*) en una red en función del incremento de la aleatoriedad de las reconexiones, como se puede observar las redes de mundo pequeño tienen tanto distancias pequeñas como agrupamiento alto. Para una comparación con otros tipos de redes ver la Figura 3.

En 1965 Price publicó un articulo [10] que es considerado un tesoro oculto [11]. En él, se analiza la topología de una red de citas entre artículos científicos —los artículos son los nodos y las citas entre ellos son las conexiones— mostrando que las distancias entre sus nodos siguen un patrón similar al de las redes de mundo pequeño, pero además posee algunos nodos altamente conectados, hoy llamados *hubs*[*], que dominan su topología, descubriendo un tipo de red de mundo pequeño hoy llamada libre de escala[*], un termino acuñado por Barabási y Albert [12] (Figura 3c). Los *hubs* no se contemplan en los modelos de redes aleatorias, por ello cuando los grupos de Barabási[13], los hermanos Faloutsos[14] y Kumar [15] compararon otras redes del mundo real, la world-wide web (WWW) y su prima el Internet, contra el modelo de redes aleatorias de

en las redes de mundo pequeño, existe un alto **grado de agrupamiento**: supongamos que existe un nodo A que se conecta con otros tres (B, C y D), el agrupamiento implica que además de las conexiones de A con sus vecinos; B, C y D también tienen conexiones entre ellos. O sea, los "amigos de mis amigos, son amigos entre si". El agrupamiento es escaso en las redes aleatorias.

[*] En las redes aleatorias todos los nodos tienen más o menos el mismo número de conexiones (*k*), resultando en una distribución Poissoniana que se puede representar con (*K*). En cambio, en las **redes libres de escala**, que son un tipo de red de mundo pequeño, hay unos cuantos nodos, llamados *hubs*, que están altamente conectados, mientras que la gran mayoría de los nodos están poco conectados. Es decir, la probabilidad p(*k*) de que un nodo de una red libre de escala esté conectado con k nodos es proporcional a $k^{-\gamma}$, aproximándose a una **ley de potencias**, por lo que (*K*) no representaría propiamente a la red. Los *hubs* no se contemplan en las redes aleatorias ni en todas las de mundo pequeño.

Erdős-Renyi, para determinar su diámetro; esto es, en cuántos *clicks* se puede ir de una página Web a otra, o de un dominio de servidores a otro, se sorprendieron al encontrar que estas redes parecían de mundo pequeño, pero al igual que la red de citas de Price, poseen *hubs* y por lo tanto son libres de escala. En un segundo articulo [16], Price fue más allá de lo empírico, al proponer un modelo por el cual se pueden generar *hubs* y redes libres de escala, basándose en el hecho de que los artículos muy citados tienden a ser más y más citados con el tiempo, en un proceso que llamó 'ventaja acumulativa'. La idea de este proceso, que es más conocido en nuestros días como 'acoplamiento preferencial con los nodos más conectados', fue retomada por el grupo de Barabási [12] para desarrollar su modelo 'BA' que fue ratificado y ampliado independientemente por Krapivsky et al. [17] y Dorogovtsev et al. [18]. El modelo BA, que es ampliamente aceptado como una de las explicaciones de la existencia de las redes libres de escala, consta de dos ingredientes: i) que la red esté creciendo; esto es, que se añadan tanto nodos como conexiones continuamente y ii) que los nuevos nodos se acoplen preferentemente con los nodos preexistentes cuya conectividad sea de por sí alta; o lo que es lo mismo, "el rico se vuelve más rico". Finalmente, el grupo de Barabási generó un segundo modelo [19,20] que trata de explicar el origen de la modularidad* en las redes libres de escala, el cual se basa en los componentes del modelo BA, pero añade un grado más de complejidad al formar grupos de nodos altamente conectados de manera sistemática emulando un crecimiento fractal† (Figura 3d). Imaginemos un triangulo cuyos vértices son nodos que se conectan a un cuarto nodo central (nodos azules en la Figura 3d), por definición los cuatro nodos están bien agrupados. Ahora tripliquemos ese grupo y conectemos los nuevos nodos (nodos verdes en la Figura 3d) con los preexistentes, recordemos que el nodo azul central está ganando cada vez más conexiones por acoplamiento preferencial. Sigamos con el proceso reiteradamente y así generaremos una red con *hubs* —como el nodo azul central—, altamente agrupada, jerárquica y modular. A la par de la formalización de este modelo en el año 2003 [20] se dio el gran salto en la Teoría de Redes cuando investigadores de diferentes disciplinas realizaron estudios empíricos mostrando que, aunque no todas las redes son libres de escala [21], muchas de ellas, tanto tecnológicas [12-15,21,22], como sociales [12,21-24] y biológicas [25-29] sí lo son, promoviendo la necesidad de un cambio de paradigma. Se tenia que dejar de pensar en las redes como sistemas aleatorios y analizarlas más desde el punto de vista de la libertad de escala y los *hubs*. En el siguiente apartado nos enfocamos en las implicaciones que estos hallazgos han tenido en el área de las redes biológicas.

---

* En algunas redes libres de escala se pueden identificar grupos de nodos altamente conectados entre ellos, y no así con el resto de la red, a esos grupos se les llama **módulos**.

† Un **fractal** es un objeto geométrico cuya estructura básica se repite en diferentes escalas y tiene auto-similaridad exacta o estadística. En el modelos de Ravasz et al. la estructura básica son los cuatro nodos iniciales.

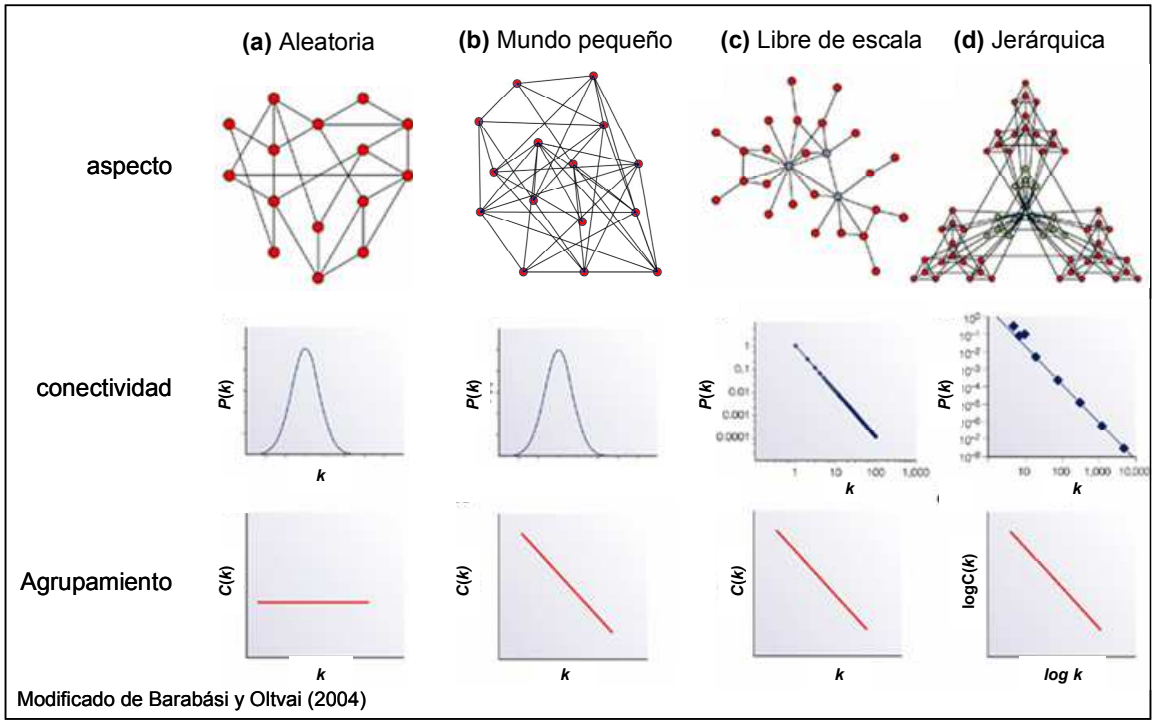


Figura 3. Tipos de redes no regulares. Los cuatro tipos de redes no regulares en los que se han enfocado los estudios de sistemas complejos. **(a)** *aleatorias*, como las del modelo Erdős-Rényi [5,6], con conectividad muy parecida en la mayoría de sus nodos y poco agrupamiento. **(b)** de *mundo pequeño*, como las del modelo Watts-Strogatz [9] con conectividad muy parecida en sus nodos y alto agrupamiento, además de distancias cortas entre sus nodos. **(c)** *libres de escala*, como las del modelo Barabási-Albert [12], con *hubs* y alto agrupamiento, se les considera un subtipo de las redes de mundo pequeño. **(d)** *jerárquicas*, como las del modelo de Ravasz et al. [19,20], con *hubs,* muy alto agrupamiento (nótese que en esta red la escala de agrupamiento es logarítmica) y módulos.

2.1.2 Redes biológicas

Consideramos que tres han sido los principales ingredientes para el *boom* de la teoría de redes en las ciencias biológicas (Figura 4). El primero, como se describió en la sección anterior, fue el hallazgo de la libertad de escala en un gran diversidad de tipos de redes, incluidas las biológicas, los temas multidisciplinarios en general son taquilleros. El segundo, es la amplia cantidad de información derivada de los estudios genómicos y proteómicos a gran escala, que puede ser manejada y entendida con redes. Y el tercero, es la influencia de algunos investigadores de renombre, que sugirieron la necesidad conjuntar los primeros dos ingredientes [2,30]. A partir del año 2000 el estudio de las redes biológicas se ha enfocado en dos aspectos [31]. El primero, es describir sus propiedades topológicas, tanto globales, como la libertad de escala y la existencia de módulos funcionales; como locales, dándose especial relevancia al agrupamiento y la formación

de motivos*. El segundo aspecto son los estudios dinámicos, en los cuales se trata de simular analíticamente o de dar seguimiento a cierto fenómeno, como el metabolismo, la trascripción génica, la señalización o la división celular. Como generalmente ocurre en la Biología, la Teoría Evolutiva ha jugado un papel indispensable en la unión y entendimiento de estos aspectos. A continuación, resumimos los avances que consideramos más sobresalientes en cada uno de ellos.

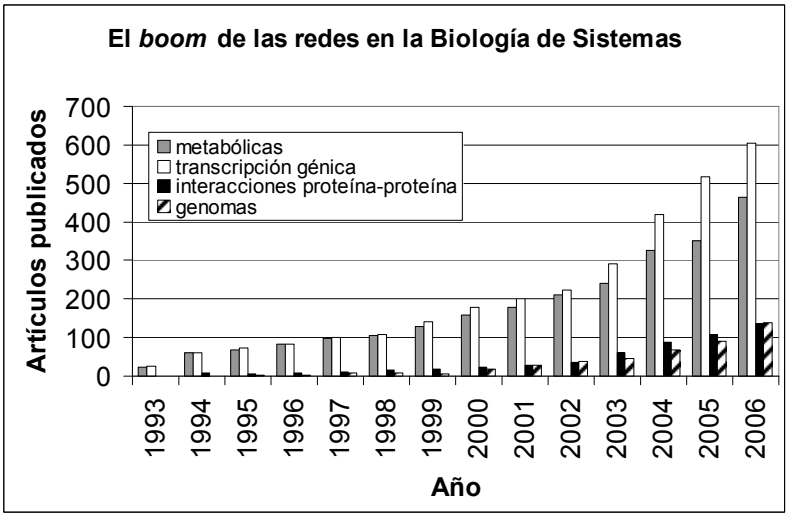


Figura 4. El *boom* de las redes en la Biología de Sistemas. Histograma que muestra el número de artículos contenidos en la base de datos PUBMED entre 1993 y 2006. Se usó la palabra "network" como se-milla de búsqueda y "metabolic OR metabolism" para redes metabólicas, "transcriptional OR transcription" para redes de regulación de la transcripción génica, y "protein-protein" para redes de interacciones proteína-proteína. Como se puede notar, en los últimos años ha habido un gran incremento en los reportes de redes biológicas. A manera de comparación se incluye el crecimiento en número de genomas completamente secuenciados en la base de datos KEGG [32].

En lo relativo a los estudios topológicos, a principios de esta década diferentes grupos caracterizaron la libertad de escala en varios tipos de redes biológicas, como las metabólicas† [22,26,28,29], las de interacciones proteína-proteína‡ [22,27], las de la regulación de la transcripción génica§ [33], y las de presencia-ausencia de dominios en las proteínas* [34]. Esto significa que hay

* En algunos tipos de redes existen nodos altamente agrupados. Se ha sugerido que su agrupamiento se debe en gran medida a una tendencia de los nodos por formar estructuras sencillas de tres o más nodos. Al determinar las frecuencias de cada una de esas estructuras, por ejemplo triángulos, en una red real y compararlas contra las esperadas en redes aleatorias con el mismo número de nodos y conexiones, podemos identificar cuáles de esas estructuras están sobre-representadas y, usando un valor de corte, establecemos que representan un **motivo**; o bien, cuando están sub-representadas constituyen un antimotivo.

† Las **redes metabólicas** se forman con reacciones bioquímicas. Sus nodos pueden ser los metabolitos que se conectan por las reacciones en que concurren. O bien, pueden ser las enzimas, por ejemplo (E1 y E2), que se conectan si algún producto de E1 es el sustrato de E2 (para más detalles ver Figura 5).

‡ Dentro la célula las proteínas forman complejos. Las técnicas a gran escala para detectar estos complejos permiten determinar qué proteínas (nodos) interactúan y con ello se reconstruyen las **redes de interacciónes proteína-proteína**.

§ Dentro de las células no todos los genes se transcriben (expresan) al mismo tiempo. Las proteínas llamadas factores de transcripción (FT) se unen a las secuencias promotoras de ciertos genes blanco (GB) activando y/o reprimiendo su transcripción. En las **redes de regulación de la transcripción génica** se establecen conexiones entre los FT y su(s) GB. Algunos FT son a la vez GB de otros FT.

pocos *hubs*, que pueden ser según el tipo de red que se trate, proteínas, metabolitos o dominios de proteínas, que interactúan funcional y/o físicamente con otros nodos, mientras que la gran mayoría de los nodos tienen una baja conectividad. Una implicación funcional de estos hallazgos la proponen Maslov y Sneppen [35] al mostrar que tanto en las redes de interacciones proteína-proteína, como en las de regulación de la transcripción génica de la levadura *Saccharomyces cerevisiae*, los *hubs* tienden a estar lejanos entre sí, a la vez que atraen nodos poco conectados. Y sugieren que gracias esto las redes biológicas pueden evitar funciones cruzadas entre *hubs*, así como la pérdida completa de la funcionalidad de la red, si se eliminara una región con muchos *hubs*. Por otro lado, dado que el modelo BA implica que los *hubs* son los nodos más antiguos de la red, un pregunta subyacente es si en las redes biológicas las proteínas más conectadas son las más antiguas. En ese sentido, los estudios con redes de interacciones proteína-proteína han sido los más reportados [36-41] señalado que en estas redes las proteínas más conectadas y agrupadas, tienden a ser más antiguas† e incluso a evolucionar más lento‡ que las poco conectadas y dispersas. Sin embargo, hay que señalar que, a excepción del estudio de Butland et. al [36], estos reportes se enfocan en las proteínas de *S. cerevisiae* y sus ortólogas† en otros eucariontes, pero al extender los análisis a relaciones filogenéticas distantes, no se encontró lo mismo. De hecho, el grupo de Li [42], determinó que usando una versión más completa de la red de *S. cerevisiae* las proteínas con ortólogos en Bacteria, que se asumen como las más antiguas, no son las más conectadas. En el caso de la red de *Escherichia coli* [36] se encontró que hay una tendencia a que las proteínas más conectadas sean las más antiguas, pero estos resultados están sesgados por la gran cantidad de proteínas ribosomales incluidas en su muestra, las cuales tiene un alto grado de agrupamiento y por tanto esta observación podría ser un efecto local.

Algo similar ocurre con la esencialidad de los genes§, estudios topológicos iniciales señalaron que los genes más conectados en la red de interacciones proteína-proteína de levadura son en mayor proporción genes esenciales [27], pero ello no implica que en todas las redes biológicas ocurra lo mismo. Aparentemente, el problema de la esencialidad Vs. conectividad es más filosófico que estadístico y en este sentido los estudios sobre la dinámica de las redes han contribuido en gran medida a su solución. Las redes biológicas no son estáticas y las proteínas

---

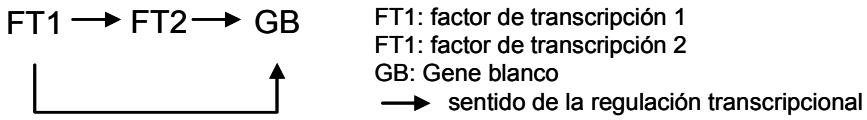
* En las **redes de dominios de proteínas** los nodos son los dominios y dos nodos se conectan si concurren en alguna proteína.

