UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
PROGRAMA DE DOCTORADO EN CIENCIAS BIOMÉDICAS
CENTRO DE CIENCIAS GENÓMICAS
CUERNAVACA, MORELOS

PREDICCIÓN Y ANÁLISIS ESTRUCTURAL DE REDES DE REGULACIÓN EN BACTERIAS A TRAVÉS DE LA INTEGRACIÓN ÓMICA PARA UNA BIOLOGÍA DE SISTEMAS COMPARATIVA

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DR. EN CIENCIAS

PRESENTA:
JUAN MIGUEL ESCORCIA RODRÍGUEZ

TUTOR PRINCIPAL
JULIO AUGUSTO FREYRE GONZÁLEZ
CENTRO DE CIENCIAS GENÓMICAS, UNAM

MIEMBROS DEL COMITÉ TUTOR

DR. ENRIQUE MERINO PÉREZ
INSTITUTO DE BIOTECNOLOGÍA, UNAM

DR. GUILLERMO GOSSET LAGARDA
INSTITUTO DE BIOTECNOLOGÍA, UNAM

CD. MX. DICIEMBRE, 2023

Prefacio

El conjunto de interacciones moleculares continuas que ocurren dentro de las células es lo que les permite responder al medio en el que se encuentran. Nuestra comprensión de dicho conjunto sigue siendo limitada debido a su gran complejidad, dada la cantidad de entidades moleculares y sus potenciales interacciones. El principio de 'divide y vencerás' es una de las estrategias más utilizadas para entender un problema complejo. Este principio hace referencia a dividir un problema complejo en partes tan pequeñas que su solución resulte obvia. Aplicar este concepto para entender un sistema complejo implicaría identificar la unidad del sistema y entenderla, para así comprender el sistema completo. Esta corriente llevó a la carrera de secuenciar el genoma humano. Siendo el gen la unidad de información biológica, conocer todos los genes nos llevaría a entender el cuerpo humano. Es como si se tratara de una gran maquinaria con una gran cantidad de componentes, la cual separáramos en cada uno de sus elementos, para así entender cómo funciona, cómo repararla, cómo mejorarla, cómo diseñarla.

¿Qué pasaría si, en lugar de encontrar tornillos y cables dentro de la máquina, encontráramos una gran 'cinta' que no parece mostrar, de manera aparente, el inicio y fin de una unidad de información? Es como una máquina de Turing, un conjunto de componentes mecánicos que 'leen' y 'modifican' una gran cinta que contiene un alfabeto muy limitado. Además, agreguemos que dichos componentes, en lugar de ser piezas rígidas con extremidades bien definidas, son piezas con alta flexibilidad cuyo estado puede ser alterado por otras piezas, teniendo como resultado una modificación en su función. Este conjunto de piezas no siempre se encuentra en cantidades constantes ni son las mismas en todo momento, sino que son codificadas dentro de la gran cinta que las piezas mismas leen. Conocer el conjunto de componentes no nos es suficiente para entender la maquinaria, así como tener el conjunto de genes no nos es suficiente para entender un organismo.

Después de la secuenciación del genoma humano y de varios organismos adicionales, se desencadenó una competencia por entender las interacciones entre estos genes. Gracias a que las restricciones tecnológicas necesarias para su investigación no eran idénticas a las requeridas para secuenciar más de 3 mil millones de bases, en esta competencia participaron más que solo un par de grupos de investigación. A pesar de los enfoques experimentales para investigar la regulación transcripcional, este trabajo se centra en el estudio *in silico* (es decir, computacional) del estrato de regulación transcripcional, donde se integran de manera indirecta los efectos provenientes de otros estratos. La investigación *in silico* nos habilita para llevar a cabo experimentos que aún no son viables de realizar de manera experimental, como el acotamiento de la estructura y su conservación en diversos organismos bacterianos.

Este trabajo se ha desarrollado con la premisa de que el examen de las interacciones moleculares, tanto a nivel global como modular, es complementario. Sin el aporte de aquellos que se han dedicado al estudio detallado de un mecanismo de regulación específico, no podríamos evaluar la fiabilidad de nuestras predicciones. Estas predicciones, a su vez, delimitarán el espacio de búsqueda para aquellos que buscan respuestas a mecanismos particulares. La integración de ambos enfoques nos permitirá identificar los principios fundamentales necesarios para comprender los aspectos esenciales de la vida.

Agradecimientos

Agradezco principalmente a mi mentor, el Dr. Julio A. Freyre González, cuya guía y constante apoyo superan los límites de este trabajo. Reconozco y agradezco su incansable esfuerzo en la formación integral de investigadores. Asimismo, extiendo mi gratitud al Dr. Enrique Merino Pérez y al Dr. Guillermo Gosset Lagarda, mi comité tutor, por su consejo invaluable y apoyo continuo a lo largo de mi travesía doctoral. A la Mtra. Mirthza Aguilar, quien despertó en mí el interés por la ciencia y a quien atribuyo el inicio de esta travesía, gracias Mirthza.

Agradezco a los colaboradores de los artículos incorporados en esta tesis, cuyas discusiones invaluables y dedicación han enriquecido este trabajo de manera significativa. Con el mismo aprecio, agradezco a los colaboradores de aquellos trabajos que están por publicarse, derivados de esta tesis, así como a aquellos proyectos que sirvieron como valiosa experiencia y donde tuve el placer de encontrar grandes amigos.

Mi reconocimiento también se extiende al comité evaluador por su valiosa retroalimentación, y a los docentes que contribuyeron a mi formación académica. A Denny, por siempre dedicar tiempo a resolver cuantas dudas tuviera respecto al doctorado.

A mi familia, especialmente a mi madre Aurelia y a mi hermana May, les agradezco profundamente por su inmensurable apoyo y motivación que me ha permitido perseguir mis sueños. A mi esposa, Angélica, por su impulso, amor incondicional y por permitirme ver el mundo a través de sus ojos, haciendo que lo infinito sea alcanzable.

A mis compañeros del laboratorio; las discusiones académicas y su apoyo invaluable han hecho que mi experiencia doctoral sea estimulante y enriquecedora. Agradezco a mis amigos, en particular a Mirthza, Julio, Karina, Luis, Carmen y Andrea, por compartir conmigo los desafíos que la academia nos presenta. A Majo, Marian, Elías, Luis y Diego de la 13, quienes contribuyeron al trabajo que sentó las bases de lo que posteriormente se convertiría en mi proyecto doctoral.

Al Programa de Doctorado de Ciencias Biomédicas, por brindarme la oportunidad de ingresar y desarrollar mi proyecto doctoral. Al CONACYT, por la beca doctoral 519023219, la cual me permitió dedicarme de tiempo completo a la ciencia.

Índice

1 Resumen

La regulación de la transcripción en bacterias, esencial para su adaptación, es mediada por factores de transcripción, codificados por genes y encargados de controlar el momento y la cantidad de transcripción génica. Comprender estas regulaciones de manera holística proporciona indicios sobre los fundamentos biológicos de la vida. La validación experimental de todas las posibles interacciones genéticas en muchos organismos no es viable debido a la complejidad combinatoria del problema. Computacionalmente, se aborda este desafío infiriendo redes a partir de datos biológicos para identificar interacciones estadísticamente probables. La comparación de estrategias se realiza con organismos modelo que tienen interacciones validadas experimentalmente. Sin embargo, la incompletez de estas redes penaliza a los métodos que identifican interacciones que suceden en la célula, pero no han sido validadas.

Este trabajo actualiza y adapta modelos de redes de regulación en bacterias para ser utilizados como estándar de oro en la predicción de redes. Se revisan y evalúan métodos para aprovechar datos genómicos y transcriptómicos en la inferencia de estas redes. Se establece el uso de predicción consenso mediante la identificación computacional de sitios de regulación con la finalidad de inferir redes de regulación globales, permitiendo una inferencia más precisa. Además, se introduce una estrategia de evaluación robusta a la incompletez de las redes experimentales, basada en valores acotados de propiedades estructurales en redes de regulación bacteriana. Esta estrategia identifica sesgos e interpretaciones erróneas en evaluaciones previas, además de examinar la utilidad de datos de expresión sintéticos, dando lugar a redes con estructuras distintas a las obtenidas con datos biológicos.

En conjunto, este trabajo contribuye al avance del conocimiento en el campo de la regulación genética bacteriana, proporcionando estrategias para una inferencia más precisa a partir de datos genómicos y transcriptómicos y una evaluación efectiva de las redes de regulación. Dichas estrategias abren nuevas perspectivas para comprender y aplicar estos procesos biológicos fundamentales.

2 Introducción

En esta sección, se ofrece una introducción general. Para tratar aspectos más específicos, se puede consultar el Anexo I, el cual contiene un artículo que profundiza en los estratos de regulación y complejidad, proporcionando además una visión introductoria sobre las limitaciones y perspectivas en el campo de la biología de sistemas. De igual manera, el anexo II profundiza en el estado actual del conocimiento sobre las redes de regulación en bacterias. Además, cada uno de los trabajos publicados que se incluyen como anexos cuentan con una introducción específica.

2.1 Regulación de la transcripción en bacterias

En bacterias, la transcripción de ácido desoxirribonucleico (ADN) a ácido ribonucleico mensajero (ARNm) se lleva a cabo por la enzima ARN polimerasa, la cual requiere estar acoplada a un factor sigma (σ) (Browning & Busby, 2004) (Fig.1A, B). Los factores σ son proteínas que secuestran las ARN polimerasas, modificando su estructura y brindándoles especificidad, por lo que actúan como reguladores de la transcripción (Burgess, Travers, Dunn, & Bautz, 1969; Gottesman, 1984, 2019).

Además de los factores σ, existen otros factores de transcripción que regulan la expresión de los genes regulados. Los factores de transcripción se clasifican, por su efecto regulatorio, en activadores y represores (Browning & Busby, 2004) (Fig.1C-D). Ciertos factores de transcripción pueden tener ambos efectos regulatorios, dependiendo del conjunto de genes que se encuentran regulando y las condiciones específicas (Rasmussen, Holst, & Valentin-Hansen, 1996).

