UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

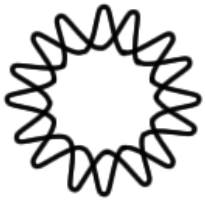
RECONSTRUCCIÓN Y ESTUDIO TOPOLÓGICO DE LAS REDES DE REGULACIÓN Y DEL METABOLISMO CENTRAL DE *BACILLUS SUBTILIS*

TESIS

QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS BIOQUÍMICAS
PRESENTA

LIC. CARLOS DANIEL VÁZQUEZ HERNÁNDEZ

DIRECTOR DE TESIS: DRA. ROSA MARÍA GUTIÉRREZ RÍOS



CUERNAVACA, MORELOS 2011

Agradecimientos

Agradezco a mis padres, Olivia Hernández y Carlos Daniel Vázquez, por su apoyo y cariño incondicionales, y cuya dedicación diaria me ha servido como inspiración permanente.

Agradezco a mi hermana, Olivia Vázquez, cuyo cariño siempre ha sido una influencia positiva en mi vida.

Agradezco a mis primos, Yadira Reyna y Juan Bernardo Diemond, sin cuyo ejemplo tal vez hubiera decidido no dedicarme a la ciencia.

Agradezco a mi tutora, la Dra. Rosa María Gutiérrez, por su guía constante derivada de su excepcional visión y por su apoyo incondicional.

Agradezco al Dr. Enrique Merino y al resto del grupo de Genómica Computacional del Instituto de Biotecnología, por su interés en este trabajo y su apoyo constante.

Agradezco a los Dres. Alfredo Martínez y Ernesto Pérez Rueda, por su revisión crítica de este trabajo.

Agradezco a los responsables del proyecto CONACyT Salud 68992, por su oportuno apoyo económico.

Reconstrucción y estudio topológico de las redes de regulación y del metabolismo de *Bacillus subtilis*

Estudiante: Carlos Daniel Vázquez H.

Maestría en Ciencias Bioquímicas

Universidad Nacional Autónoma de México

Comité Tutoral:

Dra. Rosa María Gutiérrez Ríos (tutor principal)

Dr. Alfredo Martínez Jiménez

Dr. Ernesto Pérez Rueda

Índice

Resumen

Los recientes trabajos de análisis de biología molecular han devuelto valiosa información que permite interpretar los eventos de un proceso celular respecto de otro. La mayoría de los trabajos existentes, no obstante, hacen esta comparación de manera indirecta, y los trabajos que han dado los primeros pasos hacia una unificación la llevan a cabo en un nivel muy general sin un factor dador de uniformidad para todos los elementos involucrados. El objeto de este trabajo es determinar si un traslape de procesos celulares que permita apreciar estos procesos en un solo plano informacional es factible mediante la observación de sus propiedades a gran escala y, si lo es, encontrar los parámetros que harían posible este traslape, tomando como base los procesos de regulación transcripcional y metabolismo de *Bacillus subtilis*.

Este proceso se lleva a cabo principalmente a través de la redefinición de la red metabólica con base en sus procesos enzimáticos en vez de sus metabolitos y reconstruyendo esta representación usando las unidades informacionales involucradas en las reacciones, ya sean genes u operones, las cuales se pueden adjuntar de manera intuitiva a los procesos de regulación. Se prefiere usar los procesos enzimáticos porque la representación basada en metabolitos no es adaptable a una unificación de este tipo, pues la acción de los reguladores no se da directamente sobre las moléculas procesadas en el metabolismo.

Para alcanzar esta representación, se procede primero a analizar las propiedades estadísticas de la red de regulación transcripcional y de la red metabólica basada en enzimas para *Bacillus subtilis*, por separado, a través de dos métodos de análisis topológico: el método de clustering para determinar modularidad jerárquica creado por Albert Lázlo Barabási y el método de acoplamiento de flujos, que pretende observar una red con base en los caminos lineales claramente distinguibles que en ella existen. Para la red de regulación se emplean primariamente datos de DBTBS, una base de datos de operones de *Bacillus subtilis*, y para la red metabólica se emplean principalmente datos de KEGG, usando datos tanto de los archivos XML destinados a *Bacillus subtilis* como de una curación exhaustiva de la base de datos, centrada en los datos disponibles para *Bacillus subtilis*.

Lima Méndez reporta desvíos en la distribución correspondiente a la red de metabolitos para *Escherichia coli* que se separan de la esperada bajo los supuestos de Barabási [13], y estos desvíos se hallan también en la red de regulación transcripcional de *Bacillus subtilis*. La red metabólica de enzimas de *Bacillus subtilis*, por otro lado, no sigue una distribución clara de acuerdo con el método de Barabási, lo cual lleva a probar el método de acoplamiento de flujos en esta red. Se observan rutas lineales dentro de la red, pero en la serie de archivos XML son mas abundantes y evidentes que en los datos de la curación.

Para descartar la influencia de compuestos poco relevantes al metabolismo mas puramente definido, que toma como base las transiciones entre compuestos carbonados, se prueban métodos de filtrado de compuestos. El método mas usual es el de eliminar los compuestos mas altamente conectados, que incluyen moléculas como agua y cofactores, pero este método no elimina todos los metabolitos de este tipo, y elimina también compuestos carbonados, que son de interés. Se crea un método de filtrado crudo, basado en un pesado diferencial de los átomos por elemento involucrados en cada compuesto, en un intento por producir un filtro más relevante respecto de la perspectiva biológica esperada.

Al aplicar el nuevo filtro sobre la red metabólica y analizarla por acoplamiento de flujos, se observa que el filtro produce varios positivos y negativos falsos, pero la mayoría son eliminados por el método de acoplamiento. Asimismo, el acoplamiento produce varias rutas biológicamente validas; menos que para los datos de los XML, pero mas que para la red cruda, para la cual casi no se devuelve información.

Se procede, con estos resultados, a redefinir la red de los XML de KEGG y la red de la curación, con el filtro, con base en las unidades informacionales para *Bacillus subtilis*, conectar los grupos de operones resultantes a la red de regulación, y observar la agrupación resultante. Los reguladores globales se conservan en ambas redes, lo cual muestra la estructura subyacente de la red de regulación, y tienden a coincidir en mayor número sobre los extremos de las rutas lineales encontradas.

Los resultados, en su conjunto, muestran no solo que el criterio de integración elegido es valido, sino que también muestra que la red enzimática sigue una distribución distinta a la de

metabolitos. La densidad de la red dificulta su análisis con los métodos existentes, pero la red misma es susceptible de ser separada en grupos biológicamente relevantes.

Abstract

Recent analytic works in molecular biology have returned valuable information which allows for the interpretation of a given cellular process in the context of another. Most existing developments, however, perform this comparison indirectly, and those which have given the first steps towards unification make it in a very general level which lacks a factor that can give uniformity to all involved elements. The goal of this work is to determine whether an overlapping of cellular processes which allows for their appreciation in a single informational level is feasible through the observation of large-scale properties and, if it is, identify the parameters which would make such overlap possible, based on the processes of transcriptional regulation and metabolism in *Bacillus subtilis*.

This overlap is carried out mainly through a redefinition of the metabolic network based on its enzymatic processes instead of its metabolites and rebuilding this representation using the informational units involved in the network's reactions, whether they are operons or single genes, as these can be intuitively appended to the regulatory apparatus. Enzymatic processes are preferred because a metabolite-based representation is not adaptable to the intended unification, since the molecules processed through the cell's metabolism are not the direct targets of regulatory action.

With the goal of this representation in mind, first we analyze the statistical properties of the transcriptional regulatory network and the reaction-based metabolic network for *Bacillus subtilis*, each one separately, using two methods of topological analysis: the clustering method to determine hierarchical modularity created by Albert Lázlo Barabási and the flux coupling method, which is designed to observe a network through the clearly separate linear paths it contains. The regulatory network is constructed primarily out of data from DBTBS, an operon database for *Bacillus subtilis*, while the metabolic network is created out of data from the KEGG database, with both the XML files provided for *Bacillus subtilis* and an exhaustive search of the database, centered on data available for *B. subtilis*.

Lima Méndez reports slants in the clustering distribution for the metabolite-based metabolic network of *Escherichia coli* which part significantly from the arrangement expected under the Barabási method [13], and these slants are also found in the transcriptional regulatory network for *Bacillus subtilis*. The reaction-based metabolic network for *B. subtilis*, on the other hand, does not follow a clear distribution under the Barabási method. Observing this, we attempt the flux coupling method on the metabolic network. Linear paths are detected, but the ones yielded by the XML files are more abundant and evident than the ones from the manual search data, which are almost negligible.

To help rule out the influence of compounds considered to be of less relevance to metabolism as mainly based in carbon compound transitions, some compound-based filtering methods are devised and implemented. The most usual method is removing the most connected nodes from the network, but it does not remove all undesirable compounds and removes carbonated compounds, which are of interest. We create a crude filtering method, based in a differential weighting of the element atoms involved in each compound, in an attempt to manage a filter more relevant to the biologically expected.

When the metabolic network is subjected to this filter and analyzed through flux coupling, we observe that the filter produces several false positives and negatives, but most are removed by the flux coupling method itself. The flux coupling method yields several biologically valid network paths; less than the XML data network, but many more than the unfiltered search network.

With these results, we restructure both the XML-based metabolic network and the filtered search-based metabolic network based on the available informational units for *Bacillus subtilis*, attaching the resulting groups to the regulatory apparatus. Global regulators tend to be conserved in both networks, consistently with the structure of the regulatory network itself, and tend to group in greater numbers towards the ends of the linear pathways we located.

Taken together, these results show no only that the integration criterion we devised is valid, but also that the reaction-based metabolic network has a distribution and behavior different

from those of the metabolite-based network. The network's density makes analysis with existing methods hard, but the network can still be separated in biologically relevant groups.

Introducción

El estudio de la biología en las últimas décadas se ha beneficiado de la disponibilidad de grandes cantidades de datos sobre los organismos vivos, que permiten comparaciones para expandir el conocimiento disponible. La cantidad de información que existe actualmente sobre sistemas biológicos ha dado la necesidad de complementar el estudio de estas reglas a través de diversos métodos computacionales auxiliares, como la construcción de bases de datos y los modelos computacionales. Los métodos de modelado dependen de una serie de reglas que den una perspectiva precisa sobre el nivel de detalle y el tamaño del sistema. Existen métodos muy precisos, que emplean matemáticas exactas que permiten explicar la mecánica del sistema a detalle y pueden modelar incluso moléculas, pero el crecimiento de su complejidad con la cantidad de datos los hace ineficaces al modelar sistemas de gran tamaño. Otros métodos intentan compensar esta debilidad disminuyendo el detalle con el que se analiza el sistema y aumentando la cantidad de elementos que pueden tomarse en cuenta. Entre éstos, el enfoque de redes ha resultado eficaz y su uso se ha difundido en años recientes [4]. Este método se elige para este estudio debido al relativamente gran tamaño del sistema que se desea examinar y a que la escala a la cual se piensa manejar permite hacer predicciones sobre el comportamiento del sistema a analizar en su conjunto [4].

El enfoque de redes biológicas intenta representar el sistema estudiado como un grafo según sus propiedades específicas y elucidar su comportamiento mediante métodos de análisis de grafos. Un grafo es una representación abstracta de un objeto o sistema que permite la obervación de las interacciones entre sus componentes. Tradicionalmente, se manejan métodos exclusivamente matemáticos para construir el grafo y analizar sus propiedades. Este enfoque ha permitido la revisión exitosa de sistemas complejos como el sistema regulatorio y metabólico de *Escherichia coli*, con incluso la capacidad de predecir su comportamiento bajo ciertas condiciones de crecimiento [5, 6].

La teoría de redes, al representar las interacciones entre los elementos del sistema que se desea modelar como una representación de grafo que es relativamente flexible, permite poner énfasis en la naturaleza de las interacciones que entre ellos existen. Un grafo se compone de

elementos elegidos como tales, denominados nodos, ilustrando sus interacciones como aristas; en un entorno gráfico, los nodos pueden representarse como puntos o círculos y las aristas como líneas que los conectan.

Un grafo puede construirse de distintas formas dependiendo de cómo se puede interpretar la naturaleza misma de la reacción. Si se conoce la dirección en la cual la interacción se lleva a cabo, se pueden representar como flechas de un elemento a otro. En este caso se dice que el grafo es dirigido; se dice que es no dirigido cuando esta diferencia no existe o es despreciable [6]. En algunos casos, se pueden asignar pesos a las aristas, a fin de diferenciar la naturaleza de cada conexión.

También se pueden dar distinciones de peso a los nodos, a fin de observar de manera conjunta elementos de interés y ver si una propiedad que los distingue sigue patrones o no. Sin embargo, esto implica que se toman elementos cuya naturaleza es divisible en dos clases. En este caso se dice que se maneja un grafo bipartita, cuyos elementos no son necesariamente idénticos, lo cual dificulta su análisis. Se prefiere que un grafo no considere este tipo de distinciones.

Al elaborar una red, se pueden obtener datos que revelan propiedades generales del sistema como la forma en que se distribuyen los nodos. La distribución de nodos en particular revela propiedades generales del sistema que permiten elucidar su comportamiento. Esta información es de utilidad sobre todo en sistemas grandes como las redes de organismos vivos, para los cuales es difícil determinar visualmente estas propiedades, que además pueden dar información evolutiva a través de la comparación mutua de estructuras topológicas entre especies [7].

A partir de la construcción del grafo, se pueden obtener varias propiedades del sistema dadas por las interacciones entre los elementos, que posteriormente se pueden refinar por análisis estadísticos. El número de conexiones que tiene un nodo con otros es su grado, la propiedad que más se toma como referencia para un análisis general. Un camino es una ruta que se puede tomar, a través de las aristas, de un nodo a otro, respetando las direcciones donde la haya. Los más relevantes son el camino más corto de un nodo a otro, conocido como camino mínimo, en particular al promediar los valores de todos los caminos mínimos de la red; el camino más corto

de la red; y el camino más largo de la red, conocido como diámetro. Estos caminos pueden dar una idea de qué tan fácil es llegar de un nodo a otros y, por extensión, de qué tan cercanamente conectada está la red [7].

A partir del grado se pueden determinar otras propiedades de gran relevancia. Muchas de ellas requieren que se elimine la direccionalidad de la red, pero este proceso es útil al intentar revelar propiedades generales. Por ejemplo, la distribución de conectividad o *P(k)*, donde *k* es el grado de un nodo, representa qué tan probable es que un nodo tenga un cierto grado *k*. Su mayor relevancia es distinguir entre dos tipos de estructura posibles para una red: una distribución aleatoria, en la cual todos los nodos están conectados con otros al azar y si se integrara un nuevo elemento a la red es igual de probable que se conecte con cualquiera de los existentes; y una distribución que sigue una ley de potencia (una distribución del tipo $P(k) \sim k^{-\gamma}$, donde γ es conocido como el exponente de grado), en cuyo caso la red es conocida como libre de escala. La característica principal de las redes libres de escala son elementos especiales denominados *hubs*. Un *hub* concentra en sí mismo una gran cantidad de nodos individuales, y si se integrara un nuevo nodo a la red, es significativamente más probable que se conecte con un *hub* que con cualquier otro elemento [5, 6].

Otra distribución útil es la del coeficiente de clustering, o *C(k)*. El clustering de un nodo individual es la razón entre el número de conexiones entre los vecinos de ese nodo y las máximas que pudieran darse. La *C(k)* es el clustering promedio de todos los nodos en la red por cada grado *k*. Se ha observado que las redes para las cuales $C(k) \sim k^{-\gamma}$, donde $-2<\gamma<-1$, siguen una distribución conocida como jerárquica modular. Esta distribución presenta *hubs*, pero también presenta comunidades de nodos estrechamente interconectados, que son parte de comunidades más laxamente conectadas; los *hubs* pueden entrar conectando estas últimas. Las pequeñas comunidades de nodos son conocidas como módulos [5, 6].

Se inició este enfoque en redes metabólicas [5], pero se extendió a redes de regulación, dando preferencia al inicio de la transcripción [2], y se han complementado con éxito mediante métodos experimentales diversos, además de metodologías basadas en este enfoque que incrementan la información disponible [8].

En biología, los *hubs* han sido relacionados con elementos de gran relevancia para el sistema celular, como reguladores globales en el caso de las redes de TFs o metabolitos de uso muy difundido en las redes metabólicas. Los módulos también son de gran relevancia en redes biológicas, ya que se ha demostrado que al representar una red biológica de esta forma alinean con grupos funcionales de genes o metabolitos con una gran precisión [9, 10, 7].

Los módulos en redes biológicas se pueden obtener según la estructura topológica que haya rendido la red o pueden inferirse mediante datos experimentales. Cuando existen datos informativos para formar una estructura con clusters, por ejemplo, se pueden aplicar algoritmos de clustering, los cuales ya están estandarizados para su aplicación en redes biológicas [11]. Existen varias modalidades de algoritmos de clustering, como el jerárquico aglomerativo, que pueden relacionar elementos en datos de experimentos masivos (como un microarreglo). Al aplicar estos algoritmos sobre estos datos, de nuevo se ha visto que las estructuras biológicas correlacionan con las estructuras topológicas definidas y que los nuevos datos pueden dar una mayor precisión sobre la naturaleza de los distintos módulos y su comportamiento según una cierta condición [12, 6].

Con base en observaciones recientes, se ha visto que algunas redes biológicas no siguen con exactitud los parámetros requeridos para un análisis basado en clusters [7]. De hecho, se ha observado una desviación al principio de la distribución que este método no considera [13]. Esto pone a consideración algunas debilidades del enfoque de redes, en particular el cuidado que debe ponerse cuando se intenta determinar propiedades globales de un sistema. Parte del objetivo de este trabajo es determinar la naturaleza de la desviación que se observa y ayudar a reestructurar la metodología para que refleje con mayor fidelidad la realidad biológica que se intenta retratar con estos métodos.

Uno de los más aceptados ha sido el acoplamiento de flujos, que se basa en la determinación de rutas lineales entre nodos, muy comunes en las redes metabólicas. El método determina el grado de acoplamiento que posee un nodo respecto del resto según el número de conexiones que tiene, los nodos a los cuales llevan esas conexiones y, por lo general, la dirección de la interacción. Se considera que el acoplamiento es completo cuando el nodo sólo participa en

una ruta lineal. Cuando el nodo entra en contacto con otra ruta lineal, se separan como tres rutas y el nodo se considera como parcialmente acoplado a las tres. En caso de que haya más conexiones a otras rutas, se considera que el nodo no tiene acoplamientos [3]. A la fecha se han desarrollado diversos métodos para emplear el algoritmo de acoplamiento de flujos, que refinan distintos aspectos del algoritmo para precisar mejor la información disponible [14].

Muchos otros organismos modelo estudiados con el enfoque de redes han dado información útil y novedosa. El estudio de la bacteria de suelo *Bacillus subtilis*, por ejemplo, ha dado interesantes resultados, sobre todo en su sistema de esporulación. Sin embargo, los métodos aplicados a redes no detallan todo el comportamiento de la célula; las redes de regulación génica y del metabolismo requieren análisis separados, que sólo han reflejado indirectamente el comportamiento de otros sistemas cercanamente relacionados. Por otro lado, no existe una nomenclatura estable que pueda relacionar los distintos tipos de moléculas que interactúan en el sistema, lo cual dificulta un estudio unificado de estos procesos que pueda dar más información sobre su comportamiento, sobre todo en la forma de nexos poco relevantes entre metabolitos y reacciones. [13] Parte del objetivo de este trabajo es crear parámetros para incrementar la claridad para las reacciones más relevantes, o al menos dar el primer paso en esa dirección.

El modelo a elegir para este estudio es *B. subtilis*. Esta bacteria ha sido el principal modelo de estudio para bacterias Gram-positivas. Se ha predicho que su genoma tiene del orden de 4100 genes, con un contenido de G+C de 43.5%. Actualmente se conocen las funciones de más de la mitad de los genes predichos para la bacteria, pero aún existen muchos genes cuyas funciones aún quedan por identificar. En la época en que *Escherichia coli* prevaleció como el modelo central para el estudio de las bacterias, *B. subtilis* se mantuvo en el campo por la facilidad con la cual se le puede manipular genéticamente y, sobre todo, por su capacidad de formar esporas [15, 16]. No obstante, esto ha provocado que otros sistemas que son importantes para la bacteria, como el metabolismo, hayan quedado en segundo plano frente a los estudios teóricos y experimentales que se han hecho sobre esporulación. Parte del objetivo de este trabajo es retomar estas funciones, en particular el metabolismo central, a fin de modelarlas y averiguar

qué información pueden dar sobre el funcionamiento de la bacteria como un sistema, continuando el trabajo previo que se tenía sobre esta bacteria [6].

En este trabajo proponemos generar criterios de integración de los sistemas regulatorios y metabólicos que permita estudiarlos como un sistema unificado. Primero, se comparan los principios que rigen la topología de ambos tipos de red, primero con la metodología de Barabási y después probando metodologías alternas como el acoplamiento de flujos, a fin de observar y comparar la viabilidad de cada set de datos resultante. Con esto, se comparan las propiedades obtenidas para cada red con miras a encontrar sus similitudes a nivel topológico y biológico para desarrollar un método que pueda conciliar las diferencias que puedan existir entre ambas redes. De igual forma, se prueba este método conciliatorio sobre un sistema modelo para ilustrar computacionalmente la relación entre ambos sistemas y estructurar predicciones unificadas que puedan compararse con mayor calidad con los resultados experimentales.