† Se considera que las proteínas con una amplia distribución de **ortólogos** –genes heredados de forma vertical de un ancestro a sus descendientes inmediatos– entre diversas especies son más **antiguas** que aquellas con una distribución restringida a pocas especies.

‡ En los genes que codifican para proteínas se pueden comparar las posiciones equivalentes de cada codon entre proteínas ortólogas para obtener una relación entre los cambios que afectan la secuencia de la proteína y los que no (cambios no sinónimos y sinónimos, respectivamente). Si los cambios sinónimos son muchos más que los no sinónimos se dice que la proteína **evoluciona lentamente**.

§ Se considera **gene esencial** a aquel cuya eliminación (*knock-out*) o silenciamiento provoca la muerte de la cepa mutante, bajo condiciones ambientales en las cuales la silvestre es capaz de sobrevivir.

que bajo ciertas condiciones ambientales presentan una alta conectividad, al cambiar el ambiente, modificarse los flujos metabólicos* o transcripcionales, o simplemente al no expresarse las proteínas con las que interactúan, pueden resultar pobremente conectadas [43,44]. En *E. coli*, por ejemplo, el factor transcripcional Crp no es codificado por un gene esencial [45] a pesar de ser un *hub* que regula la expresión de ~200 genes en esta bacteria [46]. Esto es probablemente porque Crp no regula la expresión de todos esos genes al mismo tiempo, ni de manera exclusiva, y otro genes pueden suplir su función bajo diferentes condiciones ambientales. De hecho, el principal motivo estructural que presentan las redes de regulación transcripcional es el llamado ciclo de alimentación predefinida (*feed-forward loop*) [47] que consta de tres nodos relacionados de la siguiente forma:



Cuando un factor transcripcional FT1 como Crp es eliminado bajo ciertas condiciones otro factor FT2 puede amortiguar su pérdida y mantener a la red trabajando. Luscombe et al. [44] encontraron que este fenómeno es muy común y que la gran mayoría de los *hubs* en las redes transcripcionales son sólo transitorios y específicos para ciertas condiciones.

Respecto a las redes metabólicas es necesario señalar que las hay de tres tipos básicos, las centradas en compuestos, las centradas en enzimas y las bipartitas que engloban a las dos primeras (Figura 5). Los análisis pioneros de Wagner y Fell [26,29] sobre la topología de las redes centradas en compuestos mostraron que los ribonucleótidos fosfatados y ciertos aminoácidos representan algunos de los nodos más conectados en este tipo de redes, apoyando las propuestas de Benner y Horowitz [48,49] sobre la naturaleza ancestral de estos metabolitos. Sin embargo, estudios más recientes de Guimerà y Amaral [50] señalan que esto pudiera no ser del todo cierto. Cabe mencionar que las redes metabólicas presentan una modularidad estructural y funcional [19,50,51] que implica que tienen grupos de nodos altamente conectados entre ellos y no así con el resto de la red, llamados módulos, y que los nodos de cada módulo tienden a presentar funciones concertadas, por ejemplo, existe el módulo de la biosíntesis de aminoácidos, el de la biosíntesis de nucleótidos, etc. Guimerà y Amaral reconstruyeron las redes metabólicas centradas en

* Una aproximación dinámica a las redes metabólicas la ofrecen **los análisis de balance de flujos** (FBA) con los cuales, conociendo la topología de la red y las constantes estequiométricas de sus reacciones, se puede simular la respuesta de la red ante ciertos estímulos iniciales, por ejemplo una determinada cantidad de sustratos, para llegar a un fin último, por ejemplo la producción de biomasa. Con esta simulación se puede predecir el **flujo metabólico** de cada reacción. También se puede simular la eliminación de ciertos nodos (enzimas y/o compuestos) obligando a que su flujo sea nulo, para determinar, por ejemplo, si un gene es esencial bajo las condiciones de la simulación.

compuestos de diversas especies y, aunque coinciden con Wagner y Fell en que los *hubs* de estas redes pudieran ser muy antiguos, encontraron otros nodos poco conectados a los que llamaron conectores no-*hubs,* que sirven de puentes entre módulos, y parecieran ser incluso más antiguos que algunos *hubs.* Del lado de las redes metabólicas centradas en enzimas y los análisis de balance de flujos metabólicos, en un influyente estudio analítico de los grupos de Hurst y Pál [43] que simuló la producción de biomasa de *S. cerevisiae* en diferentes medios de cultivo, se determinó que algunos genes que se consideran "esenciales" bajo ciertas condiciones no lo son en otras, y viceversa. Aunque casi en todas las condiciones simuladas alrededor del 45% de los genes contemplados parecen esenciales, los genes que conforman esa fracción varían mucho entre condiciones. En un estudio similar, Almaas et al. [52] determinaron que en *S. cerevisiae* apenas 33 enzimas (2.8% del total analizado) parecen "esenciales" en diversas condiciones metabólicas, y que en *E. coli* y *Helicobacter pylori* la cifras son 12 y 35%, respectivamente. Estos datos ayudan a entender los resultados de un estudio reciente de Wagner et al. [53] que muestra que si bien en *S. cerevisiae* hay una tendencia a que las enzimas muy conectadas y/o con flujos metabólicos altos cambien poco en sus secuencias, muchas de las enzimas poco conectadas y/o con flujos bajos también cambian poco, porque pudieran tener flujos altos en condiciones no analizadas. Consideramos que sería interesante determinar cuáles enzimas poco conectadas pueden servir como conectores no-*hubs* entre módulos. La idea sería determinar si por tener flujos metabólicos altos su tasa evolutiva es menor. Finalmente, otro fenómeno que provoca que en las redes metabólicas la conectividad no correlacione claramente con la tasa de cambio, ni con la esencialidad, ni con la conservación de ortólogos, es que existe una gran cantidad de enzimas y rutas alternativas que pueden amortiguar el efecto de la perdida de genes e incluso de ramas completas de las redes (Hernandez-Montes et al. *Trabajo en preparación*) [43,54]. En resumen, dado que las redes metabólicas y las de la regulación transcripcional son de los sistemas más dinámicos en la biología [43,44,50] la definición de esencialidad en estas redes no es tan sencilla como en las redes de interacciones proteína-proteína [27] o en algunas redes tecnológicas [25] ¿Qué tan esencial es lo esencial? O más aún ¿Qué tan no-esenciales son los genes "no-esenciales"? Son preguntas que tendrán que ser contestadas con estudios metabólicos a gran escala que confirmen o corrijan los resultados de análisis de balance de flujos. En el siguiente apartado continuaremos con ejemplos concretos que muestran la necesidad de tratar al metabolismo desde una perspectiva de redes.

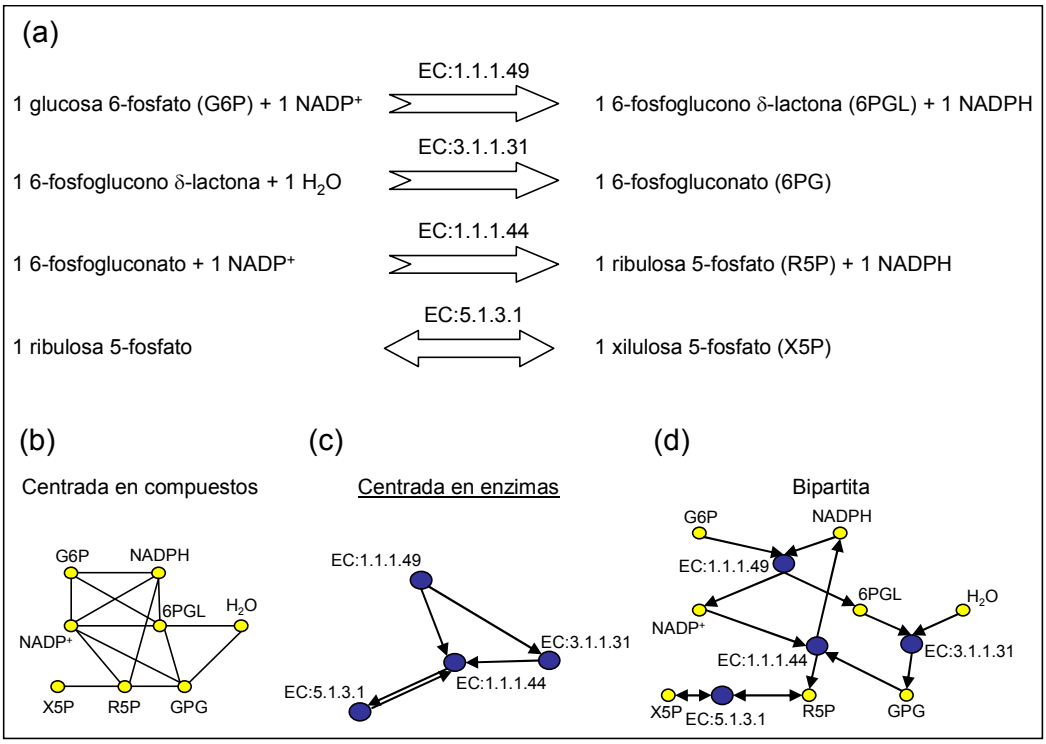


Figura 5. Diferentes tipos de redes metabólicas. (a) un grupo de reacciones del ciclo de las pentosas. **(b)** Las redes centradas en compuestos conectan a los nodos que concurren como sustratos o productos en alguna reacción (cotejar con (a)). **(c)** En las redes centradas en enzimas, que son en las que nos enfocamos en este trabajo, los nodos son las enzimas, que se conectan cuando algún producto de una enzima es el sustrato de otra. **(d)** Las redes bipartitas incluyen tanto sustratos y productos, como enzimas y son una combinación de los dos tipos de redes anteriores.

2.1.3 Redes metabólicas

Se ha definido al metabolismo como el conjunto completo de transformaciones de moléculas orgánicas catalizadas por enzimas en las células vivas, suma del catabolismo y el anabolismo [55]. Sin embargo, tradicionalmente se le ha representado como grupos de reacciones, relativamente aisladas, que sintetizan o degradan ciertos tipos de compuestos; por ejemplo, en la síntesis de algún aminoácido, la degradación de ciertos lípidos, etc., conformando las llamadas rutas metabólicas. En las secciones anteriores hemos descrito que algunas propiedades topológicas y dinámicas de las redes metabólicas emergen sólo cuando se les estudia como un todo; por ejemplo, como se muestra en la Figura 6 hay enzimas que aparentemente son distantes —en cuanto al número de pasos metabólicos que las separan— e incluso han sido clasificadas en rutas metabólicas diferentes, pero en realidad están muy cercanas una de otra. Como se mencionó anteriormente se pueden reconstruir varios tipos de redes metabólicas a partir de la misma información (Figura 5). Aunque los primeros estudios sobre las propiedades topológicas del metabolismo se enfocaron en redes centradas en compuestos, consideramos que desde un punto de vista biológico es más adecuado usar las centradas en enzimas, o cuando se requiera mayor

precisión las bipartitas. Esto responde al hecho de que, por ejemplo, en los estudios de *knock-out* se eliminan genes (proteínas) no metabolitos, y lo primero se puede simular fácilmente eliminando nodos de las redes centradas en enzimas. En cambio, eliminar nodos en las redes centradas en compuestos equivaldría, por ejemplo, a agotar toda el agua o el ATP o el NAD de una célula lo cual es muy difícil. Desafortunadamente en las redes metabólicas algunos *hubs* pueden oscurece la interpretación de los resultados [56] por lo que se les puede "filtrar" dejando solo las enzimas, sustratos, productos y cofactores que se consideran el centro del flujo metabólico [57]. Alternativamente, se pueden eliminar *hubs* de forma sistemática para determinar alguna propiedad de interés y comparar los resultados obtenidos en cada ronda de eliminación de *hubs* [56]. De esta forma varios grupos han encontrado que diversas propiedades de las redes metabólicas como la conectividad, el agrupamiento y la distribución de flujos se ajustan a una leyes de potencias (Figura 7).

La Figura 8 muestra la red metabólica centrada en enzimas de *E. coli* construida eliminando los 20 compuestos más conectados (agua, ATP, NAD, etc) (los detalles se tratarán más adelante). Como se puede observar, esta red parece compleja a juzgar por el número de nodos que la componen, por lo cual es común pensar que las especies ancestrales tuvieron menos genes que las actuales [58], y por tanto, sus redes metabólicas pudieron ser más pequeñas. Una pregunta subyacente es ¿cómo han crecido y se han diversificado las redes biológicas y en particular las metabólicas?. La duplicación génica es ampliamente reconocida como generadora de variabilidad en diversos procesos biológicos [59]. En particular, es la materia prima de enzimas que diversifican las capacidades metabólicas de las especies (Figura 9). En el siguiente apartado detallamos algunos de los trabajos empíricos y analíticos que han estudiado a la duplicación génica como generadora de la topología y la diversidad de las redes metabólicas.

2.1.4 Duplicación génica

Tras una búsqueda exhaustiva en la literatura sobre el papel de la duplicación génica en la evolución del metabolismo, percibimos que la tendencia de la comunidad [60-62] es tratar de distinguir los casos en que la retención de genes duplicados, llamados a partir de ahora **duplicados**, se puede explicar bajo los preceptos de alguno de los dos modelos tradicionales, *paso-a-paso* [63] o *de-mosaico* [64]. Estos modelos tienen dos diferencias importantes: i) el modelo *paso-a-paso* (Figura 10a) sugiere que cuando un sustrato se agota en el medio, la duplicación génica puede generar una enzima capaz de proveer ese sustrato, dando lugar a reacciones consecutivas, catalizadas por enzimas codificadas por esos duplicados. En cambio, el modelo *de-*

mosaico (Figura 10a) sugiere que la duplicación de genes que codifican para enzimas promiscuas*

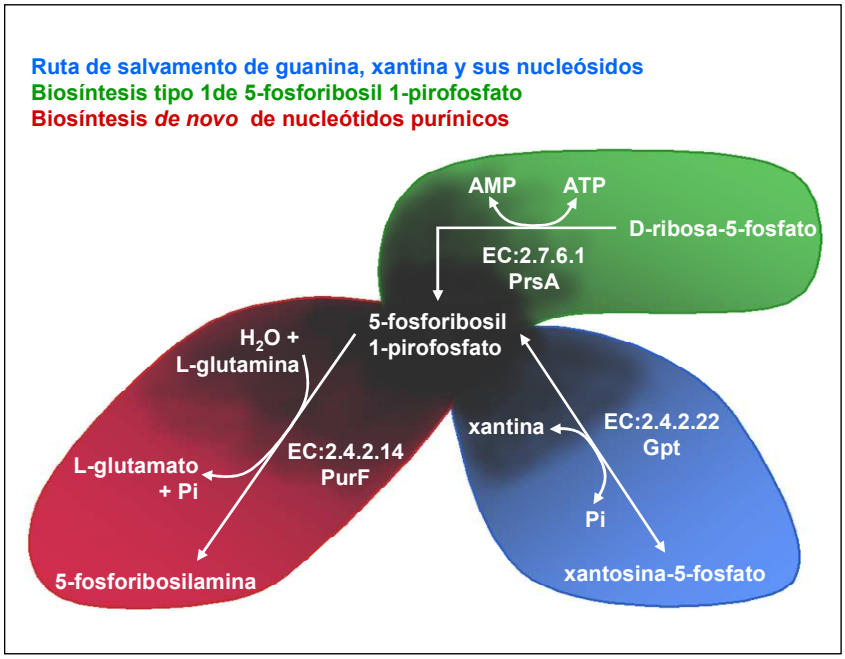
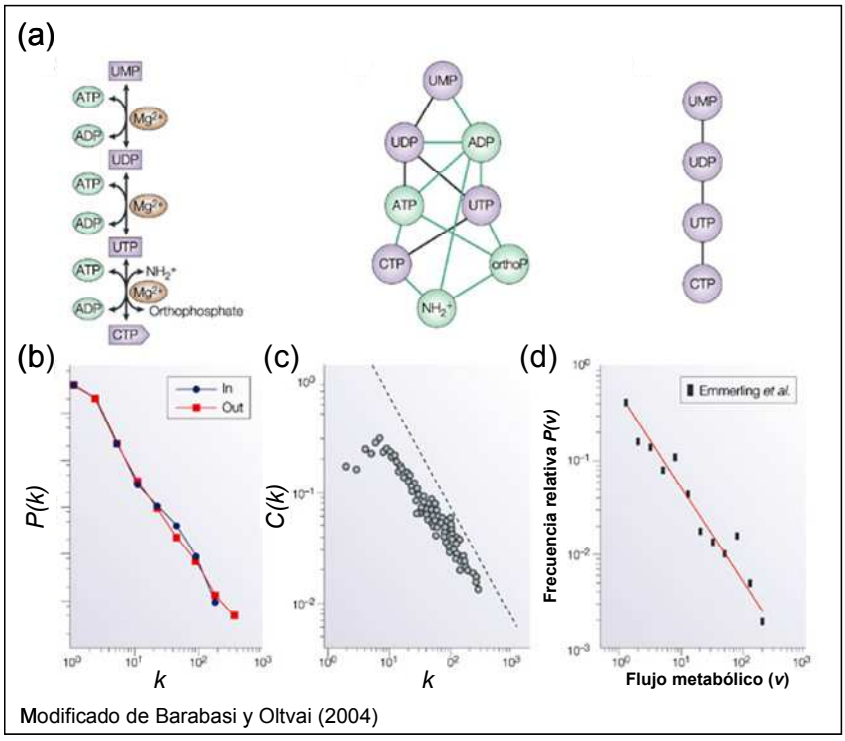


Figura 6. Una perspectiva de redes permite distinguir propiedades del metabolismo ocultas en la de rutas. Tres transferasas (PrsA, PurF y Gpt), generadas por duplicaciones génicas, catalizan reacciones consecutivas a pesar de estar clasificadas en rutas metabólicas distintas. Su cercanía metabólica es detectable sólo si se les analiza con una perspectiva de redes.