A pesar de los diversos mecanismos de regulación (Browning & Busby, 2004, 2016), en este trabajo nos centramos en los factores de transcripción que actúan por medio de su unión al ADN en regiones cercanas al inicio de la transcripción (Browning & Busby, 2016; Todeschini, Georges, & Veitia, 2014).

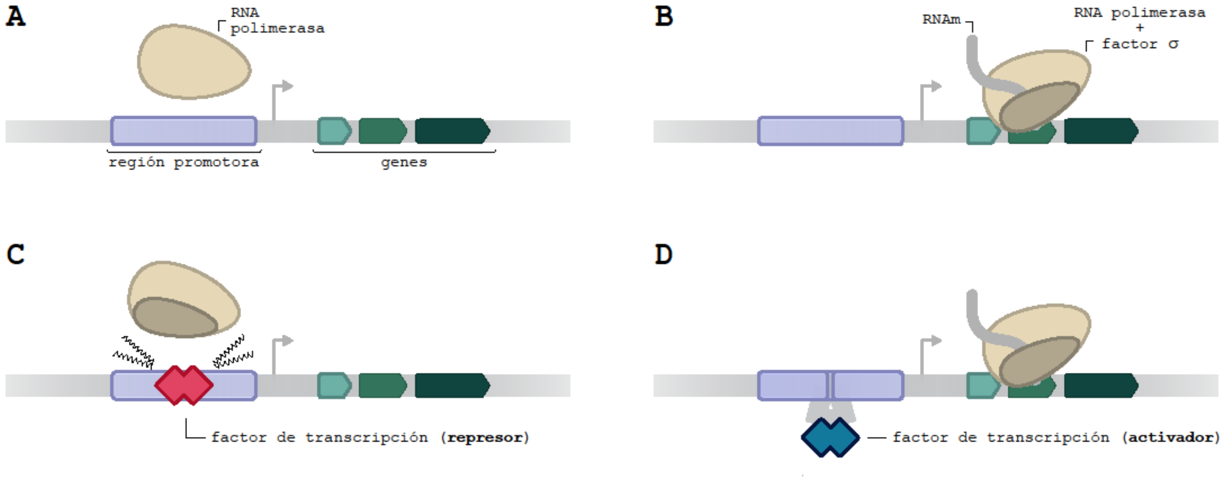

A
RNA polimerasa
región promotora
genes
B
RNAm
RNA polimerasa + factor σ
C
factor de transcripción (represor)
D
factor de transcripción (activador)

Figura 1. Representación de la transcripción. A) La ARN polimerasa es la enzima encargada de llevar a cabo la transcripción de los genes. Sin embargo, en bacterias requiere estar acoplada a un factor sigma para poder llevar a cabo su

función (B). La transcripción puede ser regulada por factores de transcripción con diversos mecanismos. Por ejemplo, los factores de transcripción pueden evitar que la ARN polimerasa se una a la región promotora, evitando la transcripción (C). También existen factores de transcripción que promueven la transcripción, por ejemplo, plegando el ADN para que la ARN polimerasa sea capaz de reconocer su sitio de unión (D).

2.2 Modelado de la regulación

Para poder estudiar los fenómenos biológicos que ocurren en la célula de manera *in silico*, es necesario generar primero modelos que representen estos fenómenos. De la misma manera que utilizamos una secuencia de caracteres [ACTG] para representar una secuencia de ADN, recurrimos a modelos matemáticos y computacionales que nos permitan representar los sitios de interacción entre las proteínas y el ADN, así como el conjunto de interacciones regulatorias que pueden ocurrir en una célula.

2.2.1 Modelado de los sitios de regulación

Los factores de transcripción poseen la capacidad de reconocer sitios específicos de unión en el ADN. Aunque también pueden unirse a regiones no específicas, lo hacen con una afinidad considerablemente menor, permitiéndoles realizar desplazamientos que les facilitan llegar a sus sitios específicos (Suter, 2020). Los factores de transcripción que regulan un número reducido de genes suelen cotranscribirse con sus genes asociados o están ubicados en proximidades, lo que les permite unirse a sus sitios correspondientes de manera casi inmediata tras su síntesis (Kolesov, Wunderlich, Laikova, Gelfand, & Mirny, 2007). Por otra parte, los factores de transcripción que regulan muchos genes se ven favorecidos de una unión a sitios con menor afinad, permitiéndoles desplazarse por el ADN para llegar a sus sitios de unión, los cuales suelen estar alejados (Kolesov et al., 2007).

En 1975, David Pribnow publicó un artículo donde identificó la región conservada -10 de los promotores en bacterias, dada una colección de 6 regiones donde únicamente dos de los 6 bases estaban conservadas (Pribnow, 1975). Esto le fue suficiente para sugerir que esa región moderadamente conservada estaba implicada en la unión de la ARN polimerasa con el ADN. Dada la pequeña cantidad de sitios y que se conocía el inicio de la transcripción de los genes posteriores al promotor, fue posible la identificación de un sitio consenso de manera manual.

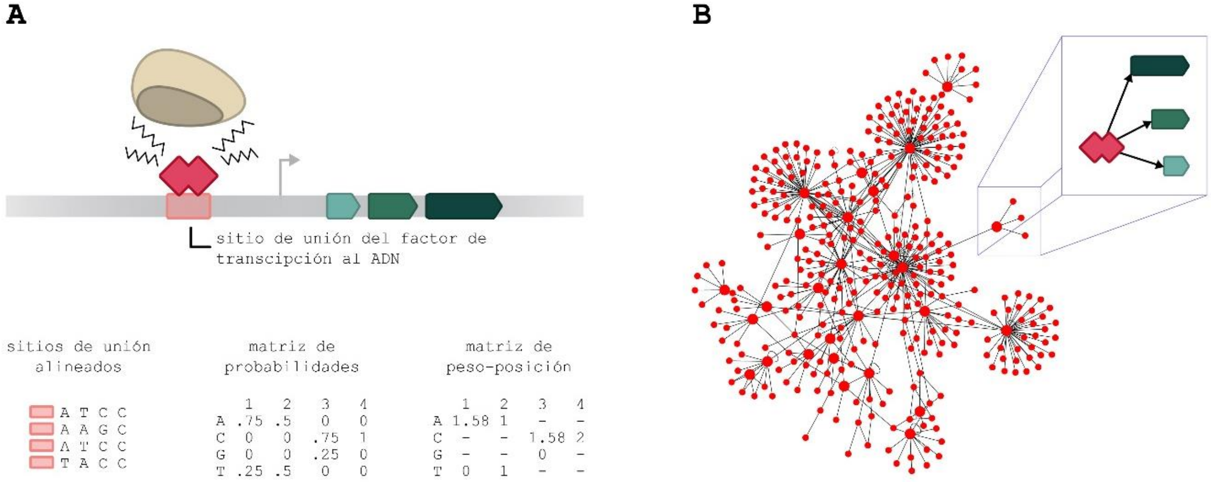
Dicho procedimiento manual no es factible cuando se tiene una gran colección de sitios, o cuando no se conoce el alineamiento de las secuencias *a priori*. Menos práctico aún, cuando el objetivo es identificar los sitios de unión de una gran cantidad de proteínas. Stormo, et al. mostraron el uso de un perceptrón para modelar la región Shine-Dalgarno, el sitio de unión del ribosoma en el ARN mensajero, a partir de una larga colección de sitios (Stormo, Schneider, Gold, & Ehrenfeucht, 1982). Dos años después, Staden publica un enfoque puramente estadístico que obtenía los pesos de la matriz a partir de las probabilidades de las bases en cada posición de los sitios observados (Staden, 1984) (Figura 2A).

2.2.2 Modelado de la red global de regulación

La regulación transcripcional es un mecanismo complejo, en el cual múltiples señales son integradas (Browning & Busby, 2004) y la complejidad del modelo depende del nivel con que pretende ser estudiado (Karlebach & Shamir, 2008). El análisis de un regulón de pequeña escala se puede modelar con un alto nivel de detalle. Sin embargo, al abordar el estudio redes de regulación a nivel del organismo, es esencial realizar una abstracción más profunda. Esto implica conservar únicamente la información más relevante, omitiendo detalles específicos que podrían introducir ruido y facilitando así el estudio del modelo en un marco temporal finito (Figura 2B).

El empleo de ciencia de redes y teoría de grafos nos permite modelar redes de regulación de gran tamaño, representando un efecto regulatorio de la proteína A sobre la región promotora del gen b, como una interacción dirigida a→b; donde a es el gen que codifica para la proteína regulatoria A. De esta manera, estamos abstrayendo los pasos que se requieren para que A se sintetice, dado que sólo buscamos representar el efecto regulatorio que existe de un gen, hacia el otro. Esta presentación gen-gen, nos permite obtener un grafo homogéneo, donde todos sus componentes representan la misma entidad biológica.

El uso de este modelo, permitió emplear análisis de teoría de redes para analizar la estructura de las redes de regulación, encontrando propiedades globales en común, tales como la jerarquía y modularidad (Barabasi & Oltvai, 2004; Freyre-González, Alonso-Pavón, Trevino-Quintanilla, & Collado-Vides, 2008; Freyre-González et al., 2013; Resendis-Antonio et al., 2005), la existencia de genes intermodulares encargados de integrar señales de distintos módulos funcionales (Freyre-González et al., 2008; Freyre-González et al., 2013) y una baja densidad, independiente del número de genes (Campos & Freyre-Gonzalez, 2019).