Antecedentes

El estudio de la biología molecular depende de las herramientas adecuadas para precisar la mecánica que los hace posibles. Herramientas como la microscopía y métodos que aprovechan las funciones de los distintos componentes, como la reacción en cadena de la polimerasa (polymerase chain reaction, PCR) han resultado cruciales para el avance de la disciplina y la obtención de resultados claros. En las últimas décadas, con la disponibilidad de refinamientos y de métodos más precisos, se ha dado una explosión de la cantidad de información que rinde la actividad experimental, y posibilidades crecientes de alcanzar un conocimiento más profundo sobre los mecanismos que emplean los seres vivos para sobrevivir y proliferar.

Las estrategias de supervivencia empleadas por los seres vivos dependen de los componentes mismos de la célula y de las interacciones que entre ellos existen. Esto se refleja en los distintos niveles de organización apreciables en los organismos vivos: las unidades moleculares básicas, como metabolitos pequeños, ácidos nucleicos (DNA, RNA), proteínas con distintas funciones, y otros componentes, interactúan de distintas formas, permitiendo interpretar estas relaciones como sistemas de distinta complejidad. Pueden ser modelos sencillos, como los objetivos de uno o dos factores transcripcionales o una ruta metabólica sencilla, o entes más elaborados como comunidades funcionales o una vía de señalización con varias posibilidades. Eventualmente esta perspectiva de síntesis puede llevarse a complejos que abarquen gran parte de la célula, para analizar estos componentes a gran escala. En esta lógica se basa la perspectiva del modelado; comprender el comportamiento de las moléculas de la célula como partes de una especie de mecanismo, con partes funcionales claramente definidas (*infra*).

Definición del sistema de interés

El modelo a elegir para este estudio es la bacteria de suelo *Bacillus subtilis*. *B. subtilis* ha sido el principal modelo de estudio para bacterias Gram-positivas. Es un bacilo (bacteria de forma alargada) que crece mejor a 25-35°C y tiene una pared celular consistente de peptidoglicanos, especialmente mureína. Se ha predicho que su genoma tiene del orden de 4100 genes, con un contenido de G+C (guanina y citosina, nucleótidos cuya abundancia relativa determina la estabilidad térmica del genoma) de 43.5%. La bacteria copia su genoma desde un solo origen

usando dos complejos de replicación paralelos. Actualmente se conocen las funciones de más de la mitad de los genes predichos para *B. subtilis*, pero aún existen muchos genes cuyas funciones aún quedan por identificar.

B. subtilis se puede mover con un flagelo, aunque necesita sintetizar surfactantes para adherirse y moverse. Así mismo, tiene muchos genes destinados a la producción de antibióticos, lo cual le permite segregarlos en grandes cantidades. Pese a esto, su actividad antibiótica no es dañina para los humanos; ésta tiende a funcionar para preservar su ventaja al colonizar raíces, con las que forma relaciones mutualistas benéficas, resupliendo nutrientes en el suelo e impidiendo la colonización de patógenos. Cabe señalar, no obstante, que *B. subtilis* posee también propiedades de putrefacción de sustancias orgánicas, lo cual ayuda a su rol de proveedor de nutrientes en el suelo. Tradicionalmente, se le considera un aerobio estricto, al igual que a todos los microorganismos de suelo [16], pero se ha demostrado que esto no siempre es cierto, pues puede funcionar bajo condiciones fermentativas, usando iones nitrito o nitrato como aceptores de electrones [6].

En la época en que *Escherichia coli* prevaleció como el modelo central para el estudio de las bacterias, *B. subtilis* se mantuvo en el campo por la facilidad con la cual se le puede manipular genéticamente y, sobre todo, por su capacidad de formar esporas [15, 16], capacidad que le permite sobrevivir los estreses y falta de alimento comunes en su entorno. Principalmente por esta capacidad se secuenció su genoma, completado en 1997.

No obstante, esto ha provocado que otros sistemas que son importantes para la bacteria, como el metabolismo, hayan quedado en segundo plano frente a los estudios teóricos y experimentales que se han hecho sobre esporulación. Sólo hasta años recientes se han empezado a tomar en cuenta, debido en parte a estudios de relevancia para la industria alimentaria. Una cepa de *B. subtilis*, antes conocida como *Bacillus natto*, es un componente esencial para la elaboración de ciertos productos de soya fermentada [17], lo cual ha motivado estudios para incrementar el conocimiento sobre el metabolismo de esta bacteria.

Con el modelo que se ha elegido para este trabajo, cabe considerar las estrategias de construcción que éste adopta. Las estrategias de estructuración interna adoptadas por los

procariotes son más simples en su estructura y, por ende más sencillas de plasmar en un modelo. Muchas funciones que son más complejas que llevan a cabo los eucariotes no existen en los procariotes o toman formas más sencillas, lo cual permite tomar menos supuestos para construir un modelo acorde con los eventos que ocurren en la célula viviente. Este estudio se centra principalmente en el metabolismo y la regulación de la expresión génica, con énfasis en el inicio de la transcripción.

Breve panorama del metabolismo

El metabolismo es el conjunto de las reacciones químicas que ocurren en un organismo vivo. Normalmente se consideran como metabolismo a las reacciones que manejan la estructura de los bloques más generales de la célula (principalmente las moléculas pequeñas, incluyendo también a veces azúcares mayores), designando los procesos de síntesis no genéricos, como los de los ácidos nucleicos, los azúcares complejos y las proteínas, como procesos separados. Se clasifica comúnmente el metabolismo en dos grandes secciones: el catabolismo, que reduce compuestos complejos a otros más simples para obtener energía, y el anabolismo, que toma compuestos simples y los reconstituye como las construcciones complejas que constituyen el organismo. Los componentes comúnmente asociados con el metabolismo son los compuestos sujetos a transformación, denominados metabolitos, y los agentes que llevan a cabo estas transformaciones, conocidos como enzimas. Las enzimas son por lo general proteínas, pero se sabe que también existen RNAs que ejecutan funciones de este procesamiento, como los RNAs ribosomales y otros RNAs conocidos como ribozimas**.** Una de las características más notables del metabolismo es su elevado grado de conservación, sobre todo en el caso de sistemas dedicados a la producción de energía, como el ciclo de los ácidos tricarboxílicos (tricarboxylic acid cycle: TCA) y los generadores de ATP. La conservación es lo suficientemente elevada como para hacer probable que estas rutas hayan aparecido relativamente temprano en la evolución y se hayan difundido entre diversos organismos [18].

El metabolismo mantiene a otros subsistemas de un organismo vivo proveyendo los bloques básicos para las estructuras que éstos emplean, y se mantiene a sí mismo acoplando el proceso catabólico al anabólico, a fin de producir la energía que se requiere para producir las

estructuras más complejas prevalentes en la célula. En otros términos, el metabolismo mantiene la elevada complejidad de un organismo vivo tomando la energía de estructuras ordenadas que lo rodean, creando así el desorden que los hace satisfacer las leyes de la termodinámica. La producción de energía es, así, el proceso más crucial y prevalente en el metabolismo de los organismos vivos. Esta energía puede generarse o almacenarse a partir de procesos especializados, como la fosforilación oxidativa o la fotosíntesis.

El anabolismo se reserva para crear estructuras complejas, y comprende todos los procesos de síntesis de moléculas, e incluso los procesos de producción de proteínas y ácidos nucleicos, a los cuales se hará referencia más adelante. El enfoque más aceptado para entender el anabolismo es dividirlo según los compuestos que se generan a través de él: carbohidratos, lípidos, aminoácidos para generar proteínas, y nucleótidos para generar ácidos nucleicos, entre otras funciones. Existen también otros procesos asociados, como la absorción se compuestos de nitrógeno o azufre y su integración a las rutas metabólicas comunes, que también deben considerarse para entender el metabolismo.

La forma canónica de representar el metabolismo es tomar como elementos principales los metabolitos y a las enzimas, implícitas como las reacciones que los procesan, como las uniones entre ellos. Esto se debe, primariamiente, a que los metabolitos son los componentes más conservados del metabolismo, incluso más que las enzimas, y por ello es más sencillo establecer comparaciones entre organismos y organizar el metabolismo mismo en fragmentos, facilitando su estudio en gran medida desde los primeros intentos de comprender su funcionamiento. Las metodologías utilizadas para estudiar el metabolismo son, por ello, inseparables del metabolito como concepto. Sólo en años recientes se han empezado a tomar como suplemento los métodos genéticos que manejan las proteínas involucradas, al hacer accesibles éstas a la observación de las metodologías disponibles.

El metabolismo provee a todos los demás procesos de los componentes que necesitan para operar. Provee a los sistemas de transporte de sus componentes y de la energía que muchos necesitan para funcionar, así como lo hace con la señalización, el movimiento celular

(motilidad), la reproducción, como división celular o cualquier otra de sus formas, y el segundo subsistema celular relevante a este trabajo: la regulación de la expresión génica.

Breve panorama de la regulación génica

La expresión de los genes, que son los fragmentos considerados como informacionales en el DNA (que acarrean información codificada de un efector), necesita llevarse a cabo de forma diferencial de acuerdo con las circunstancias que rodean a la célula, a fin de adaptarla a las variaciones posibles de su entorno y economizar la producción de los efectores en sí mismos. La base de la expresión génica es el mecanismo de la transcripción y traducción de la información contenida en el ácido desoxirribonucleico (deoxyribonucleic acid: DNA) a una molécula funcional, generalmente una proteína, usando un intermediario de ácido ribonucleico (ribonucleic acid: RNA) que es referido como transcrito o, cuando toma su forma informativa final, RNA mensajero (messenger RNA: mRNA), teniendo como nombre transcripción el proceso mediante el cual es producido. Así mismo, se llama traducción al proceso de producción de una proteína a partir de un mRNA, que se lleva a cabo mediante los ribosomas, complejos de RNA y proteína que construyen las cadenas de péptido que son la base de las proteínas ayudados por RNAs de transferencia (transfer RNAs: tRNAs) que acarrean un aminoácido y lo ligan en el orden codificado en el mRNA. El patrón en el cual se entiende que normalmente se asocia de manera constante un solo tipo de aminoácido a un sólo tipo de tRNA se conoce como código genético, la principal herramienta que se usa para entender cómo está la información de cada proteína codificada en el DNA. Recientemente se han hecho diversos estudios sobre la flexibilidad del código genético, con el fin de entender y manipular la mecánica que lo constituye [46].

Tanto transcripción como traducción toman como base esta información inicial en el DNA. Los genes contienen información no sólo sobre la molécula efectora que representan, sino también sobre los sitios donde debe asociarse la RNA polimerasa correspondiente para formar el intermediario de RNA, además de otros modificadores. En esta información accesoria descansan los fundamentos de la regulación de la expresión génica.

La regulación de la expresión génica es el conjunto de los mecanismos que emplea la célula para controlar la producción de las moléculas que ejecutan funciones de manera directa, que

en su mayoría son los distintos tipos de proteínas, incluyendo también algunos RNAs, como los ribosomales. Estas moléculas ejercen su influencia sobre las directamente funcionales controlando la producción de éstas últimas o de otras de su misma especie que a su vez modulan de forma directa o indirecta a las funcionales. Existen distintos modos en los cuales estas moléculas de control, llamadas regulatorias, pueden llevar a cabo esta tarea. En los eucariotes existen sofisticados mecanismos de corte del transcrito luego de la transcripción, y luego de su transporte fuera de la membrana nuclear. Éstos no existen en los procariotes, donde la transcripción y la traducción se llevan a cabo de forma acoplada; a medida que se crea un mRNA, en él se arman varios ribosomas que traducen la información. El principal momento en el cual se lleva a cabo la regulación en los procariotes en en el comienzo de la transcripción.

La RNA polimerasa interactúa con diversas moléculas en formas que determinan si un gen se transcribe o no. En primera instancia, depende de la presencia de una proteína que se asocia al DNA cerca del sitio de inicio de la transcripción, precisamente en el sitio -35 (35 nucleótidos antes del sitio de inicio, generalmente un sitio rico en adenina y timina, conocido como caja TATA), conocida como factor sigma, el cual estabiliza la unión de la polimerasa misma con el DNA. Existen diversos factores sigma que funcionan en distintas condiciones; el sigma mejor descrito es el housekeeping (sigma 70 en *E. coli*), funciona la mayoría del tiempo y sobre casi todos los genes, mientras que otros sigmas son más estables y prevalentes en condiciones diferentes, como el sigma de choque térmico (sigma 32 en *E. coli*) o el sigma de fase estacionaria (sigma 38 en *E. coli*).

De forma análoga, existen proteínas asociables al DNA, alternas a los factores sigma, que alteran la posibilidad de unión de la polimerasa al DNA, incluso cuando el factor sigma está presente. Estas proteínas son conocidas como factores transcripcionales, o TFs (transcription factors). Los TFs pueden ya sea promover o bloquear la unión de la RNA polimerasa al estar unidos al DNA en sus respectivos sitios. Cuando el TF promueve la unión de la polimerasa, se dice que es un activador, y cuando la bloquea se dice que es un represor. Se considera que la primera descripción rigurosa de un TF en funcionamiento es la del operón *lac* en *E. coli*, descubierto por Jacques Monod, el cual actúa como un represor del operón (un conjunto de

genes contiguos que quedan bajo el control de una sola señal regulatoria) que bloquea la transcripción hasta que se asocia con moléculas de lactosa, que impiden su asociación con el operón y posibilitan la producción de las proteínas efectoras. Este represor, codificado por el gen *lacI*, tiene un nivel de expresión constitutivo (siempre activo) y actúa cooperativamente con el funcionamiento de CRP (cAMP receptor protein, antes carbon represison protein), el cual ayuda al operón a responder a los niveles de adenosín monofosfato cíclico (AMP cíclico; cyclic adenosine monophosphate, cAMP), los cuales reflejan los niveles de glucosa de forma indirecta. El efecto neto es que el operón se transcribe cuando la glucosa se ha agotado pero hay lactosa disponible [48].

Dependiendo de las interacciones que un TF tenga con sus propios reguladores, existen diversas formas de lograr un efecto regulatorio. Se puede activar un operón ya sea promoviendo la acción de un activador o bloqueando la de un represor, y se puede reprimir reprimiendo un activador o activando un represor. Es usual hallar operones controlados por un arreglo más bien complejo de TFs con distintos efectos, lo cual permite que un operón responda a varias condiciones.

Existen algunos TFs que no son coordinados por otro TF de manera directa. Este es el caso de los TFs que siempre mantienen un nivel relativamente alto de expresión, conocidos como constitutivos, o de TFs que son regulados por otros mecanismos, como RNAs regulatorios. También se da el caso de TFs regulados por una molécula alterna, como un metabolito. Este es el caso del operón *lac* ya mencionado, pero también es el caso de otros reguladores. El regulador transcripcional CRP en *E. coli* es modulado por su asociación directa con cAMP, con lo cual la variación de concentración molécula reprime de manera indirecta los genes de proceso catabólico cuando el factor no está asociado a la molécula. CRP es también un TF que afecta el sistema regulatorio de *E.coli* a gran escala, afectando una gran variedad de procesos, lo cual hace TFs con comportamiento similar importantes para la célula en su conjunto [49].

Los primeros modelos de la regulación transcripcional fueron sencillos, abarcando sistemas pequeños. Más tarde se desarrollarían perspectivas a más gran escala (*infra*), principalmente debido al desarrollo de las metodologías experimentales dedicadas a la exploración de este

proceso. Algunas como los ensayos de footprinting, análisis de retardo en gel y la desactivación artificial de un gen regulatorio son capaces de revelar el comportamiento de un solo TF con gran precisión, mientras que otros procedimientos, como los microarreglos y los ensayos de inmunoprecipitación de cromatina (chromatin immunoprecipitation, ChIP) dan una perspectiva más amplia del contexto funcional del TF.

Definición del enfoque

Las metodologías recientes han generado una explosión de información que da una oportunidad sin precedente de analizar la regulación y el metabolismo de los organismos vivos en detalle y a gran escala. No obstante, esta información se ha acumulado a un ritmo que hace poco factible su recopilación y análisis a través de métodos tradicionales. Las máquinas de cómputo han dado una solución a este problema, facilitando la conservación, el acceso y la circulación de la información disponible, así como su visualización, integración y procesamiento. Se han construido y adaptado diversas herramientas computacionales con este fin, como es el caso de construcción de bases de datos con distintos cúmulos de información o distintos enfoques dados a su interpretación. Por ejemplo, existen bases de datos como la Enciclopedia de Genes y Genomas de Kioto (Kyoto Encyclopedia of Genes and Genomes, KEGG) [19], que está dedicada a la información disponible sobre metabolismo en distintas especies, y DBTBS [20], una base de datos dedicada a las relaciones entre TFs y operones en *B. subtilis*. De forma similar, se han desarrollado métodos computacionales que facilitan visualizar datos e interpretarlos, así como ayudar en la generación de modelos e hipótesis de trabajo que pueden contrastarse con la información disponible [11]. Existen varios de estos métodos, como modelos basados en ecuaciones diferenciales para lograr cálculos de flujo relativos a la concentración de moléculas relacionadas, o sofisticadas simulaciones del comportamiento de la proteína con los metabolitos correspondientes [4].

Varios métodos de modelado se ejemplifican en la figura 1. Estos métodos de modelado se diferencian entre sí según varios criterios que afectan el tamaño del sistema que son capaces de analizar y la calidad del análisis que rinden. Se prefieren métodos si requieren menos datos o son más rápidos, pero también según su fidelidad a la realidad biológica y su capacidad de

integrar datos para facilitar inferencias. Debido a que varias de estas propiedades a veces son mutuamente excluyentes en su aplicación, a veces se deben ignorar algunas propiedades favorables con el objeto de obtener otras. Es posible visualizar esto como una lente, en la cual se sacrifica visión de conjunto por detalle, o viceversa. Por un lado, existen métodos que dan visión sobre detalles pequeños que afectan al sistema, como son las simulaciones estocásticas o modelos de ecuaciones diferenciales, que rinden un comportamiento muy fiel al de las moléculas involucradas, pero necesitan enormes cantidades de información y a menudo son inviables para analizar sistemas incluso de un tamaño mediano [4].

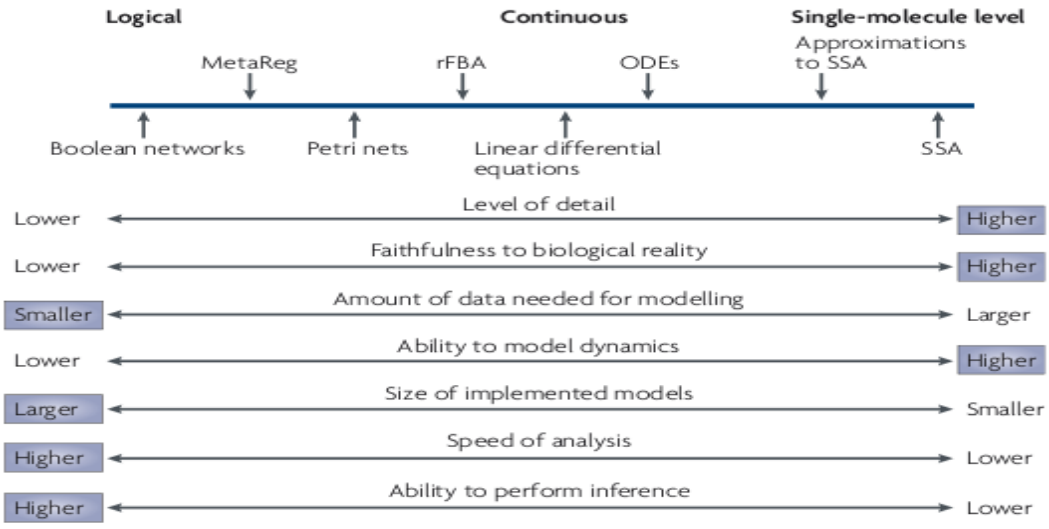


Figura 1. Resumen de varios métodos de modelado computacional, junto con su especialización y capacidades. El enfoque de redes está marcado como redes booleanas. Tomado de Karlebach, Shamir, 2008.

Existen otros métodos que están pensados para analizar sistemas de gran tamaño a gran escala, que al sacrificar el nivel de detalle y el realismo del comportamiento dan margen para cálculos relativamente más rápidos y una visión de conjunto que permite inferir el comportamiento a grandes rasgos. Algunos permiten el análisis de ciertos detalles de la dinámica, como los Petri nets, aunque entorpecen algo de la visión de conjunto para lograrlo. El modelo más capaz de ilustrar propiedades a gran escala es el modelo de redes, que necesita poco procesamiento para rendir propiedades generales de sistemas muy grandes. Este enfoque ha resultado muy eficaz para este objetivo, y su uso de ha difundido en años recientes. [4]

La principal razón para elegir este modelo en este trabajo es la escala del sistema que se desea analizar. El metabolismo de una bacteria es un sistema muy vasto a pesar de que se trata de un organismo relativamente simple; incluso cuando se hacen cortes para simplificar el análisis, se necesitan varias otras regiones para apreciar el contexto en el cual funciona el subsistema. De forma análoga, se planean hacer predicciones a grandes rasgos de este sistema y relacionarlos con la regulación del la expresión génica a nivel del inicio de la transcripción (conocido también sólo como regulación transcripcional), otro sistema de gran tamaño, lo cual hace necesario recurrir a este modelo para generar datos y predicciones adecuados a la escala del sistema resultante [4].