Modificado de Barabasi y Oltvai (2004)

Figura 7. Propiedades libres de escala en las redes metabólicas. (a) Construcción de una red centrada en compuestos (en medio) a partir de los datos de reacciones bioquímicas (izquierda), y su tratamiento para dejar solo el flujo metabólico central (derecha). **(b)** Con este tipo de redes se determinó que la conectividad del metabolismo es libre de escala, tanto en las conexiones que entran a los nodos (In) como en las que salen de ellos (Out). **(c)** De manera similar se encontró que la distribución del agrupamiento de nodos (C) en estas redes es libre de escala. **(d)** Con análisis de balance de flujos se encontró que en el metabolismo hay unas cuantas reacciones con un alto flujo metabólico y que la gran mayoría de ellas tienen un flujo bajo, siguiendo una ley de potencias.

* Las **enzimas promiscuas** son aquellas capaces de catalizar diferentes reacciones metabólicas usando el mismo dominio funcional.

mosaico (Figura 10) sugiere que la duplicación de genes que codifican para enzimas promiscuas* puede generar duplicados que se especializan en alguna de las catálisis del enzima original, con lo que las enzimas generadas según este modelo pueden ser más lejanas† que en el modelo *paso-a-paso*. ii) Dado que el modelo *paso-a-paso* involucra reacciones consecutivas, se considera [61] que puede generar reacciones químicamente diferentes (RQD), preservando las propiedades de unión por el tipo de sustrato. En contraste, se ha sugerido [61] que el modelo *de-mosaico* puede generar enzimas que catalizan reacciones químicamente similares (RQS), aun cuando actúen sobre sustratos de diferentes tipos. Una manera sencilla de determinar si dos reacciones son químicamente similares o no es comparando los primeros dos dígitos de sus número enzimáticos (EC:a.b.-.-) [60,62,65]. Varios autores [60,62,65] han usado las diferencias entre estos modelos para tratar de determinar su contribución en evolución del metabolismo, señalando al modelo *de-mosaico* como el predominante. A un par de años de iniciado este proyecto Light y Kraulis [65], en un estudio empírico que trató de cuantificar la duplicación génica en la red metabólica de *E. coli*, sugirieron que el origen de cuatro ligasas (MurC, MurD, MurE y MurF) de la biosíntesis del peptidoglicano, se explica con el modelo *de-mosaico*. Por las mismas fechas, nuestros resultados preliminares sugirieron que en realidad ambos modelos lo explican parcialmente (Figura 10b), porque estas ligasas actúan consecutivamente, explicando su origen con el modelo *paso-a-paso,* pero también catalizan reacciones químicamente similares, acorde al modelo *de-mosaico*. Procedimos a determinar qué tan comunes podrían ser los casos como éste y encontramos que, como se detalla más adelante, son tanto numerosos como diversos, lo cual representa un problema mayor con la forma en que se ha analizado la influencia de la duplicación génica en la evolución del metabolismo. Por lo cual, decidimos que no bastaba con extender la búsqueda de duplicados de estudios empíricos previos [60,62,65], hacia otras especies o usando algoritmos más poderosos, sino que debíamos introducir una perspectiva que ayudara a evitar este problema.

Del lado de los estudios analíticos sobre la evolución de las redes biológicas, los hay de dos tipos: i) los modelos basados en la duplicación génica que tratan de recrear la libertad de escala tanto en redes de interacciones proteína-proteína‡ [66-69], como en las de la regulación de la transcripción génica [70,71] y en las metabólicas [72]. Y ii) los que tratan de explicar la modularidad

---

* Las **enzimas promiscuas** son aquellas capaces de catalizar diferentes reacciones metabólicas usando el mismo dominio funcional.

† En este trabajo, la **distancia entre dos enzimas** es una medida del número de pasos metabólicos que separan las reacciones catalizadas por esas enzimas.

‡ En los modelos sobre la evolución de redes metabólicas por duplicación génica y divergencia se escoge sistemáticamente un nodo al azar y se duplica con una probabilidad (*D*), heredando sus conexiones al nuevo nodo. Luego, según el modelo se puede simular la divergencia haciendo que algún(os) nodos se reconecten al azar con una probabilidad (*R*) o que simplemente pierdan o ganen conexiones con una

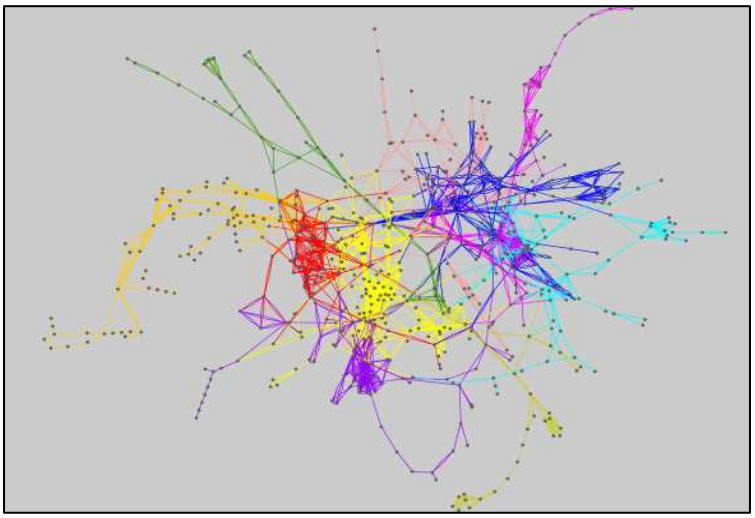


Figura 8. Red metabólica centrada en enzimas de *E. coli*. Los colores representan módulos estructurales identi-ficados con un agrupamiento jerárquico de los nodos. Al igual que en las redes centra-das en compuestos [19,50], se ha identificado que estos módulos reflejan grupos de enzimas con funciones concertadas. En esta red la porción amarilla, por ejemplo, corresponde al meta-bolismo de aminoácidos.

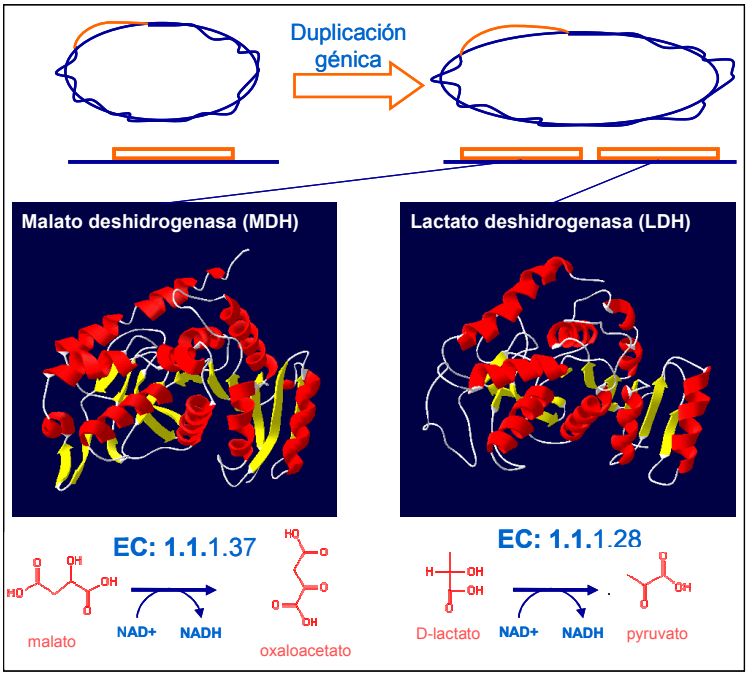
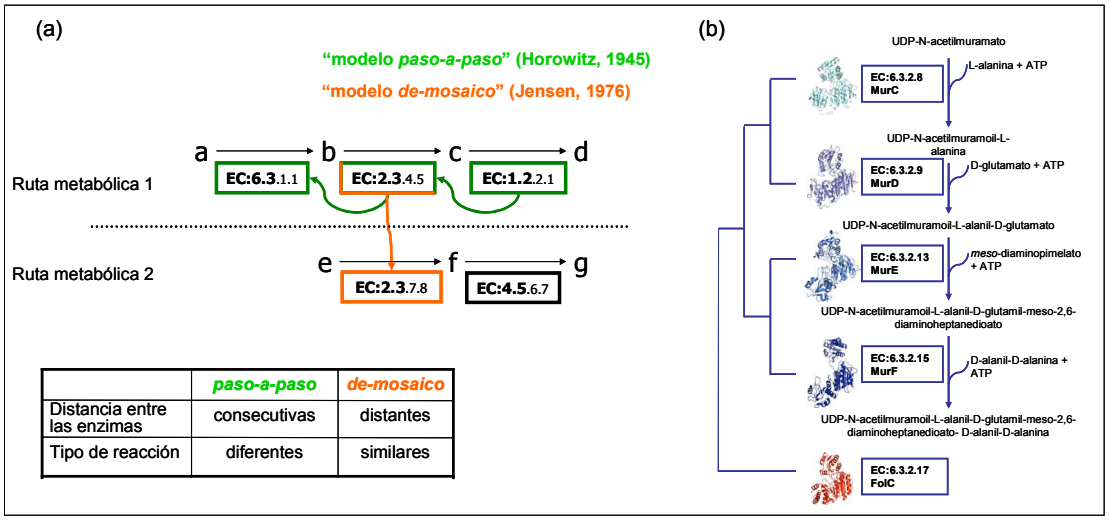


Figura 9. Duplicación génica. Una región del DNA de un organismo es duplicada. Con el tiempo, las sequencias de los genes duplicados divergen y pueden especializarse en la catálisis de alguna de las fun-ciones del gene ancestral (sub-funcionalización); o bien, ad-quirir nuevas funciones (neo-funcionalización), que pueden ser variantes de la ancestral o innovaciones. Por ejemplo, se ha sugerido [74] que las malato y lactato deshidrogenasas provie-nen de una subfuncionalización, sus estructuras y las catálisis que realizan son muy parecidas, como lo indican sus números EC. Aunque con el tiempo se han formado dos grupos (MDHs y LDHs), algunas MDHs pueden degradar tanto malato como D-lactato.

jerárquica diversos tipos de redes basándose en la duplicación sistemática de motivos estructurales [19,20,73]. En general, todos estos modelos recrean la libertad de escala de las redes reales, pero a excepción de los modelos de Goh et al. [67] y Pfeiffer et al. [72], en ellos no se toman en cuenta restricciones funcionales de las redes que tratan de abstraer [68] por lo que son válidos sólo parcialmente. Aunque se dice que algunos de estos modelos simulan el crecimiento de cierto tipo de redes, por ejemplo las de interacciones proteína-proteína [66,68,69], pudieran ser validos para cualquier tipo de red, pero al ser tan generales no podemos verificar que sean 'funcionalmente

probabilidad (N). El proceso se realiza reiteradamente hasta que la red alcanza el tamaño deseado y se pueden manipular las probabilidades D, R y N para obtener redes con libertad de escala y agrupamiento similares a las redes reales.

correctos' (i.e. qué tanto recrean la fuerza de los flujos o la estequiometría de las redes metabólicas). De manera similar, los modelos que simulan el origen de la modularidad en las redes pueden ser válidos tanto para redes metabólicas [19], como tecnológicas y de interacciones entre personas [20], e incluso para redes abstractas sin un representante específico en el mundo real [73], pero no se sabe qué tanto las redes resultantes son potencialmente funcionales. En contraste, los modelos que incluyen restricciones funcionales específicas [67,72] no solo tienen una funcionalidad potencial innata, sino que los parámetros de libertad de escala y agrupamiento que logran simular se ajustan mejor a los de las redes reales.



	paso-a-paso	*de-mosaico*
Distancia entre las enzimas	consecutivas	distantes
Tipo de reacción	diferentes	similares

Figura 10. Modelos tradicionales de la evolución del metabolismo por duplicación génica. (a) Las letras minúsculas representan compuestos transformados por distintas enzimas, cuyos números EC se enmarcan con colores que indican homología. Las flechas en verde muestran el sentido de la duplicación génica según el modelo *paso-a-paso* y la color naranja, lo propio para el modelo *de-mosaico*. La tabla de la parte inferior resume las propiedades que se ha sugerido [61] que deben reunir los duplicados generados por cada uno de estos modelos. Sin embargo, hay dos combinaciones que no se consideran: los pasos consecutivos-químicamente similares (ver panel b) y los pasos distantes-químicamente diferentes. **(b)** El origen de cuatro ligasas (MuC, MurD, MurE y MurF) de la biosíntesis de peptidoglicano puede explicarse parcialmente por ambos modelos, el *paso-a-paso* y el *de-mosaico*. Una quinta ligasa (FolC) está ampliamente distribuida en los tres dominios celulares, mientras que las otras cuatro ligasas están restringidas a Archaea y Bacteria, por lo tanto FolC pudiera ser el ancestro de la familia.

3. Planteamiento

A la luz de que las redes metabólicas son sistemas complejos y que los modelos actuales sobre su evolución no captan las propiedades funcionales que las caracterizan, en este trabajo nos planteamos determinar el papel de la duplicación génica en el crecimiento y diversificación de las redes metabólicas. Para ellos se necesita diseñar una estrategia que combine la teoría de redes con las herramientas bioinformáticas más poderosas a la fecha para detectar duplicados, y de esa forma tratar de determinar la relevancia de algunas propiedades topológicas y funcionales específicas de las redes metabólicas sobre la retención de duplicados.

4. Objetivo

Determinar de manera empírica la contribución de la duplicación génica en la evolución de las rutas metabólicas.

5. Hipótesis

Estudios previos han mostrado que algunas propiedades de las redes metabólicas, como la similitud química entre las reacciones y la distancia (número de pasos metabólicos) entre las enzimas que las catalizan influyen sobre la retención de duplicados. También se ha detectado una tendencia entre nodos cercanos de las redes metabólicas por agruparse en módulos; por lo tanto, esperamos encontrar una mayor retención de duplicados dentro de los módulos de estas redes que entre ellos. Los modelos existentes sobre la evolución de las redes biológicas que incorporan restricciones funcionales particulares a cada tipo de red tienen un mejor funcionamiento que los modelos generales, consideramos que si logramos abstraer algunas restricciones bioquímicas de las redes metabólicas y las incorporamos en los modelos sobre su evolución, las simulaciones obtenidas nos permitan entender mejor las presiones selectivas que actúan en el crecimiento y diversificación de las redes metabólicas por duplicación génica.

6. Justificación y alcance

Tratar de definir la contribución de los modelos *paso-a-paso* y *de-mosaico* en la evolución metabólica es conceptualmente deficiente, por lo que en este trabajo dejamos de lado sus nombres y nos enfocarnos en el mecanismo que hay detrás de ellos, la duplicación génica.