	1	2	3	4
A	.75	.5	0	0
C	0	0	.75	1
G	0	0	.25	0
T	.25	.5	0	0

	1	2	3	4
A	1.58	1	-	-
C	-	-	1.58	2
G	-	-	0	-
T	0	1	-	-

Figura 2. Modelos de la unión de proteínas al ADN y de la regulación global de la transcripción. A) Los factores de transcripción reconocen sitios específicos moderadamente conservados. Esto permite obtener una representación de las posibles regiones de ADN a las que pueden unirse. En el ejemplo mostrado en la parte inferior del panel A, a partir de los sitios de unión se calcula la matriz de probabilidades para cada uno de los sitios y nucleótidos. Es decir, dados los sitos conocidos, la probabilidad de que el nucleótido en la primera posición sea una adenina (A) es de 0.75 y 0.25 de que sea una timina (T). Posteriormente, las matrices de probabilidades se pueden transformar en una matriz de peso posición como el logaritmo del cociente entre la probabilidad obtenida para una base en un sitio específico y una probabilidad esperada. Dicha probabilidad esperada es la probabilidad de encontrar esa base en el genoma del organismo. Para el ejemplo de la ilustración se consideró que las 4 bases tienen las mismas probabilidades, 0.25. Ocupando el logaritmo base 2 para representar el resultado en bites, para A en la primera posición obtenemos que log2(0.75/0.35) = ~ 1.584. En la ilustración se enmascaran las probabilidades cero. Lo común es usar una pseudocuenta, un valor muy cercano a cero, para evitar valores infinitos. B) La unión de todas las interacciones regulatorias que suceden en la célula nos permite obtener la red de regulación global. Se usa como ejemplo una representación gráfica de la red 100226_v2019_sA22-DBSCR15_eStrong (Escorcia-Rodríguez, Tauch, & Freyre-González, 2020; Zorro-Aranda, Escorcia-Rodríguez, Gonzalez-Kise, & Freyre-González, 2022).

2.2.2.1 Ciencia de redes

La ciencia de redes parte de la teoría de grafos, que, a pesar de estar estrechamente relacionados, la ciencia de redes se enfoca en los sistemas reales, mientras que la teoría de grafos se emplea comúnmente para hacer referencia a las representaciones matemáticas de dichas redes (Barabási & Pósfai, 2016). La teoría de grafos nace con la demostración, por parte de Leonhard Euler, de que un problema no tenía solución (Euler, 1741). Años después, durante una década, a partir de 1959, los matemáticos Paul Erdős y Alfréd Rényi publican una serie de artículos, sentando las bases de la ciencia de redes moderna por medio del estudio de un modelo aleatorio de redes, actualmente conocido como redes Erdős-Rényi (Barabási & Pósfai, 2016).

Fue hasta 1999, cuando Barabási junto con dos de sus postdoctorados publican un artículo caracterizando la red de internet (Albert, Jeong, & Barabási, 1999), que se comenzaron a estudiar redes del mundo real a gran escala. Fue el mismo laboratorio quienes también identificaron los primeros fundamentos de las redes moleculares en el transcurso del primer lustro del siglo XXI (Barabasi & Oltvai, 2004; Jeong, Mason, Barabasi, & Oltvai, 2001; Jeong, Tombor, Albert, Oltvai, & Barabasi, 2000). Posteriormente, numerosos trabajos han investigado la estructura de las redes de regulación y su conservación en distintos organismos (Freyre-González et al., 2008; Freyre-González et al., 2013; Freyre-González & Tauch, 2017).

2.3 Inferencia de redes de regulación génica

Debido a la incompletez de las redes de regulación, aún en organismos modelo bacterianos (Escorcia-Rodríguez, Tauch, & Freyre-González, 2020), se ha optado por su inferencia computacional a partir de datos genómicos y transcriptómicos.

2.3.1 Inferencia a partir de datos genómicos

Antes de que la inferencia de redes de regulación a escala organismo se popularizara, la inferencia de interacciones regulatorias utilizando matrices de peso, secuencias y anotaciones de genomas ya era una práctica común. Sin embargo, el proceso era prácticamente artesanal.

El fundamento de la inferencia de interacciones regulatorias se basa en que, genes que son regulados por un mismo factor de transcripción, comparten un sitio de unión para dicho factor en su región promotora (D'Haeseleer, 2006b). Dicha región, al estar ocupada frecuentemente por el factor de transcripción, es menos propensa a modificaciones, en comparación con las regiones flanqueantes. Por lo que, al tener una colección de secuencias que contienen el sitio de unión, dicho sitio de unión va a estar estadísticamente sobrerrepresentado, es decir, un motivo (D'Haeseleer, 2006b). Para incrementar la significancia estadística, lo ideal es que todas las secuencias a estudiar contengan el sitio y el menor número de bases flanqueantes (D'Haeseleer, 2006a).

En la práctica, no sabemos a priori el sitio de regulación. Sin embargo, sabemos de datos experimentales que los sitios de regulación en bacterias comúnmente se encuentran en regiones cercanas al inicio de la transcripción (Robison, McGuire, & Church, 1998). Por lo que podemos tener una aproximación en la que los sitios de regulación se encuentran estadísticamente sobrerrepresentados en nuestro conjunto de secuencias promotoras (Figura 2B). En el anexo III se describen de manera detallada los métodos empleados para la inferencia a partir de la identificación de sitios de regulación.

Hasta la fecha, no existe un organismo para el cual se conozca su red de regulación completa (Escorcia-Rodríguez et al., 2020). Sin embargo, existen organismos modelos para los cuales se ha estudiado una cantidad considerable de interacciones regulatorias. Podemos usar lo conocido para trasladar la información regulatoria a organismos cercanos bajo la premisa que los sitios de regulación están más conservados que las regiones flanqueantes y que son menos propensas a mutaciones, dada su función que la célula requiere (Novichkov et al., 2013; Rodionov et al., 2011).

2.3.2 Inferencia a partir de datos transcriptómicos

La transcriptómica es el estudio del conjunto de ARN en una célula. Dado que el resultado de la regulación transcripcional es la ausencia o presencia de ARN mensajero, es intuitivo hacer uso de la transcriptómica para la inferencia de interacciones regulatorias. La base fundamental de los métodos de inferencia que utilizan datos de expresión implica la construcción de una matriz que comprende los datos de expresión génica del organismo en diversas condiciones. Este proceso incluye el preprocesamiento de los datos con el objetivo de eliminar el ruido ocasionado por lotes, garantizando así la calidad y confiabilidad de la información obtenida (Zhang, Parmigiani, & Johnson, 2020).

Posteriormente, se aplican estrategias estadísticas, probabilísticas, algoritmos de aprendizaje automático o combinaciones de los mencionados, con el propósito de identificar los perfiles de expresión génica que presentan una relación más significativa entre sí que el resto y que supera lo que se podría esperar al azar (Marbach et al., 2012). Los resultados obtenidos por métodos individuales incluyen una alta tasa de falsos positivos, por lo que comúnmente se integran los resultados de diversos métodos, dándole más importancia a las interacciones mejor evaluadas por los métodos utilizados (Marbach et al., 2012). El anexo IV contiene una revisión detallada del estado del arte acerca de las metodologías empleadas para la predicción de redes de regulación utilizando datos de transcripción.

Independientemente de la estrategia empleada, la inferencia de redes de regulación nos brinda la oportunidad de explorar la intrincada red de interacciones que tiene lugar en un organismo y cómo esta red se distingue o compara con la de otros organismos. Además, nos permite formular hipótesis sobre la función de los genes basándonos en su regulación. Este enfoque resulta especialmente valioso cuando no es factible obtener una anotación funcional confiable únicamente a partir de relaciones genéticas.

2.4 Relaciones Genéticas

En un trabajo seminal, Walter M. Fitch diferenció proteínas homólogas de análogas y definió la ortología y la paralogía como una subclasificación de la homología (Fitch, 1970). Definió a los ortólogos como genes homólogos resultantes de un evento de especiación y a los parálogos como resultantes de un evento de duplicación (Fitch, 1970). Otro término importante de relaciones genéticas son los xenólogos, genes adquiridos mediante transferencia horizontal de genes (Koonin, 2005), relaciones esenciales en el proceso evolutivo de las bacterias (Arnold, Huang, & Hanage, 2022).

Entre dos organismos pueden existir distintos tipos de ortología que dependen de las historias evolutivas de sus genes, estas pueden ser 1:1, m:1, 1:n o m:n. Siendo 1:1 aquellas donde un solo gen es ortólogo de un gen único en el otro organismo. Por otro lado, las relaciones m:1, 1:n y m:n involucran la existencia de co-ortología donde más de un gen en un organismo son ortólogos con uno o más genes en el otro organismo (Altenhoff, Glover, & Dessimoz, 2019).

2.4.1 Asociación funcional a las relaciones de homología.

Una de las aplicaciones más amplias de la identificación de ortólogos es la inferencia de función de proteínas no caracterizadas. Una gran cantidad de trabajos se han centrado en la conservación de la función en genes ortólogos para estudiar la teoría que establece que los ortólogos tienden a conservar su función más que los parálogos. La idea de que los parálogos tienden a tener funciones diferentes surge como resultado de trabajos previos donde se ha observado innovación funcional y subfuncionalización de genes duplicados (Conant & Wolfe, 2008). Sin embargo, el mal uso de la definición de ortólogos y parálogos en términos de conservación de funciones (Gerlt & Babbitt, 2000) en lugar de la definición original de Fitch, en términos de evolución (Fitch, 2000), generó varias complicaciones en la comunidad (Jensen, 2001).