El enfoque de redes; teoría de grafos

El enfoque de redes ataca el problema de la representación de un sistema biológico de gran tamaño a través de la reducción y la simplificación, a fin de obtener una perspectiva sencilla que pueda dar información integrable de las relaciones entre los elementos, aunque éstos sean subsistemas complejos a su vez. Lo que se hace es encontrar parámetros que puedan unificar las complejidades del sistema para reducirlo a una representación de red que pueda rendir las propiedades generales que se busca analizar e interpretar. El método usual para lograr esta integración es reducir el sistema a un grafo, que da una red de elementos que es fácil de analizar a esta escala.

Un grafo es, definido simplemente, una representación abstracta de un objeto o sistema que permite la observación de las interacciones entre sus componentes. La fracción de mayor

interés cuando se ocupa este enfoque no son los elementos en sí, pues éstos han sido resumidos como entes individuales, sino las relaciones e interacciones que entre ellos existen, que existen a la escala que se busca ver [21].

Un grafo se define formalmente como un par ordenado G = (V,E), donde V es un conjunto de nodos o vértices y E un conjunto de aristas, que son subconjuntos de dos elementos de V representables como un par no ordenado de los nodos respecto al arista. Esto quiere decir que un grafo se compone de nodos y aristas, de forma tal que un nodo puede concentrar en sí mismo varias aristas, pero cada arista conecta solamente a dos nodos, los cuales no tienen ninguna otra relación entre sí. Tradicionalmente los grafos no comprenden conexiones de un nodo a sí mismo o más de un arista para un par de nodos. En un entorno gráfico, los nodos pueden representarse como puntos o círculos y las aristas como líneas que los conectan; esto facilita un despliegue visual que es acorde a sus propiedades [21].

Un grafo puede tener distintas propiedades según la propiedad del grafo mismo que se desea revisar y la naturaleza del sistema que se pretende representar mediante él. A continuación se describen algunas de las formas en que se pueden construir estas distinciones.

A veces la dirección de un nodo a otro es una parte importante de la naturaleza de la interacción, y reflejarla en el grafo refleja las propiedades de la red en sí misma. Se llama dirigido a un grafo que contiene una direccionalidad clara, y no dirigido al caso en el cual la direccionalidad es poco visible o no existe [2]. Por ejemplo, si se toman como elementos personas en una fiesta y las aristas como un saludo de mano, el grafo puede considerarse como no dirigido, pues ambas personas se dan la mano entre sí. Si se toman las aristas para decir si una persona conoce a otra, se puede decir que el grafo es dirigido, pues una persona puede saber de otra sin que ésta le haya visto o sepa de ella. En algunos casos, se pueden asignar pesos a las aristas, a fin de diferenciar la naturaleza de cada conexión. Por ejemplo, si se toman como nodos ciertas ciudades y como aristas los caminos entre ellas, a las aristas del grafo puede agregárseles un peso por la distancia de estos caminos. En la figura 2 se puede apreciar un ejemplo de como se puede trasladar un sistema de transcripción-traducción a una representación de grafo.

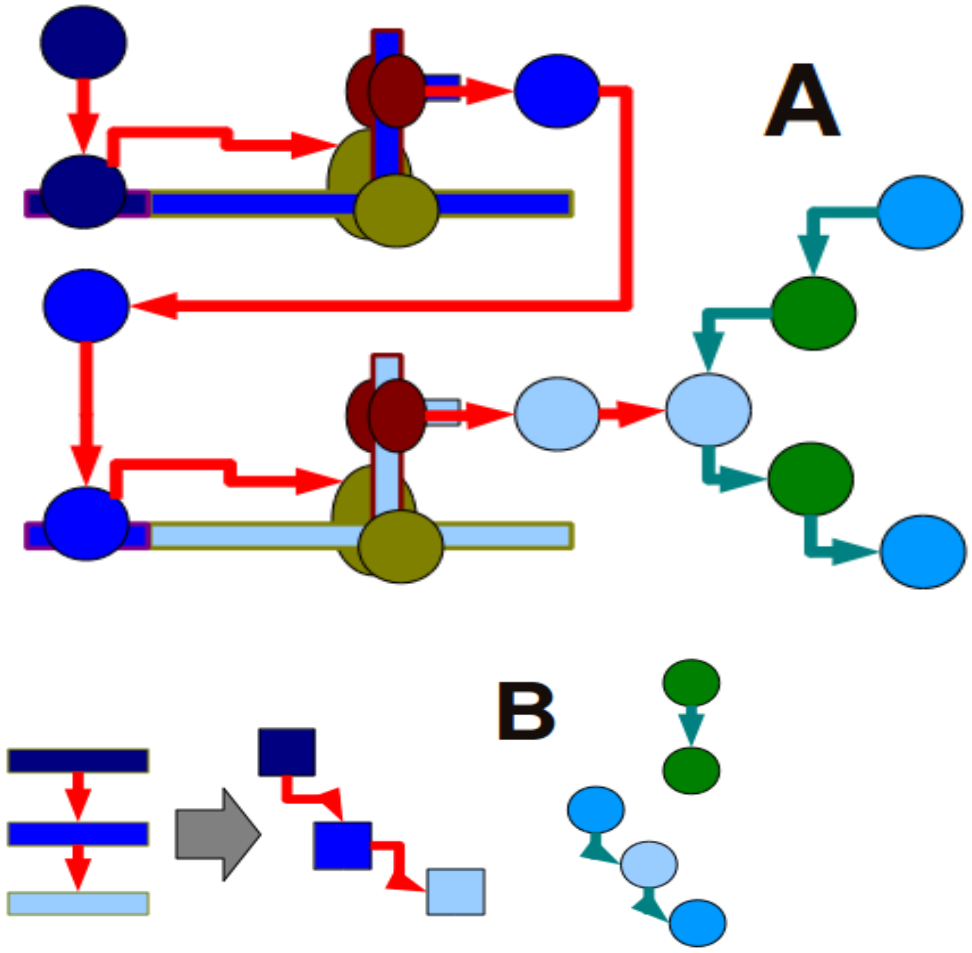


Figura 2. Esquematización de cómo puede simplificarse un sistema biológico para ser representado como un grafo. (A) Esquema que representa los mecanismos regulatorios que preceden la producción de una enzima de interés (La enzima es el círculo del color más claro), en flechas rojas, y a su inserción en una ruta metabólica, en flechas verdes. Los metabolitos son los círculos verdes. Los círculos azules a la derecha son otras enzimas de la ruta. Los círculos azules de la izquierda son reguladores transcripcionales; las secuencias que codifican las proteínas (rectángulos. línea amarilla: DNA; línea roja: RNA; línea negra: estructura primaria) y respectivos sitios de unión (morado) son representados por su respectivo color de relleno. La RNA polimerasa está en amarillo. El ribosoma está en rojo. (B) Distintas formas de resumir el esquema anterior para elaborar un grafo. A la izquierda se resume la regulación transcripcional. A la derecha, se resume el metabolismo. Normalmente se toman los metabolitos como criterio, al ser éstos más constantes como marcadores del sistema. No obstante, también se pueden tomar las enzimas para construir un grafo, tomando como interacciones la conexión entre enzimas que dan los metabolitos mismos.

En algunas ocasiones, se puede construir un grafo usando más de un tipo de nodo. Esto puede usarse para comparar datos que se pueden agrupar en entes de distinta índole. Un ejemplo de esto es si se busca analizar, a nivel de nodo, tanto el rol de los metabolitos involucrados en los procesos de un organismo vivo como el de las enzimas que los procesan. Conservan un nivel que los hace comparables (ambos interactúan claramente usando al otro como intermediario) pero son necesariamente distintos. A este tipo de representación se le llama grafo bipartita (de dos partes). En algunos casos se puede hablar de grafos multipartitas, de más de dos clases de elemento. No obstante, la mayoría de los estudios de redes se han centrado en analizar perspectivas que no requieran este tipo de representación, pues los grafos multipartitas se organizan de formas que los hacen menos susceptibles de análisis por sus propiedades estadísticas. Estas propiedades son más visibles cuando sólo se define un solo tipo de elemento a usar.

El análisis de grafos fue, en sus primeras épocas, casi por completo exclusivo de las matemáticas, y por ello se destinaban métodos estrictamente formulados desde las matemáticas para la construcción y el análisis de estas estructuras. La aplicación del análisis de grafos en la biología se popularizó de forma relativamente reciente, cuando se aplicó al sistema regulatorio y metabólico de *Escherichia coli* [2]. Desde entonces, el enfoque ha ganado terreno hasta convertirse en una de las herramientas auxiliares más utilizadas en el análisis de sistemas biológicos.

La razón por la cual se ha buscado tanto analizar propiedades estadísticas de las redes biológicas (grafos que intentan representar el comportamiento de un sistema biológico) es porque éstas dan pautas generales del comportamiento del sistema que son fácilmente trasladables a la función biológica. Al paso de los años se han descubierto diversas propiedades y estructuras en las redes biológicas que, cuando se les compara con el funcionamiento real del sistema, rinden información relevante o embonan bien con diversas funciones observables experimentalmente.

Tener propiedades estadísticas o estructurales como una herramienta base es clave en vastas redes biológicas, que por su mismo tamaño y complejidad dificultan un análisis inmediato de

todas sus propiedades por ser éstas poco observables en una representación gráfica. Es necesario tener a mano otros métodos que ayuden en la reorganización de la información involucrada de forma tal que el análisis, visual o de otro tipo, se facilite. Por otro lado, una posibilidad que siempre ha despertado gran interés es comparar estos elementos no sólo dentro del modelo individual a estudiar, sino con modelos provenientes de varios organismos, a fin de detectar posibles patrones evolutivos en la construcción de estas redes [2, 8, 6].

Por lo general, las redes que se analizan en biología son grafos dirigidos o mixtos (una combinación de elementos con y sin dirección, que puede usarse para reflejar la presencia de interacciones sin sentido fijo, como las reacciones reversibles) con aristas sin peso cuantitativo, aunque se les dan características según su naturaleza (positiva, negativa, etc.) que regularmente no se consideran relevantes para el análisis estadístico. Se inició este enfoque en redes metabólicas [22], pero se extendió a redes de regulación, dando preferencia al inicio de la transcripción [2] en un enfoque que resume el ciclo de vida de una proteína para enfocarse en la influencia de su transcripción en otros elementos (figura 2); en el caso de la regulación transcripcional, por ejemplo, al obviar toda la maquinaria de transcripción-traducción y asumir que siempre está activa, se puede reducir el sistema a las influencias transcripcionales que tiene cada elemento y construir el grafo respecto de éstas.

Además de recurrir al análisis computacional para establecer la relevancia de este enfoque, se han complementado los resultados disponibles con éxito mediante métodos experimentales diversos, además de metodologías basadas en este enfoque que incrementan la información disponible [8]. Uno de los más relevantes es la construcción de microarreglos, ya mencionado. Estos métodos tienden a complementarse con metodología de análisis individual para definir los detalles de elementos particulares y afinar las perspectivas generales.

Las propiedades del grafo que se pueden elucidar con estas herramientas abren la posibilidad de estudiar la organización del sistema a partir del arreglo de sus elementos apreciando construcciones entre varias escalas. Al realizar el análisis estadístico, se espera obtener parámetros que den una idea inicial sobre el comportamiento del sistema como un todo. La más utilizada de estas propiedades es el grado del nodo, es decir, la cantidad de aristas que se

conectan a él. En la mayoría de las redes que reflejan sistemas reales, el grado de los nodos tiende a no ser homogéneo, y puede organizarse con el fin de observar su comportamiento general [5, 2].

Otra propiedad frecuentemente usada es el camino. Un camino es una ruta que se puede tomar, a través de las aristas, de un nodo a otro, respetando las direcciones donde la haya. Los caminos se toman a consideración porque dan una idea general de qué tan densamente conectada está la red; la densidad relativa de conexiones es mayor si, en general, es más fácil llegar de un nodo cualquiera a cualquier otro. Los más relevantes son el camino más corto de un nodo a otro, conocido como camino mínimo, ya sea al promediar los valores de todos los caminos mínimos de la red o al tomarlos para ver la distribución de su comportamiento; el camino más corto de la red; y el camino más largo de la red, conocido como diámetro [5, 2].

La razón por la cual el grado del nodo ha sido tan usado como propiedad unificadora es que permite el acceso a distribuciones relativas al número de conexiones que dan propiedades generales de primera instancia para el comportamiento de la red y, por extensión del sistema que modela. En muchas ocasiones, se prefiere eliminar la direccionalidad de la red a analizar pues incluir la dirección crea problemas relativos de representación, sobre todo al comparar entradas con salidas, que son difíciles de modelar satisfactoriamente; además, por lo general se considera que estas propiedades no se ven demasiado alteradas por la dirección de la interacción [5, 2].

Las propiedades basadas en el grado que han permanecido como relevantes se describen en la figura 3. Una de estas propiedades es la distribución de conectividad o *P(k)*, donde *k* es el grado de un nodo, la cual representa qué tan probable es que un nodo cualquiera tenga un cierto grado *k*. Se usa para comparar la influencia relativa que distintos rangos de grado ejercen sobre la estructura. En general, interesa distinguir entre dos tipos de red cuya organización inherente es distinta, así como las consecuencias biológicas de esta organización. El supuesto central detrás de esta perspectiva es que las redes biológicas adquieren elementos nuevos a través de duplicaciones génicas o adquisiciones externas como la transferencia horizontal, y que el contexto de los nuevos elementos depende de su asociación con los ya existentes. Por ejemplo,

al duplicar un gen, se duplican también los sitios regulatorios y funcionalmente se duplica también la influencia; a su vez, elementos adquiridos de fuera empiezan como piezas levemente asociadas con el resto, y van adquiriendo relevancia funcional y sistémica a medida que su uso se va haciendo necesario para el organismo y se refina su regulación. Las reconexiones que puedan resultar de, por ejemplo, mutaciones al azar, rara vez modifican todo este entorno de golpe; cambian una característica (sitio regulatorio, afinidad, etc.) a la vez. Por ende, una conexión tiende a conservarse entre más esencial sea para el organismo en sí, los elementos más esenciales tienden a tener una influencia mayor en los demás y éstos son estadísticamente más propensos a ampliarla y conservarla sobre los elementos nuevos.

El primer tipo de red con el que se pretende comparar para ver este supuesto en acción sigue una distribución aleatoria, en la cual todos los nodos están conectados con otros al azar y si se integrara un nuevo elemento a la red es igual de probable que se conecte con cualquiera de los existentes. Lo que esto implica es que ningún grado tiene una influencia diferencial sobre la estructura inherente de la red, lo cual hace a las redes aleatorias útiles como controles en varios análisis de redes.

El segundo tipo de red deja una distribución que sigue una ley de potencia (una distribución del tipo $P(k) \sim k^{-\square}$, donde □es conocido como el exponente de grado), en cuyo caso la red es conocida como libre de escala. La característica principal de las redes libres de escala son nodos especiales denominados *hubs*. Un *hub* concentra en sí mismo una cantidad de aristas individuales mucho mayor que el nodo promedio (y por ende más rutas a otros nodos), lo cual refleja su importancia para la red en su conjunto. Siguiendo el supuesto arriba mencionado, si se integrara un nuevo nodo a la red, es significativamente más probable que se conecte con un *hub* que con cualquier otro elemento. En biología, los *hubs* han sido relacionados con elementos de gran relevancia para el sistema celular, como reguladores globales en el caso de las redes de regulación transcripcional o metabolitos de uso muy difundido en las redes metabólicas [2]. En específico, podemos hablar reguladores como CRP en *E. coli* o CcpA en *Bacillus subtilis*, o de metabolitos como agua o cofactores como ATP o CoA en el caso del metabolismo, que constantemente surgen como *hubs* en distintos organismos.

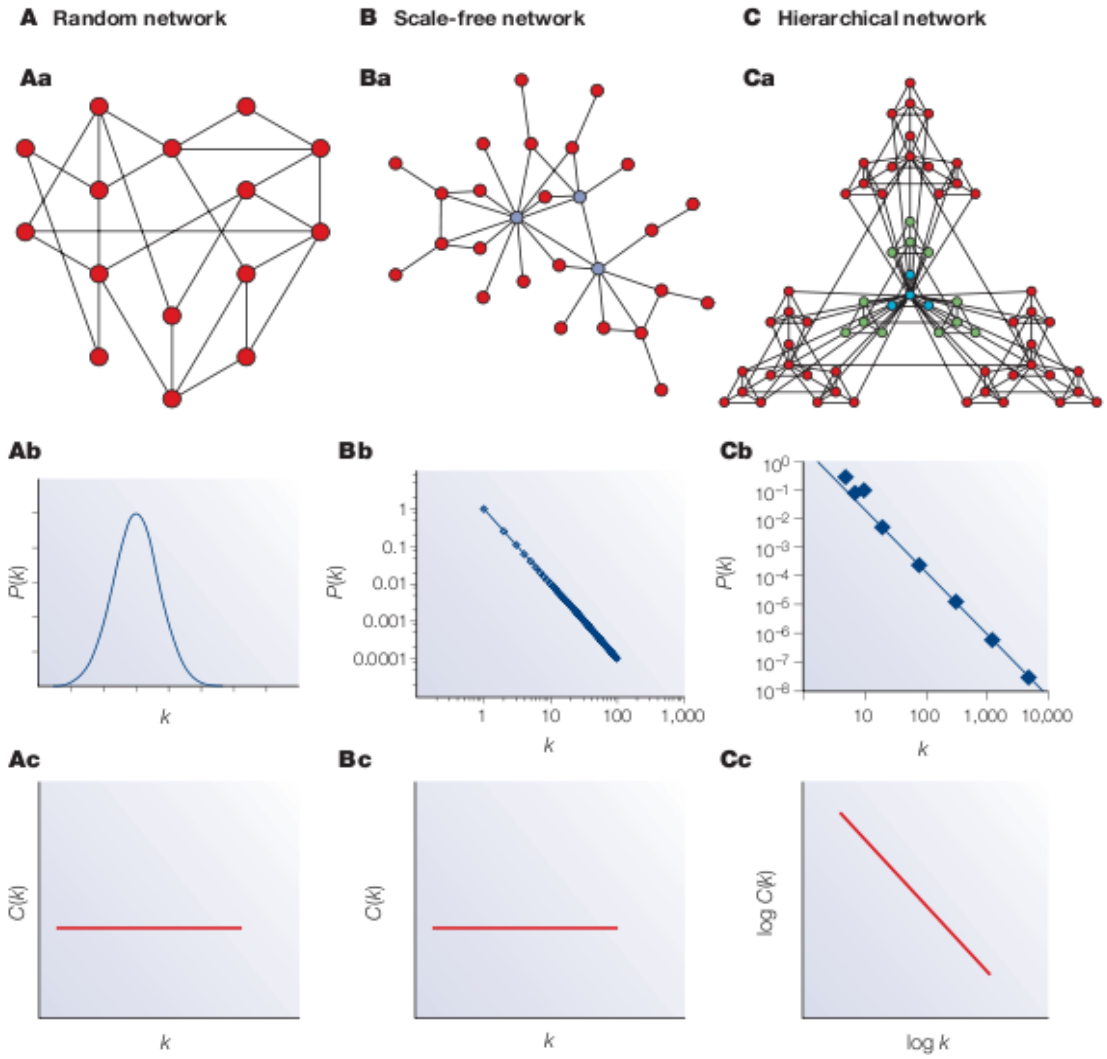


Figura 3. Comparación de los tipos de modelo de red originalmente propuestos por Barabási *et al*. La red aleatoria puede identificarse por una P(k) con distribución normal y una C(k) de distribución uniforme. La red libre de escala puede identificarse por una P(k) con distribución de ley de potencia y una C(k) de distribución uniforme. La red jerárquico-modular puede identificarse por una P(k) y una C(k) de distribución de ley de potencia. [2].

Otra distribución útil es la del coeficiente de clustering, o *C(k)*. El clustering de un nodo individual es la razón entre el número de conexiones entre los vecinos de ese nodo y las máximas que pudieran darse, lo cual refleja qué tan unida es la comunidad de nodos que rodea al nodo de interés. Como ejemplo, en una red social, el clustering sería la proporción que es el número de relaciones sociales que sostienen entre sí los conocidos de un cierto individuo del número máximo de relaciones sociales que pudieran darse entre ellos. Una razón alta refleja que la persona forma parte de una comunidad estrecha, y una razón baja refleja a alguien que conoce muchas personas que entre sí tienen contacto limitado o nulo.

La *C(k)* es el clustering promedio de todos los nodos en la red por cada grado *k*. Se ha observado que las redes para las cuales $C(k) \sim k^{-\gamma}$, donde $-2<\gamma<-1$, siguen una distribución conocida como jerárquica modular (figura 3). Esta distribución presenta *hubs*, pero también presenta comunidades de nodos estrechamente interconectados, que son parte de comunidades progresivamente menos conectadas; los *hubs* pueden entrar conectando a las comunidades progresivamente más grandes con cada vez mayor prevalencia. Las pequeñas comunidades de nodos son conocidas como módulos. Los módulos también son de gran relevancia en redes biológicas, ya que se ha demostrado que al representar una red biológica de esta forma alinean con grupos funcionales de genes o metabolitos con una gran precisión [5, 2].