Con este fin, reconstruimos las redes metabólicas centradas en enzimas de distintas especies y comparamos las secuencias de sus enzimas para detectar duplicados. Los resultados enfatizan la influencia de dos propiedades de las redes metabólicas sobre la retención de duplicados: la similitud química de las reacciones y la distancia (número de pasos metabólicos) entre las enzimas que las catalizan. La perspectiva de redes empleada en este trabajo permite reconciliar los modelos *paso-a-paso* y *de-mosaico*, llevándolos de antagónicos a complementarios. También descubrimos de manera empírica que en las redes metabólicas existe un acoplamiento bioquímico preferente entre los tipos de reacciones que mejora significativamente los modelos sobre la evolución del metabolismo.

7. Resultados y Discusión

En este trabajo hemos cuantificado exhaustivamente la retención de genes duplicados en las redes metabólicas de diversas especies. Para ello, empleamos una estrategia que combina la teoría de grafos con poderosos algoritmos bioinformáticos para detectar duplicados. El primer paso, técnicamente hablando, fue la reconstrucción de las redes metabólicas de distintas especies. La Tabla I muestra que las propiedades generales de las redes reconstruidas —como son el número de nodos, las conexiones entre ellos y el diámetro de las redes— varían de una red a otra, incluso comparando las redes de la misma especie, por ejemplo las de *E. coli* (EcoCyc vs. EcoKegg). Esto es importante dado que refleja que, una vez que las bases de datos BioCyc [75,76] y KEGG [32] son construidas de forma distinta, la información que proveen también es diferente. Sin embargo, como se verá más adelante, nuestros resultados sobre la retención de duplicados no son afectados cualitativamente por estas diferencias.

Tabla I. Propiedades generales de las redes analizadas

	Hubs eliminados ╫	Número de nodos	Número de conexiones	Distancia mínima promedio §	Diámetro ‡
EcoKegg	0 hubs	833	4,485	6.01	26
	10 hubs	820	2,831	7.59	26
	20 hubs	804	2,410	8.87	27
	30 hubs	784	2,110	10.62	31
EcoCyc	0 hubs	1,076	59,936	2.52	8
	10 hubs	1,022	9,324	3.79	12
	20 hubs	976	4,473	6.01	16
	30 hubs	942	3,500	6.92	20
RefKegg	0 hubs	2,605	19,147	6.94	33
	10 hubs	2,587	12,996	7.76	37
	20 hubs	2,575	11,499	8.26	38
	30 hubs	2,548	10,322	8.70	38
MetaCyc	0 hubs	1,056	62,282	2.50	7
	10 hubs	1,002	9,275	3.73	10
	20 hubs	964	4,230	5.99	21
	30 hubs	944	3,214	6.92	23

╫ Cada red se reconstruyo eliminando gradualmente el número de compuestos altamente conectados (*hubs*) indicados en este campo (ver Materiales y Métodos).

§ Se obtuvieron las distancias mínimas (MPL, del inglés *minimal path length*) que separan a cada par de nodos dentro de una red. El valor aquí referido, corresponde al promedio de esas distancias.

‡ El diámetro de una red es la mayor de las distancias mínimas entre sus nodos. E1s decir, representa la distancia entre los nodos más separados en cada red.

7.1 El acoplamiento bioquímico preferente entre los tipos de reacciones en las redes metabólicas refleja una restricción funcional

Se ha demostrado que el metabolismo sigue reglas lógicas que implican que las reacciones y sus flujos no ocurren al azar, sino de manera específica en el tiempo y el espacio [77,78]. Dado que las redes metabólicas centradas en enzimas que reconstruimos constan de pares de reacciones consecutivas, lo primero que hicimos fue determinar si estos pares reflejan algunas de esas reglas. Para ello, determinamos las frecuencias de cada tipo de par de reacciones basándonos en sus números enzimáticos (EC:a.b.-.- → EC:w.x.-.-) y comparamos dichas frecuencias contra las esperadas al azar (Figura 11). Para calcular los valores esperados empleamos un grupo de modelos nulos, que son versiones reconectadas al azar de las redes reales, preservando el grado de conectividad de cada nodo de la red original, como lo sugieren Maslov y Sneppen [35] (ver Materiales y Métodos). Encontramos que algunos tipos de reacciones tienden a ser consecutivos, implicando que existen restricciones funcionales en el orden de las reacciones de las redes metabólicas, por ejemplo, para sintetizar peptidoglicano (Figura 10a), poliisoprenoides, AMP→ ADP → ATP, etc. Hemos llamado a este fenómeno **acoplamiento bioquímico preferente entre los tipos de reacciones.** Durante la preparación del artículo derivado de este trabajo [79] (Apéndice 1), el grupo de Palsson [80] describió que, al comparar como concurren los metabolitos en las reacciones en general, también estos se acoplan preferentemente en las redes metabólicas, complementándose con nuestros hallazgos. Es de especial importancia la diagonal de color rosa en la Figura 11, puesto que en ella se muestran en verde a amarillo, los casos en que los modelos *paso-a-paso* y *de-mosaico* potencialmente estarían en conflicto conceptual, como en el caso de las ligasas de la biosíntesis del peptidoglicano (flecha rosa en la Figura 11). Tomando en cuenta que estos casos no son anecdóticos, sino vastos y diversos, fue que decidimos dejar de lado, temporalmente, los nombres de esos modelos y enfocarnos en entender el papel de la duplicación génica en función de dos propiedades de las redes: i) la **similitud química entre las reacciones** (usando los primeros dos dígitos de los números enzimáticos) y ii) la **distancia entre las enzimas** (numero mínimo de pasos metabólicos entre ellas). Los siguientes párrafos versan alrededor de los hallazgos más relevantes derivados de estos análisis.

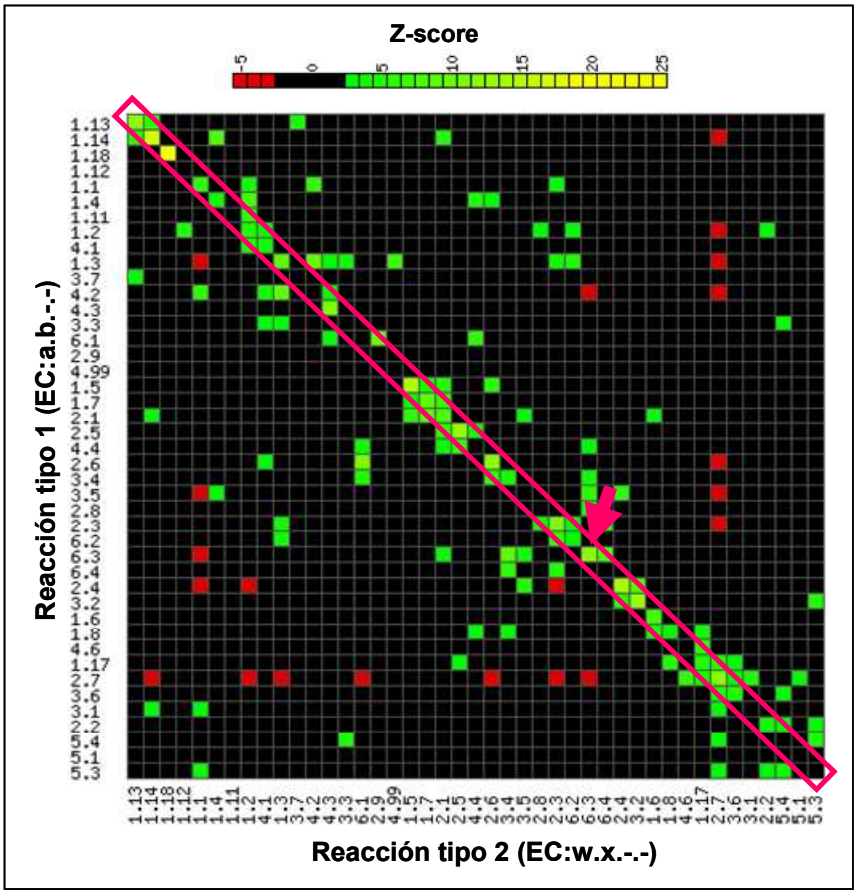


Figura 11. Acoplamiento bioquímico preferente entre los tipos de reacciones en las redes metabólicas. La frecuencia de cada par de tipos de reacciones consecutivas (EC:a.b.-.- → EC:w.x.-.-) observado en la red metabólica de *E. coli* (EcoKegg) se comparó contra las frecuencias esperadas, usando un grupo de modelos nulos tipo Maslov-Sneppen. Los pares de reacciones sobrerepresentados ($P < 0.001$) en esta red se muestran en verde-a-amarillo, mientras que los pares subrepresentados se muestran en rojo. La diagonal en rosa resalta los pares en que la retención de duplicados se ajusta tanto al modelo *paso-a-paso*, como al *de-mosaico*, como en el caso de las ligasas de la biosíntesis de peptidoglicano (Figura 3b). Se obtuvieron resultados equivalentes usando otras bases de datos, de *E. coli* (EcoCyc) y multiespecíficas (MetaCyc y RefKegg) (ver Materiales y Métodos).

7.2 Influencia de la similitud química entre las reacciones sobre la retención de duplicados

Hemos calculado que la tasa de retención de duplicados entre RQS es alrededor de seis veces mayor que entre RQD (Figura 12a). Esto es, *cuando dos reacciones tienen los primeros dos dígitos de sus números enzimáticos iguales (EC:a.b.-.- vs. EC:a.b.-.-) hay mayor probabilidad de que las enzimas que las catalizan provengan de una duplicación génica, que si sus números son diferentes*. Este hallazgo era esperado, dado que, en general, cuando un gene es duplicado, sus descendientes tienden a conservar la función del gene ancestral, en este caso las propiedades catalíticas de la enzima ancestral [81,82], correlacionando con reportes previos [60,62,65]. Sin embargo, un análisis del significado estadístico de estos resultados, usando modelos nulos tipo Maslov-Sneppen (círculos rojos en la Figura 12), mostró que tanto la retención de duplicados entre RQS,

como entre RQD cercanas es mayor de lo esperado al azar (Z-score > 3, $P < 0.001$). Esto, deja en manifiesto que si bien las frecuencias por sí solas pueden ser informativas, también lo es su significado estadístico, que es algo que en estudios previos se había omitido [62]. Las principales implicaciones de este hallazgo son que la retención de duplicados es un proceso que no ocurre al azar, y que es importante no sólo para generar variantes en el metabolismo (RQS), sino también innovaciones (RQD).

7.3 Influencia de la distancia entre las enzimas sobre la retención de duplicados

Además de la retención de duplicados que generan tanto RQS como RQD, la Figura 12 muestra un *incremento significativo (Z-score > 3, P < 0.001) en la retención de duplicados entre enzimas que catalizan reacciones cercanas*. Por cercanas, nos referimos a las reacciones consecutivas o con pocos pasos metabólicos que las separan. Este fenómeno había sido reportado previamente [60,65]; sin embargo, hasta donde sabemos, no se había descrito alguna propiedad del metabolismo que pudiera explicar dicho comportamiento. Cuando encontramos este fenómeno en todas las redes metabólicas analizadas (EcoCyc, EcoKegg, MetaCyc y RefKegg), consideramos que debíamos dedicar mayores esfuerzos a tratar de entenderlo. Por lo cual, lo primero que hicimos fue analizar otros tipos de redes biológicas, de forma similar a las metabólicas, para determinar si este fenómeno es universal, o bien, característico del metabolismo.

Analizamos dos tipos de redes, una de regulación de la transcripción génica [47] y otra de interacciones proteína-proteína [36], ambas de *E. coli*. En el caso de la red transcripcional encontramos que la retención de duplicados es independiente de la distancia entre los factores de transcripción y las proteínas reguladas (Figura 12c). De hecho, un hallazgo interesante fue que los modelos nulos tipo Maslov-Sneppen mostraron que la retención de duplicados en esta red es globalmente menor a lo que se esperaría por azar. Mientras analizábamos estos resultados, dos grupos [83,84] reportaron que en términos evolutivos, la función original de los factores de transcripción se pierde fácilmente, y que otros fenómenos, como la convergencia funcional, pudieran explicar mejor como se han ensamblado las redes transcripcionales. En lo referente a las redes de interacciones proteína-proteína encontramos, al igual que en las redes metabólicas, una mayor retención de duplicados entre nodos cercanos, en este caso, entre proteínas que interactúan físicamente o con pocos nodos entre ellas (Figura 12d). Un análisis más detallado nos permitió determinar que las interacciones enzima-enzima son en realidad las que provocan este sesgo, puesto que lo propio para las interacciones entre proteínas no-enzimáticas no supera lo esperado

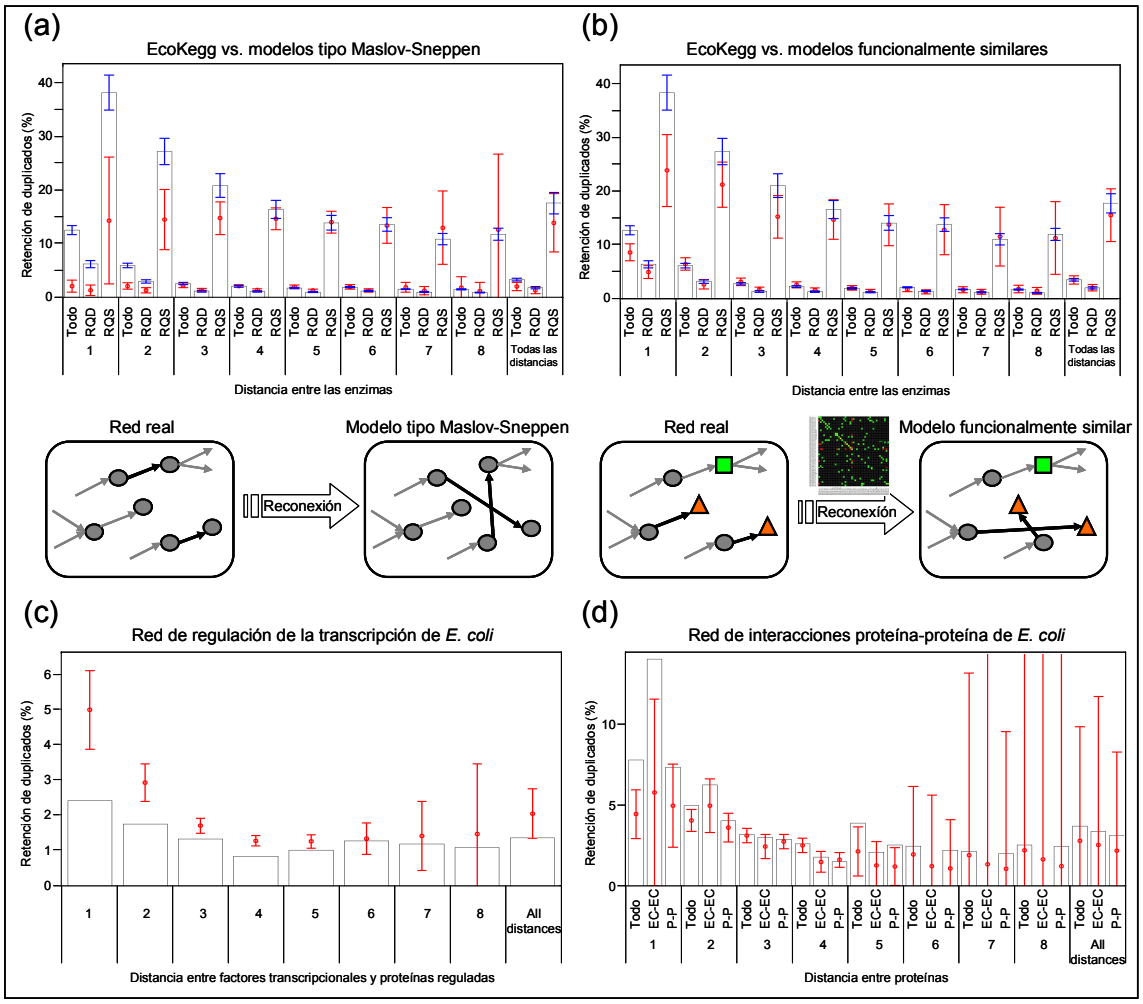


Figura 12. Influencia de la similitud química y la distancia entre las reacciones sobre la retención de duplicados. **(a)** Frecuencias de retención de duplicados (histograma) observadas en EcoKegg para el conjunto completo de reacciones (Todo), y para los subconjuntos de reacciones químicamente diferentes (RQD) y químicamente similares (RQS) en función de la distancia entre las enzimas (numero de pasos metabólicos). Las líneas azules indican tres desviaciones estándar (σ) de estas frecuencias, obtenidas con muestreos aleatorios (ver ver Materiales y Métodos). Los puntos rojos representan los valores esperados, ± 3 σ, usando un grupo de modelos nulos tipo Maslov-Sneppeñ El proceso de reconexión para generar estos modelos se esquematiza en la parte inferior y se detalla en Materiales y Métodos. **(b)** Un procedimiento similar que en (a) se realizó para construir los modelos nulos funcionalmente similares, con los cuales determinamos la influencia del acoplamiento bioquímico preferente entre los tipos de reacciones sobre la retención de duplicados. Nótese el incremento en la retención de duplicados esperados comparando los puntos rojos a distancias cortas en (b) contra sus equivalentes en (a). **(c)** Retención de duplicados en una red de regulación transcripcional como función de la distancia (número de interacciones regulatorias) entre factores de transcripción y proteínas (genes) reguladas. **(d)** Retención de duplicados en una red de interacciones proteína-proteína. Se muestra el conjunto de todas las interacciones (Todo) y los subconjuntos de interacciones enzima-enzima (EC-EC) y de proteína-proteína no-enzimáticas (P-P). En (c) y (d) los puntos rojos corresponden a los promedios esperados, usando modelos tipo Maslov-Sneppen.