Nehrt et al. realizaron un estudio a gran escala de la conjetura de ortología utilizando ortólogos y parálogos de humanos y ratones, así como datos de expresión, y descubrió que los parálogos son un mejor recurso para predecir la función genética (Nehrt, Clark, Radivojac, & Hahn, 2011). Llegaron a la conclusión de que, más que la secuencia, el contexto celular era el factor más importante en la evolución/conservación de la función de las proteínas (Nehrt et al., 2011). Las conclusiones de este trabajo estaban en contra del modelo estándar. Como resultado, un grupo independiente volvió a analizar los mismos casos estudiados por Nehrt et al. y encontraron sesgos no considerados en su análisis debido a la preferencia de un tipo particular de experimentos en cada organismo y en la anotación de ontología genética (Thomas et al., 2012). Otro grupo más discutió el sesgo en el uso de anotaciones GO para validar la conjetura de ortología y utilizó datos de RNA-Seq de nueve eucariotas para mostrar que la similitud de expresión entre ortólogos es significativamente mayor que entre parálogos (X. Chen & Zhang, 2012). Además, Rogozin et al. encontraron que la mayor similitud de expresión entre parálogos era resultado del alto ruido, siendo mayor la correlación entre ortólogos que entre parálogos (Rogozin, Managadze, Shabalina, & Koonin, 2014).

Estos trabajos sugieren el potencial sesgo a tener en cuenta al momento de usar anotaciones funcionales, en gran parte por el uso de ortologías para llevar a cabo dichas anotaciones. Asimismo, se deben considerar las implicaciones evolutivas del tipo homología al inferir la funcionalidad (Gabaldon & Koonin, 2013). Además, la ortología no implica conservación de funciones, ni la conservación de funciones implica ortología (Gabaldon & Koonin, 2013).

3 Planteamiento del problema

Para comprender un organismo, es crucial entender sus componentes y cómo interactúan. A pesar de décadas de secuenciación genómica, la comprensión de las interacciones entre genes sigue siendo desafiante. Identificar estas interacciones mediante métodos experimentales implica evaluar la combinatoria de elementos. Aun limitándonos a los factores de transcripción que potencialmente pueden regular el genoma (~7% (Perez-Rueda, Collado-Vides, & Segovia, 2004)), tenemos ft*n potenciales interacciones, donde ft es el número de factores de transcripción y n es el número de genes. Por lo que, para un organismo de ~5,000 genes tendríamos 1.75 millones (350*5,000) de posibles interacciones, de las cuales, únicamente el ~1% se espera que realmente ocurran en la célula (Campos & Freyre-Gonzalez, 2019).

Aunque las tecnologías de alto rendimiento permiten identificar objetivos en todo el genoma para un factor de transcripción, esto requiere una serie de experimentos costosos en términos monetarios y temporales. Además, explorar la combinatoria de interacciones estará limitado al conjunto de interacciones que ocurren bajo una condición específica. Si queremos conocer el conjunto completo de interacciones entre los genes, necesitamos hacer el mismo experimento en todas las condiciones posibles, o por lo menos tantas como sea posibles. Este es el esfuerzo de la comunidad científica en el último medio siglo, iniciado por el trabajo seminal de Jacob y Monod en 1961 (Jacob & Monod, 1961) y acelerado por el surgimiento de las tecnologías de alto rendimiento. Dicho esfuerzo nos ha permitido tener un compendio de redes de regulación en bacterias, todas incompletas (Escorcia-Rodríguez et al., 2020).

El conjunto de interacciones que tiene lugar en un organismo refleja un caso específico de reacciones fisicoquímicas que posibilitan su existencia. Este conjunto puede variar significativamente entre diferentes organismos. La comprensión de las redes de regulación en una amplia variedad de organismos nos brinda la oportunidad de explorar tanto las similitudes como las diferencias fundamentales entre ellos, revelando los mecanismos que les permiten existir, adaptarse al entorno y reproducirse. Esto requeriría analizar todas las posibles interacciones, en una gran cantidad de condiciones, en una gran cantidad de organismos.

Para avanzar el campo de estudio de la regulación, una gran cantidad de metodologías para inferir redes de regulación han sido propuestas (Marbach et al., 2012). Dando el desequilibrio en el tamaño de los conjuntos de interacciones que suceden y que no suceden en la célula, identificar ese ~1% no es tarea trivial, dado que es un problema indeterminado (Siegenthaler & Gunawan, 2014), y ese ~1% es la probabilidad de encontrar una verdadera interacción al azar. Esto se ve reflejado en el poco poder predictivo de los métodos (Marbach et al., 2012).

La incompletez de los estándares de oro actuales complican la evaluación de las predicciones, dado que interacciones que ocurren en la célula y están siendo predichas, i.e., verdaderos positivos, son incorrectamente clasificados como falsos positivos (Figura 3). Esta es una de las razones por las no hay un solo método que se desempeñe mejor siempre, y con todos los organismos (Marbach et al., 2012). Cada estándar de oro tiene un nivel de completez distinto y diferentes sesgos en el estudio de sus interacciones. Por ejemplo, dado el interés en la producción de antibióticos, el estudio de la regulación de *Streptomyces coelicolor* está más enfocado en el metabolismo secundario, en comparación con un organismo como *Escherichia coli*.

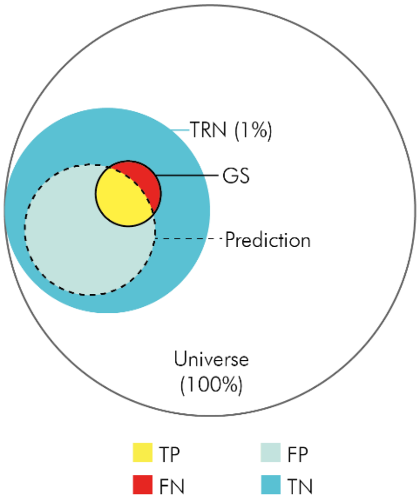


Figura 3. Representación del efecto de la incompletez del estándar de oro en la evaluación de predicciones. En este ejemplo, la predicción es un subconjunto de las interacciones que suceden en la célula, pero dada la incompletez del estándar de oro, ésta es altamente penalizada con falsos "falsos positivos". TRN representa la red de regulación transcripcional que existe en la célula, GS el estándar de oro, Prediction la predicción, Universe el universo de potenciales interacciones donde todos los genes interactúan entre ellos, TP verdaderos positivos, FP falsos positivos, FN falsos negativos y TN verdaderos negativos.

Evaluar predicciones mediante anotaciones funcionales es una alternativa poco útil para redes causales, dado que el enriquecimiento de funciones para un conjunto de genes no es indicativo del factor de transcripción responsable. Además, la conservación de función se presta a debate en genes con historias evolutivas complejas, las cuales comunes en bacterias (Koonin, 2005), y la presencia de reguladores globales complica la asociación de funciones específicas a grupos regulados, ya que abarcan gran parte del genoma, y de la red, con múltiples funciones asociadas.

4 Objetivos y sus antecedentes

4.1 Objetivo general

Desarrollar estrategias computacionales para la predicción y evaluación de redes de regulación transcripcional en bacterias que consideren la incompletez y desequilibrio de los datos.

4.2 Objetivos específicos

4.2.1 Tener un estándar de oro actualizado que nos permita evaluar las predicciones con organismos modelo.

Las interacciones regulatorias en organismos bacterianos son comúnmente reportadas en artículos científicos mas no depositadas en bases de datos. Existen bases de datos específicas para cierto organismo que recolectan la información relevante a la regulación transcripcional. Tal es el caso de RegulonDB (Tierrafria et al., 2022) para *E. coli*, Subtiwiki (Pedreira, Elfmann, & Stulke, 2022) y DBTBS (Sierro, Makita, de Hoon, & Nakai, 2008) para *Bacillus subtilis*, CoryneRegNet (M. T. D. Parise et al., 2020) para *Corynebacterium glutamicum*, RegulomePA (Galan-Vasquez, Luna-Olivera, Ramirez-Ibanez, & Martinez-Antonio, 2020) para *Pseudomonas aeruginosa*. Sin embargo, la información contenida en las bases de datos, así como el nivel de confianza de las interacciones dada su validación experimental, se encuentran de manera heterogénea, dificultando el desarrollo de análisis con fines comparativos con múltiples organismos (Escorcia-Rodríguez et al., 2020). Además, múltiples interacciones se encuentran aún dispersas en la literatura, teniendo poca redundancia con lo ya reportado en las bases de datos organismo-específicas (Escorcia-Rodríguez et al., 2020).

Abasy Atlas se desarrolló previamente en el laboratorio con la finalidad de facilitar una biología de sistemas comparativa, homogeneizando la información incluida en bases de datos organismo-específicas y literatura, y brindando anotaciones a nivel de sistemas, tales como la identificación de reguladores globales, genes de la maquinaria basal, genes intermodulares y módulos de genes destinados a funciones específicas (Ibarra-Arellano, Campos-Gonzalez, Trevino-Quintanilla, Tauch, & Freyre-Gonzalez, 2016). Sin embargo, gracias al uso de tecnologías de alto rendimiento, el número de nuevas interacciones reportadas ha crecido con mayor exponenciación recientemente, por lo que es necesario actualizar las redes de regulación constantemente. Para ello es necesario automatizar la actualización de la base de datos, así como la actualización de los archivos requeridos para las anotaciones de genes. De igual manera, esta primera versión de Abasy, al igual que las bases de datos organismo-específicas, conservan sólo las versiones más recientes de las redes de regulación. Tener un histórico de las redes con distintos niveles de completez nos permitiría estudiar el efecto de dicha incompletez sobre la inferencia de redes y su evaluación.

4.2.2 Desarrollar una estrategia para inferir redes de regulación a partir del genoma y evaluar su rendimiento.

Previos trabajos han usado la conservación de sitios de regulación por medio de estrategias computacionales para identificar motivos, secuencias estadísticamente sobrerrepresentadas, para identificar sitios de unión de los factores de transcripción al ADN (D'Haeseleer, 2006a, 2006b; McGuire, Hughes, & Church, 2000).