Los módulos en redes biológicas se pueden obtener según la estructura topológica que haya rendido la red o puede inferirse mediante datos experimentales. Cuando existen datos informativos para formar una estructura con clusters, por ejemplo, se pueden aplicar algoritmos de clustering, los cuales ya están estandarizados para su aplicación en redes biológicas [11]. Existen varias modalidades de algoritmos de clustering, como el jerárquico aglomerativo, que pueden relacionar elementos en datos de experimentos masivos (como un microarreglo).

Al aplicar los algoritmos de clustering sobre estos datos, de nuevo se ha visto que las estructuras biológicas correlacionan con las estructuras topológicas definidas. Los módulos topológicos son estructuras que constantemente muestran coexpresión en los datos de

experimentos masivos, y por ello de alta relevancia biológica. Los nuevos datos pueden dar una mayor precisión sobre la naturaleza de los distintos módulos y su comportamiento según una cierta condición [6].

Con base en observaciones recientes, se ha visto que algunas redes biológicas no siguen con exactitud los parámetros requeridos para un análisis basado en clusters [2]. De hecho, se ha observado una desviación al principio de la distribución que este método no considera [13], ilustrada en la figura 4. Esto pone a consideración algunas debilidades del enfoque de redes, en particular el cuidado que debe ponerse cuando se intenta determinar propiedades globales de un sistema. El asumir supuestos sobre la estructura global de una red es un asunto delicado que no puede ser atacado sin una base sólida de información relevante.

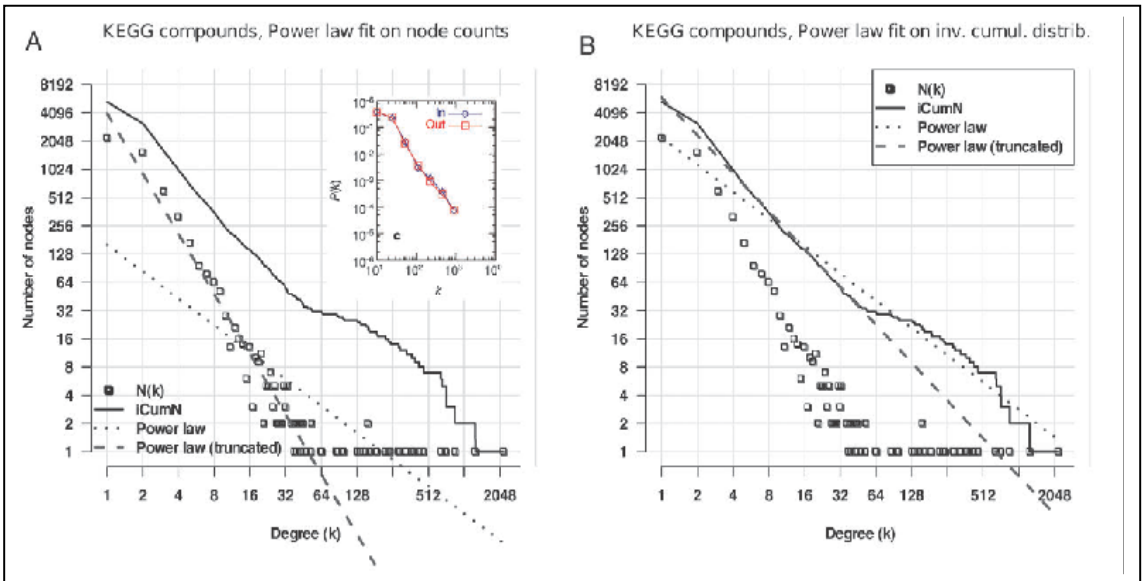


Figura 4. Ajuste de las distribuciones de grado en datos de KEGG. La abscisa es el grado k. La ordenada es la P(k). N(k): distribución normal. iCumN: acumulada inversa. Power law: ajuste de toda la N(k) a una ley de potencia. Power law (truncated): ajuste de la N(k) al no tomar en cuenta la cola de la extrema derecha. El inciso pequeño en A muestra la distribución originalmente provista para justificar la existencia de la distribución. Tomado de [13].

Debido a la aparición de resultados de este tipo, se ha buscado determinar la naturaleza de la desviación que se observa y ayudar a reestructurar o complementar la metodología para que refleje con mayor fidelidad la realidad biológica que se intenta retratar con estos métodos. Uno de los más recientes ha sido el acoplamiento de flujos, que se basa en la determinación de rutas lineales entre nodos, muy comunes en las redes metabólicas [14]. Debido a que este método no busca hallar clusters, como es el sentido aceptado del término, sino que toma como referencia una hipótesis canónica sobre la estructura de la red (la constitución de la red misma como un conjunto de rutas lineales interconectadas), el problema de la relación de rutas se puede atacar sin una dependencia directa de las interpretaciones basadas en el agrupamiento según la acumulación relativa de conexiones a las cuales están sujetos los métodos de clusterización. Cabe aquí distinguir el método de acoplamiento de flujos del análisis estequiométrico de flujos metabólicos, establecido por Varma y Palsson en los años 90 para la interpretación del metabolismo central de *Escherichia coli* [53]. Este último hace uso del conocimiento de la estequiometria en cada una de las reacciones de una red metabólica determinada para la estimación de flujos. El método de acoplamiento de flujos es un método topológico que analiza a la red metabólica asignando un valor de aproximación a una ruta lineal a cada nodo según el número de conexiones.

El método determina el grado de acoplamiento que posee un nodo respecto del resto según el número y naturaleza de las conexiones que tiene. Se considera que el acoplamiento es completo cuando el nodo sólo participa en una ruta lineal. Cuando el nodo entra en contacto con otra ruta lineal, se separan como tres rutas y el nodo se considera como parcialmente acoplado a las tres. En caso de que haya más conexiones a otras rutas, se considera que el nodo no tiene acoplamientos [3].

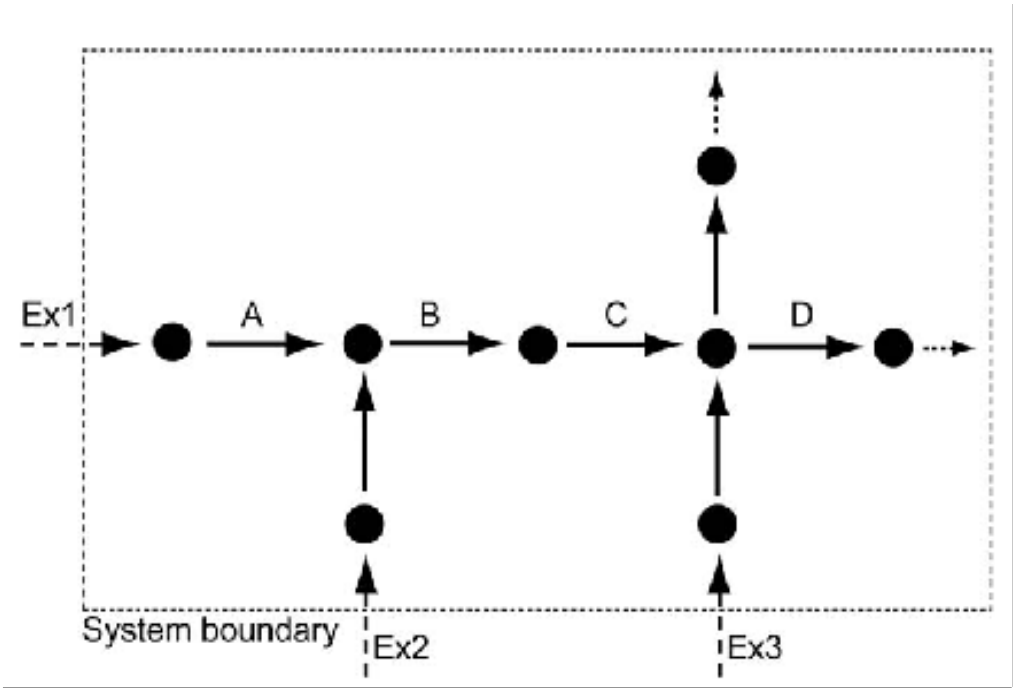


Figura 5. El acoplamiento de flujos se basa en hallar rutas lineales definiéndolas con base en su relación con nodos más conectados a los cuales se conectan. En este ejemplo, el nodo entre A y B es un acoplamiento parcial, entre B y C acoplamiento completo y entre C y D sin acoplamiento. Imagen de PMID: [3].

A la fecha se han desarrollado diversos métodos para emplear el algoritmo de acoplamiento de flujos. Uno emplea supuestos sobre cómo la actividad de un flujo influencia la actividad de otro en vez de los supuestos direccionales originalmente planteados. Una de las ventajas de este método es que hace una distinción precisa entre flujos unidireccionales y flujos reversibles [2]. Otro emplea un método analítico optimizado que aproxima la solución más viable desde un conjunto reducido del total posible. La ventaja que se proclama sobre el método anterior es que no se requiere reconfigurar la red (dividiendo flujos reversibles en dos conexiones, por ejemplo) para calcular las relaciones de conectividad, lo cual incrementaría el número de variables a procesar, sino sólo el tipo de reacción según su reversibilidad [2]). Ambos algoritmos han mostrado que el acoplamiento de flujos es una alternativa lo suficientemente viable para aplicarla al análisis de redes metabólicas, a grado tal que se han iniciado esfuerzos para lograr su estandarización, centrándose en su uso para grafos bipartitas [1].

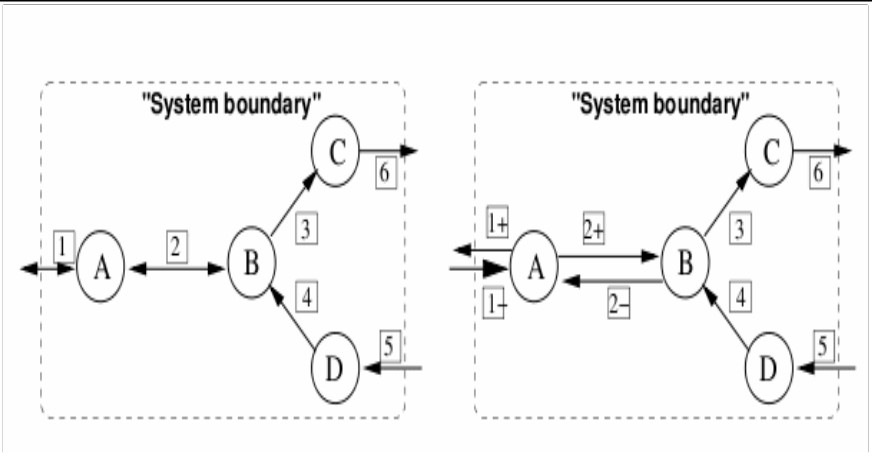


Figura 6. Una debilidad posible del método de Burgard *et al*. es que necesita descomponer las reacciones reversibles en dos reacciones de un sentido, lo cual podría introducir inconsistencias como, si hay un flujo de cero en la reacción 2, la reacción 2- pudiera necesitar un flujo negativo en ciertas circunstancias. La desventaja del método de David es que es más complejo de implementar. [1].

Planteamiento del problema y justificación

Desde los inicios de los estudios de redes, se ha tomado como modelo a *E. coli*, pues la información existente sobre esta bacteria funge como un respaldo confiable para la interpretación de resultados. Muchos otros organismos modelo estudiados con este enfoque han dado información útil y novedosa; el estudio de *Bacillus subtilis*, por ejemplo, ha dado interesantes resultados, sobre todo en su sistema de esporulación. Otros casos productivos han sido en su mayoría procariotes [1], pero la levadura (*Saccharomyces cerevisiae*) también ha rendido resultados valiosos [23]. Pese a esto, los métodos aplicados a redes no detallan todo el comportamiento de la célula; las redes de regulación génica y del metabolismo requieren análisis separados, que sólo han reflejado indirectamente el comportamiento de otros sistemas cercanamente relacionados. Una de las herramientas más útiles en estudio de este tipo son las metodologías de clusterización, que, como se mencionó, dependen de interpretaciones a gran escala que han resultado sujetas a error y son poco trasladables al contexto de otros sistemas.

Los primeros pasos hacia la comprensión unificada de regulación y metabolismo han descansado en comparaciones indirectas: tomar uno de los sistemas, estudiarlo, y transferir las conclusiones a la comprensión prevalente del otro, como trasladar resultados y conclusiones en regulación a metabolismo [6]. Un ejemplo de esto es lo realizamos en un estudio previo, en donde a partir de datos de micro arreglos y la red de regulación transcripcional en *B. subtilis*, generamos módulos que resultaron encontrarse funcionalmente relacionados pero que en muchas ocasiones toman secciones diferentes del metabolismo, no necesariamente conectadas entre ellas a nivel de rutas. Un ejemplo de ello es el módulo comprendido por fuentes de carbono y regulado por el factor transcripcional global CcpA que agrupa a casi el total de genes los involucrados en la degradación de fuentes de carbono y que metabólicamente se observan como secciones independientes del metabolismo. Así mismo, encontramos que el sistema PTS que regula la toma de glucosa en este organismo se obtiene a través de la organización regulatoria como un módulo desconectado del de toma de fuentes de carbono [6]. Otros estudios, como el de Goelzer *et al.*, han dado los primeros pasos hacia una unificación formal, tomando estructuras regulatorias genéricas (como TRAP) y metabolitos relevantes, ligándolos

según los efectos en la producción del metabolito o la función reguladora por medio de compuertas lógicas que reflejan los requerimentos del cambio [24].

Por otro lado, el estudio unificado de estos sistemas se entorpece por la carencia de parámetros que permitan determinar las relaciones entre los distintos tipos de molécula que participan en estos sistemas con la claridad que un estudio unificado necesita. Mucho de esta dificultad yace en el metabolismo mismo, para el cual se emplean criterios de remoción del ruido altamente arbitrarios que tienen poca consideración por lo que se desea ver en el análisis. De hecho, otro de los principales puntos que se consideran actualmente [13] es que, a pesar de que muchos metabolitos que abarcan lugar de *hub* tienen funciones altamente relevantes, generan conexiones irrelevantes entre elementos, incrementando considerablemente el ruido que se genera dentro de la red. El ejemplo más inmediato de esto es el agua, que crea conexiones poco relevantes entre compuestos carbonados que de otro modo caen en rutas separadas debido a su rol en la hidratación e hidrólisis. Los estudios disponibles de la red de regulación manejan únicamente factores transcripcionales, agregando sigmas a lo sumo, pero cabe la posibilidad de que se caiga en un escenario similar si se desea agregar otros tipos de molécula. Es necesario generar criterios que permitan resaltar las reacciones que más interesantes resultan al estudiar el metabolismo, a fin de adquirir mayor claridad en su estudio.

Hipótesis

Creemos que es posible generar un método teórico que permita la integración de redes regulatorias y de metabolismo, a fin de elucidar sus propiedades estadísticas y obtener arreglos modulares usando como modelo biológico el metabolismo de *Bacillus subtilis* y su regulación a nivel del inicio de la transcripción.

Objetivo General

Generar criterios que permitirían el estudio unificado de las redes de regulación y metabolismo de *Bacillus subtilis*.

Objetivos específicos

- Construir el modelo biológico de *Bacillus subtilis* con base en información de bases de datos y literatura reciente y confiable.

- Evaluar las propiedades estadísticas de la red de regulación por factores transcripcionales y factores sigma.

- Evaluar las propiedades estadísticas de la red correspondiente al metabolismo.

- Encontrar un método que permita integrar la información de ambas redes

- Evaluar el tipo de distribución y la formación de módulos en la red integral.

Metodología

Fuentes de datos

La red de metabolismo se construye con datos de KEGG, una base de datos de metabolismo en donde se incluyen las rutas descritas hasta la fecha en *B. subtilis*, organismo que utilizaremos como modelo de estudio [19]. De esta fuente, tomamos los archivos formato XML, denominados KGML, sobre los cuales se efectúa una curación manual de todas las interacciones metabólicas disponibles para *B. subtilis*. De este proceso de extracción se generan dos sets de datos: uno con datos crudos construidos a partir de los archivos en KGML de KEGG, y que en el resto del texto será llamado del XML, y un segundo archivo también procedente de los datos crudos de KEGG sobre el cual se realizó una curación manual, referido por este nombre por el resto del texto.

La red de regulación se construye usando datos de DBTBS, una base de datos que contiene operones de *Bacillus subtilis* y su regulación a nivel de inicio de la transcripción [20] y SubtiList, una base de datos de genes de *B. subtilis* organizados a través de sus clasificaciones funcionales [15]. En ambos casos obtuvimos archivos en texto plano que fueron tratados con programas escritos en el lenguaje de programación Perl.

Un set adicional fue utilizado en este estudio, tomado del proceso de curación realizado por *Goelzer et al*. [24], en vías involucradas en el metabolismo central y que sobre el cual se realizarán análisis topológicos, que a su vez complementarán los archivos obtenidos de KEGG y DBTBS.

Construcción de las redes regulatorias y metabólicas de B. subtilis.

Regulación: La red de regulación se estructura tomando como elementos factores transcripcionales, factores sigma y los genes regulados, de forma tal que se tienen los elementos que conforman la interacción y su signo (positivo, negativo).

Metabolismo: Se forma la red metabólica tomando como elementos los metabolitos o enzimas participantes y elementos celulares que interactúan directamente con productos del metabolismo. Se genera en primera instancia un grafo bipartita, representado con un formato que ilustre las conexiones entre reacciones de KEGG y metabolitos, que después es procesado para obtener elementos de un solo tipo, ya sea metabolitos, reacciones o proteínas involucradas, en un formato similar al usado para la red de regulación (figura 11).

Para ambas redes se generan también tablas que representan la longitud de los caminos entre nodos como la longitud del camino más largo entre los caminos mínimos (diámetro).

Determinación de las propiedades estadísticas

Se emplean programas hechos en casa que siguen el método delineado por Barabási *et al*. para la construcción de módulos con base en las distribuciones de *P(k)* y *C(k)* [2]. Se calculan las

propiedades para cada nodo y de ellas se infiere la distribución general, la cual se analiza en detalle para determinar los módulos según el grado de mutua conexión de los nodos.

Filtros a las conexiones

Red Regulatoria. Para la red de regulación se genera, para este trabajo, un filtro basado en los datos de DBTBS para las evidencias más confiables, usando los denominativos de la base de datos para separar los tipos de evidencia existentes. Se toman como fiables los datos provenientes de experimentos verificables (footprinting, asociación de proteínas) y se desechan los que no proveen evidencia directa sobre la interacción entre genes, como los arreglos y las evidencias estrictamente computacionales. Además, se genera set de datos adicional para los datos de DBTBS que no incluye el factor sigma A.

Red Metabólica. Para esta red, se construyó un filtro arbitrario enfocado en el parecido que el peso de un metabolito tiene con otro, en función de los elementos que lo conforman y el número de átomos de cada elemento que en cada metabolito está representado. De este modo se asignó el valor más alto se asignó al carbono (las razones de esta elección serán descritas como parte de los resultados y discusión), que cuenta las conexiones según la similitud de los compuestos involucrados. El filtro asigna a cada tipo de átomo un valor numérico: 10 para carbono, 8 para nitrógeno, 6 para azufre, 4 para fósforo, 2 para oxígeno, 1 para hidrógeno y 2 para otros átomos. Sumando los valores por átomo en la fórmula del compuesto, se calcula el peso del compuesto, el cual se compara con el peso de los compuestos con los cuales está conectado por una reacción mediante dos medidas: una medida de similitud total, que divide el peso menor entre el peso mayor; y una medida de verosimilitud de paso, que resta el peso menor del peso mayor. Las conexiones con una similitud total menor al 60% (a la cual el método rinde un 5% de falsos positivos) son ignoradas, y la medida de verosimilitud se conserva para comparar casos notables. Los archivos de elemento único se reconstruyen a partir de los resultados de este filtro.

Análisis para acoplamiento de flujos

El método se aplica sobre las redes metabólicas procedentes de los datos de XML y sobre los datos de la curación. Todos los nodos de grado 2 o menor se toman como totalmente acoplados, los nodos de grado 3 se toman como semiacoplados (acoplamiento parcial) y los nodos de grado 4 o mayor se consideran faltos de acoplamiento (figura 5). El método toma este resultado y estructura las rutas lineales a partir de los nodos con grado 1 y 2 que están conectados entre sí, tomando los grados 3 y mayores como limítrofes. Posteriormente el método toma estas rutas y determina cuales están conectadas por un nodo límite en sus dos extremos, asignando estas estructuras como ojos. Luego, devuelve las rutas lineales y los ojos, rutas lineales conectadas por su inicio y su final a otras rutas lineales, llamadas así por las formas de abertura que introducen en las rutas lineales (figura 8). Finalmente, resume estas estructuras en la red origen como conexiones sencillas de un nodo a otro. El método está diseñado para ser recursivo, a fin de poder resumir todas las rutas lineales existentes en la red.

Método de integración

A partir de los datos generados por el filtro de metabolitos y las rutas acopladas, se eliminan de la red las conexiones metabolito-reacción-metabolito del grafo bipartita que constituyen una conexión suprimida por el filtro y de este recorte se toman los datos para una red de sólo reacciones.