al azar (Z-score < 3, $P > 0.001$). Lo que hace sentido con lo hallado en las redes metabólicas, considerando que las interacciones físicas enzima-enzima tienden a reflejar reacciones cercanas o componentes de complejos multiméricos. Por lo anterior, propusimos que las leyes que gobiernan las interacciones substrato-enzima-producto son diferentes, y tras una duplicación génica son retenidas en mayor medida, que las que operan en las interacciones proteína-DNA y proteína-proteína no-enzimáticas. Una posible explicación de esto es que las superficies involucradas en las interacciones metabólicas son más pequeñas (involucran sustratos pequeños) que las superficies de interacción proteína-proteína o proteína-DNA. Esta hipótesis correlaciona con los hallazgos de otros grupos, que describen que, en términos evolutivos, las interacciones proteína-DNA se pierden rápidamente [83,84], en cambio, las interacciones proteína-proteína, y en mayor grado las enzima-enzima, se conservan [85]. En síntesis, con estos resultados entendimos que *las redes metabólicas presentan una alta retención de duplicados entre enzimas cercanas. Y que este fenómeno es característico de redes que involucran enzimas*. Aunque no sabíamos el por qué de este fenómeno, sí sabíamos que la respuesta podía estar en las propiedades que distinguen a las redes metabólicas de otro tipo de redes.

La pregunta subyacente fue entonces: ¿Qué distingue a las redes metabólicas de otro tipo de redes que pueda explicar una mayor retención de duplicados entre enzimas cercanas? Como se describió anteriormente, las redes metabólicas muestran un acoplamiento bioquímico preferente entre los tipos de reacciones que las conforman. Así que construimos un nuevo grupo de modelos nulos "funcionalmente similares" que mantienen, además del grado de conectividad, las proporciones de cada tipo de pares de reacciones de la red original (ver Materiales y Métodos). La retención de duplicados simulada con los modelos tipo Maslov-Sneppen (círculos rojos en la Figura 12a) muestra un comportamiento independiente de la distancia entre las enzimas. En cambio, usando los modelos "funcionalmente similares" (círculos rojos en la Figura 12b) se observa un incremento en la retención de duplicados entre enzimas cercanas, ajustándose mejor a lo que ocurre en las redes reales. Con esto, demostramos que ***el incremento en la retención de duplicados entre enzimas cercanas se explica, en parte, por el acoplamiento bioquímico preferente entre los tipos de reacciones, lo que es probablemente la mayor contribución de este trabajo al entendimiento de la evolución de las redes metabólicas.*** Dado que este acoplamiento preferente no ocurre en las redes transcripcionales, ni en las redes de interacciones proteína-proteína (no enzimáticas), es entendible que en ellas tampoco se observara el sesgo en la retención de duplicados entre nodos cercanos.

7.4 Influencia de la modularidad de las redes sobre la retención de duplicados

Se ha reportado que las redes metabólicas poseen una modularidad estructural, por la cual conjuntos de nodos tienden a estar más conectados entre ellos que con otros de la red, y que cada módulo puede contener una o varias rutas metabólicas funcionalmente relacionadas [19,51]. Esto es que: un módulo puede encargarse del metabolismo de aminoácidos, otro de los lípidos, otro de los nucleótidos, etc. Como los nodos de un módulo están altamente conectados, las distancias entre ellos son relativamente pequeñas [86]. Lo cual, nos llevo a platear la hipótesis de que, dada una alta retención de duplicados entre enzimas cercanas (Figura 12), podíamos esperar una mayor retención de duplicados entre las rutas metabólicas de un mismo módulo, que entre las rutas de diferentes módulos. Para abordar esta hipótesis detectamos los módulos de las redes analizadas (Figura 13a), agrupando jerárquicamente sus nodos en función de la distancia entre ellos (ver Materiales y Métodos). Y luego, usando una medida de distancia evolutiva (DE), que indica la tasa de retención de duplicados entre dos rutas metabólicas, del mismo o de diferentes módulos (ver Materiales y Métodos), calculamos (DE) para todas-contra-todas las rutas metabólicas de cada red. Cabe señalar que (DE) no es la distancia entre dos enzimas (número de pasos metabólicos), sino una medida de qué tanto dos rutas metabólicas retienen duplicados entre ellas. Nuestra medida (DE) fue modificada del la que originalmente el grupo de Doolittle usó para determinar la distancia evolutiva entre especies, en función de su contenido de dominios proteicos [87]. La Figura 13b y c muestra que las rutas metabólicas del mismo módulo tienden a presentar, significativamente (Z-score > 3, $P < 0.001$) una menor (DE), o sea, una mayor retención de duplicados, que las rutas de diferentes módulos. Por ejemplo, la red metabólica completa de *E. coli* presenta alrededor de un 15% de retención de duplicados entre RQS (Figura 12a); en cambio, al extraer el módulo encargado del metabolismo de aminoácidos de esta red (porción amarilla en las Figura 13a y b), y calcular la retención de duplicados dentro de él, encontramos que el porcentaje de retención de duplicados es de un 50%. Este incremento, nos habla de los duplicados retenidos preferentemente dentro de cada módulo, como las transferasas dependientes de pirodoxal-fosfato del metabolismo de aminoácidos. Estamos llevando a cabo un estudio más profundo para entender la evolución de este módulo (Hernandez-Montes G. et al. *Trabajo en proparación*). A la fecha, podemos relacionar una alta retención de duplicados dentro de cada módulo con el hecho de que las reacciones de cada módulo emplean sustratos y cofactores similares, por ejemplo, el piridoxal-fosfato en el metabolismo de aminoácidos. En síntesis, sugerimos que *la capacidad de las redes metabólicas para crecer modularmente, por duplicación génica, está íntimamente relacionada con dos factores: i) la cercanía entre las enzimas que conforman cada módulo y, ii) la conservación de las interfases enzima-tipo de sustrato privativas*

a cada módulo. Como se verá en el siguiente apartado, esto ha tenido gran relevancia en la evolución de las rutas metabólicas que involucran sustratos y reacciones químicamente similares. Queda como una perspectiva cuantificar la similitud de los sustratos de cada módulo y cotejarla con la tasa de retención de duplicados. Desafortunadamente, hasta donde sabemos, actualmente no existe una clasificación de los tipos de sustratos, análoga a los números ezimáticos, que nos permitan llevar a cabo este análisis.

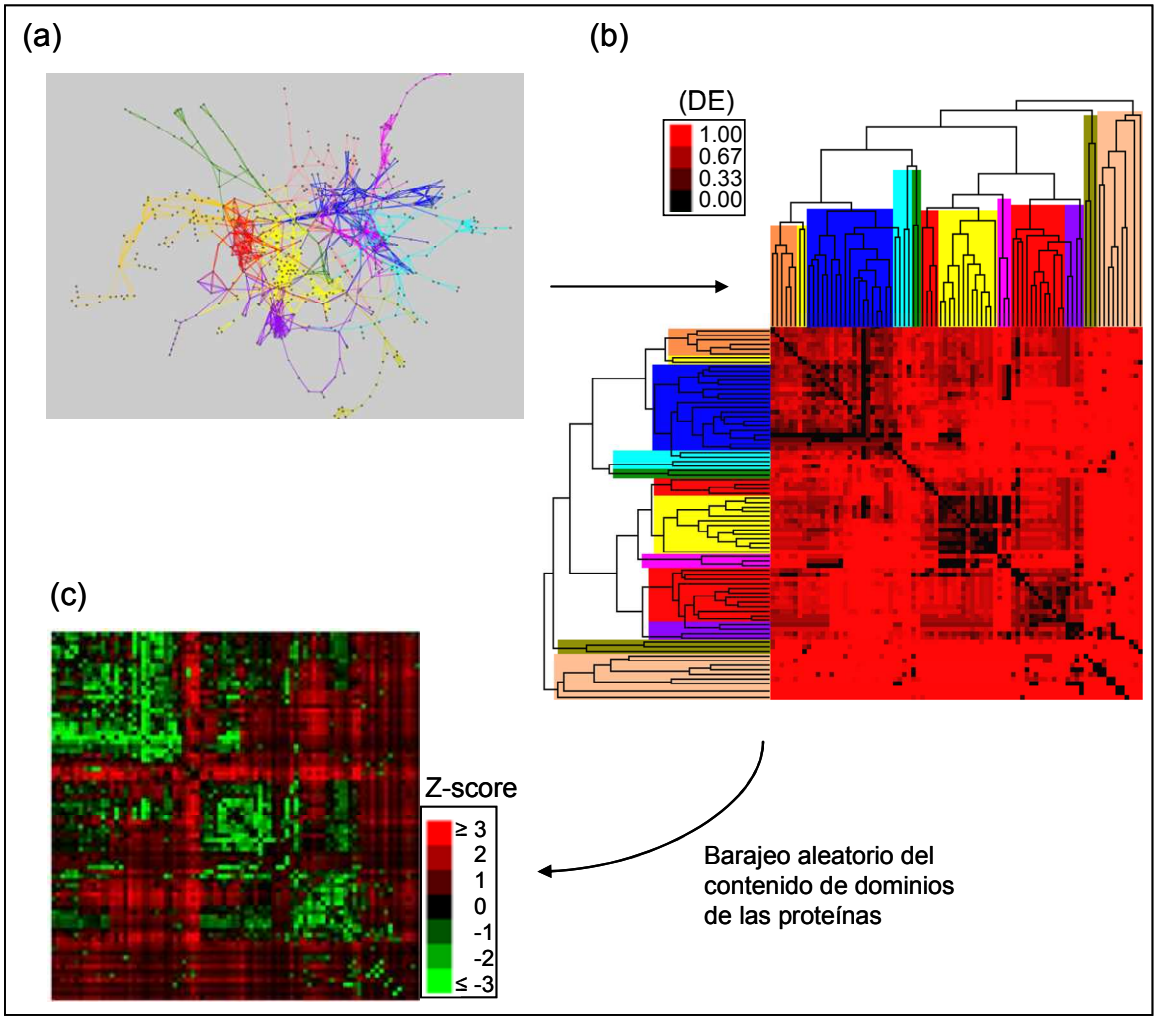


Figura 13. Influencia de la modularidad de las redes sobre la retención de duplicados entre las rutas metabólicas. (a) Se realizó un agrupamiento jerárquico para identificar los módulos en EcoKegg. Cada color representa un módulo; por ejemplo, la porción amarilla corresponde al metabolismo de amino ácidos. **(b)** Se realizó una comparación pareada de todas-contra-todas las rutas metabólicas (ramas de los árboles) dentro de cada módulo y entre módulos, para determinar la tasa de duplicados que retienen entre si. A mayor retención de duplicados, menor valor de (DE) (ver Materiales y Métodos). **(c)** Los valores observados de (DE) se compararon contra los esperados, usando un conjunto de 1000 modelos nulos, en los que se barajeo el contenido de dominios de las proteínas (ver Materiales y Métodos), mientras que la red permaneció intacta. Un valor de Z-score ≤ -3 implica que hay más retención de duplicados entre las rutas en cuestión que lo esperado al azar.

7.5 Retención de duplicados como grupos y como entidades separadas

Hemos mostrado que hay dos factores importantes en el crecimiento de las redes metabólicas, la cercanía de las enzimas y la conservación del tipo de sustrato. En un estudio previo [88], determinamos que las rutas de biosíntesis y degradación de ácidos grasos (β-oxidación), son similares bioquímicamente hablando. Como se muestra en la Figura 14a, en realidad una ruta es la imagen especular de la otra, utilizan sustratos y cofactores parecidos, por lo que sus números EC son también parecidos, aunque actúan en sentido inverso y el acarreador de grupos acilos es distinto, la biosíntesis emplea una proteína acarreadora de acilos, mientras que la degradación usa coenzima A. Pues bien, cuando comparamos de forma pareada las enzimas que catalizan RQS equivalentes en estas rutas, encontramos que cada par parece provenir de una duplicación génica. Por lo anterior, proponemos que las rutas contemporáneas de biosíntesis y degradación de ácidos grasos se originaron con la duplicación de los genes de una ruta ancestral, que se han retenido como grupo. Es posible que, dado que los acarreadores de acilos distinguen a las rutas actuales, el flujo de la ruta ancestral, para sintetizar o degradar ácidos grasos, fuera dependiente de la concentración de la proteína acarreadora de acilos y coenzima A en el medio.

Para tener una primera aproximación de la generalidad de esta observación en otras rutas metabólicas químicamente similares, comparamos todas-contra-todas las enzimas que catalizan RQS en pares de reacciones consecutivas. Los resultados de este análisis (Figura 14b) indican que alrededor del 15% de las enzimas tienen al menos un duplicado dentro de la redes metabólicas a la que pertenecen. De ese porcentaje, una tercera parte se ha retenido como grupos de duplicados (escenario III en la Figura 14b), como en el metabolismo de ácidos grasos, y dos tercios como entidades separadas (escenario II en la Figura 14b). Interesantemente, la retención de ambos, grupos y entidades separadas, es significativamente mayor que lo esperado al azar (Z-scores > 50, $P < 0.001$), mientras que los casos en que no se detectó retención de duplicados, son menos de lo esperado (Z-scores < -20, $P < 0.001$). En resumen, estos datos muestran que *la similitud química entre reacciones es importante no solo para retener duplicados separados, sino grupos de ellos que actúan consecutivamente, generando incluso rutas metabólicas completas*. Por tanto, sugerimos que los modelos que traten de explicar la evolución de las redes metabólicas deben contemplar tanto la retención de duplicados separados, como de grupos de duplicados.