La identificación computacional de estos sitios de unión nos ha permitido ampliar nuestro conocimiento más allá de lo que hemos aprendido con técnicas experimentales. En el contexto de la inferencia de redes regulatorias, podemos tomar un regulón con varios TG y usar sus secuencias previas para predecir los motivos de unión de su regulador con herramientas de descubrimiento de motivos *de novo* (D'Haeseleer, 2006b). Posteriormente, podemos utilizar estos motivos para construir un modelo del sitio de unión (Figura 2B) e identificar nuevos sitios de unión, y por ende nuevas dianas, en el mismo organismo (expansión de regulón) con herramientas de escaneo de sitios de unión. También podemos utilizar estos modelos para escanear las regiones reguladoras de un organismo filogenéticamente relacionado con regulador y dianas ortólogas y reajustar los motivos de unión.

Análisis tales como el análisis de regulogs (Alkema, Lenhard, & Wasserman, 2004) y huella filogenética ("phyllogenetic footprinting") (Blanchette, Schwikowski, & Tompa, 2002) emplean estos enfoques. Regulogs son grupos de genes corregulados que comparten un sitio de regulación conservado en múltiples organismos (Alkema et al., 2004) y "phyllogenetic footprinting" es la identificación de secuencias conservadas en genes ortólogos en múltiples especies (Blanchette et al., 2002). Ambas son metodologías ampliamente usadas en genómica comparativa, sin embargo, el resultado son grupos de genes que son corregulados, mas no se conoce el factor de transcripción responsable, por lo que el resultado no es una red causal.

En el ámbito de transferencia de interacciones, RegTransBase (Cipriano et al., 2013) era una base de datos que reportaba redes de transcripción curadas y predichas con RegPredict (Novichkov et al., 2010), una herramienta computacional que permitía la transferencia de interacciones. Sin embargo, ninguna de las dos herramientas se encuentra disponible en la actualidad. Prodoric (Dudek & Jahn, 2022) es una base de datos de sitios de regulación predichos usando genómica comparativa, utilizando un solo método para la identificación *de novo* de los sitios de regulación y construcción de matrices.

4.2.3 Explorar la incorporación de datos transcriptómicos para la inferencia de redes de regulación y evaluar su rendimiento.

El uso de la transcriptómica, cuantificación de los transcritos, para estudiar la regulación de la transcripción es una conexión lógica. Por ello existe una gran cantidad de métodos con la finalidad de inferir redes de regulación a partir de datos de transcripción (Marbach et al., 2012). Sin embargo, adicional a la baja probabilidad de identificar las verdaderas interacciones que suceden en la célula discutido en el planteamiento del problema, hay múltiples interacciones moleculares que son abstraídas en una red de regulación y que pueden verse reflejadas en los datos de transcripción, dando lugar a una pobre consistencia entre los datos de transcripción y las redes de regulación (Larsen, Rottger, Schmidt, & Baumbach, 2019; D. Parise et al., 2021).

De igual manera, en múltiples trabajos evaluando el rendimiento de las herramientas para inferir redes de regulación globales a partir de datos de transcripción, se ha encontrado que no hay un sola herramienta o estrategia que tenga el mejor rendimiento en todos los casos evaluados (S. Chen & Mar, 2018; Marbach et al., 2012). Es por ello por lo que ha prevalecido la estrategia de "sabiduría de las masas", un enfoque basado en el consenso de un conjunto de predicciones individuales para obtener una única predicción con mejor rendimiento que las individuales, aun cuando el beneficio puede ser pequeño o inexistente (Marbach et al., 2012).

4.2.4 Desarrollar una estrategia que permita la evaluación de inferencia de redes de regulación, considerando el desequilibrio de los conjuntos positivos/negativos y la incompletez de los estándares de oro.

En cuando a la consideración del desequilibrio en los tamaños de las potenciales predicciones positivas y negativas, en un ámbito más general, el área bajo la curva dada por la precisión y la tasa de recuperación del estándar de oro ha demostrado ser más informativa cuando se evalúan conjuntos de datos desequilibrados, en comparación con el área bajo la curva de la curva entre la tasa de falsos positivos y verdaderos positivos (Saito & Rehmsmeier, 2015). De igual manera, el coeficiente de correlación de Matthew ha demostrado ser mejor métrica que el score F1 bajo las mismas condiciones de desequilibrio del tamaño de los conjuntos (Chicco & Jurman, 2020). Dichos trabajos contribuyen a una mejor evaluación dado el desequilibrio de los datos, mas no atacan el problema de un estándar de oro incompleto.

Los principios relacionados con la estructura global de una red reguladora transcripcional, como la estructura jerárquica y modular (Barabasi & Oltvai, 2004) se han dilucidado mediante modelos de redes para representar el conjunto global de interacciones potenciales en una célula (Barabasi & Oltvai, 2004). Podemos hacer uso de esas propiedades estructurales y otras como la existencia de genes intermodulares (Freyre-González et al., 2008; Freyre-González et al., 2013) y una baja densidad (Campos & Freyre-Gonzalez, 2019) para evaluar la similitud de una red predicha a la del espacio acotado que conocemos en las redes de regulación validadas experimentalmente. Para esto, requerimos previamente estudiar el espacio de valores acotados para un conjunto de propiedades (Costa, Rodrigues, Travieso, & Villas Boas, 2007), el cual, no existe hasta la fecha.

5 Resultados y discusión

5.1 Estándar de oro para evaluar la calidad de las predicciones

Actualizamos la base de datos Abasy Atlas (Escorcia-Rodríguez et al., 2020). En dicha actualización, se trabajó en modelos de representación de las redes listos para ser usados como estándar de oro en las evaluaciones de predicción de redes. Es posible utilizar una representación que incluya complejos regulatorios para inferencias de factores de transcripción que actúan como complejos heteroméricos, así como una representación sin complejos regulatorios, para la identificación de interacciones entre genes, útil para la evaluación de predicciones con métodos basados en datos de expresión. En el mismo artículo se plantea la posibilidad de usar la estructura de la red como indicio de la calidad de las redes. De igual manera, con base en un trabajo previo (Campos & Freyre-Gonzalez, 2019), se reporta un modelo para inferir el número de interacciones que se espera en una red completa, dado el tamaño del genoma. Más información sobre la actualización de Abasy Atlas se encuentra reportada en el artículo en el anexo II.

5.2 Inferencia con datos genómicos y un primer enfoque para la evaluación estructural de las redes

En el anexo III se encuentra el artículo "*Corynebacterium glutamicum* Regulation beyond Transcription: Organizing Principles and Reconstruction of an Extended Regulatory Network Incorporating Regulations Mediated by Small ARN and Protein-Protein Interactions" (Escorcia-Rodríguez, Tauch, & Freyre-Gonzálex, 2021). En el cual aplicamos una primera aproximación de la metodología de inferencia con datos genómicos y estudiamos la conservación de interacciones regulatorias en tres organismos modelo, a través de un enfoque conservativo para minimizar falsos positivos. En el mismo trabajo se hace una primera exploración de un conjunto de propiedades estructurales a nivel global y el cómo cambian dichas propiedades conforme la red se vuelve más completa debido a nuevos experimentos estudiando la regulación en *C. glutamicum*. Como parte del mismo trabajo, se evalúa el efecto de incluir interacciones proteína-proteína con efectos regulatorios consecuentes, e interacciones regulatorias mediadas por RNAs pequeños.

5.3 Incorporación de datos transcriptómicos y su evaluación para contrastar su poder predictivo con las inferencias basadas en datos genómicos

En el anexo IV se encuentra el artículo "Improving gene regulatory network inference and assessment: The importance of using network structure" (Escorcia-Rodríguez et al., 2023). Donde hacemos una evaluación de métodos de inferencia que ocupan datos transcriptómicos como punto de partida. El anexo V, correspondiente al artículo "Curation, inference, and assessment of a globally reconstructed gene regulatory network for *S. coelicolor*" (Zorro-Aranda, Escorcia-Rodriguez, Gonzalez-Kise, & Freyre-Gonzalez, 2022), aplicamos inferencia de redes con métodos basados en datos transcriptómicos y genómicos, así como la integración de ambos para inferir la red de *S. coelicolor.* En este trabajo se hace una primera evaluación del uso de las propiedades estructurales globales como indicio de la calidad de las redes predichas, comparándola con métodos de evaluación estándar basados en presencia o ausencia de las interacciones. Finalmente se discute la plasticidad de los componentes modulares y jerárquicos que constituyen a la red de *S. coelicolor*, comparándola con la red regulatoria de *C. glutamicum.*

El reto de predecir redes de regulación a partir de datos de expresión continúa abierto y nuevas propuestas siguen surgiendo cada día, aun cuando partiendo sólo del genoma podemos llegar a obtener mejores resultados (Zorro-Aranda et al., 2022). Enfoques más recientes hacen uso de la integración de otros datos ómicos como la proteómica y enfoques de optimización del poder predictivo (Patel et al., 2023; Rychel et al., 2021), tratando de encontrar el camino entre los datos transcriptómicos y la red de regulación. Además, continúan surgiendo métodos que utilizan el enfoque de "sabiduría de las masas" para obtener una mejor predicción a partir de datos transcriptómicos, agrupando metodologías con distintas finalidades (Shen, Coruzzi, & Shasha, 2023). El presente trabajo demuestra que la integración de predicciones de metodologías con objetivos similares es más efectiva que la integración de predicciones provenientes de metodologías con propósitos distintos (Anexo IV) (Escorcia-Rodríguez et al., 2023). A pesar de eso, el desempeño de las predicciones sigue siendo modesto (Escorcia-Rodríguez et al., 2023).

La aparente falta de coherencia entre las redes reguladoras transcripcionales y los datos de expresión génica (Larsen et al., 2019; D. Parise et al., 2021) radica en los amplios mecanismos moleculares que contribuyen a los patrones de expresión génica observados (Freyre-González et al., 2022). Encontrar coherencia entre los datos de expresión génica y las redes reguladoras de la transcripción habría resuelto el desafío de la inferencia de redes. Las redes reguladoras transcripcionales se validan con experimentos bajo ciertas condiciones, frecuentemente *in vitro*, la unión de esas interacciones conduce a una red global; el conjunto potencial de regulación de la transcripción en una célula. Sin embargo, sólo un subconjunto de esas interacciones ocurre bajo una condición particular, pudiendo actuar de manera cooperativa o antagonista, resultando en las respuestas observadas en los datos de expresión genética.