A cada proteína que participa en las reacciones obtenidas como producto del paso anterior, se le asigna el gen o genes que la codifican. Los genes son ubicados en sus operones (con el fin de hacerla mas compacta y visualmente más fácil de interpretar) y relacionados con sus factores de transcripción y factores sigma que los regulan. El resultado de esto es una red unida que resume las conexiones entre reacciones como conexiones entre operones usando como referencia las correspondencias arriba descritas entre genes, reacciones y operones. Estas conexiones pueden entonces anexarse a las regulatorias que ya existen y plegarse como un solo set de datos, el cual se analiza tomando como referencia principal los grupos formados por el acoplamiento.

Visualización

Los archivos nodo origen a nodo destino para las redes generadas son visualizables por Cytoscape [25]; las visualizaciones de red son creadas alimentando el archivo de texto y reacomodando los nodos de forma automática por la opción y files, por el arreglo específico definido como orgánico, que tiene por objeto organizar los nodos de la red de forma tan natural, como sea posible.

Resultados y Discusión

Construcción y análisis topológico de la redes de metabolismo y regulación.

El enfoque de redes, al ser muy general, necesita definiciones de datos precisas y acertadas para reflejar en el modelo un comportamiento realista. Al seleccionar y separar los datos, se decidió que era preferible tomar evidencias que fueran altamente confiables, prefiriendo las que tienen un respaldo experimental sólido sobre las que no fueran traducibles directamente a un fenómeno biológicamente medible, como las estrictamente computacionales o las altas en ruido. Las bases de datos marcan el tipo de evidencia más fuerte que se tiene en el momento para un gen particular, hecho que se aprovechó para seleccionar las evidencias. Se desecharon evidencias puramente computacionales como las búsquedas por homología de secuencia y los alineamientos, así como evidencias provenientes de experimentos de arreglos sin respaldo de alguna técnica experimental más puntual, como los ensayos de asociación de proteínas. Al analizar la red de regulación transcripcional, se aplicó este enfoque al set de datos proveniente de DBTBS para comparar el efecto de la exclusión de evidencias sobre la estructura de la red. En total, el filtro establece como válidas 862 de las 911 conexiones entre elementos establecidas en el set de datos original. Esta red tiene también una conectividad promedio de 2.92 y una máxima de 337, lo cual indica que es una red de baja densidad para la cual es fácil hacer estudios topológicos. Así mismo, se usa como punto de comparación un archivo que no incluye el factor sigma A, del cual se ha dicho que puede ser una fuente de sesgo al analizar las propiedades de la red por ser prácticamente ubicuo, a fin de establecer otro punto de comparación para ver cuánto se alteran las propiedades de la red con la selección de nodos. Estos filtros se aplican para obtener, al menos en primera instancia, una fuente de datos coherente y limpia que permita los análisis posteriores con éxito.

Al iniciar el análisis de las fuentes de datos de metabolismo, se analizaron los archivos planos disponibles en KEGG para el metabolismo de *B. subtilis*. Durante la revisión, se encontró que el archivo de la base de datos tiende a eliminar las relaciones entre elementos que introducen conexiones en las rutas metabólicas con el fin de preservar la visión dada previamente por esta estructura. Los casos más influyentes son metabolitos moneda, que introducen alta

ambigüedad por su alta frecuencia entre reacciones, pero en algunos casos se encontró que esta eliminación también se da en metabolitos importantes pero que no están tan difundidos, lo cual podría modificar la estructura de la red en formas poco visibles. Un ejemplo es la reacción 4.1.3.27, la antranilato sintasa (producto del gen *trpE* en *Bacillus subtilis*), que requiere L-glutamina y libera L-glutamato y piruvato. El piruvato está muy difundido y en ciertos casos se le podría tomar como un metabolito moneda, pero los dos aminoácidos mencionados no llegan a ese rango, pese a que también concentran un flujo importante, y pudieran brindar información adicional sobre el comportamiento de la red. Pese a este hecho, ninguno de los tres compuestos mencionados es tomado en cuenta para la reacción en los archivos planos de KEGG. Por ello, revisamos los criterios con base en los cuales se eliminan las relaciones entre metabolitos y enzimas con el fin de precisarlos y elaboramos criterios propios los cuales fueron comparados con los archivos originales de KEGG.

La primera tarea que se da con los datos curados de todos los grupos de datos es hallar la estructura de la red. Para lograr esto se prefiere hallar criterios particulares que mejor definan la estructura de cada red por separado, en vez de contemplar la posibilidad de un criterio universal de manera inmediata. Primero, se prueba la red de regulación transcripcional con los criterios establecidos por Barabási [2], con programas hechos en casa que siguen rigurosamente estos criterios, tanto con el set de datos de Goelzer *et al*. como con el proveniente de DBTBS (con y sin las evidencias establecidas como desechables). Los resultados de este proceso se ilustran en la tabla 1, que muestra los valores promedio de *P(k)* y *C(k)* de cada red. Así mismo, en la figura 7 podemos observar las distribuciones obtenidas para cada red. Se nota que, a pesar de que el proceso arroja valores similares de *P(k*) y *C(k)* para todos los sets de datos y razonables según los supuestos del procedimiento ($\gamma \approx 1$), el ajuste no es perfecto y se detectan las desviaciones notadas por Lima-Méndez y van Helden mostrado en la figura 4 de la sección introductoria. Al principio de tres de las cuatro distribuciones analizadas, se nota una curva ascendente en la distribución que no es compatible con una distribución potencial, cuya naturaleza se encuentra poco descrita en la literatura. La desviación propuesta se hace más evidente cuando se observan los datos a escala log-log, lo cual revela el nivel de desviación que tiene un punto particular de la línea de tendencia, que se aprecia como una

recta. Se procede a determinar el comportamiento de estos grupos de nodos a fin de comprender mejor la estructura de la red y posiblemente elaborar un criterio que pueda describirla con mayor exactitud.

	DBTBS			Goelzer
	Original	sin *sigA*	filtro de evidencias	
P(k)	$0.09k^{-1.17}$	$0.09k^{-1.23}$	$0.08k^{-1.14}$	$0.25k^{-0.98}$
C(k)	$2.93k^{-1.32}$	$1.65k^{-1.2}$	$2.95k^{-1.34}$	$0.66k^{-0.79}$

Tabla 1. Datos de distribuciones generales de la red de regulación, tomando encuenta varios sets de datos. El set de datos Goelzer resume los datos de regulación de Goelzer *et al*. El set de datos DBTBS original toma todos los datos de DBTBS. El set de datos DBTBS sin sigA suprime la influencia de *sigA* en la estructura de la red. El set de datos DBTBS filtro de evidencias toma en cuenta a *sigA*, pero suprime conexiones de acuerdo al filtro de evidencias para la red de regulación propuesto en la sección de métodos.

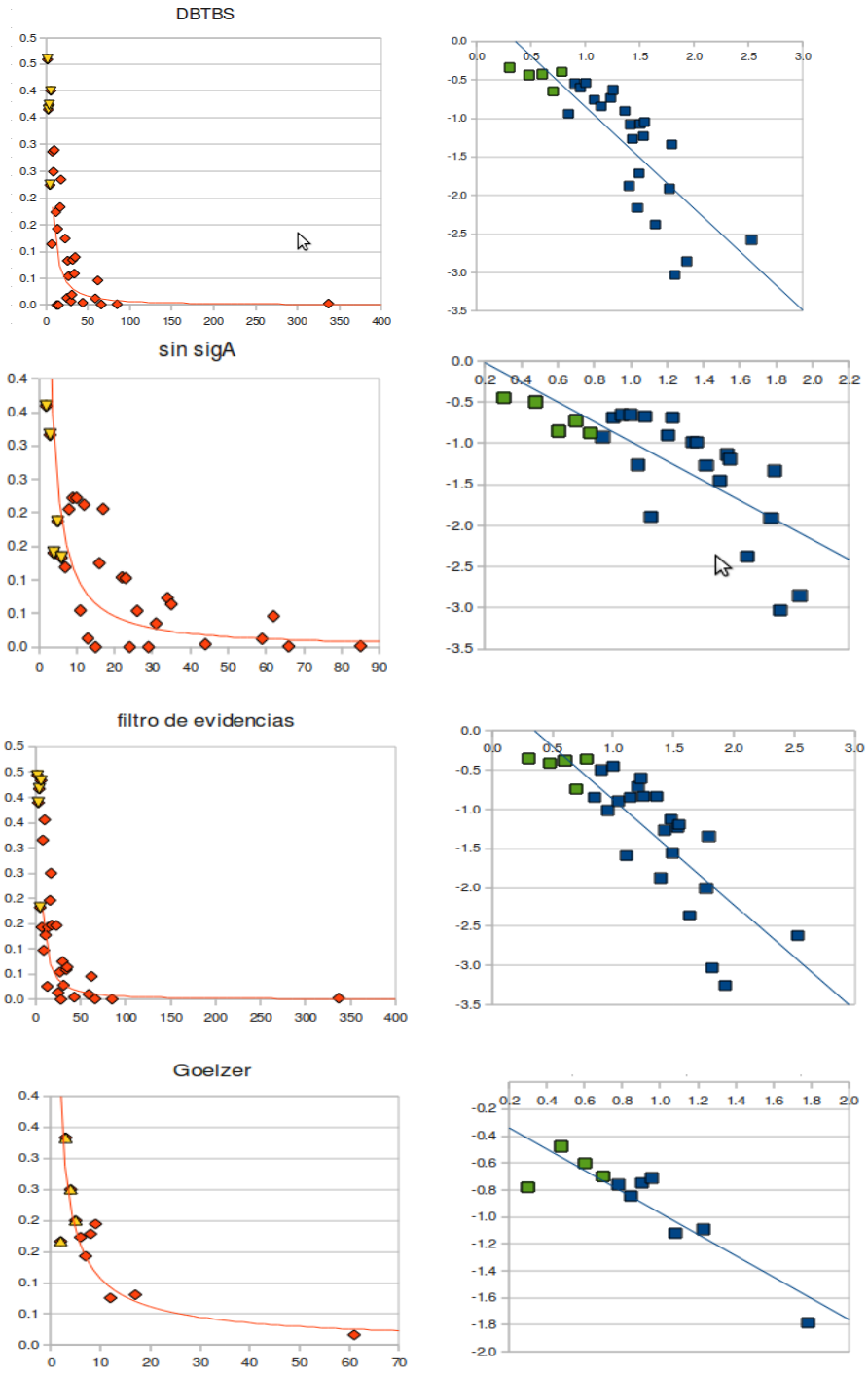


Figura 7. Las graficas describen el delineado de los resultados de las distribuciones para cada grupo de datos obtenidas con distintos filtros o fuentes de la red de regulación de *B. subtilis* descritos en la tabla 1. Para cada una de ellas, en el caso de los gráficos de la Izquierda, en el eje horizontal tenemos representada: la conectividad (*k*); Eje vertical: *C(k)*; Amarillo remarca el inicio de la distribución de la *C(k)*. Para el caso de los gráficos de la derecha: Eje horizontal: log(*k*); Eje vertical: log *C(k)*; Verde remarca el inicio de la distribución de la *C(k)*. Las líneas son las líneas de tendencia.

En cuanto a la estructura de la red metabólica, se analizaron inicialmente 2 redes. La primera conteniendo los datos directos de los archivos XML de KEGG y aquella sobre la cual se insertaron las relaciones omitidas en los archivos planos (red curada), podemos observar en ambos casos que los valores de *P(k)* y *C(k)* distan mucho de aproximarse a los propuestos para describir una red libre de escalar con propiedades jerárquico modulares. La distribución como se puede observar en la figura 8a y b, no se ajusta a una ley de potencias, razón por la cual observamos que el método propuesto por Barabási es inadecuado para describir la estructura de la red. Lo cual nos motivo para probar otros métodos para el análisis de grafos. El método seleccionado fue el acoplamiento de flujos como un enfoque alterno que pueda definir mejor la estructura de la red.

No obstante, las observaciones hechas por Lima *et a*l [13], relacionados con las propiedades estadísticas de la red de metabolitos indicaban que era necesario prescindir de alguna relaciones poco informativas y sobre representadas en la red metabólica. Un ejemplo de ello son el agua o cofactores conectando a elementos de rutas poco relacionadas desde la perspectiva del flujo de carbono. En la tabla 2, mostramos a los primeros 20 compuestos y sus valores de conectividad. Como podemos observar, sí usamos como criterio de eliminación parámetros esencialmente estadísticos, como la supresión de los elementos más conectados, tendríamos que descartar compuestos carbonados que no pueden ser razonablemente eliminados como moléculas menos relevantes, piruvato o Acetil-CoA, ya que son metabolitos que participan en rutas metabólicas como intermediarios cruciales. Con estos resultados se procede a generar el criterio de limpia para la red metabólica. El objeto del filtro es implementar un criterio de eliminación de interacciones menos significativas que sea más acorde a la biología que un criterio meramente estadístico. Con este objetivo en mente, se asume que las reacciones de mayor relevancia para el estudio del metabolismo son, en primer lugar, las de compuestos carbonados, con la participación de otros átomos (nitrógeno, azufre, oxígeno, fósforo) como secundaria.

Molecula	k	n	c
h2o	524	7902	0.057667888
hplus	394	7571	0.097790005
nadph	248	3843	0.125473423
nadpplus	244	3881	0.130911421
co2	177	1710	0.109784284
nadh	172	3008	0.204542364
pyruvate	168	1534	0.109352723
nadplus	167	3032	0.218743236
udp-l-rhamno	136	2897	0.315577342
dtdp-l-mycar	131	2581	0.303112155
diphosphate	124	779	0.102150538
nh3	117	970	0.142941350
ch3-r	113	3439	0.543457649
petunidin-3-(p	113	3439	0.543457649
r	113	3439	0.543457649
coa	112	704	0.113256113
dtdp-d-desos	107	657	0.115852583
dtdp-3-methy	107	714	0.125903721
norgalanthan	104	635	0.118558626
l-glutamate	100	1235	0.249494949

Tabla 2. Las 20 moléculas con mayor número de conexiones para la red de la curación. Los metabolitos altamente relacionados con el metabolismo de carbono están marcados en amarillo.

Como primera aproximación, se maneja un pesado diferencial burdo, que coloca en primer lugar de relevancia al carbono, al ser el elemento de mayor relevancia en los flujos metabólicos; después a nitrógeno y azufre, por su relevancia en la síntesis de varios compuestos (principalmente aminoácidos; con azúcares complejos en el caso de nitrógeno); y fósforo, crucial como punto de cesión energético, dejando a oxígeno e hidrógeno hasta el final, pues éstos participan más a nivel de cofactores. Se procede a una implementación de este filtro con un programa hecho en casa sobre la red de compuestos, con un glosario de compuestos de la misma base de datos para extraer de él los datos de intercambio de átomos para cada compuesto. Se otorga a cada átomo un peso fijo, y el peso del compuesto se define como la suma de los pesos de todos los átomos que lo componen. Para determinar la significancia se crean tres criterios simples:

Restar el peso menor del peso mayor. Esto refleja la cantidad de átomos que tendrían que cederse en la reacción. Entre más grande sea esta diferencia, menor relevancia tiene.

Dividir el peso menor entre el peso mayor. Esto da una medida de la similitud mutua entre ambos compuestos. Entre mayor sea ésta, más factible se considera el paso.

Ilustrar la cesión relativa de átomos en detalle, como la cantidad de átomos de cada tipo que faltan o sobran en alguno de los compuestos. Esto ayuda a proporcionar una perspectiva más amplia sobre los átomos que se están moviendo. La relevancia puede entonces determinarse según el tipo de átomo al que se le de mayor preferencia.

Se elige dar preferencia al segundo criterio como un método de determinar el valor absoluto de la diferencia entre ambos compuestos.

Una ventaja secundaria de este método es que potencialmente puede centrar la atención en el flujo de otros átomos según se desee, a fin de elucidar las propiedades de estos flujos. La figura 11 proporciona una mirada breve del producto de este filtro luego de la aplicación del acoplamiento. La mayoría de los nodos de la red están concentrados en rutas lineales, dando una minoría de elementos unificadores. Esto también se da consistentemente entre otros tipos de red (*supra*), lo cual demuestra que el enfoque de acoplamiento de flujos es más conveniente en general para representar redes metabólicas. Una vez aplicado el filtro se reconstruye la red de metabólica. Sobre esta red se recalculan las propiedades estadísticas con el fin de observar si encontramos cambios en la distribución que nos indiquen, la posibilidad de obtener por el método clásico de Barabási un organización jerárquico modular. Los resultados, en la tabla 1-figura 8c, muestran que, a pesar del filtro, la red sigue mostrando un comportamiento no jerárquico y que no es posible obtener relaciones modulares por este método. Por esta razón se eligió el acoplamiento de flujos como método alternativo.

Criterios de integración

Para obtener datos claros para el algoritmo de acoplamiento y la posterior inclusión de la regulación, se procede a establecer criterios de mapeo lo suficientemente estrictos para

asegurar la mayor fidelidad a la realidad biológica alcanzable por el modelo. Se establecen como sigue:

a. Los elementos deben ser del mismo orden. Se requiere esto para establecer con precisión la naturaleza del sistema a analizar. Es por esto que en el metabolismo se optó por usar reacciones primero en vez de las proteínas involucradas de manera directa, pues las proteínas pueden actuar solas o en conjunto, y no siempre de la misma manera. En la regulación, así mismo, se pueden tomar en cuenta no sólo TFs, sino factores sigma o incluso RNAs como elementos alternos, lo cual hace importante encontrar qué tan iguales pueden considerarse estos elementos. Se prefiere asociar TFs con sigmas, debido a que ambos son proteínas y siguen el mismo mecanismo de acción.

b. Las relaciones entre los elementos deben ser tan homogéneas como sea posible. Así como debe haber un solo tipo de elemento, se prefiere que exista un solo tipo de conexión entre los elementos de la red, lo cual simplifica la representación que se quiere lograr y define la estructura misma de la red. Es posible construir redes que tengan distinciones entre conexiones, siempre y cuando compartan un rasgo unificador. Una red de regulación transcripcional, por ejemplo, distingue entre regulaciones positivas y negativas, pero incluye únicamente interacciones de regulación transcripcional (sólo regulador - regulado). Un factor sigma no puede bloquear la transcripción de un gen, pero esto puede obviarse si simplemente se consideran todas sus interacciones como positivas. Si dos factores se asocian al mismo sitio, estos aparecen conectados únicamente al destinatario de la regulación. Pudieran estar conectados entre sí, para indicar su estado como correguladores del gen del sitio regulado, pero no lo están. Se excluye esa información explícita sobre corregulación no solo porque es redundante e infla artificialmente el número de conexiones, sino porque introduce un tipo de interacción demasiado distinto del que se desea manejar, que es la influencia en la transcripción por asociación del factor al sitio regulatorio. Podría elaborarse una red conectando por participación en situaciones de corregulación a esencialmente los mismos nodos, pero su estructura y propiedades serían distintas de las de la red de regulación transcripcional. De forma similar, se prefiere tener un único tipo de arista cuando no se puede

tener un solo tipo de nodo. En el caso de un grafo bipartita metabolito - enzima, los metabolitos y las enzimas se manejan como nodos siendo que se trata de tipos distintos de molécula. El factor unívoco son las aristas, que siempre son conexiones entre un metabolito y una enzima, pero no entre nodos de la misma clase.

c. Las relaciones entre los elementos deben ser uno a uno; minimizar la redundancia incrementa la claridad, y evita una falsa sobrerrepresentación respecto a las conexiones que algunos nodos pueden tener.

La forma en la cual se plantea integrar la red metabólica y la regulatoria como una sola bajo estos criterios de mapeo consiste en lo siguiente: primero, se toma la red metabólica y se define con base en las reacciones, quedando la influencia de los metabolitos implícita en las relaciones de dador o receptor de sustrato entre las reacciones. Fue sobre esta representación que se aplicó el método de acoplamiento, y los grupos funcionales se definen principalmente con base en los grupos obtenidos a este nivel.

Después se procede a redefinir cada nodo de esta red como los operones que participan en la reacción, se trate uno o varios. Los operones que quedan conectados entre sí son los que participan dentro de una misma reacción y los que están relacionados por cesiones de sustrato entre las reacciones, lo cual indica que quedarían conectados los operones que codifican para isozimas o para subunidades de la misma enzima.

El resultado es una red de interacciones metabólicas entre operones, la cual se puede integrar intuitivamente a la red de regulación; sin embargo, el mecanismo de la conexión deja de ser exclusivamente enzimático para incluir el mecanismo regulatorio, que es muy distinto. La red unificada, entonces, queda definida no como una red de influencias moleculares específicas, sino como una red de flujo de información entre operones, con las aristas definiendo el tipo de información que se transmite.

Para desarrollar este método, se deben tener los métodos mas adecuados para procesar debidamente los sets de datos. Las metodologías disponibles de acoplamiento de flujos [14, 3] requieren que haya en la red una distinción clara sobre la reversibilidad de las reacciones. Sin

embargo, en los archivos planos de KEGG no se aclara esta distinción con la precisión esperada. La curación automática de la base de datos tampoco resolvió esta limitante de forma satisfactoria, por lo cual se optó por modificar el algoritmo como sigue:

Se emplean los identificadores de reacción en vez de genes o enzimas para eliminar la redundancia causada por isozimas o el agrupamiento de subunidades. En el caso de las enzimas con múltiples subunidades, como la piruvato deshidrogenasa, incluir las proteínas individuales genera conexiones idénticas para todas las subunidades, independientemente de su rol dentro de la enzima completa, inflando artificialmente el número de nodos y conexiones. Tomar las reacciones como entidades primarias resuelve este conflicto, pues tienden a ser únicas.