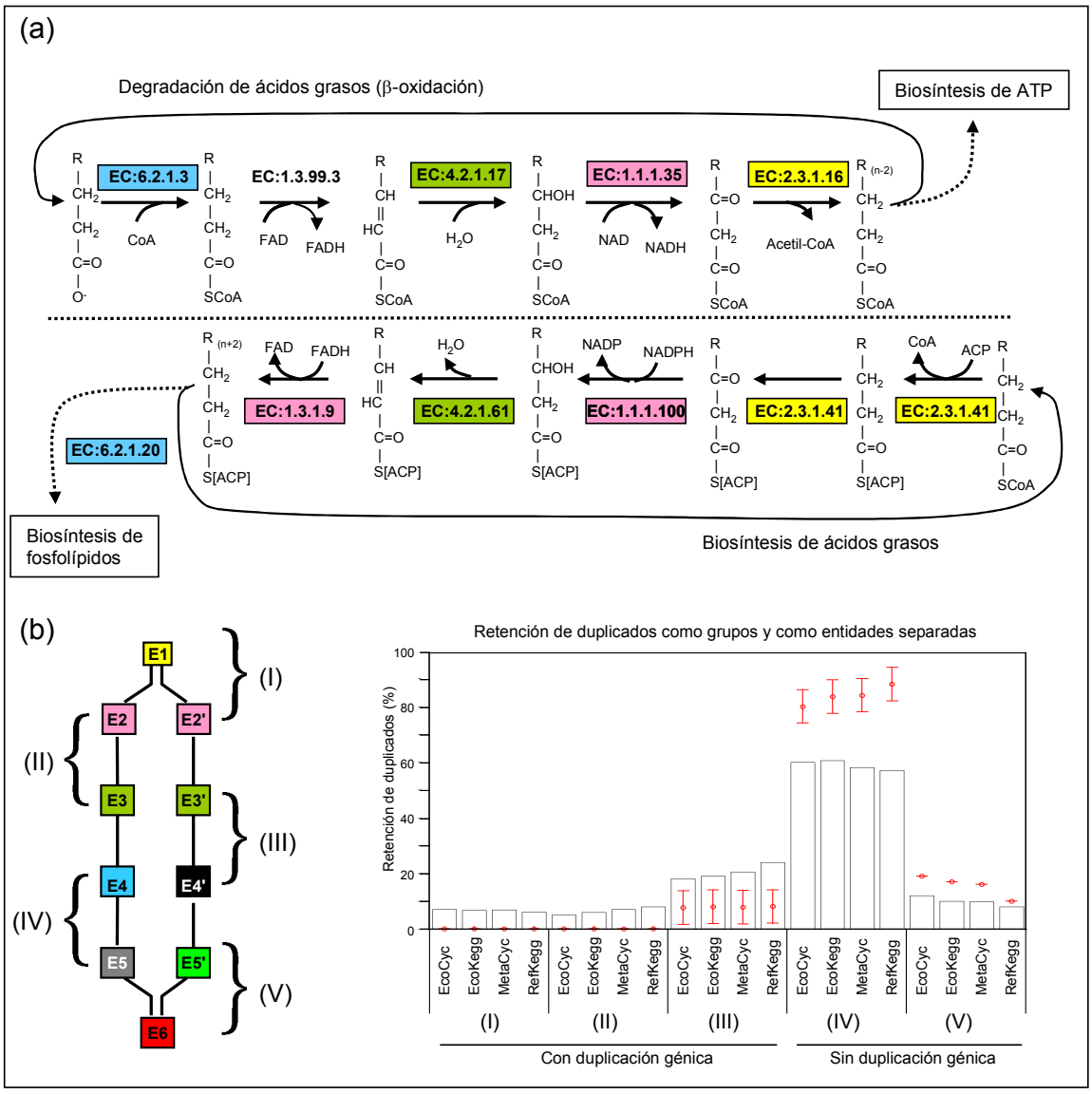


Figura 14. Retención de duplicados como grupos y como entidades separadas. (a) Las rutas de biosíntesis y degradación de ácidos grasos ilustran la retención de duplicados como grupos. Los colores de las cajas de los números EC indican que enzimas son duplicados. **(b)** Retención de duplicados en reacciones consecutivas. Se analizaron los cinco escenarios posibles (panel izquierdo). Los colores de las cajas indican homología. El número y letra indican el orden de las reacciones. Los escenarios (I) y (V) tienen una reacción en común, seguida o precedida por otras dos. En (I), se detectó duplicación génica, en (V) no se detectó. Los escenarios (II, III y IV) involucran pares de reacciones en dos ramas de la red. En (II) ambos pares son duplicados, en (III) solo un par es duplicado, y en (IV) ninguno de los pares es duplicado. En este diagrama se puede observar que un par puede participar en más de un escenario, río arriba y abajo del flujo de la red. El histograma de la derecha muestra las frecuencias observadas para cada escenario en cada una de las cuatro bases de datos analizadas en este trabajo, reconstruidas eliminando los 20 *hubs* más conectados. Y corresponden a la comparación de todos-contra-todos los pares de reacciones consecutivas, tanto RQS como RQD. Los puntos rojos muestran los valores esperados, ± 3 σ, usando un grupo de 1000 modelos nulos tipo Maslov-Sneppen.

7.6 Controles de algunas propiedades de las redes y las estrategias de detección de duplicados

En las redes metabólicas existen compuestos altamente conectados, como el agua, ATP, NAD, etc. llamados *hubs*, que pueden oscurecer el entendimiento de las propiedades topológicas de las redes [56]. Una de las principales propiedades sobre las que los *hubs* influyen es la distancia entre los nodos, al servir como atajos entre los nodos, de otro modo, distantes. Dado que, en este trabajo analizamos la influencia de la distancia entre los nodos sobre la retención de duplicados, un control imperante fue evaluar el papel de los *hubs* en nuestros hallazgos. Los resultados presentados en las secciones anteriores derivan de la red de *E. coli*, según ECOCYC, reconstruida eliminando los 20 *hubs* más conectados. En el Apéndice 2 se muestra que la eliminación gradual de *hubs*, durante la reconstrucción de esta y otras redes metabólicas, no cambia nuestros hallazgos cualitativamente. Es decir, la alta retención de duplicados entre enzimas cercanas se mantiene significativa (Z-score > 3, $P < 0.001$).

Un segundo control fue evaluar si la alta retención de duplicados entre enzimas cercanas, está restringida a porciones de la red; o bien, si está distribuida por toda ella. Para evaluar esto, hicimos 10,000 muestreos aleatorios, cada uno representando el 50% de la red original, y para cada muestra repetimos los cálculos de la retención de duplicados. Los resultados, que se muestran como el promedio de los valores obtenidos de los muestreos (líneas azules en la Figura 12a y b), nos permitieron concluir que no hay un sesgo significativo entre los valores obtenidos para la red completa y los muestreos. Por lo tanto, la alta retención de duplicados entre enzimas cercanas no está restringida a porciones de la red, sino que es global.

El tercer control se deriva del hecho de que algunas enzimas son multidominio y al compararlas se pueden obtener falsos positivos. Supongamos que existen dos enzimas E1 y E2, y que E1 tiene un solo dominio (d1); en cambio, E2 tiene dos dominios (d2 y d3). Ahora, supongamos que queremos saber si d1 y d2 provienen de una duplicación génica, si simplemente comparamos las secuencias completas de E1 y E2 podríamos obtener un resultado en el que, en efecto E1 y E2 sean homólogas, pero no en sus dominios d1 y d2, sino en d1 y d3. Con lo que tendríamos un resultado falso positivo. Los análisis iniciales de este proyecto mostraron que la cantidad de falsos positivos podría ser del mismo orden de magnitud que los verdaderos positivos, por lo que para evitar estas confusiones, al inicio del proyecto, recortamos las secuencias de las enzimas analizadas en este trabajo en dominios funcionales, según sus anotaciones en la base de datos SWISS-PROT [89]. Dado que este procedimiento fue manual, consumió alrededor de un semestre el poder delimitar los 4,537 dominios con que se realizó el estudio, y están disponibles en la versión electrónica del artículo derivado de este trabajo [79]. Aun después de esto, quedaba la posibilidad de que los dominios funcionales no anotados en Swiss-

Prot sesgaran nuestros hallazgos. Para determinar si era el caso, realizamos dos controles aumentando la astringencia del método usado para detectar duplicados. El primero fue restringir las comparaciones a aquellas enzimas con un solo dominio (Apéndice 3a), y el segundo consideró como duplicados solo a aquellas enzimas que compartieran todos sus dominios homólogos (Apéndice 3b). En conjunto, estos controles mostraron que, aun aumentando la astringencia del método para detectar duplicados, la alta retención de duplicados entre enzimas cercanas permaneció significativa (Z-score > 3, $P < 0.001$).

El cuarto control se derivó del método empleado para la búsqueda de homología (duplicados). A lo largo del proyecto, usamos el programa HMMER para comparar los 4,537 dominios contra una batería de modelos ocultos de Markov, que representan familias y superfamilias de dominios homólogos, según las bases de datos PFAM y SUPERFAMILY (ver Materiales y Métodos). Aunque el 95% de los 4,537 dominios tienen al menos una región representada en estas bases de datos, otras de sus regiones pudieran no estarlo, y no sabíamos que tanto ello sesgaría nuestros hallazgos. Para controlar lo anterior, empleamos dos algoritmos de búsqueda de homología, BLAST, que ayuda a detectar homología cercana, y su contraparte iterativa, PSI-BLAST, que detecta homología remota. Si bien, estos algoritmos detectan entre 10 y 20% menos homólogos que HMMER + PFAM + SUPERFAMILY [90], permiten comparar la totalidad de las secuencias de los dominios. El principal hallazgo de este control (Apéndice 3c y 3d) es que la alta retención de duplicados entre enzimas cercanas permaneció significativa (Z-score > 3, $P < 0.001$) aun considerando solamente los homólogos cercanos.

Finalmente, es necesario mencionar que el criterio que empleamos para determinar si dos reacciones son químicamente similares usa los primeros dos dígitos de sus números enzimáticos (EC:a.b.-.-). El primer dígito se refiere a uno de los seis posibles tipos de reacción en general, el segundo se refiere al tipo de enlace sobre el cual ocurre la reacción, el tercero describe al cofactor empleado, y el cuarto es específico para el sustrato transformado. Por lo cual, como estándar se usan los primeros dos dígitos como indicadores de similitud química [60,62,65]. Pero era necesario saber si siendo más laxos, usando solo el primer dígito (EC:a.-.-.-), o más estrictos, usando los tres primeros (EC:a.b.c.-) se afectaban las conclusiones derivadas de nuestros resultados. Como era de esperarse, las tasa de retención de homólogos se modificó dependiendo del criterio de similitud química (Apéndice 3e y 3f), pero la alta retención de duplicados entre enzimas cercanas permaneció significativa (Z-score > 3, $P < 0.001$).

En conjunto, estos controles corroboran nuestro hallazgo inicial de que *en las redes metabólicas de diversas especies, y construidas de diferente forma, se puede detectar una alta retención de duplicados entre enzimas cercanas*. Consideramos que otros controles, usando diferentes algoritmos, y con una depuración más detallada de las bases de datos de secuencias y

dominios, podrían ayudarnos a precisar la tasa de retención de duplicados, pero no cambiarían cualitativamente este hallazgo.

8. Conclusiones

En este trabajo hemos usado una perspectiva de redes, centradas en enzimas, para determinar las frecuencias y el significado estadístico de la retención de duplicados en el metabolismo de varias especies, usando diferentes tipos de modelos nulos. *Colectivamente, nuestros resultados enfatizan la importancia de dos propiedades de las redes metabólicas sobre la retención de duplicados: i) la distancia entre las enzimas y, ii) la similitud química entre las reacciones.* Específicamente, encontramos una alta retención de duplicados entre enzimas cercanas (Figura 12a y b), y *demostramos que el acoplamiento bioquímico preferente entre los tipos de reacciones de las redes metabólicas explica parcialmente este fenómeno*. Los análisis similares en las redes de regulación de la transcripción y de interacciones proteína-proteína mostraron que dicho comportamiento es exclusivo de redes que involucran relaciones enzimáticas, por lo cual proponemos que las leyes que gobiernan las interacciones sustrato-enzima-producto, son diferentes de las que rigen a las interacciones proteína-DNA y proteína-proteína no-enzimáticas (Figura 12c y d). Esto se refleja en una mayor retención de duplicados dentro de los módulos funcionales de las redes, que entre ellos (Figura 13b y c). Adicionalmente, nuestros resultados muestran que la retención de duplicados genera significativamente tanto variantes metabólicas (RQS), como innovaciones (RQD). Un efecto sinérgico de la distancia y la similitud química entre las reacciones, que parecen ser los factores más importantes en la retención de duplicados, promueve que en RQS consecutivas exista la tasa de retención de duplicados más alta del metabolismo (~35%) (Figura 12a). Además, hemos encontrado que los duplicados se han retenido significativamente tanto en grupos como en entidades separadas, un fenómeno que, hasta donde sabemos no había sido explorado.

Es común oír que cierto caso de duplicación génica "es originada por el modelo de Horowitz (*paso-a-paso*) o por el modelo de *patchwork* (*de-mosaico*)" [60-62,65]; sin embargo, estos son modelos, no mecanismos. Un modelo es una herramienta conceptual que abstrae los componentes principales de un fenómeno para entenderlo, por lo que estas expresiones, como tales, son erróneas. Y más allá de dificultades semánticas, usando una perspectiva de redes, hemos mostrado que la duplicación génica puede y debe ser considerada como un solo proceso, que depende en gran medida de la distancia entre las enzimas y la similitud química entre las reacciones. Con lo cual reconciliamos estos modelos, al pasarlos de antagónicos a complementarios.

dominios, podrían ayudarnos a precisar la tasa de retención de duplicados, pero no cambiarían cualitativamente este hallazgo.

8. Conclusiones

En este trabajo hemos usado una perspectiva de redes, centradas en enzimas, para determinar las frecuencias y el significado estadístico de la retención de duplicados en el metabolismo de varias especies, usando diferentes tipos de modelos nulos. *Colectivamente, nuestros resultados enfatizan la importancia de dos propiedades de las redes metabólicas sobre la retención de duplicados: i) la distancia entre las enzimas y, ii) la similitud química entre las reacciones.* Específicamente, encontramos una alta retención de duplicados entre enzimas cercanas (Figura 12a y b), y *demostramos que el acoplamiento bioquímico preferente entre los tipos de reacciones de las redes metabólicas explica parcialmente este fenómeno*. Los análisis similares en las redes de regulación de la transcripción y de interacciones proteína-proteína mostraron que dicho comportamiento es exclusivo de redes que involucran relaciones enzimáticas, por lo cual proponemos que las leyes que gobiernan las interacciones sustrato-enzima-producto, son diferentes de las que rigen a las interacciones proteína-DNA y proteína-proteína no-enzimáticas (Figura 12c y d). Esto se refleja en una mayor retención de duplicados dentro de los módulos funcionales de las redes, que entre ellos (Figura 13b y c). Adicionalmente, nuestros resultados muestran que la retención de duplicados genera significativamente tanto variantes metabólicas (RQS), como innovaciones (RQD). Un efecto sinérgico de la distancia y la similitud química entre las reacciones, que parecen ser los factores más importantes en la retención de duplicados, promueve que en RQS consecutivas exista la tasa de retención de duplicados más alta del metabolismo (~35%) (Figura 12a). Además, hemos encontrado que los duplicados se han retenido significativamente tanto en grupos como en entidades separadas, un fenómeno que, hasta donde sabemos no había sido explorado.

Es común oír que cierto caso de duplicación génica "es originada por el modelo de Horowitz (*paso-a-paso*) o por el modelo de *patchwork* (*de-mosaico*)" [60-62,65]; sin embargo, estos son modelos, no mecanismos. Un modelo es una herramienta conceptual que abstrae los componentes principales de un fenómeno para entenderlo, por lo que estas expresiones, como tales, son erróneas. Y más allá de dificultades semánticas, usando una perspectiva de redes, hemos mostrado que la duplicación génica puede y debe ser considerada como un solo proceso, que depende en gran medida de la distancia entre las enzimas y la similitud química entre las reacciones. Con lo cual reconciliamos estos modelos, al pasarlos de antagónicos a complementarios.

Las redes biológicas comparten propiedades topológicas generales, como la libertad de escala y la modularidad. Es más, algunas de esas propiedades han sido detectadas en redes tecnológicas y sociales [19,28,29,47,91]. Nuestros resultados coinciden con estudios previos [72,91,92] al sugerir que el siguiente paso en el modelado del origen y la evolución de las redes debe considerar no solo las propiedades que éstas tienen en común, sino aquellas que las distinguen. En particular, hemos mejorado los modelos nulos para el análisis de las redes metabólicas incluyendo una restricción funcional metabólica, el acoplamiento bioquímico preferente entre los tipos de reacciones (Figura 12a y b).

9. Perspectivas

La principal perspectiva de este proyecto es la implementación de un modelo analítico sobre el crecimiento y evolución de las redes metabólicas que contemple el acoplamiento bioquímico preferente entre los tipos de reacciones, tomando como base los modelos presentados por Goh et al. [67] y Pfeiffer et al. [72]. Si bien hemos mejorado los modelos nulos empleados en este trabajo, con un modelo analítico podríamos pasar de lo descriptivo a lo predictivo. Quedan por resolver algunas vertientes que surgieron a los largo de este proyecto: i) determinar si existe alguna correlación entre otras restricciones biológicas —como la similitud entre sustratos, y las interfases de unión de sustratos Vs. catalíticas de las proteínas— y la retención de duplicados. ii) determinar su existen conectores no-*hubs* en las redes metabólicas centradas en enzimas y/o bipartitas y comparar sus marcas evolutivas contra las de *hubs* y nodos pocos conectados. La idea es tener una imagen más clara de cómo influyen los flujos metabólicos en la evolución de las redes metabólicas. iii) determinar la influencia de otros procesos que originan versatilidad en el metabolismo, como las duplicaciones génicas masivas y la transferencia horizontal, sobre la evolución de las redes metabólicas.