5.4 Uso de la estructura de las redes para evaluar la calidad de las predicciones

Aplicando métodos de evaluación estándar y nuestra propuesta basada en propiedades estructurales de la red, en el artículo incluido en el anexo IV, identificamos sesgos y malas interpretaciones en evaluaciones anteriores, donde métodos para inferir redes de coexpresión y métodos para inferir redes de regulación han sido evaluados por igual, sin tomar en cuenta la diferencia de estructura que se llega a obtener dependiendo del objetivo buscado. En el mismo trabajo se investiga la utilidad de datos de expresión sintéticos para llevar a cabo las evaluaciones, resultando en redes estructuralmente diferentes a las obtenidas con datos de expresión reales, remarcando la importancia del uso de datos reales en la evaluación de las predicciones.

El anexo VI corresponde al artículo "*Rhizobium etli* CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network" (Taboada-Castro et al., 2022). En este trabajo colaboramos llevando a cabo una evaluación de estructural de las redes inferidas para *Rhizobium etli*, organismo para el cual no se cuenta con un estándar de oro de su red global, mostrando la utilidad del enfoque de evaluación para evaluar la calidad global de las redes. En dicho artículo, los demás autores hicieron inferencias de la red de regulación de *Rhizobium etli*. Al comparar la estructura de las redes predichas con las redes curadas de *E*. coli y *B. subtilis* y modelos aleatorios Erdős-Rényi, se demostró que las redes predichas tenían una estructura más similar a las redes biológicas que a los modelos aleatorios.

5.5 Estudio de la conservación de la función sin anotaciones funcionales

Dado que la genómica comparativa es un aspecto importante al momento de transferir interacciones entre organismos, la elección del tipo correcto de homología potencialmente influirá en la predicción obtenida. En el anexo VII se encuentra el artículo "Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty" (Escorcia-Rodriguez, Esposito, Freyre-Gonzalez, & Moreno-Hagelsieb, 2022) donde investigamos la conjetura funcional de ortólogos y parálogos, a través de un enfoque libre de anotaciones funcionales. En su lugar, utilizamos diversas definiciones de ortología y analizamos la proporción de sustituciones no sinónimas a sinónimas (dN/dS) como indicador de la divergencia funcional.

Los resultados indican que, independientemente de la definición de ortología utilizada, los ortólogos tienden a ser más funcionalmente estables que los parálogos, con valores de dN/dS más bajos. Además, las diferencias en las tasas de dN/dS eran más evidentes a altas identidades de secuencia y sugerían que la divergencia funcional de los parálogos ocurre relativamente temprano después de la duplicación génica. En conclusión, el estudio respalda la elección de ortólogos como enfoque adecuado para la genómica comparativa debido a su mayor estabilidad funcional en comparación con los parálogos.

6 Conclusiones

Se estableció un estándar de oro para evaluar la calidad de las predicciones mediante la actualización de Abasy Atlas, permitiendo su uso como referencia en la evaluación de predicciones de redes en bacterias. La versatilidad de utilizar representaciones con y sin complejos regulatorios permite su aplicación a diferentes enfoques de inferencia, ya se partiendo de datos genómicos o transcriptómicos.

Se emplearon datos genómicos para inferir y estudiar las redes de regulación de *C. glutamicum*. Se exploraron propiedades estructurales globales y se analizó su evolución a medida que la red se vuelve más completa. También se evaluó el impacto de incluir interacciones proteína-proteína y regulaciones mediadas por ARNs pequeños en la inferencia de redes, encontrando que las redes que incluyen interacciones mediadas por ARNs pequeños tienen una estructura atípica en comparación con redes limitadas a las interacciones mediadas por factores de transcripción que se unen al ADN.

Se emplearon y evaluaron métodos computacionales que utilizan datos transcriptómicos o genómicos como punto de partida, así como la integración de ambos enfoques en la inferencia de la red de *S. coelicolor*. Se emplearon propiedades estructurales globales y métodos estándar basados en la presencia o ausencia de interacciones para evaluar la calidad de las redes predichas. Ambos enfoques de evaluación mostraron a las predicciones basadas únicamente en datos genómicos como las más aproximadas a las redes validadas experimentalmente.

La evaluación de la estructura de las predicciones con diversos métodos basados en transcriptómica desveló sesgos y malas interpretaciones en evaluaciones previas, donde métodos para inferir redes de coexpresión y regulación se evaluaron de igual manera sin considerar la diferencia de estructura obtenida dependiendo del tipo de red esperada. Se observó que las conclusiones varían entre el uso de datos sintéticos y biológicos, tanto en términos estadísticos como estructurales. Además, se identificó a los métodos que infieren coexpresión, como una mejor alternativa a los métodos que infieren causalidad cuando el objetivo es inferir regulones para reguladores locales.

Finalmente, dada la importancia de considerar la divergencia funcional en la transferencia de interacciones entre organismos, se exploró la conservación de la función a través un enfoque libre de anotaciones funcionales. La elección de ortólogos se respaldó como enfoque adecuado, ya que demostraron ser más estables funcionalmente en comparación con los parálogos.

En conjunto, este las conclusiones derivadas de este trabajo ofrecen valiosas aportaciones al campo de la inferencia de redes de regulación genética en bacterias, resaltando la relevancia de considerar la estructura de las redes, la integración de datos, y la elección adecuada de enfoques y estándares para evaluar con precisión las predicciones de las redes de regulación genética. De igual manera, se provee una estrategia para la inferencia integrativa a partir de datos genómicos, la integración de datos transcriptómicos y un nuevo enfoque para la evaluación de la predicción de redes con base en su estructura.

7 Limitaciones del estudio

La inferencia de redes de regulación a partir del genoma requiere conocimiento previo sobre la regulación transcripcional en el organismo de interés o en algún organismo, preferiblemente cercano. Esto permite explorar la regulación de la transcripción en genes que no han sido investigados anteriormente, así como caracterizar el papel regulador de factores que no han sido previamente identificados en el organismo, siempre y cuando se conozca la función reguladora de proteínas ortólogas a dicho factor de transcripción en otros organismos.

Es importante destacar que la especificidad con la que los factores de transcripción se unen al ADN varía ampliamente. Este trabajo no incluye la consideración del número esperado de genes regulados dada la especificidad del sitio de regulación, una variable que podría explorarse en investigaciones futuras para mejorar las predicciones.

En cuanto a la inferencia de redes de regulación a partir de datos transcriptómicos, se observa un bajo poder predictivo. Esto podría deberse a la limitación inherente de inferir redes causales de regulación desde datos transcriptómicos, ignorando las múltiples capas de regulación intermedias, tales como las redes de interacciones entre proteínas y las redes metabólicas, las cuales podrían ser integradas en trabajos posteriores

Al considerar la integración de genómica y transcriptómica, la integración directa de la predicción por votación de los métodos no parece ser la mejor opción, ya que los resultados pueden tener un rendimiento incluso peor que solo con datos genómicos. Además, tener en cuenta el tipo de regulador (local o global) puede favorecer la integración, buscando una mayor correlación entre genes de un regulón local que entre genes de un regulón global.

Finalmente, al evaluar las predicciones basadas en su estructura, es esencial destacar que dos redes de regulación pueden presentar una estructura idéntica con identificadores de nodos aleatorizados. Este enfoque proporciona una evaluación de la similitud global entre dos redes, pero, si el objetivo es identificar regulaciones transcripcionales específicas, se recomienda combinarla con una evaluación estadística siempre que sea posible para obtener una visión más completa y precisa. En ausencia de esta combinación, se puede determinar si una predicción es mejor de lo esperado al azar mediante la utilización de modelos aleatorios y la comparación con redes experimentales de otros organismos.

8 Referencias

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the World-Wide Web. *Nature, 401*(6749), 130-131. doi:10.1038/43601

Alkema, W. B., Lenhard, B., & Wasserman, W. W. (2004). Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res, 14*(7), 1362-1373. doi:10.1101/gr.2242604

Altenhoff, A. M., Glover, N. M., & Dessimoz, C. (2019). Inferring Orthology and Paralogy. *Methods Mol Biol, 1910*, 149-175. doi:10.1007/978-1-4939-9074-0_5

Arnold, B. J., Huang, I. T., & Hanage, W. P. (2022). Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol, 20*(4), 206-218. doi:10.1038/s41579-021-00650-4

Barabási, A.-L. s., & Pósfai, M. r. (2016). *Network science*. Cambridge, United Kingdom: Cambridge University Press.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet, 5*(2), 101-113. doi:10.1038/nrg1272

Blanchette, M., Schwikowski, B., & Tompa, M. (2002). Algorithms for phylogenetic footprinting. *J Comput Biol, 9*(2), 211-223. doi:10.1089/10665270252935421

Browning, D. F., & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nat Rev Microbiol, 2*(1), 57-65. doi:10.1038/nrmicro787

Browning, D. F., & Busby, S. J. (2016). Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol, 14*(10), 638-650. doi:10.1038/nrmicro.2016.103

Burgess, R. R., Travers, A. A., Dunn, J. J., & Bautz, E. K. (1969). Factor stimulating transcription by RNA polymerase. *Nature, 221*(5175), 43-46. doi:10.1038/221043a0

Campos, A. I., & Freyre-Gonzalez, J. A. (2019). Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. *Sci Rep, 9*(1), 3618. doi:10.1038/s41598-019-39866-z

Chen, S., & Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics, 19*(1), 232. doi:10.1186/s12859-018-2217-z

Chen, X., & Zhang, J. (2012). The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol, 8*(11), e1002784. doi:10.1371/journal.pcbi.1002784

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. doi:10.1186/s12864-019-6413-7