Se obvia la dirección de la red, basando el acoplamiento en el número de conexiones. Una de las características más importantes de los métodos detallados de acoplamiento de flujos es resolver el acoplamiento total de las rutas incluyendo nodos en los que convergen varias rutas, con una salida o sin salidas [14]. Esto, por desgracia, requiere la dirección, la cual no se tiene con el set de datos disponible; esto obliga a tomar el número de conexiones como principal parámetro en la forma discutida en los antecedentes, pues no existe uno mejor.

Se buscan así mismo rutas lineales conectadas por su inicio y su final a otras para observar su influencia en el acoplamiento e integrarlas a las rutas lineales a las cuales se conectan. Estas rutas paralelas se denominan “ojos”, por las formas de abertura que introducen en las rutas lineales. La definición es cercana a la definición de los ciclos descrita en otras referencias [47], pero distinguir su presencia en una ruta lineal y la carencia de direccionalidad obligan a usar un concepto análogo. Un ejemplo de cómo pueden visualizarse estas construcciones se ve en la figura 9. Para ahondar en el análisis, se recurre a una segunda corrida del algoritmo de acoplamiento para descubrir más rutas lineales reduciendo las rutas lineales obtenidas a interacciones entre dos nodos. La segunda corrida no detecta más rutas lineales, teniendo la mayoría de los elementos entre una y tres conexiones.

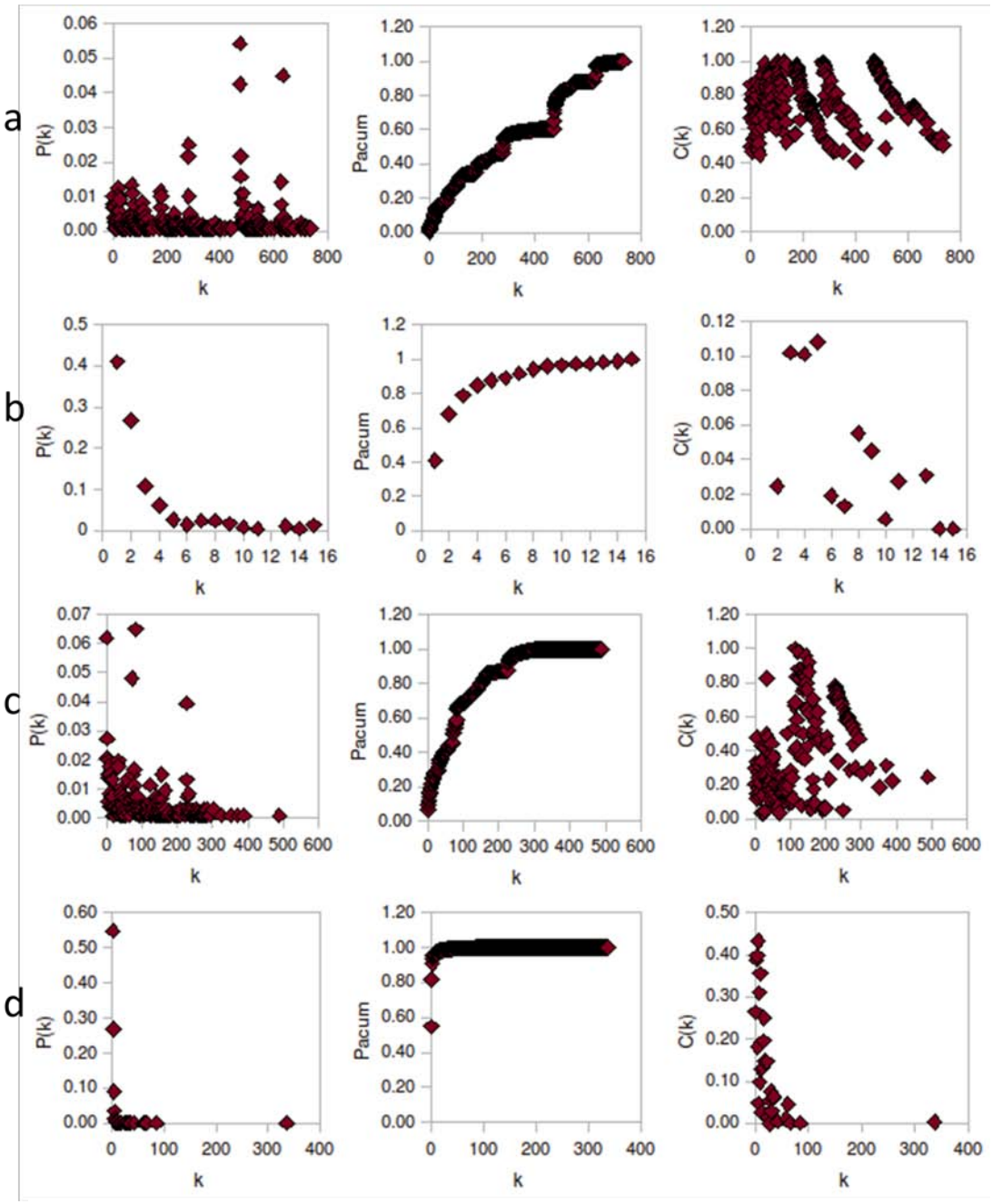


Figura 8. Distribuciones de obtenidas con el método de Barabási para las redes metabólicas usadas en este trabajo, basando la red en reacciones. **a -** archivo de la curación, sin procesos adicionales. **b -** archivo del XML de KEGG. **c -** archivo de la curación, con el filtro de metabolitos. **d -** Como punto de comparación, se ilustran las distribuciones de la red de DBTBS (la red regulatoria de *B. subtilis*, basada en genes), que se sabe es una red jerárquico-modular.

La red para el archivo del XML tiene 451 nodos y 666 conexiones. El acoplamiento encuentra 52 rutas, 8 de las cuales forman ojos, con 4 ojos que introducen caminos alternos en los ojos ya existentes. Se trata así mismo de una red muy poco densa, con una conectividad promedio de 2.81 y una máxima de 15. Las rutas lineales cubren una gran variedad de funciones. La vasta mayoría de rutas son de tres nodos, siendo la más larga la ruta de síntesis de nucleótidos que lleva a ATP, de 12 elementos. Los sectores que abarcan más rutas son ácidos grasos, con 9 rutas; y ciclo de pentosas, con 5 rutas. Estas rutas, en su mayoría, son fragmentos de subsistemas que normalmente serían considerados como sectores formales del metabolismo, lo cual refleja que el acoplamiento de flujos es menos efectivo en una representación basada en enzimas que en una basada en metabolitos o en un grafo bipartita.

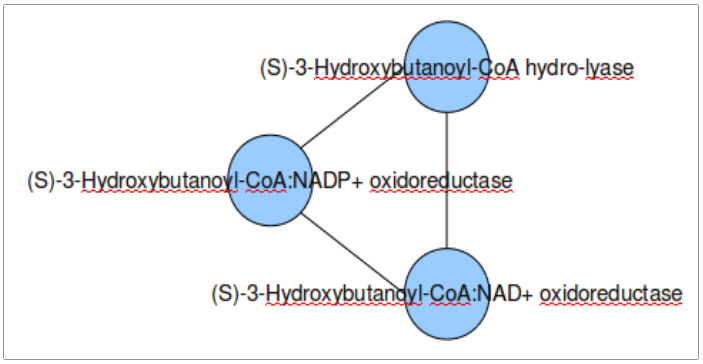


Figura 9. Ejemplo de un grupo de compuestos que forman una estructura de ojo. Cada elemento representa una enzima, conectada con otras enzimas por medio de recibir un metabolito de las demás. El ojo, como el ciclo, se caracteriza por su estructura cerrada, pero se le da ese nombre para resaltar el hecho de que se puede reducir como una fracción de una ruta lineal en un grafo.

Con 1199 nodos y 178224 conexiones, además de una conectividad promedio de 297.29 y una máxima de 736, la red para la curación aplicada es mucho más grande y mucho más densa. No obstante, sólo tiene 26 rutas lineales, 6 de las cuales forman ojos con 2 fungiendo como caminos alternos en otros ojos. Esto se debe al masivo incremento de la densidad de la red, el

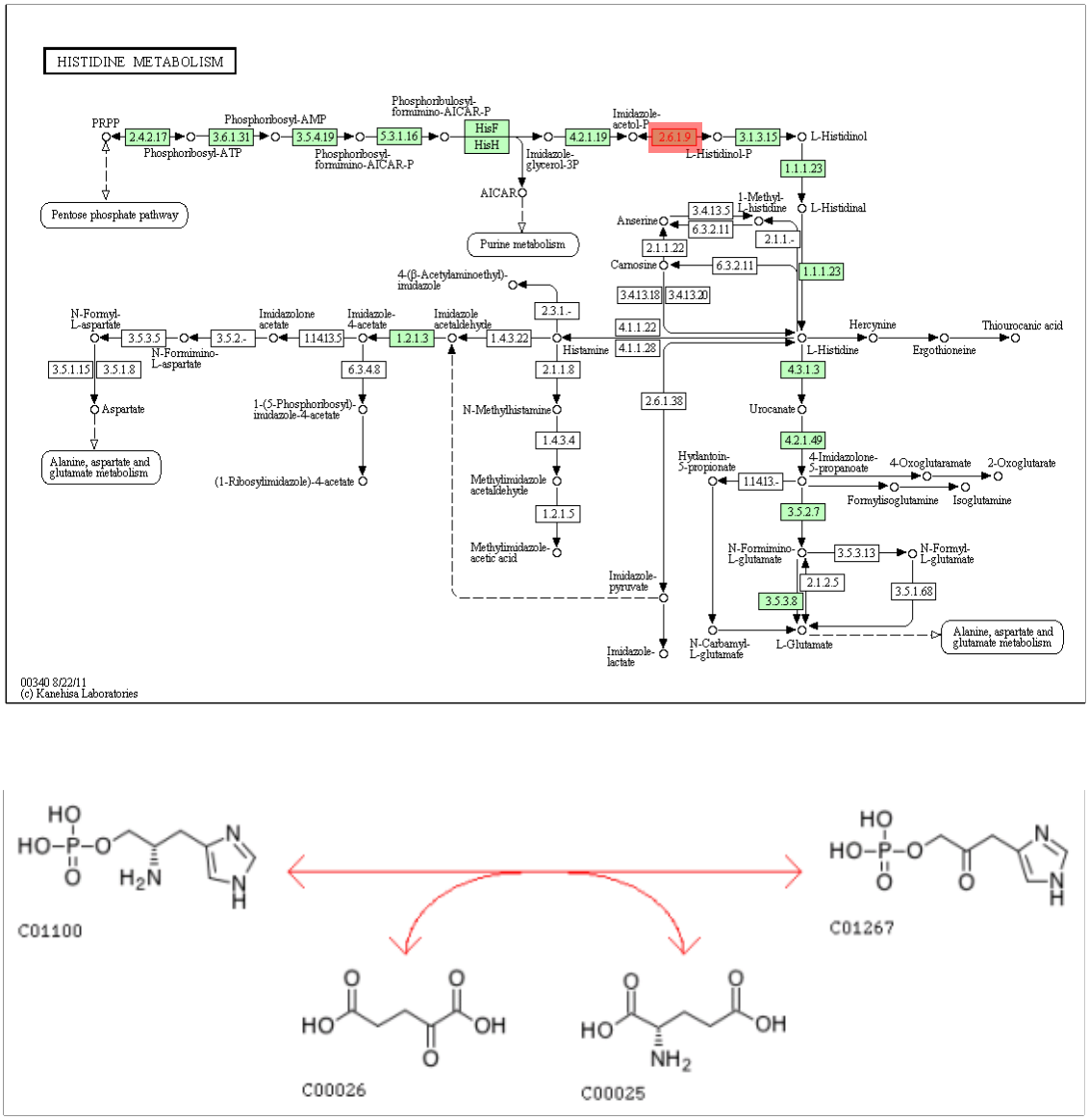


Figura 10. Ilustraciones de la histidinol fosfato aminotransferasa, otra reacción con eventos de pérdida de acoplamiento que no se ilustran inmediatamente en los mapas de KEGG. Arriba:. En el mapa, la reacción indicada está marcada en

cual incrementa también la posibilidad de cortes intermedios entre los caminos que quedarían marcados como rutas lineales en el archivo plano. Esto refleja cómo la robustez misma de la red dificulta su observación como un arreglo de rutas claramente separables. Las rutas lineales encontradas, como es de esperar según los datos, traslapan a veces con las rutas de la red anterior, pero quedan como fragmentos entrecortados de las rutas más largas de ésta, y todas son de tres elementos. Algunas, no obstante, agregan información nueva que de otro modo hubiera quedado sin examinar, lo cual da más valor a este set de datos. Ejemplos de esto son el ya mencionado de la antranilato sintasa, la histidinol fosfato aminotransferasa (que depende de un paso de 2-oxoglutarato a glutamato) y la 6-fosfo-5-deshidro-2-desoxi-D-gluconato aldolasa, en la cual se obvia la generación de glicerona fosfato, que se sabe participa en algunos procesos de metabolismo de lípidos en *E. coli* [26] y podría conectar varias rutas que existen en *Bacillus subtilis* a través de la función de la triosa fosfato isomerasa, una enzima de la ruta de glicólisis/gluconeogénesis. La figura 10 ilustra estas diferencias; cabe notar que los archivos XML están expresados en la forma de los mapas, similar al mapa de arriba en la figura 10.

Una característica importante de las rutas encontradas es que varias se componen de tres enzimas que interactúan con un solo compuesto. Eventualmente esto se detectaría como un ojo una vez las rutas sean definidas. Esto sucede porque el compuesto crea conexiones mutuas entre las tres reacciones, de las cuales el método de acoplamiento crea una ruta. Esto ilustra el hecho de que la representación de red que se elija para un sistema particular determina sus propiedades topológicas, hecho que debe tomarse en cuenta para determinar que representación es preferible.

Las rutas lineales a veces comparten nodos extremos, los cuales conectan las rutas directamente a pesar de que no existe un acoplamiento. Uniendo los extremos de las rutas y los ojos podemos construir grupos más largos y significativos, que llamaremos módulos metabólicos. La cantidad de módulos metabólicos obtenidos para cada red se ilustra en la tabla 2.

	Nodos totales	Aristas totales	Conectividad promedio	Módulos totales (rutas totales)
Red XML	451	666	2.81	52(67)
Red de curación	1199	178224	297.29	5(5)
Red de curación, con filtro	1063	64702	85.54	15(26)

Tabla 2. Comparación de datos numéricos para las redes metabólicas en este trabajo. La última columna compara el total de módulos metabólicos obtenidos con el total de rutas obtenidas sin unir los extremos de las rutas y los ojos.

Análisis topológico y acoplamiento de flujos sobre la red filtrada

Una vez aplicado el filtro, la red de la curación cambia significativamente. Tiene un total de 1063 nodos, pero ahora con 64702 conexiones, una conectividad promedio de 85.5 y una conectividad máxima de 488. Es una disminución notable en la densidad de la red, pero es aún más densa que la red del XML, lo cual muestra que el impacto en las reacciones de las conexiones dadas por los metabolitos cuya influencia en la red se recorta es muy grande y vale la pena tomarlo en cuenta en este análisis. La ruta más larga, síntesis de GTP, consta de 7 reacciones; las demás son de 4 o 3 elementos.

Un análisis de las conexiones sometidas al filtro revela que el parámetro de cociente entre pesos de compuesto sigue una distribución bimodal, con alta abundancia de conexiones entre compuestos muy similares en peso y entre compuestos muy distintos en peso (figura 11). La debilidad de este filtro crudo se hace notable; las conexiones de peso distinto pueden incluir conexiones válidas entre compuestos muy distintos en tamaño, como entre piruvato y acetil coenzima A. No obstante, que estas conexiones tienden a llevar a metabolitos que son por sí mismos altamente conectados, las cuales, serán desechadas posteriormente por el método de acoplamiento, como continuadores válidos de rutas lineales. Esta observación implica que las

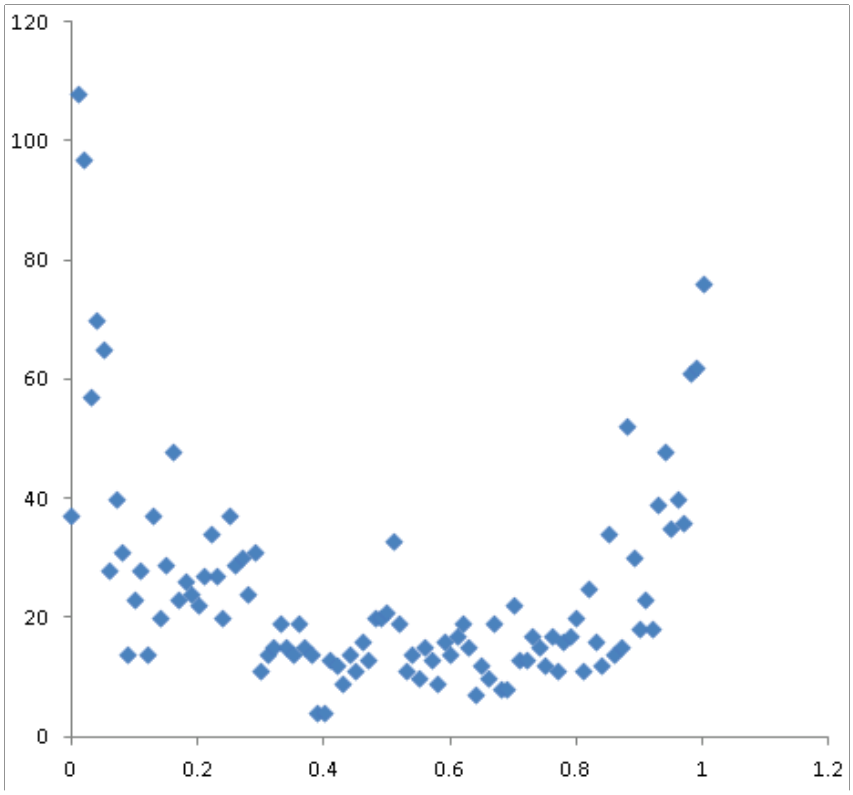


Figura 11. La distribución de los cocientes de las conexiones en la red, con base en la cual se aplica el filtro de metabolitos. El filtro elimina todo valor < 0.6.

debilidades del filtro tienen poco impacto en el resultado neto del método, lo cual hace al filtro relativamente confiable como una medida preliminar para la validez de las conexiones.

Sobre esta red se recalculan las propiedades estadísticas con el fin de observar si encontramos cambios en la distribución que nos indiquen, la posibilidad de obtener por el método clásico de Barabási un organización jerárquico modular. Los resultados se muestran en la tabla 1- figura 8c, que muestran, que a pesar del filtro la red sigue mostrando un comportamiento no jerárquico y que no es posible obtener relaciones modulares por este método.

Organización modular desde la perspectiva del metabolismo

Usando el acoplamiento de flujos como método de análisis gráfico de la red, se procede a fabricar dos redes: una para los datos extraídos por la curación exhaustiva de KEGG y otra basada en el archivo XML definido para la red. El grafo de la curación es procesado como sigue: se genera el grafo unívoco de compuestos, al cual se le aplica el filtro, que se usa sobre el grafo bipartita; el grafo recortado se usa para extraer el grafo unívoco de reacciones, que después de identificar a las reacciones acopladas, es asociado a los operones según correspondan como se señala en los métodos. En cuanto al grafo del XML, se procede directamente a la construcción de las reacciones, pues este archivo ya está estructurado para incluir sólo ciertos metabolitos, por lo cual aplicar el filtro es innecesario.

Integración de la red de regulación a los módulos metabólicos

A partir de los módulos generados para cada red, incluimos elementos regulatorios en el modelo del metabolismo. Se piensa iniciar con criterios similares a los usados para mapear la red metabólica, incluyendo principios de reducción propios a las redes regulatorias, como la organización de operones y de regulones más complejos. Uno de los posibles criterios que se puede usar para lograr esto es una red basada en operones. Un operón es controlado en general por los mismos reguladores para todos los genes, lo cual permite centrar la perspectiva que se tiene en el control que tiene el flujo de la regulación en el del metabolismo y viceversa.

Es menos preciso, pero lo que se pierda en exactitud se puede compensar en manejabilidad y simplicidad.

Usando este enfoque, se procede a fabricar dos redes derivadas de los elementos resultantes del acoplamiento, una para los datos extraídos por la curación exhaustiva de KEGG, incluyendo el efecto del filtro de metabolitos, y otra basada en el archivo XML definido para la red. La generación del grafo que muestra los resultados de las redes centradas en el análisis sobre las rutas lineales, está ordenado en 4 capas (figura 12-esta corresponde a la imagen de la ruta de histidina). La capa interior, está compuesta de los TFs y factores sigma que regulan de manera directa a las rutas acopladas, la siguiente capa se compone de los operones que codifican para las enzimas que forma la ruta lineal, la capa siguiente lleva a las rutas que compone al operon según la notación del archivo de reacciones de KEGG y la capa más externa contiene los nombres de los compuestos que formaron la ruta lineal.