10. Materiales y Métodos

10.1 Reconstrucción de las redes

Se reconstruyeron las redes metabólicas de dos tipos de bases de datos, BioCyc v8.0 [75,76] y KEGG v0.4 [32] De cada una, se obtuvieron las porciones de *E. coli*, llamadas aquí EcoCyc y EcoKegg, respectivamente; así como, los mapas que incluyen información de diversas especies, llamadas aquí MetaCyc y RefKegg, respectivamente. Estas bases de datos contienen la información de cada reacción, como son: sustratos, productos, cofactores, reversibilidad, y el número EC de la enzima que la cataliza. Con esto establecimos las conexiones entre los números enzimáticos (nodos) de cada red, usando la siguiente regla: Si la reacción R1 produce un compuesto o libera un cofactor que es tomado por la reacción R2, se estableció una conexión que va del número EC de R1 al de R2. Si ambas reacciones son reversibles, se estableció una segunda conexión entre los nodos, pero en sentido inverso. La información de cada reacción en EcoCyc y MetaCyc se obtuvo de los siguientes archivos: *reactions.dat* (sustratos, productos y cofactores), *enzrxns.dat* (reversibilidad) y *reactionlinks.dat* (números EC). Mientras que lo propio para EcoKegg y RefKegg se obtuvo de los archivos *xml* de KEGG, usando sus secciones: *reaction* (sustratos, productos, cofactores y reversibilidad), y *entries id* (números EC). Para cada red se detectaron los compuestos más conectados (*hubs*) y las conexiones establecidas solamente por *hubs* fueron eliminadas gradualmente, considerando los diez, veinte y treinta compuestos más conectados como *hubs*.

10.2 Detección de duplicados

Se obtuvieron las secuencias de las enzimas de las redes descritas más arriba, usando sus números EC como semillas contra las bases de datos ECOCYC [75], UNIPROT [93], BRENDA [94] y KEGG [32]. Las secuencias se recortaron manualmente según las anotaciones de dominios funcionales de SWISS-PROT [89] para evitar resultados falsos positivos derivados de la comparación de enzimas multifuncionales. El conjunto final contiene 4,534 secuencias representando 1,527 números EC completamente anotados y 348 anotados parcialmente. Los duplicados se detectaron comparando cada una de estas secuencias contra los modelos ocultos de Markov (HMMs)* de familias y superfamilias de dominios homólogos de las bases de datos

* *Modelos ocultos de Markov (HMMs):* son matrices que especifican la probabilidad de que ocurra cada uno de los eventos posibles, en este caso cada uno de los veinte aminoácidos y un *gap*, en cada posición (n) de una secuencia de eventos. Las probabilidades se obtienen de un alineamiento múltiple de secuencias homólogas previamente construido. Además, incluyen la probabilidad de que cada posición

PFAM-A [95] y SUPERFAMILY [90]. Para las comparaciones se uso el programa HMMER [96], considerando un valor de $E \leq 0.001$ como significativo. Las enzimas cuyas secuencias contienen uno o más dominios homólogos en común se consideraron duplicados. La base de datos PFAM-A contiene una amplia colección de HMMs de familias de proteínas, construidos y depurados manualmente con base en la similitud de las secuencias de las proteínas. La base de datos SUPERFAMILY consta de conjuntos de HMMs que representan familias de dominios homólogos, según los criterios de homología estructural de SCOP [97]. Dado que, en términos evolutivos, la secuencia cambia más rápido que la estructura terciaria de las proteínas, SUPERFAMILY permite detectar homólogos más distantes que los de PFAM, pero PFAM es independiente de las bases de datos de estructuras tridimensionales. Adicionalmente, se realizaron comparaciones pareadas de las 4,534 secuencias usando los programas BLAST [98] y PSI-BLAST [98], que emplean matrices posición-específicas*, con cinco interacciones como máximo y considerando un valor de $E \leq 0.001$ como significativo.

Para cada una de las redes metabólicas reconstruidas, se construyo una matriz de adyacencia en la que cada par de nodos (i,j) pueden o no estar conectados directamente ($i,j = 1$ ó $i,j = 0$, respectivamente). Esta matriz contiene todas las reacciones con sustratos/productos conocidos, incluyendo las que carecen de enzimas (genes) asignadas. Implementamos el algoritmo de Floyd-Warshall [99] en un programa "caminador" de redes, que determina la distancia mínima entre cada par de nodos (i,j). Las distancias obtenidas y los duplicados detectados con HMMs, se usaron para determinar si existe alguna relación entre la distancia de las enzimas y la retención de duplicados. Mientras que los primeros dos dígitos de los números enzimáticos (EC:a.b.-.-) hicieron lo propio para analizar la influencia de la similitud química sobre la retención de duplicados.

10.3 Influencia de la modularidad de las redes sobre la retención de duplicados

Las distancias entre los nodos, obtenidas con el "caminador", se usaron para construir una matriz de asociaciones normalizadas de cada par de nodos (i,j) con la función $(1/\text{distancia}_{(i,j)}^2)$. Esta matriz se uso para agrupar jerárquicamente los nodos de cada red, usando la

---

(n) sea precedida (n-1) y sucedida (n+1) por cada uno de los eventos posibles antes mencionados. Estas dos probabilidades "no se ven" en el alineamiento, ni en las matrices tradicionales, como las PSSMs (ver más abajo), de ahí la denominación de "ocultos".

* *Matrices posición-específicas (PSSMs)*: resumen la probabilidad de que ocurra cada uno de los veinte amino ácidos en cada posición de una secuencia. Las probabilidades pueden ser tomadas de alineamientos múltiples de familias de proteínas homólogas o, en el caso de PSI-BLAST, de manera iterativa, cada que una familia es enriquecida con nuevos miembros.

τ de Kendall implementada en el programa CLUSTER v3.0 [100], y detectar con ello sus módulos. También se uso la correlación de Spearman, obteniéndose resultados similares. Para determinar la retención de duplicados dentro y entre los módulos detectados, se calculó la distancia evolutiva (DE) entre cada par de rutas metabólicas —como se les define en sus respectivas bases de datos— con la formula:

$$(DE) = A' / (A' + AB)$$

En donde A' es el número de enzimas de la ruta más pequeña (rA) sin duplicados en la otra ruta (rB) y AB es el número de enzimas de rA con duplicados en rB. En un extremo, cuando todas las enzimas de rA tienen duplicados en rB el valor de (DE) converge en cero. Mientras que, cuando las dos rutas no comparten duplicados, el valor de (DE) converge en uno.

10.4 Modelos nulos y pruebas estadísticas

Para determinar si la alta retención de duplicados entre enzimas cercanas se limita a una región de la red o está dispersa por toda ella, para cada red realizamos 10,000 muestreos aleatorios, cada uno de los cuales tomó al azar la mitad de la red original y calculó la retención de duplicados como se describió antes. Por otro lado, para saber si las frecuencias observadas de retención de duplicados son significativamente diferentes de lo esperado al azar, construimos dos tipos de modelos nulos. En el primero, como lo recomiendan Maslov y Sneppen [35], se reconecta al azar a cada red original, preservando el grado de conectividad de cada nodo (conexiones de entrada y salida). Para ello, se eligen dos pares de nodos al azar; por ejemplo (E1 → E2) y (E5 → E6), y se reconectan, quedando (E1 → E6) y (E5 → E2) (ver parte baja de la Figura 13a). Este proceso se repite hasta que la red ha sido completamente reconectada, generando un modelo nulo. Se generaron 10,000 de estos modelos y en cada uno se determinó la retención de duplicados, como se describió anteriormente. En el segundo tipo de modelos, "funcionalmente similares", se retiene tanto grado de conectividad de cada nodo, como la frecuencia de cada tipo de pares de reacciones (ver parte baja de la Figura 13b), restringiendo la reconexión con el acoplamiento bioquímico preferente entre los tipos de reacciones. Esto es, siguiendo el ejemplo de arriba, se escogen dos pares de nodos al azar, pero estos se reconectan si y solo si, E1 y E5 catalizan RQS (EC:a.b.-.-), y E2 y E6 catalizan RQS (EC:w.x.-.-). Cuando esta restricción no se cumple, se escogen otros dos pares de nodos al azar y los anteriores se "regresan" al grupo original para ser escogidos posteriormente. Este proceso se repite hasta que la red ha sido completamente reconectada, generando un modelo nulo. Se generaron 10,000 de estos modelos y en cada uno se determinó la retención de duplicados, como se describió anteriormente. En ocasiones, al final del

proceso, quedan pares de nodos cuya reconexión generaría combinaciones ya existentes en los modelos nulos. Esos pares, al igual que en el último par, no apareado, de redes con un número non de pares de nodos, que a lo sumo representan en conjunto el 5% de la red original, se omiten del modelo nulo, por lo que este último puede quedar con algunas conexiones menos que la red original.

Usamos la prueba de *Z-score* (Z_i) para determinar el significado estadístico de las frecuencias de determinado atributo (i) en cada red real; por ejemplo, la retención de duplicados entre enzimas a cierta distancia, como sigue:

$$Z_i = Nreal_i - < Nazar_i > / std(Nazar_i)$$

En donde $Nreal_i$ es la frecuencia del atributo (i) en la red real. $< Nazar_i >$ es el promedio de las frecuencias del atributo (i) en los modelos nulos, y $std(Nazar_i)$ es su desviación estándar.

Un valor de Z-score ≥ 3 implica que la frecuencia del atributo (i) en la red real es significativamente mayor que lo esperado al azar ($P < 0.001$). En contraste, un Z-score ≤ -3, indica que el atributo (i) está subrepresentado.

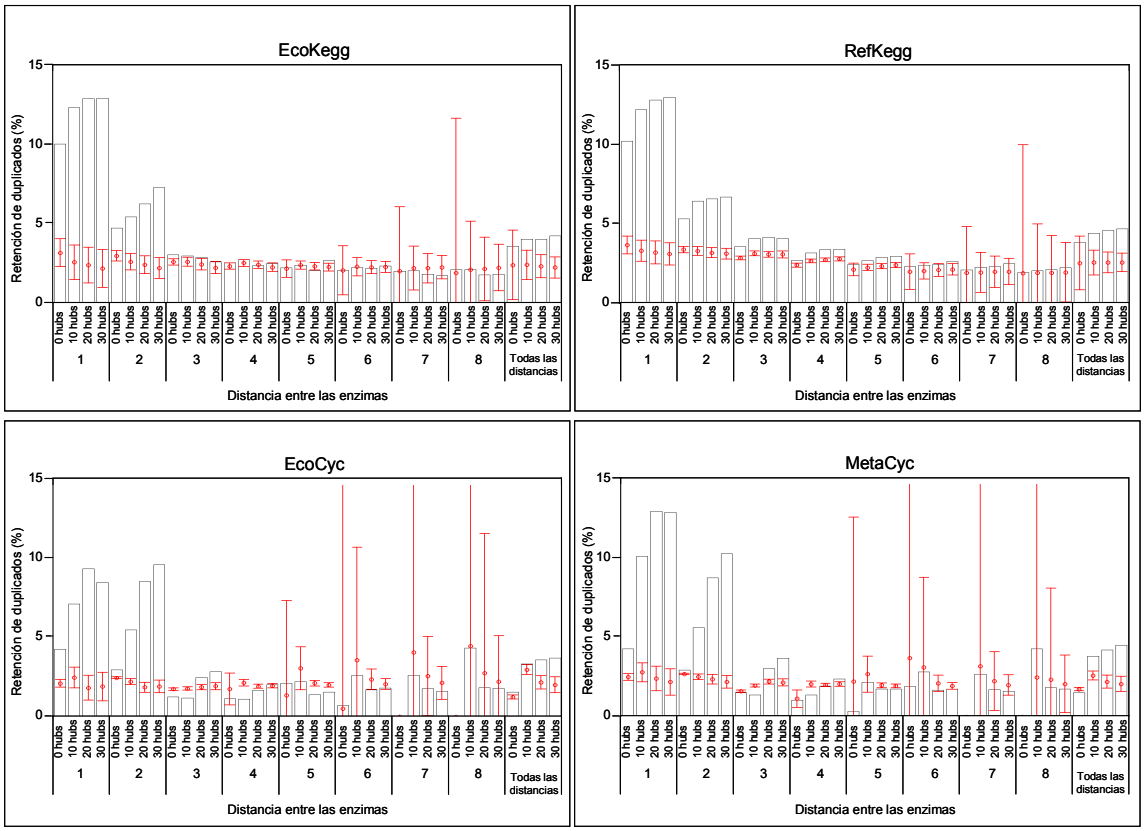
Para determinar el significado estadístico de la distancia evolutiva (DE) entre pares de rutas metabólicas, comparamos los valores observados contra los esperados al azar usando 1000 modelos nulos. Estos modelos mantuvieron las redes intactas, y lo que cambiamos al azar fue el contenido de dominios (PFAM + SUPERFAMILY) de cada secuencia. En este caso, un Z-score ≤ -3 implica la (DE) entre las dos rutas en cuestión es menor de lo esperado al azar ($P < 0.001$), o sea, que esas dos rutas han retenido entre si más duplicados de lo esperado.