Cipriano, M. J., Novichkov, P. N., Kazakov, A. E., Rodionov, D. A., Arkin, A. P., Gelfand, M. S., & Dubchak, I. (2013). RegTransBase--a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics, 14*, 213. doi:10.1186/1471-2164-14-213

Conant, G. C., & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet, 9*(12), 938-950. doi:10.1038/nrg2482

Costa, L. d. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167-242. doi:10.1080/00018730601170527

D'Haeseleer, P. (2006a). How does DNA sequence motif discovery work? *Nat Biotechnol, 24*(8), 959-961. doi:10.1038/nbt0806-959

D'Haeseleer, P. (2006b). What are DNA sequence motifs? *Nat Biotechnol, 24*(4), 423-425. doi:10.1038/nbt0406-423

Dudek, C. A., & Jahn, D. (2022). PRODORIC: state-of-the-art database of prokaryotic gene regulation. *Nucleic Acids Res, 50*(D1), D295-D302. doi:10.1093/nar/gkab1110

Escorcia-Rodriguez, J. M., Esposito, M., Freyre-Gonzalez, J. A., & Moreno-Hagelsieb, G. (2022). Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty. *PeerJ, 10*, e13843. doi:10.7717/peerj.13843

Escorcia-Rodríguez, J. M., Gaytan-Nuñez, E., Hernandez-Benitez, E. M., Zorro-Aranda, A., Tello-Palencia, M. A., & Freyre-González, J. A. (2023). Improving gene regulatory network inference and assessment: The importance of using network structure. *Front Genet, 14*, 1143382. doi:10.3389/fgene.2023.1143382

Escorcia-Rodríguez, J. M., Tauch, A., & Freyre-González, J. A. (2020). Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization. *Comput Struct Biotechnol J, 18*, 1228-1237. doi:10.1016/j.csbj.2020.05.015

Escorcia-Rodríguez, J. M., Tauch, A., & Freyre-González, J. A. (2021). Corynebacterium glutamicum Regulation beyond Transcription: Organizing Principles and Reconstruction of an Extended Regulatory Network Incorporating Regulations Mediated by Small RNA and Protein-Protein Interactions. *Microorganisms, 9*(7). doi:10.3390/microorganisms9071395

Euler, L. (1741). Solutio Problemat is ad Geometriam Situs Pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae, 8*, 14.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool, 19*(2), 99-113. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/5449325

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends Genet, 16*(5), 227-231. doi:10.1016/s0168-9525(00)02005-9

Freyre-González, J. A., Alonso-Pavón, J. A., Trevino-Quintanilla, L. G., & Collado-Vides, J. (2008). Functional architecture of Escherichia coli: new insights provided by a natural decomposition approach. *Genome Biol, 9*(10), R154. doi:10.1186/gb-2008-9-10-r154

Freyre-González, J. A., Escorcia-Rodríguez, J. M., Gutiérrez-Mondragon, L. F., Martí-Vértiz, J., Torres-Franco, C. N., & Zorro-Aranda, A. (2022). System Principles Governing the Organization, Architecture, Dynamics, and Evolution of Gene Regulatory Networks. *Front Bioeng Biotechnol, 10*, 888732. doi:10.3389/fbioe.2022.888732

Freyre-González, J. A., Manjarrez-Casas, A. M., Merino, E., Martinez-Nunez, M., Perez-Rueda, E., & Gutierrez-Rios, R. M. (2013). Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis. *BMC Syst Biol, 7*, 127. doi:10.1186/1752-0509-7-127

Freyre-González, J. A., & Tauch, A. (2017). Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: Novel insights from a dataset with a high genomic coverage. *J Biotechnol, 257*, 199-210. doi:10.1016/j.jbiotec.2016.10.025

Gabaldon, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nat Rev Genet, 14*(5), 360-366. doi:10.1038/nrg3456

Galan-Vasquez, E., Luna-Olivera, B. C., Ramirez-Ibanez, M., & Martinez-Antonio, A. (2020). RegulomePA: a database of transcriptional regulatory interactions in Pseudomonas aeruginosa PAO1. *Database (Oxford), 2020*. doi:10.1093/database/baaa106

Gerlt, J. A., & Babbitt, P. C. (2000). Can sequence determine function? *Genome Biol, 1*(5), REVIEWS0005. doi:10.1186/gb-2000-1-5-reviews0005

Gottesman, S. (1984). Bacterial regulation: global regulatory networks. *Annu Rev Genet, 18*, 415-441. doi:10.1146/annurev.ge.18.120184.002215

Gottesman, S. (2019). Trouble is coming: Signaling pathways that regulate general stress responses in bacteria. *J Biol Chem, 294*(31), 11685-11700. doi:10.1074/jbc.REV119.005593

Ibarra-Arellano, M. A., Campos-Gonzalez, A. I., Trevino-Quintanilla, L. G., Tauch, A., & Freyre-Gonzalez, J. A. (2016). Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database (Oxford), 2016*. doi:10.1093/database/baw089

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol, 3*, 318-356. doi:10.1016/s0022-2836(61)80072-7

Jensen, R. A. (2001). Orthologs and paralogs - we need to get it right. *Genome Biol, 2*(8), INTERACTIONS1002. doi:10.1186/gb-2001-2-8-interactions1002

Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature, 411*(6833), 41-42. doi:10.1038/35075138

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., & Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature, 407*(6804), 651-654. doi:10.1038/35036627

Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol, 9*(10), 770-780. doi:10.1038/nrm2503

Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand, M. S., & Mirny, L. A. (2007). How gene order is influenced by the biophysics of transcription regulation. *Proc Natl Acad Sci U S A, 104*(35), 13948-13953. doi:10.1073/pnas.0700672104

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet, 39*, 309-338. doi:10.1146/annurev.genet.39.073003.114725

Larsen, S. J., Rottger, R., Schmidt, H., & Baumbach, J. (2019). E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res, 47*(1), 85-92. doi:10.1093/nar/gky1176

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., . . . Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nat Methods, 9*(8), 796-804. doi:10.1038/nmeth.2016

McGuire, A. M., Hughes, J. D., & Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res, 10*(6), 744-757. doi:10.1101/gr.10.6.744

Nehrt, N. L., Clark, W. T., Radivojac, P., & Hahn, M. W. (2011). Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol, 7*(6), e1002073. doi:10.1371/journal.pcbi.1002073

Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., . . . Rodionov, D. A. (2013). RegPrecise 3.0--a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics, 14*, 745. doi:10.1186/1471-2164-14-745

Novichkov, P. S., Rodionov, D. A., Stavrovskaya, E. D., Novichkova, E. S., Kazakov, A. E., Gelfand, M. S., . . . Dubchak, I. (2010). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res, 38*(Web Server issue), W299-307. doi:10.1093/nar/gkq531

Parise, D., Parise, M. T. D., Kataka, E., Kato, R. B., List, M., Tauch, A., . . . Baumbach, J. (2021). On the Consistency between Gene Expression and the Gene Regulatory Network of Corynebacterium glutamicum. *Netw Syst Med, 4*(1), 51-59. doi:10.1089/nsm.2020.0014

Parise, M. T. D., Parise, D., Kato, R. B., Pauling, J. K., Tauch, A., Azevedo, V. A. C., & Baumbach, J. (2020). CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks. *Sci Data, 7*(1), 142. doi:10.1038/s41597-020-0484-9

Patel, A., McGrosso, D., Hefner, Y., Campeau, A., Sastry, A. V., Maurya, S., . . . Palsson, B. O. (2023). Proteome allocation is linked to transcriptional regulation through a modularized transcriptome. *bioRxiv*. doi:10.1101/2023.02.20.529291

Pedreira, T., Elfmann, C., & Stulke, J. (2022). The current state of SubtiWiki, the database for the model organism Bacillus subtilis. *Nucleic Acids Res, 50*(D1), D875-D882. doi:10.1093/nar/gkab943

Perez-Rueda, E., Collado-Vides, J., & Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem, 28*(5-6), 341-350. doi:10.1016/j.compbiolchem.2004.09.004

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc Natl Acad Sci U S A, 72*(3), 784-788. doi:10.1073/pnas.72.3.784

Rasmussen, P. B., Holst, B., & Valentin-Hansen, P. (1996). Dual-function regulators: the cAMP receptor protein and the CytR regulator can act either to repress or to activate transcription depending on the context. *Proc Natl Acad Sci U S A, 93*(19), 10151-10155. doi:10.1073/pnas.93.19.10151

Resendis-Antonio, O., Freyre-Gonzalez, J. A., Menchaca-Mendez, R., Gutierrez-Rios, R. M., Martinez-Antonio, A., Avila-Sanchez, C., & Collado-Vides, J. (2005). Modular analysis of the transcriptional regulatory network of E. coli. *Trends Genet, 21*(1), 16-20. doi:10.1016/j.tig.2004.11.010

Robison, K., McGuire, A. M., & Church, G. M. (1998). A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. *J Mol Biol, 284*(2), 241-254. doi:10.1006/jmbi.1998.2160

Rodionov, D. A., Novichkov, P. S., Stavrovskaya, E. D., Rodionova, I. A., Li, X., Kazanov, M. D., . . . Gelfand, M. S. (2011). Comparative genomic reconstruction of transcriptional networks controlling central metabolism in the Shewanella genus. *BMC Genomics, 12 Suppl 1*(Suppl 1), S3. doi:10.1186/1471-2164-12-S1-S3

Rogozin, I. B., Managadze, D., Shabalina, S. A., & Koonin, E. V. (2014). Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biol Evol, 6*(4), 754-762. doi:10.1093/gbe/evu051

Rychel, K., Decker, K., Sastry, A. V., Phaneuf, P. V., Poudel, S., & Palsson, B. O. (2021). iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res, 49*(D1), D112-D120. doi:10.1093/nar/gkaa810

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One, 10*(3), e0118432. doi:10.1371/journal.pone.0118432