Una observación interesante pero no sorpresiva dado los resultados del análisis topológico de la red de regulación transcripcional, es que *sigA*, conocido como el factor sigma maestro (house keeping) de *B. subtilis,* presenta una k=782 (se conecta con el 46.5% de los genes de la red), seguido sólo por SigE con una k=169 (10% de los genes), lo cual se hace evidente al analizar la redes estudiadas. En cada caso, *sigA* conecta a 28 grupos de los 33 encontrados en la red XML y 9 de 15 en la red curada y filtrada, además de conectarse al grupo más grande de la red de la curación sin el filtro. En el caso de las rutas desconectadas, del grupo regulado por *sigA*, es posible que aun no estén caracterizados los sitios para *sigA* o, en el caso del paso glutamato-ornitina para la red del XML, el grupo de glucarato para la red curada y filtrada y el grupo de estreptomicina para la red curada sin filtro, existen sitios para sigmas alternativos como *sigL*, *sigB*/*sigD* y *sigK*, respectivamente. Esta observación es acorde al hecho de que *sigA* se caracteriza por una estructura con un dominio de asociación al DNA que le da mayor flexibilidad en el reconocimiento de promotores, a diferencia del resto de los factores sigma en *B. subtilis*.

Con el fin de definir con mayor nitidez a la forma en que los módulos metabólicos están regulados procedimos a quitar los siguientes reguladores más ampliamente conectados de todas las redes *ccpA*, *codY* y *sigE*.

Figura 12. Pérdida de acoplamiento. la figura es de la red del xml. nodos: negro, regulador; borde verde, operón; borde morado, reacción; borde amarillo, metabolito. aristas: amarillo, metabolismo; verde, transcripción; azul claro, regulación +; azul oscuro, regulación -. Los nodos que no tienen texto de color del borde no aparecen en el archivo de la curación. (Los detalles de las reacciones están en el apéndice B.)

Descripción de la remoción de factores de transcripción más altamente conectados en la red XML

En la red del XML, la remoción de estos FTs ayudó a dar una visión intuitiva del agrupamiento por rutas (figura 13). Al final se obtiene una visión clara de las rutas como grupos de operones

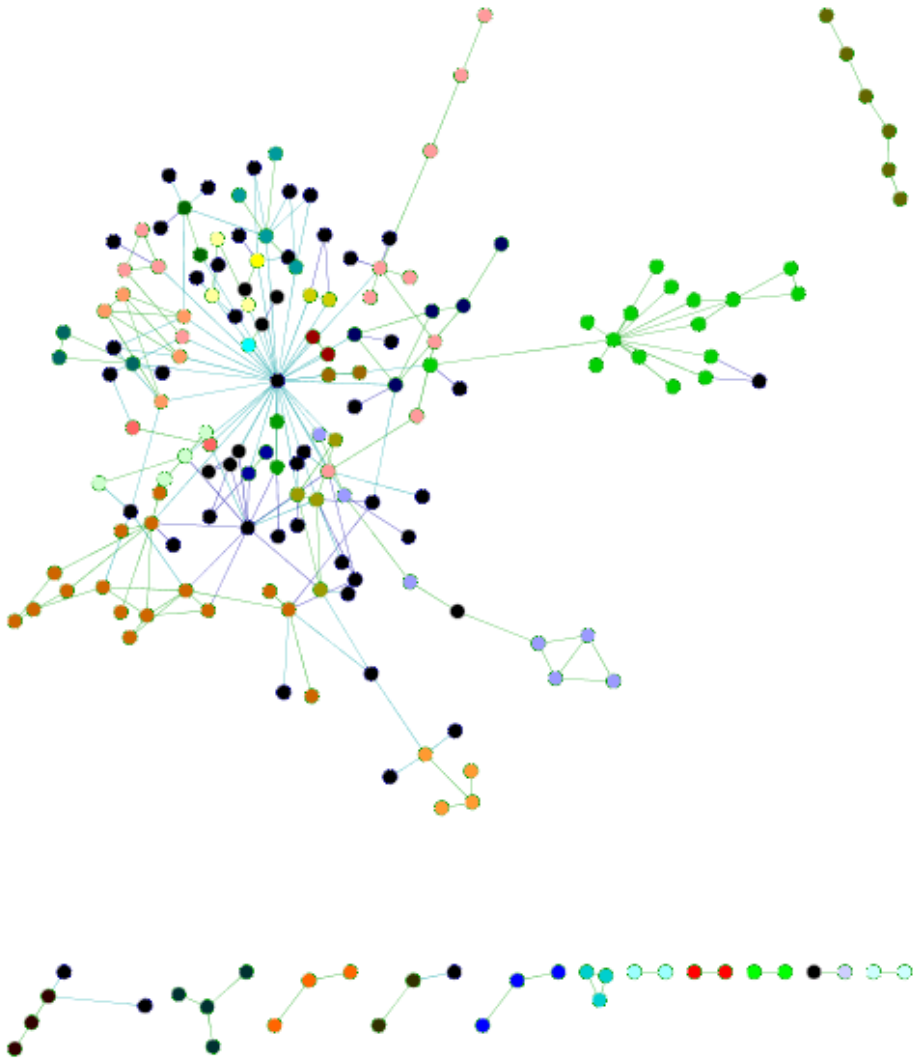


Figura 13. Imagen de la red del XML sin modificación. La red está densamente conectada, con pocas rutas que carecen de regulación transcripcional.

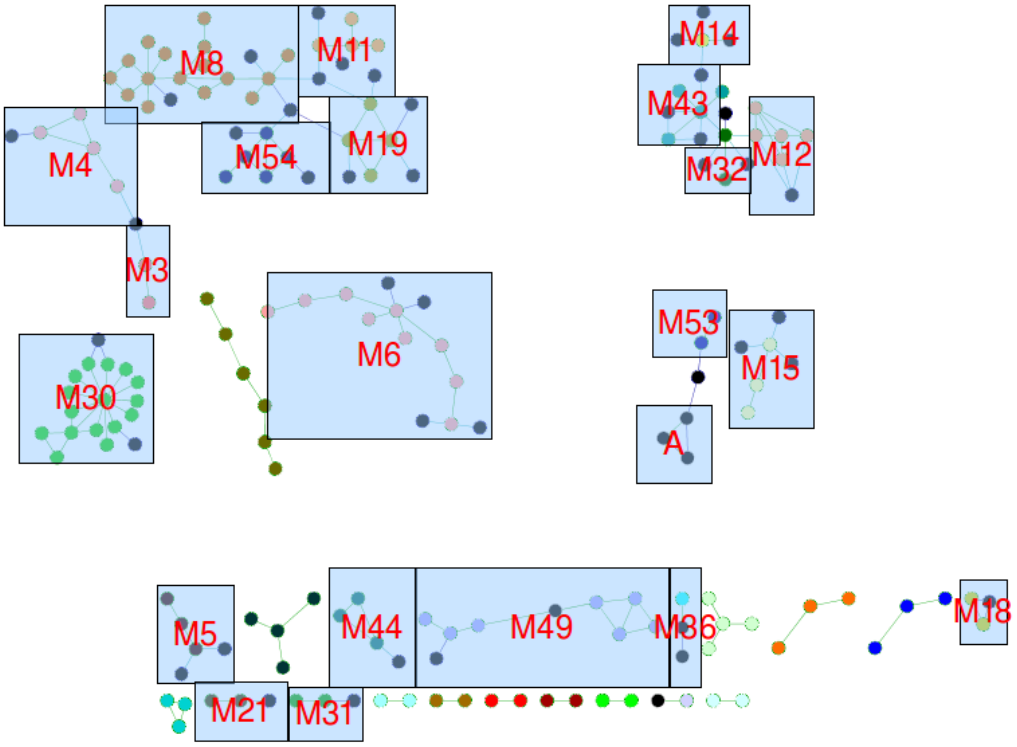


Figura 14. Imagen de la red del XML cuando se ha suprimido *sigA, ccpA, codY* y *sigE* de la red. En algunos sectores, siguen existiendo reguladores intermedios, pero los grupos se separan un poco más y ya son diferenciables de manera intuitiva. Es en este punto cuando se hace evidente la constitución jerárquica de la red de regulación y su efecto sobre los resultados del proceso. Marcados en cajas están los módulos con regulación transcripcional conocida. Los detalles de los módulos están en el apéndice A.

(figura 14). Cabe notar que todos los genes eliminados son considerados reguladores globales [44-45], lo cual confirma que la regulación de los elementos del metabolismo que rompen el acoplamiento es compleja y puede involucrar reguladores de amplio espectro. Así mismo, se aprecia el hecho de que el elemento de jerarquización de funciones está dado más por la regulación que por el metabolismo, como lo muestra la necesidad de eliminar capas de reguladores para apreciar gráficamente los grupos.

Aplicando la metodología de integración, la red del XML rinde un total de 33 grupos, de los cuales 20 tienen regulación transcripcional conocida (Apéndice A). Estos se organizan como sigue:

-Conectados por *sigL* y *tnrA*: biosíntesis de ácidos grasos (M8, regulado por *iolR*, *bkdR* y *tnrA*), paso glutamato – ornitina (M11, regulado por *rocR*, *ahrC* y *sigL*), síntesis de porfirinas (M54, regulado por *glnR*, *perR* y *tnrA*) y uso de acetoína (M19, regulado por *acoR*, *alsR*, *rex*, *tnrS*, *tnrA* y *sigL*). Todos estos grupos son consistentes respecto de la información esperada para *B. subtilis*.

-Los operones de *tenA* y *sacB* participan como integrantes de ruta en los fragmentos de metabolismo de pirimidinas (M32, regulado por *tenI*, *aprE*, *sacB* y *nprE*) y absorción de sucrosa (M43, regulado por *sacY*, *sacT*, *degU* y *tenA*), pero al mismo tiempo fungen como conectores regulatorios en dos casos: metabolismo de folato (M12, regulado por *tenA* y *phoP*) y precursores de terpenoides (M14, regulado por *hpr*, *ipi* y *degU*), con *sacB* interactuando con *degU*. Estos grupos son consistentes respecto de la información esperadas, salvo el modulo 14, cuyas reacciones están ligadas por una isomerasa que mueve grupos fosfato.

- Conectados por *sigB*: síntesis de NAD (M4, regulado por *yrxA* y *sigB*) y manejo de succinato (M3, regulado por *sigB*). Estos dos grupos también son consistentes respecto del metabolismo.

- Conectados por *ccpB*: uso de gluconato (M53, regulado por *ccpB*) y uso de xilosa (A, regulado por *ccpB* y *xylR*), partes del proceso de conversión pentosa – gluconato. Además de

ser consistentes dentro de su propia ruta metabólica, estos grupos están cercanamente relacionados funcionalmente, lo cual se confirma con la corregulación.

- Por separado: histidina – aromáticos, de 3-deshidroxishikimato a corismato (M6, regulado por *mtrB*, *hutP* y *abrB*); síntesis de porfirinas, de uroporfirinógeno III a protohemo (regulado por *fnr*, *resD* y *arfM*); síntesis de lisina (regulado por *sigG* y *spoVT*); procesamiento de fosfatidil serina (regulado por *sigX*); rescate de metionina (regulado por *yrzC*); paso reductor de síntesis de ácidos grasos (*fabHA*, *fapR*; regulado por *ylpC*); uso de dipicolinato (regulado por *sigK*); pentosas – glucuronato (regulado por *exuR*). Los grupos con más reacciones son el de histidina y aromaticos, porfirinas y rescate de metionina. Los demás son grupos de máximo tres operones. Dentro de la ruta metabolica, todos son consistentes.

Descripción de la remoción de factores de transcripción más altamente conectados en la red curada

La red de la curación rinde 15 grupos, de los cuales 11 tienen regulación transcripcional conocida. Estos se organizan como sigue:

- Conectados por *sigB*: uso de glucarato (regulado por *sigB* y *sigD*) y síntesis de desoxipirimidinas (regulado por *sigB*).

- Conectados por *ccpA* y *codY*: síntesis de isoleucina-valina (regulado por *rex*, *alsR*, *tnrA*, *tnrS*, *ccpA* y *codY*) y síntesis de histidina (regulado por *hutP*, *abrB*, *ccpA* y *codY*).

- Por separado: biosíntesis de pterinas (regulado por *sigE*, *tenA* y *phoP*), síntesis de xantina (regulado por *pucH*, *ureA*, *ywoE*, *yurH*, *pucJ* y *guaD*), síntesis de porfirinas (regulado por *arfM*, *fnr* y *resD*), metabolismo de espermidina (regulado por *yqzB*), síntesis de aromáticos (regulado por *mtrB*), síntesis de lisina (regulado por *sigK*).

Todos estos grupos son pequeños, de en su mayor parte tres operones, mostrados en detalle en la tabla 3. La representación, como ya se ha hecho notar, modifica la estructura de los grupos que se encuentran. En la representación por operones, las rutas lineales que se

encuentran bajo la representación de reacciones se manifiestan como grupos compactos más similares a los módulos funcionales que se manejan comúnmente, con varias conexiones entre nodos. Como representación gráfica, la representación por operones permite observar los grupos que se forman por las rutas, mientras que el grafo de varios componentes permite observar el sistema dando prioridad a los tipos de elementos que en verdad se conectan, permitiendo apreciar mejor las transiciones.

Es importante mencionar que, en los grupos que no poseen reguladores transcripcionales, aparecen posibles inconsistencias. Un ejemplo es la aparición de los genes *cypB* y *yetO*, conectados metabólicamente e involucrados en el metabolismo de hormonas esteroides. La ruta metabólica parece ilustrar sólo un paso de la degradación, inconexo, que ambos genes llevan a cabo, además de estar colocados en el metabolismo de otros compuestos. Esto puede deberse a que la presencia de homólogos de estos transcritos en *B. subtilis* sea ilustrado en la base de datos con poca distinción de otros tipos de reacción, sin aparecer en los archivos de XML. Esto indica que la curación de la base de datos también esta estructurada de antemano para lidiar con este tipo de dificultades.

Otro fenómeno notable es que los nodos que representan reguladores globales marcados en otras referencias [44] tienden a conservarse en ambas redes, lo cual implica no sólo que el método es confiable, sino que los reguladores globales tienden a interactuar con las rutas lineales observadas por el método, con lo cual se puede inferir que son puntos cruciales para la regulación, ayudando a la célula a activar o reprimir a estas rutas desde el punto de vista transcripcional en condiciones variadas. En la mayoría de los casos los reguladores globales tienden a concentrarse en los puntos extremos de las rutas lineales, con reguladores locales o ningún regulador en medio, así como existe un traslape neto de 5 rutas entre ambas redes. Estas rutas conservadas son: isoleucina-valina, aminoácidos aromáticos, histidina, folato-nucleótidos y porfirinas. A continuación se discuten los detalles sobre estas rutas.

Las rutas de aminoácidos aromáticos e histidina existen como un solo grupo según los datos del archivo del XML pese a que se constituye de rutas separadas, debido a que ambas están ligadas por el gen *hisC* (la histidinol fosfato aminotransferasa), en el operón *trpEDCFBA-hisC-tyrA-aroE*,

que codifica mayoritariamente las partes finales de la síntesis de triptofano. En el archivo de la curación estas rutas están claramente separadas por eventos de pérdida de acoplamiento (infra). La ruta de histidina es mayoritariamente parte de la ruta degradativa de la histidina, codificada por el operón autoactivable *hut*, que es reprimido por *ccpA* (el regulador central del metabolismo de carbono) y *codY* (que sensa la abundancia de nutrientes mediante su capacidad de asociarse con GTP [27, 28], y activado por *abrB* (un factor relacionado a la transición a esporulación desde el crecimiento vegetativo y a la síntesis de antibióticos, que trabaja en conjunto con *spo0A*, un regulador crucial de la esporulación [29]. El operón *trpEDCFBA-hisC-tyrA-aroE* es a su vez reprimido por *mtrB* (en su forma final como la proteína TRAP [30]).

La ruta de uso de isoleucina-valina es rica en enzimas con funciones altamente flexibles, que conectan la síntesis de ambos aminoácidos y la de leucina, lo cual dificulta el que el método de acoplamiento encuentre rutas lineales. Además, existen varios casos de rotura de acoplamiento por metabolitos poco esperados (*supra*), como es el caso de la aminotransferasa que completa la síntesis de los tres aminoácidos en cuestión (codificada por *bcd*), que depende de una transición entre 2-oxoglutarato y glutamato. Por ende, tanto la fuente de datos del XML como la fuente de la curación devuelven poca información con el método de acoplamiento. La ruta que se representa como lineal es el principio de la síntesis, efectuada por la acetolactato sintasa; los elementos lineales de reacción son los pasos de síntesis que se dan para producir isoleucina y valina, que tienen un precursor único cuya síntesis depende de tiamina difosfato. La enzima está codificada dentro de los operones *alsSD* (por *alsS*) e *ilvBHC-leuABCD* (por *ilvB* e *ilvH*). Ambos operones son activados por *ccpA*. El operón *alsSD* es activado por *alsR* (el regulador local) [31] y el operón de la lactato deshidrogenasa [32] y desactivado por *rex* (antes *ydiH*) que trabaja con *ndh* para mantener la estabilidad de la cadena respiratoria modulando la proporción NADH/NAD+ [33]. El operón *ilvBHC-leuABCD*, por su parte, es activado por el tRNA de leucina en su estado no cargado [50] y reprimido por *tnrA* (un factor altamente involucrado en la degradación de compuestos nitrogenados que compite con *glnR* por sitio [34] y *codY*. El archivo del XML, por medio del grafo bipartita, asocia esta ruta con una fracción de la función de la piruvato deshidrogenasa (codificada por *pdhA*, *pdhB*, *acoA* y *acoB*) mediante el uso de

tiamina difosfato, la cual es considerada como un factor secundario o un cofactor [35], lo cual indica que el filtro de metabolitos pudo no haberlo quitado como significativo. No obstante, el archivo de la curación toma en cuenta que la formación del precursor de tiamina difosfato desprende piruvato como producto secundario, con lo cual esta conexión es eliminada por el procesamiento de la red.

La ruta de folato comprende el inicio de la síntesis, desde GTP hasta 7,8-dihidropteroato, dos pasos antes de la formación completa del folato. El acoplamiento se rompe en el dihidropteroato por la necesidad de incluir glutamato para sintetizar de él dihidrofolato, el precursor inmediato del folato. La ruta de folato no mantiene conexiones tanto por la linealidad de la ruta metabólica como por la unidad de los operones. La liberación de formato en la segunda reacción de la ruta la cortaría antes de la conexión con GTP, pero se conserva porque las primeras cuatro reacciones son llevadas a cabo en *Bacillus subtilis* por el producto de *folE*, que no tiene regulación transcripcional conocida. La síntesis continúa con el producto de *phoA*, *phoD* y *phoB*, activado por *phoR* (el regulador local [36]) y *tenA* (un regulador de enzimas degradativas que requiere la función de *degS* [37]), además de *sigE* (un factor sigma específico para la esporulación [36]). Completan el grupo *folB*, *folK* y *sul*, del operón *pabBAC-sul-folBK-yazB-yacF-lysS*, que es reprimido por TRAP [38]. Se nota también un artefacto del filtro de metabolitos: el grupo incluye, en el archivo de la curación, a *ndk* como parte del grupo y productor de GTP, pese al hecho de que es una fosfotransferasa.

El grupo de porfirinas es un fragmento breve de la ruta, desde uroporfirinógeno III al protohemo. Aquí la conexión se debe nuevamente a la constitución de los operones. El acoplamiento se rompe en la coproporfirinógeno III oxidasa (codificada por *hemN* y *hemZ*), que requiere adenosilmetionina y expulsa metionina, pero las dos reacciones posteriores son codificadas por el operón *hemEHY*, que también codifica la precedente. La regulación transcripcional se halla en *hemZ*, con tres activadores: *arfM* (un modulador de respiración anaerobia [39, 40]), *fnr* (el cual sensa oxígeno mediante un complejo [4Fe-4S] de 3 cisteínas y opera genes de asimilación de nitrato [39, 41]) y *resD* (parte del sistema de dos componentes ResD-ResE, necesario para mantener la actividad de FNR en condiciones anaerobias, con un

nexo recientemente descubierto con NsrR, que opera el metabolismo de óxido nítrico [42, 39, 43]).

Se aprecia, en general, la existencia de conexiones alternas entre varios de los grupos detectados. Por ejemplo, las síntesis de folato y aminoácidos aromáticos están correguladas por TRAP [51]; se aprecian los nexos directos con la esporulación a través de los grupos de histidina (con *abrB* y *spo0A*) y folato (con *sigE*); se pueden hacer inferencias sobre la influencia del oxígeno sobre el funcionamiento del grupo de isoleucina-valina (dependiente de cadena respiratoria) y el de porfirinas (activado por condiciones anaerobias) [52].

Se aprecia también la aparición de un módulo que no cabe entre los anteriores, designado como módulo A. Varios de los componentes parecen ser artefactos, pero los componentes significativos, ilustrados en la tabla 3, llevan a cabo la reacción de la arginasa y la glutaminasa, lo cual puede llevar a designarlo como un módulo de metabolismo de nitrógeno. Es posible que la inclusión de estos elementos se deba a correlaciones efectuadas por el método de acoplamiento que no se consideraron anteriormente o por el mismo método de integración.

Uno de los efectos del incremento de conexiones para la red de la curación es cómo se estima, al usar la curación, la separación o pérdida de rutas localizables por el método de acoplamiento. Este efecto es apreciable en la ruta de histidina (figura 12), la cual aparece completa en la red del XML, pero en la red de la curación aparece como dos grupos separados. Esto se debe a la información que la curación agrega; el histidinol y la histidinol fosforilasa rompen el acoplamiento en la red de la curación. Esto muestra el efecto de los metabolitos sobre la estructura de la red; el elemento central es la conservación del uso de los metabolitos y la robustez de la estructura enzimática necesaria para mantener esta conservación.