11. Literatura consultada

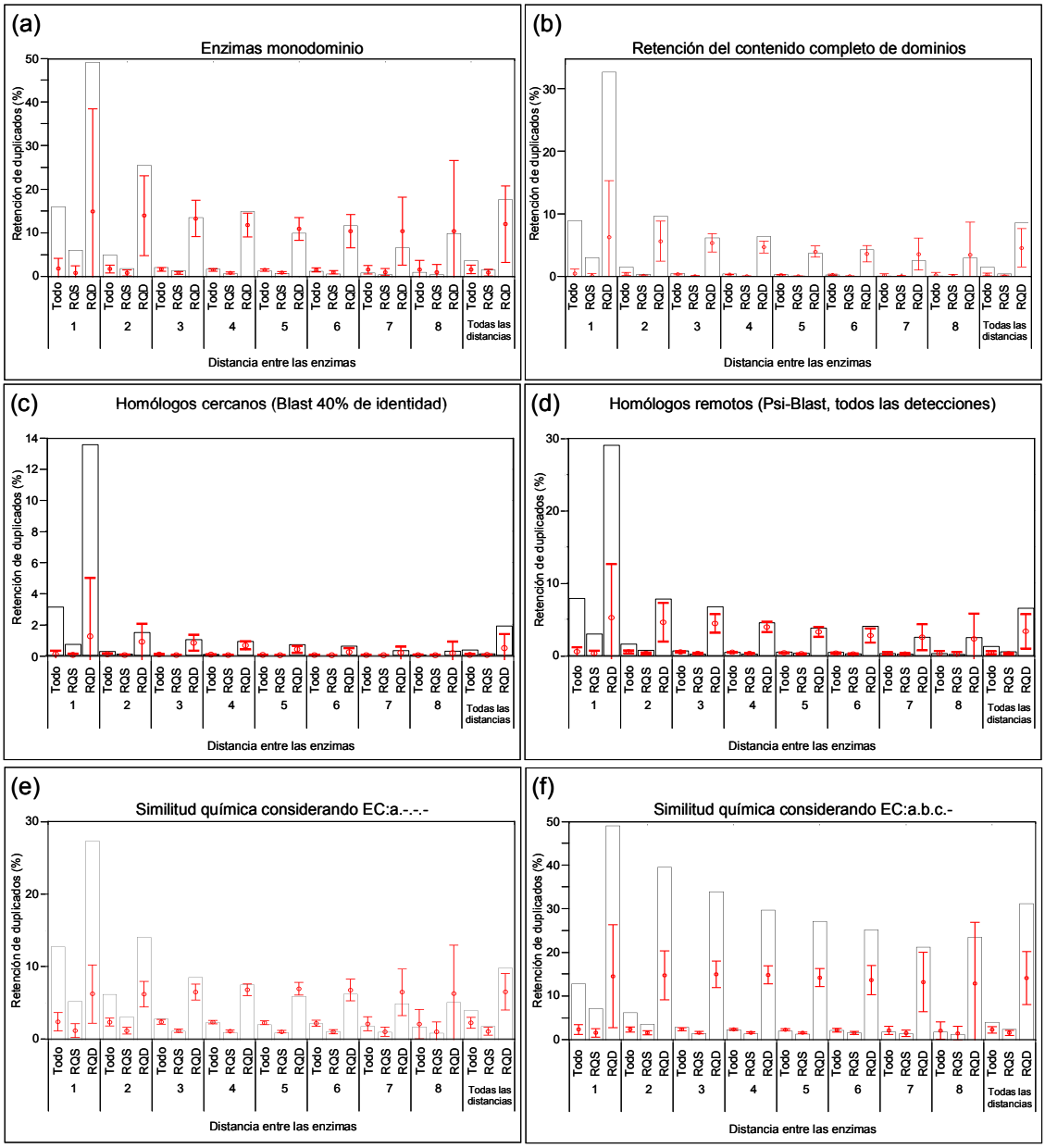
1. Egeland, J.A. et al. (1987). **Bipolar affective disorders linked to DNA markers on chromosome 11**. *Nature* 325, 783-7.
2. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. (1999). **From molecular to modular cell biology**. *Nature* 402, C47-52.
3. Noble, D. (2006). *The music of life: biology beyond the genome*, 176 (Oxford University Press, USA).
4. Solomonoff, R. & Rapoport, A. (1951). **Connectivity of random nets**. *Bull Math Biophysics* 13, 107-117.
5. Erdos, P. & Renyi, A. (1959). **On random graphs**. *Publ Math* 6, 290-297.
6. Erdos, P. & Renyi, A. (1960). **On the evolution of random graphs**. *Publ Math Inst Hungar Acad Sci* 5, 17-61.
7. Pool, I.d.S. & Kochen, M. (1978). **Contacts and Influece**. *Social Networks* 1, 5-51.
8. Travers, J. & Milgram, S. (1969). **An experimetal study of the small world problem**. *Sociometry* 32, 425-443.
9. Watts, D.J. & Strogatz, S.H. (1998). **Collective dynamics of 'small-world' networks**. *Nature* 393, 440-2.
10. Price, D.J.d.S. (1965). **Networks of Scientific Papers**. *Science* 149, 510-5.
11. Newman, M.E.J., Barabási, A.-L. & Watts, D.J. (2006). *The structure and dynamics of networks*, x, 582 p. (Princeton University Press, Princeton, N.J. ; Oxford).
12. Barabasi, A.L. & Albert, R. (1999). **Emergence of scaling in random networks**. *Science* 286.
13. Albert, R., Jeong, H. & Barabasi, A.L. (1999). **Diameter of the world-wide web**. *Nature* 401, 130-131.
14. Faloutsos, M., Faloutsos, P. & Faloutsos, C. (1999). **On power-law relationship of the internet topology**. *Comp Comm Rev* 29, 251-262.
15. Broder, A. et al. (2000). **Graph structure in the Web**. *Comput Networks* 33, 309-320.
16. Price, D.J.d.S. (1976). **A general theory of bibliometric and other cumulative disadvantage processes**. *J Am Soc Inform Sci* 27, 292-306.
17. Krapivsky, P.L., Redner, S. & Leyvraz, F. (2000). **Connectivity of growing random networks**. *Phys Rev Lett* 85, 4629-4632.
18. Dorogovtsev, S.N., Mendes, J.F.F. & Samukhin, A.N. (2000). **Structure of Growing Networks with Preferential Linking**. *Phys Rev Lett* 85, 4633-4636.
19. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. & Barabasi, A.L. (2002). **Hierarchical organization of modularity in metabolic networks**. *Science* 297, 1551-5.
20. Ravasz, E. & Barabasi, A.L. (2003). **Hierarchical organization in complex networks**. *Phys Rev E Stat Nonlin Soft Matter Phys* 67, 026112.
21. Amaral, L.A., Scala, A., Barthelemy, M. & Stanley, H.E. (2000). **Classes of small-world networks**. *Proc Natl Acad Sci U S A* 97, 11149-52.
22. Goh, K.I., Oh, E., Jeong, H., Kahng, B. & Kim, D. (2002). **Classification of scale-free networks**. *Proc Natl Acad Sci U S A* 99, 12583-8.
23. Liljeros, F., Edling, C.R., Amaral, L.A., Stanley, H.E. & Aberg, Y. (2001). **The web of human sexual contacts**. *Nature* 411, 907-8.
24. Newman, M.E. (2001). **The structure of scientific collaboration networks**. *Proc Natl Acad Sci U S A* 98, 404-9.
25. Albert, R., Jeong, H. & Barabasi, A.L. (2000). **Error and attack tolerance of complex networks**. *Nature* 406, 378-82.
26. Fell, D.A. & Wagner, A. (2000). **The small world of metabolism**. *Nat Biotechnol* 18, 1121-2.
27. Jeong, H., Mason, S.P., Barabasi, A.L. & Oltvai, Z.N. (2001). **Lethality and centrality in protein networks**. *Nature* 411, 41-2.
28. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. & Barabasi, A.L. (2000). **The large-scale organization of metabolic networks**. *Nature* 407, 651-4.
29. Wagner, A. & Fell, D.A. (2001). **The small world inside large metabolic networks**. *Proc Biol Sci* 268, 1803-10.
30. Venter, J.C. et al. (2001). **The sequence of the human genome**. *Science* 291, 1304-51.
31. Barabasi, A.L. & Oltvai, Z.N. (2004). **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 5, 101-13.
32. Kanehisa, M. & Goto, S. (2000). **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 28, 27-30.
33. Hughes, A.L. & Friedman, R. (2005). **Gene duplication and the properties of biological networks**. *J Mol Evol* 61, 758-64.
34. Wuchty, S. (2001). **Scale-free behavior in protein domain networks**. *Mol Biol Evol* 18, 1694-702.
35. Maslov, S. & Sneppen, K. (2002). **Specificity and stability in topology of protein networks**. *Science* 296, 910-3.
36. Butland, G. et al. (2005). **Interaction network containing conserved and essential protein complexes in Escherichia coli**. *Nature* 433, 531-7.
37. Ekman, D., Light, S., Bjorklund, A.K. & Elofsson, A. (2006). **What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?** *Genome Biol* 7, R45.
38. Hahn, M.W. & Kern, A.D. (2005). **Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks**. *Mol Biol Evol* 22, 803-6.
39. Qin, H., Lu, H.H., Wu, W.B. & Li, W.H. (2003). **Evolution of the yeast protein interaction network**. *Proc Natl Acad Sci U S A* 100, 12820-4.

40. Wagner, A. (2003). **How the global structure of protein interaction networks evolves**. *Proc Biol Sci* 270, 457-66.
41. Wuchty, S., Oltvai, Z.N. & Barabasi, A.L. (2003). **Evolutionary conservation of motif constituents in the yeast protein interaction network**. *Nat Genet* 35, 176-9.
42. Prachumwat, A. & Li, W.H. (2006). **Protein function, connectivity, and duplicability in yeast**. *Mol Biol Evol* 23, 30-9.
43. Papp, B., Pal, C. & Hurst, L.D. (2004). **Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast**. *Nature* 429, 661-4.
44. Luscombe, N.M. et al. (2004). **Genomic analysis of regulatory network dynamics reveals large topological changes**. *Nature* 431, 308-12.
45. Baba, T. et al. (2006). **Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection**. *Mol Syst Biol* 2, 2006 0008.
46. Martinez-Antonio, A. & Collado-Vides, J. (2003). **Identifying global regulators in transcriptional regulatory networks in bacteria**. *Curr Opin Microbiol* 6, 482-9.
47. Shen-Orr, S.S., Milo, R., Mangan, S. & Alon, U. (2002). **Network motifs in the transcriptional regulation network of Escherichia coli**. *Nat Genet* 31, 64-8.
48. Benner, S.A., Ellington, A.D. & Tauer, A. (1989). **Modern metabolism as a palimpsest of the RNA world**. *Proc Natl Acad Sci U S A* 86, 7054-8.
49. Morowitz, H.J. (1999). **A theory of biochemical organization, metabolic pathways, and evolution**. *Complexity* 4, 39-53.
50. Guimera, R. & Nunes Amaral, L.A. (2005). **Functional cartography of complex metabolic networks**. *Nature* 433, 895-900.
51. von Mering, C. et al. (2003). **Genome evolution reveals biochemical networks and functional modules**. *Proc Natl Acad Sci U S A* 100, 15428-33.
52. Almaas, E., Oltvai, Z.N. & Barabasi, A.L. (2005). **The activity reaction core and plasticity of metabolic networks**. *PLoS Comput Biol* 1, e68.
53. Vitkup, D., Kharchenko, P. & Wagner, A. (2006). **Influence of metabolic network structure and function on enzyme evolution**. *Genome Biol* 7, R39.
54. Kitami, T. & Nadeau, J.H. (2002). **Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication**. *Nat Genet* 32, 191-4.
55. Nelson, D.L., Cox, M.M. & Lehninger, A.L. (2005). *Lehninger principles of biochemistry*, 1 v. (various pagings) (W.H. Freeman, New York).
56. Ma, H. & Zeng, A.P. (2003). **Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms**. *Bioinformatics* 19, 270-7.
57. Horne, A.B., Hodgman, T.C., Spence, H.D. & Dalby, A.R. (2004). **Constructing an enzyme-centric view of metabolism**. *Bioinformatics* 20, 2050-5.
58. Kunin, V. & Ouzounis, C.A. (2003). **The balance of driving forces during genome evolution in prokaryotes**. *Genome Res* 13, 1589-94.
59. Ohno, S. (1970). *Evolution by gene duplication*, (Springer, New York).
60. Alves, R., Chaleil, R.A. & Sternberg, M.J. (2002). **Evolution of enzymes in metabolism: a network perspective**. *J Mol Biol* 320, 751-70.
61. Gerlt, J.A. & Babbitt, P.C. (2001). **Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies**. *Annu Rev Biochem* 70, 209-46.
62. Teichmann, S.A. et al. (2001). **The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli**. *J Mol Biol* 311, 693-708.
63. Horowitz, N.H. (1945). **On the evolution of biochemical synthesis**. *Proc Natl Acad Sci USA* 31, 153-7.
64. Jensen, R.A. (1976). **Enzyme recruitment in the evolution of new function**. *Annu Rev Microbiol* 30, 409-25.
65. Light, S. & Kraulis, P. (2004). **Network analysis of metabolic enzyme evolution in Escherichia coli**. *BMC Bioinformatics* 5, 15.
66. Berg, J., Lassig, M. & Wagner, A. (2004). **Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications**. *BMC Evol Biol* 4, 51.
67. Goh, K.-I., Kahng, B. & Kim, D. (2005). **Evolution of the Protein Interaction Network of Budding Yeast: Role of the protein family compatibility constraint**. *J Korean Phys Soc* 46, 551-555.
68. Pastor-Satorras, R., Smith, E. & Sole, R.V. (2003). **Evolving protein interaction networks through gene duplication**. *J Theor Biol* 222, 199-210.
69. Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. (2003). **Modeling of protein interaction networks**. *Complexus* 1, 38-44.
70. Bhan, A., Galas, D.J. & Dewey, T.G. (2002). **A duplication growth model of gene expression networks**. *Bioinformatics* 18, 1486-93.
71. Wagner, A. (1994). **Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization**. *Proc Natl Acad Sci U S A* 91, 4387-91.
72. Pfeiffer, T., Soyer, O.S. & Bonhoeffer, S. (2005). **The evolution of connectivity in metabolic networks**. *PLoS Biol* 3, e228.
73. Kashtan, N. & Alon, U. (2005). **Spontaneous evolution of modularity and network motifs**. *Proc Natl Acad Sci U S A* 102, 13773-8.

74. Madern, D. (2002). **Molecular evolution within the L-malate and L-lactate dehydrogenase super-family**. *J Mol Evol* 54, 825-40.
75. Karp, P.D. et al. (2002). **The EcoCyc Database**. *Nucleic Acids Res* 30, 56-8.
76. Krieger, C.J. et al. (2004). **MetaCyc: a multiorganism database of metabolic pathways and enzymes**. *Nucleic Acids Res* 32, D438-42.
77. Tu, B.P., Kudlicki, A., Rowicka, M. & McKnight, S.L. (2005). **Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes**. *Science* 310, 1152-8.
78. Zaslaver, A. et al. (2004). **Just-in-time transcription program in metabolic pathways**. *Nat Genet* 36, 486-91.
79. Diaz-Mejia, J.J., Perez-Rueda, E. & Segovia, L. (2007). **A network perspective on the evolution of metabolism by gene duplication**. *Genome Biol* 8, R26.
80. Becker, S.A., Price, N.D. & Palsson, B.O. (2006). **Metabolite coupling in genome-scale metabolic networks**. *BMC Bioinformatics* 7, 111.
81. Lynch, M. & Katju, V. (2004). **The altered evolutionary trajectories of gene duplicates**. *Trends Genet* 20, 544-9.
82. Aharoni, A. et al. (2005). **The 'evolvability' of promiscuous protein functions**. *Nat Genet* 37, 73-6.
83. Lozada-Chavez, I., Janga, S.C. & Collado-Vides, J. (2006). **Bacterial regulatory networks are extremely flexible in evolution**. *Nucleic Acids Res* 34, 3434-45.
84. Madan Babu, M., Teichmann, S.A. & Aravind, L. (2006). **Evolutionary dynamics of prokaryotic transcriptional regulatory networks**. *J Mol Biol* 358, 614-33.
85. Sharan, R. et al. (2005). **Conserved patterns of protein interaction in multiple species**. *Proc Natl Acad Sci U S A* 102, 1974-9.
86. Resendis-Antonio, O. et al. (2005). **Modular analysis of the transcriptional regulatory network of E. coli**. *Trends Genet* 21, 16-20.
87. Yang, S., Doolittle, R.F. & Bourne, P.E. (2005). **Phylogeny determined by protein domain content**. *Proc Natl Acad Sci U S A* 102, 373-8.
88. Diaz-Mejia, J.J. & Lazcano, A. (2002). **The origin and evolution of fatty acids metabolic pathways: a genomic perspective**. *10th Meeting of the International Society for the Study of Origin of Life* Oaxaca, Mexico.
89. Boeckmann, B. et al. (2003). **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 31, 365-70.
90. Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 313, 903-19.
91. Milo, R. et al. (2004). **Superfamilies of evolved and designed networks**. *Science* 303, 1538-42.
92. Artzy-Randrup, Y., Fleishman, S.J., Ben-Tal, N. & Stone, L. (2004). **Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks"**. *Science* 305, 1107; author reply 1107.
93. Apweiler, R. et al. (2004). **UniProt: the Universal Protein knowledgebase**. *Nucleic Acids Res* 32, D115-9.
94. Schomburg, I. et al. (2004). **BRENDA, the enzyme database: updates and major new developments**. *Nucleic Acids Res* 32, D431-3.
95. Bateman, A. et al. (2004). **The Pfam protein families database**. *Nucleic Acids Res* 32, D138-41.
96. Eddy, S.R. (1996). **Hidden Markov models**. *Curr Opin Struct Biol* 6, 361-5.
97. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995). **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 247, 536-40.
98. Altschul, S.F. et al. (1997). **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 25, 3389-402.
99. Lipschutz, S. (1987). *Data Structures*, (McGraw-Hill).
100. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998). **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 95, 14863-8.



Apéndice 1. Influencia de la eliminación de hubs sobre la retención de duplicados en las redes metabólicas de distintas especies. Los resultados mostrados en el texto principal corresponden a la red EcoKegg, reconstruida eliminando los 20 *hubs* más conectados. Aquí se presentan los resultados de las otras redes (EcoCyc, MetaCyc y RefKegg), reconstruidas sin eliminar *hubs*, así como eliminando 10, 20 y 30 *hubs*. El histograma representa la retención de duplicados en cada red y los puntos rojos los valores esperados, ± 3 □, calculados usando 1000 modelos nulos tipo Maslov-Sneppen.



Apéndice 2. Controles del contenido de dominios, el método de detección de duplicados y el criterio de similitud química. Originalmente, EcoKegg incluyó 541 números EC. En (a) se incluyeron los 291 números EC con un solo dominio. Originalmente, se consideraron duplicados a las enzimas que tuvieran al menos un dominio homólogo, en (b) se consideraron duplicados solo aquellos casos en que la enzima con más dominios tuviera homólogos de todos dominios de la enzima con menos dominios. Originalmente, se utilizó HMMER+PFAM+SUPERFAMILY para detectar duplicados, en (c) se buscaron homólogos cercanos con BLAST (valor de E = 0.001 y 40% de identidad como mínimo) y en (d) se buscaron homólogos distantes con PSI-BLAST (mismos parámetros y 5 iteraciones). Originalmente se consideraron RQS aquellas con los mismos dos primero dígitos del número EC (EC:a.b.-.-), en (e) se considero solo el primer dígito (EC:a.-.-.-), y en (f) se consideraron los tres primeros (EC:a.b.c.-). En todas las figuras, los histogramas son las frecuencias de las redes reales y los puntos rojos son los promedios, , ± 3 σ, calculados usando 1000 modelos nulos tipo Maslov-Sneppen.