Módulo	Gen	Grupo	Operón	Primer gen	Regulación
1	*aldX*	uso de glucarato	*aldX*	*aldX*	
1	*aldY*	uso de glucarato	*aldY*	*aldY*	*sigB*
1	*dhaS*	uso de glucarato	&		
1	*garD*	uso de glucarato	&		
1	*gudD*	uso de glucarato	&		
1	*ycbC*	uso de glucarato	*ycbC*	*ycbC*	
1	*yfmT*	uso de glucarato	*yfmTS*	*yfmT*	*sigD*
1	*ywdH*	uso de glucarato	*ywdH*	*ywdH*	
2	*aroF*	síntesis de aromáticos	&		
2	*aroK*	síntesis de aromáticos	*aroK*	*aroK*	
2	*trpE*	síntesis de aromáticos	*trpEDCFBA-hisC-tyrA-aroE*	*trpE*	*-mtrB, sigA*
3	*alsS*	síntesis de isoleucina-valina	*alsSD*	*alsS*	*alsR, -ydiH, ccpA*
3	*ilvB*	síntesis de isoleucina-valina	*ilvBHC-leuABCD*	*ilvB*	*trnS, -tnrA, ccpA, -codY, sigA*
4	*pucA*	uso de xantina (hacia urea)	*pucABCDE*	*pucA*	*pucR-, sigA*
4	*pucR*	uso de xantina (hacia urea)	*pucRJKLM*	*pucR*	*pucH, ureA, ywoE, yurH, pucJ, guaD*
5	*hom*	síntesis de histidina (desde aspartato, homoserina)	&		
5	*spoVFA*	síntesis de histidina (desde aspartato, homoserina)	*spoVFB-asd-dapG-dapA*	*spoVFA*	*sigK*
6	*hisD*	síntesis de histidina (pasos finales)	&		
6	*hutP*	síntesis de histidina (pasos finales)	*hut*	*hutP*	*abrB, hutP, -ccpA, -codY, sigA*
7	*gapB*	rescate de metionina (con espermidina)	*gapB-speD*	*gapB*	*-yqzB, sigA*
7	*speE*	rescate de metionina (con espermidina)	*speE-speB*	*speE*	*sigA*
8	*cypB*	metabolismo de esteroides (testosterona)	&		
8	*yetO*	metabolismo de esteroides (testosterona)	&		
9	*ganA*	metabolismo de glucosilceramida	&		
9	*ybbD*	metabolismo de glucosilceramida	&		
9	*yesZ*	metabolismo de glucosilceramida	&		
10	*folE*	síntesis de folato (inicio desde GTP)	&		
10	*ndk*	síntesis de folato (inicio desde GTP)	&		
10	*pabB*	síntesis de folato (inicio desde GTP)	*pabBAC-sul-folBK-yazB-yacF-lysS*	*pabA*	*sigA*
10	*phoA*	síntesis de folato (inicio desde GTP)	*phoA*	*phoA*	*phoP, tenA, sigA*
10	*phoB*	síntesis de folato (inicio desde GTP)	*phoB-ydhF*	*phoB*	*sigE, phoP, sigA*
10	*phoD*	síntesis de folato (inicio desde GTP)	*phoD*	*phoD*	*phoP, sigA*
11	*cdd*	síntesis de esteroides (estrona)	*cdd-era*	*cdd*	*sigA, sigB*
11	*deoD*	síntesis de esteroides (estrona)	&		
11	*pupG*	síntesis de esteroides (estrona)	&		
12	*hemE*	síntesis de porfirinas	*hemEHY*	*hemE*	
12	*hemZ*	síntesis de porfirinas	*hemZ*	*hemZ*	*arfM, fnr, resD*
12	*lepA*	síntesis de porfirinas	*lepA-hemN-hrcA-grpE-dnaKJ-yqeTUV*	*lepA*	*sigA*
A	*ansA*	asparaginasa/glutaminasa	*ansAB*	*ansA*	*ansR-, sigA*
A	*ybgJ*	asparaginasa/glutaminasa	*ybgJ-ybgH*	*ybgJ*	*ycbA, sigA*
A	*ansZ*	asparaginasa/glutaminasa	&		
A	*ylaM*	asparaginasa/glutaminasa	&		

Tabla 3: Resumen de los grupos encontrados usando el procedimiento de integración sobre el set de datos de la curación, centrada en los operones involucrados. La tabla asigna colores para los genes involucrados según los colores asignados a ellos en la figura 13. Los genes directamente involucrados están marcados por su nombre. La columna primer gen indica el gen del operón que interactúa con otros a nivel regulatorio, ya sea como regulador o como sitio. Se nombra la regulación transcripcional conocida para cada operón. Los genes con el símbolo (&) en la columna operón no están acumulados en operón con otros genes.

Conclusiones

Este trabajo buscó generar un criterio de integración para observar las redes de regulación y metabolismo de *Bacillus subtilis*, y compararlo con los criterios existentes para verificar su validez. Esto requirió lograr una reconstrucción eficaz de los posibles agrupamientos para los elementos de la red.

Se encontró que el criterio elegido – asumir las reacciones que se dan dentro del metabolismo como los operones que las codifican dentro de la red transcripcional - puede conciliar con eficacia los agrupamientos individuales para la regulación transcripcional y el metabolismo como un solo sistema, lo cual lo hace un enfoque prometedor para ampliar la perspectiva del análisis de redes.

Cabe resaltar que los dos métodos utilizados para crear los agrupamientos dentro de la red – el método de clustering de Barabási y el acoplamiento de flujos – son menos eficaces de lo que se esperaba para este rol. El método de Barabási es inadecuado para el análisis de la red de metabolismo basada en reacciones, pues ésta no tiene la estructura que el método requiere y asume. El acoplamiento de flujos es mucho más efectivo, sobre todo para una red basada en metabolitos, pues la estructura de la red es natural para esta aplicación. En el caso de una red de reacciones sólo es efectivo cuando la red está construida de antemano para incluir sólo los compuestos más esenciales para preservar la sucesión directa de la cadena de reacciones según el flujo de carbono, ante la perspectiva de perder la información que contribuyen las salidas secundarias de las reacciones que también son compuestos carbonados. Cuando se incluyen definiciones más refinadas de las reacciones involucradas, el método pierde eficacia debido a que la red misma pierde nitidez. Una forma de superar esta limitante necesita tomar en cuenta la robustez que la red metabólica ha desarrollado para mantener su propia estabilidad. Intentar esto fue, precisamente, el objeto de incluir el filtro de reacciones basado en metabolitos. Este filtro es bastante crudo, pero es aceptable como un enfoque preliminar. Trabajos futuros podrían profundizar en este aspecto del análisis, generando un filtro más efectivo y universal.

Una resultado muy importante es que de acuerdo con el tamaño, el tipo de distribuciones obtenidas y el arreglo de los grupos discretos obtenidos por este método, el metabolismo no

tiene una forma que sea claramente mapeable por un criterio general, sino que se centra en la formación de un estado de robustez que permita mantener los flujos basados en metabolitos. Se puede definir esta robustez como la capacidad de relacionar los componentes de la red de varias formas, que en el metabolismo es interpretable como el incremento de opciones para producir los metabolitos que se requieren, lo cual es crucial para todo organismo vivo, sobre todo para una bacteria de suelo como *Bacillus subtilis*. La función de jerarquización y establecimiento de prioridades está dada casi por completo por el bloque regulatorio. La jerarquía es visible en la red integrada mediante el proceso de remoción de los factores más conectados, empezando con sigA y removiendo otra capa como se hizo en la red del XML de KEGG, a fin de revelar los grupos metabólicos con reguladores locales.

Respecto del análisis funcional, se encuentra que el método de integración no sólo confirma con éxito los resultados de los análisis elaborados para determinar las funciones de los grupos formados por las rutas de acoplamiento, sino que los hace mucho más evidentes de lo que se lograría con análisis independientes de regulación y metabolismo. Al poder apreciarse directamente la influencia de los reguladores sobre las rutas metabólicas y los puntos en los cuales se ejerce ésta, se puede detectar con mayor precisión el efecto que tiene un regulador sobre sus blancos y la respuesta de éstos respecto de los grupos que los rodean. De forma similar, el método de integración hace evidente que los reguladores globales concentran su influencia en puntos de las rutas metabólicas que involucran roturas en interacciones y rutas más simples, además de metabolitos con un elevado número de posibles destinos.

Perspectivas

Con los resultados disponibles, el método desarrollado en este trabajo resulta prometedor para análisis topológicos de los elementos de la célula viva, pero puede rendir resultados aún más significativos con algunas adiciones. El trabajo próximo sobre esta metodología habrá de enfocarse en incrementar la precisión del método y ampliar su perspectiva de manera preliminar.

El primer paso para mejorar el método es dar mejoras a las metodologías involucradas. El filtro de conexiones metabólicas, en particular, es muy crudo y produce varios artefactos. Un cambio

significativo podría darse al complementar el pesado de átomos con otros criterios que definan con más precisión a las moléculas involucradas, siendo un enfoque basado en la estructura el más adecuado. Un enfoque estructural daría un parámetro muy poderoso para distinguir los fragmentos moleculares que se intercambian en las reacciones y justificaría los pasos metabólicos con gran precisión. El pesado atómico, en si mismo, vale la pena como una parte del filtro, pues permitiría centrar la observación de las conexiones con base en un tipo de átomo específico.

Otra forma de afinar el método sería estandarizar los métodos en medida de lo posible, siguiendo los desarrollos recientes. El acoplamiento de flujos, en particular, está siendo estandarizado de manera importante, por lo que vale la pena llevar a cabo una curación intensiva de los datos disponibles que permita esta aplicación, sobre grafos bipartitas que incluyen dirección de conexiones [1]. Así mismo, pueden buscarse, o incluso idearse, otros métodos de agrupamiento que puedan establecer grupos de elementos metabólicos de manera más natural para la red metabólica de reacciones, incluyendo todas las interacciones que implica el uso de los metabolitos por cada enzima, y también dentro de la red integrada en su totalidad.

Una vez se lleven a cabo estas mejoras, se puede ampliar la perspectiva del método para englobar a otros organismos y fenómenos, o se puede hacer esta ampliación a medida que se mejora el método a fin de asegurar que se pueda aplicar de forma general. Este trabajo surgió como una extensión de un análisis sobre *Bacillus subtilis* [6], lo cual es parte del por qué se eligió este organismo para trabajar. Otros organismos, como *Escherichia coli* y *Saccharomyces cerevisiae*, tienen información muy completa que sería de gran ayuda para refinar este método. De igual forma, se puede extender el criterio de integración u otros fenómenos dentro de la célula, como interacciones proteína-proteína, transducción de señales o represión de factores transcripcionales por metabolito. En *E. coli*, el trabajo hecho en *B. subtilis* se puede trasladar con relativa facilidad, al ser ambos organismos bacterianos; en *S. cerevisiae*, no obstante, se tendrían que hallar formas de reconocer fenómenos que son exclusivos de los eucariotes, como la compartamentalización celular.

Apéndices

Ambos apéndices se encuentran en la dirección web.

Apéndice A: Constitución detallada de los grupos metabólicos crudos en todas las redes analizadas.

Apéndice B: Nombres de las reacciones incluidas en este trabajo, ordenadas por su respectivo identificador en KEGG.

Referencias

1. David L, Marashi SA, Larhlimi A, Mieth B, Bockmayr A: **FFCA: a feasibility-based method for flux coupling analysis of metabolic networks**. *BMC Bioinformatics* 2011, **12:**236.

2. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5:**101-113.

3. Notebaart RA, Teusink B, Siezen RJ, Papp B: **Co-regulation of metabolic genes is better explained by flux coupling than by network distance**. *PLoS Comput Biol* 2008, **4:**e26.

4. Karlebach G, Shamir R: **Modelling and analysis of gene regulatory networks**. *Nat Rev Mol Cell Biol* 2008, **9:**770-780.

5. Balazsi G, Barabasi AL, Oltvai ZN: **Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli***. *Proc Natl Acad Sci U S A* 2005, **102:**7841-7846.

6. Vazquez CD, Freyre-Gonzalez JA, Gosset G, Loza JA, Gutierrez-Rios RM: **Identification of network topological units coordinating the global expression response to glucose in *Bacillus subtilis* and its comparison to *Escherichia coli***. *BMC Microbiol* 2009, **9:**176.

7. Ravasz E, Barabasi AL: **Hierarchical organization in complex networks**. *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67:**026112.

8. Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO: **Reconstruction of biochemical networks in microorganisms**. *Nat Rev Microbiol* 2009, **7:**129-143.

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

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

11. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95:**14863-14868.

12. Gutierrez-Rios RM, Freyre-Gonzalez JA, Resendis O, Collado-Vides J, Saier M, Gosset G: **Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli***. *BMC Microbiol* 2007, **7:**53.

13. Lima-Mendez G, van HJ: **The powerful law of the power law and other myths in network biology**. *Mol Biosyst* 2009, **5:**1482-1493.

14. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD: **Flux coupling analysis of genome-scale metabolic network reconstructions**. *Genome Res* 2004, **14:**301-312.

15. Moszer I, Jones LM, Moreira S, Fabry C, Danchin A: **SubtiList: the reference database for the *Bacillus subtilis* genome**. *Nucleic Acids Res* 2002, **30:**62-65.

16. Sonoshein AL HJLR: ***Bacillus subtilis* and its closest relatives.** In *Bacillus subtilis: From Cells to Genes and from Genes to Cells*. 1st edition. Edited by AL HJLRWDCA; 2011:1-6.

17. Feng J, Liu X, Xu ZR, Lu YP, Liu YY: **Effect of fermented soybean meal on intestinal morphology and digestive enzyme activities in weaned piglets**. *Dig Dis Sci* 2007, **52:**1845-1850.

18. Smith E, Morowitz HJ: **Universality in intermediary metabolism**. *Proc Natl Acad Sci U S A* 2004, **101:**13168-13173.

19. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2010, **38:**D355-D360.

20. Sierro N, Makita Y, de HM, Nakai K: **DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information**. *Nucleic Acids Res* 2008, **36:**D93-D96.

21. Gross JL YJ: *Handbook of Graph Theory*: CRC; 2003.

22. Edwards JS, Palsson BO: **The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities**. *Proc Natl Acad Sci U S A* 2000, **97:**5528-5533.

23. Farkas IJ, Wu C, Chennubhotla C, Bahar I, Oltvai ZN: **Topological basis of signal integration in the transcriptional-regulatory network of the yeast, *Saccharomyces cerevisiae***. *BMC Bioinformatics* 2006, **7:**478.

24. Goelzer A, Bekkal BF, Martin-Verstraete I, Noirot P, Bessieres P, Aymerich S, Fromion V: **Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis***. *BMC Syst Biol* 2008, **2:**20.

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

26. Guimera R, Nunes Amaral LA: **Functional cartography of complex metabolic networks**. *Nature* 2005, **433:**895-900.

27. Handke LD, Shivers RP, Sonenshein AL: **Interaction of *Bacillus subtilis* CodY with GTP**. *J Bacteriol* 2008, **190:**798-806.

28. Shivers RP, Sonenshein AL: **Activation of the *Bacillus subtilis* global regulator CodY by direct interaction with branched-chain amino acids**. *Mol Microbiol* 2004, **53:**599-611.

29. Fisher SH, Strauch MA, Atkinson MR, Wray LV, Jr.: **Modulation of *Bacillus subtilis* catabolite repression by transition state regulatory protein AbrB**. *J Bacteriol* 1994, **176:**1903-1912.

30. Yakhnin AV, Babitzke P: **NusA-stimulated RNA polymerase pausing and termination participates in the *Bacillus subtilis trp* operon attenuation mechanism *in vitro***. *Proc Natl Acad Sci U S A* 2002, **99:**11067-11072.

31. Renna MC, Najimudin N, Winik LR, Zahler SA: **Regulation of the *Bacillus subtilis alsS*, *alsD*, and *alsR* genes involved in post-exponential-phase production of acetoin**. *J Bacteriol* 1993, **175:**3863-3875.

32. Cruz RH, Hoffmann T, Marino M, Nedjari H, Presecan-Siedel E, Dreesen O, Glaser P, Jahn D: **Fermentative metabolism of *Bacillus subtilis*: physiology and regulation of gene expression**. *J Bacteriol* 2000, **182:**3072-3080.

33. Gyan S, Shiohira Y, Sato I, Takeuchi M, Sato T: **Regulatory loop between redox sensing of the NADH/NAD(+) ratio by Rex (YdiH) and oxidation of NADH by NADH dehydrogenase Ndh in *Bacillus subtilis***. *J Bacteriol* 2006, **188:**7062-7071.

34. Tojo S, Satomura T, Morisaki K, Yoshida K, Hirooka K, Fujita Y: **Negative transcriptional regulation of the *ilv-leu* operon for biosynthesis of branched-chain amino acids through the *Bacillus subtilis* global regulator TnrA**. *J Bacteriol* 2004, **186:**7971-7979.

35. Muller YA, Schulz GE: **Structure of the thiamine- and flavin-dependent enzyme pyruvate oxidase**. *Science* 1993, **259:**965-967.

36. Abdel-Fattah WR, Chen Y, Eldakak A, Hulett FM: ***Bacillus subtilis* phosphorylated PhoP: direct activation of the E(sigma)A- and repression of the E(sigma)E-responsive phoB-PS+V promoters during pho response**. *J Bacteriol* 2005, **187:**5166-5178.

37. Pang AS, Nathoo S, Wong SL: **Cloning and characterization of a pair of novel genes that regulate production of extracellular enzymes in *Bacillus subtilis***. *J Bacteriol* 1991, **173:**46-54.

38. Yakhnin H, Zhang H, Yakhnin AV, Babitzke P: **The trp RNA-binding attenuation protein of *Bacillus subtilis* regulates translation of the tryptophan transport gene *trpP* (*yhaG*) by blocking ribosome binding**. *J Bacteriol* 2004, **186:**278-286.

39. Homuth G, Rompf A, Schumann W, Jahn D: **Transcriptional control of *Bacillus subtilis* hemN and hemZ**. *J Bacteriol* 1999, **181:**5922-5929.

40. Marino M, Ramos HC, Hoffmann T, Glaser P, Jahn D: **Modulation of anaerobic energy metabolism of *Bacillus subtilis* by arfM (ywiD)**. *J Bacteriol* 2001, **183:**6815-6821.

41. Reents H, Gruner I, Harmening U, Bottger LH, Layer G, Heathcote P, Trautwein AX, Jahn D, Hartig E: ***Bacillus subtilis* Fnr senses oxygen via a [4Fe-4S] cluster coordinated by three cysteine residues without change in the oligomeric state**. *Mol Microbiol* 2006, **60:**1432-1445.

42. Geng H, Zuber P, Nakano MM: **Regulation of respiratory genes by ResD-ResE signal transduction system in *Bacillus subtilis***. *Methods Enzymol* 2007, **422:**448-464.

43. Kommineni S, Yukl E, Hayashi T, Delepine J, Geng H, Moenne-Loccoz P, Nakano MM: **Nitric oxide-sensitive and -insensitive interaction of *Bacillus subtilis* NsrR with a ResDE-controlled promoter**. *Mol Microbiol* 2010, **78:**1280-1293.

44. Alejandra A Mayela Manjarrez Casas: **Análisis Topologico de la Red de Regulacion Transcripcional de *Bacillus subtilis*.** Universidad Nacional Autónoma de México; 2009.

45. Freyre-Gonzalez JA, Alonso-Pavon JA, Trevino-Quintanilla LG, Collado-Vides J: **Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach**. *Genome Biol* 2008, 9:R154.

46. Turanov AA, Lobanov AV, Fomenko DE, Morrison HG, Sogin ML, Klobutcher LA, Hatfield DL, Gladyshev VN: **Genetic code supports targeted insertion of two amino acids by one codon.** *Science* 2009, Jan 9;323(5911):259-61.

47. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of *Escherichia coli*.** *Nat Genet*. 2002 May;31(1):64-8.

48. Oehler S, Eismann ER, Krämer H, Müller-Hill B: **The three operators of the *lac* operon cooperate in repression.** *EMBO J*. 1990 Apr;9(4):973-9.

49. Cooper TF, Remold SK, Lenski RE, Schneider D: **Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli.** *PLoS Genet*. 2008 Feb;4(2):e35.

50. Garrity DB, Zahler SA: **Mutations in the gene for a tRNA that functions as a regulator of a transcriptional attenuator in *Bacillus subtilis*.** *Genetics*. 1994 Jul;137(3):627-36.

51. Babitzke P: **Regulation of transcription attenuation and translation initiation by allosteric control of an RNA-binding protein: the *Bacillus subtilis* TRAP protein.** *Curr Opin Microbiol*. 2004 Apr;7(2):132-9.

52. Nakano MM, Zhu Y: **Involvement of ResE phosphatase activity in down-regulation of ResD-controlled genes in *Bacillus subtilis* during aerobic growth.** *J Bacteriol.* 2001 Mar; 183(6):1938-44.
53. Varma A, Palsson BO: **Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use.** *Nat Biotech* 1994 12, 994 - 998 doi:10.1038/nbt1094-994