Shen, B., Coruzzi, G., & Shasha, D. (2023). EnsInfer: a simple ensemble approach to network inference outperforms any single method. *BMC Bioinformatics, 24*(1), 114. doi:10.1186/s12859-023-05231-1

Siegenthaler, C., & Gunawan, R. (2014). Assessment of network inference methods: how to cope with an underdetermined problem. *PLoS One, 9*(3), e90481. doi:10.1371/journal.pone.0090481

Sierro, N., Makita, Y., de Hoon, M., & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res, 36*(Database issue), D93-96. doi:10.1093/nar/gkm910

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res, 12*(1 Pt 2), 505-519. doi:10.1093/nar/12.1part2.505

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res, 10*(9), 2997-3011. doi:10.1093/nar/10.9.2997

Suter, D. M. (2020). Transcription Factors and DNA Play Hide and Seek. *Trends Cell Biol, 30*(6), 491-500. doi:10.1016/j.tcb.2020.03.003

Taboada-Castro, H., Gil, J., Gomez-Caudillo, L., Escorcia-Rodriguez, J. M., Freyre-Gonzalez, J. A., & Encarnacion-Guevara, S. (2022). Rhizobium etli CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network. *Front Microbiol, 13*, 947678. doi:10.3389/fmicb.2022.947678

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., Blake, J. A., & Gene Ontology, C. (2012). On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS Comput Biol, 8*(2), e1002386. doi:10.1371/journal.pcbi.1002386

Tierrafria, V. H., Rioualen, C., Salgado, H., Lara, P., Gama-Castro, S., Lally, P., . . . Collado-Vides, J. (2022). RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in Escherichia coli K-12. *Microb Genom, 8*(5). doi:10.1099/mgen.0.000833

Todeschini, A. L., Georges, A., & Veitia, R. A. (2014). Transcription factors: specific DNA binding and specific gene regulation. *Trends Genet, 30*(6), 211-219. doi:10.1016/j.tig.2014.04.002

Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform, 2*(3), lqaa078. doi:10.1093/nargab/lqaa078

Zorro-Aranda, A., Escorcia-Rodriguez, J. M., Gonzalez-Kise, J. K., & Freyre-Gonzalez, J. A. (2022). Curation, inference, and assessment of a globally reconstructed gene regulatory network for Streptomyces coelicolor. *Sci Rep, 12*(1), 2840. doi:10.1038/s41598-022-06658-x

9 Anexos

I. System Principles Governing the Organization, Architecture, Dynamics, and Evolution of Gene Regulatory Networks

Freyre-González et al. (2022)

PERSPECTIVE
published: 12 May 2022
doi: 10.3389/fbioe.2022.888732

System Principles Governing the Organization, Architecture, Dynamics, and Evolution of Gene Regulatory Networks

Julio A. Freyre-González[1]*, *Juan M. Escorcia-Rodríguez*[1], *Luis F. Gutiérrez-Mondragón*[1,2], *Jerónimo Martí-Vértiz*[1], *Camila N. Torres-Franco*[1] and *Andrea Zorro-Aranda*[1,3]

[1] *Regulatory Systems Biology Research Group, Program of Systems Biology, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Cuernavaca, México,* [2] *Undergraduate Program in Genomic Sciences, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Cuernavaca, México,* [3] *Department of Chemical Engineering, Universidad de Antioquia, Medellín, Colombia*

OPEN ACCESS

Edited by:
Jean Marie François,
Institut Biotechnologique de Toulouse (INSA), France

Reviewed by:
Grant Bowman,
University of Wyoming, United States
Hyungyu Lim,
University of California, San Diego, United States

***Correspondence:**
Julio A. Freyre-González
jfreyre@ccg.unam.mx

Specialty section:
This article was submitted to Synthetic Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 03 March 2022
Accepted: 27 April 2022
Published: 12 May 2022

Citation:
Freyre-González JA, Escorcia-Rodríguez JM, Gutiérrez-Mondragón LF, Martí-Vértiz J, Torres-Franco CN and Zorro-Aranda A (2022) System Principles Governing the Organization, Architecture, Dynamics, and Evolution of Gene Regulatory Networks. Front. Bioeng. Biotechnol. 10:888732. doi: 10.3389/fbioe.2022.888732

Synthetic biology aims to apply engineering principles for the rational, systematical design and construction of biological systems displaying functions that do not exist in nature or even building a cell from scratch. Understanding how molecular entities interconnect, work, and evolve in an organism is pivotal to this aim. Here, we summarize and discuss some historical organizing principles identified in bacterial gene regulatory networks. We propose a new layer, the concilion, which is the group of structural genes and their local regulators responsible for a single function that, organized hierarchically, coordinate a response in a way reminiscent of the deliberation and negotiation that take place in a council. We then highlight the importance that the network structure has, and discuss that the natural decomposition approach has unveiled the system-level elements shaping a common functional architecture governing bacterial regulatory networks. We discuss the incompleteness of gene regulatory networks and the need for network inference and benchmarking standardization. We point out the importance that using the network structural properties showed to improve network inference. We discuss the advances and controversies regarding the consistency between reconstructions of regulatory networks and expression data. We then discuss some perspectives on the necessity of studying regulatory networks, considering the interactions' strength distribution, the challenges to studying these interactions' strength, and the corresponding effects on network structure and dynamics. Finally, we explore the ability of evolutionary systems biology studies to provide insights into how evolution shapes functional architecture despite the high evolutionary plasticity of regulatory networks.

Keywords: gene regulatory networks, organization, functional architecture, system principles, hierarchy, consistency, incompleteness, evolution

INTRODUCTION

Synthetic biology aims to apply engineering principles for the rational, systematical design and construction of biological systems displaying functions that do not exist in nature or even building a cell from scratch (Abil and Danelon, 2020). To fulfill these ambitious goals, we not only need to understand how the various entities within a cell interact but also to identify the principles governing how the cellular systems interconnect, work, and evolve, as these are design cornerstones underpinning a successful rational design.

Whereas studying the whole set of molecular interactions across the different layers (e.g., transport, gene regulation, protein-protein interactions, metabolism, etc.) in a cell is necessary, it is not fully possible nowadays as current knowledge of the networks integrating the different layers is limited, and the integration of heterogeneous networks poses problems not yet solved. We thus focus on gene regulation as it is the key process that controls and integrates signals from all the other layers to cope with the environment.

Advances in understanding the inner workings of small regulatory circuits (i.e., network motifs) have provided good foundations to develop small synthetic circuits, but an understanding of the system principles governing the large-scale organization of complex biological networks is still elusive. However, these principles are pivotal to understanding how the organization of gene regulatory networks (GRNs) governs its possible dynamic outcomes (Ruklisa et al., 2019) and to enabling the successful integration of newly designed systems into the preexisting circuitry of molecular interactions in a chassis.

THE BASIC ORGANIZATIONAL LAYER, COUPLED GENES: THE OPERON

In 1960, Jacob *et al.* proposed the first genetic organizational level in the cell as a “unit of coordinated expression”, the operon (Jacob et al., 1960). This functional unit plays a key role in the hypothesis of the operator, explaining the polar effect occurring because of some mutations affecting the induction of enzymes needed to metabolize lactose in *Escherichia coli*. An operon comprises a set of adjacent genes that are regulated as a unit and co-transcribed into a single polycistronic mRNA (Jacob and Monod, 1961) (**Figure 1A**, top left). Genes composing an operon are usually functionally related (de Daruvar et al., 2002; Osbourn and Field, 2009) as they collaborate to attain a specific physiological function, although they commonly possess different biochemical activities. However, there are also cases of operons comprising genes without any apparent functional relation. In these cases, genes may be required in the same environmental conditions despite being involved in different pathways (Osbourn and Field, 2009), as if a special element, responsible for integrating, at the promoter level, disparate physiological responses, was possibly lurking there. While the operon solves the problem of co-regulating functionally related genes diminishing gene expression noise and ensuring more precise stoichiometry (Osbourn and Field, 2009), it has some limitations. First, some cellular processes involve too many genes. For example, anaerobic respiration in *E. coli* comprises more than 150 genes. An operon containing all these genes would encode a huge transcript whose transcription and processing, if possible, would be inefficient. Besides, these dozens of genetic products must be, not only expressed, but also precisely coordinated in time and quantity, something that an operon is unable to achieve.

COORDINATING TIMING AND STOICHIOMETRY OF UNCOUPLED GENES: THE REGULON

A single regulatory protein may affect various promoters shaping what is defined as a regulon as was defined by Maas in 1964 (Maas, 1964). This organization enables the coordination of operons that are physically scattered throughout the genome. There are two types of regulons: simple and complex. Simple regulons are the set of genes, operons, or both regulated by a specific regulatory protein (Maas, 1964), whereas complex regulons are defined as the set of genes, operons, or both regulated by the same set of (two or more) regulatory proteins (Gutierrez-Rios et al., 2003) (**Figure 1A**, top right). As genes composing an operon are usually functionally related, the same holds for the operons controlled by a simple regulon. Besides, the expression of genes composing a regulon is not strictly coordinated, thus allowing variations in quantity and timing of synthesized products. These variations depend upon the concerted action of the respective promoters for each gene or operon in the regulon and the corresponding binding sites for their regulatory proteins. While regulons solve the organizational problems posed by operons, they open a new problem. How to control a single complex function that requires the coordinated expression of different regulons?

THE POWER OF DECENTRALIZED GLOBAL COORDINATION: THE MODULON

The integration of single regulatory circuits into complex networks led Susan Gottesman to propose the existence of global regulatory proteins controlling these global networks in 1984 (Gottesman, 1984). In her seminal paper, she also provided a set of diagnostic criteria to identify this kind of regulator: 1) global regulators control a large number of genes, 2) the regulated genes are involved in more than one metabolic pathway, and 3) global regulators coordinate gene expression in response to a common need. Four years later, Iuchi and Lin defined the modulon as the set of operons, regulons, or both modulated—hence the word modulon, which has no relation

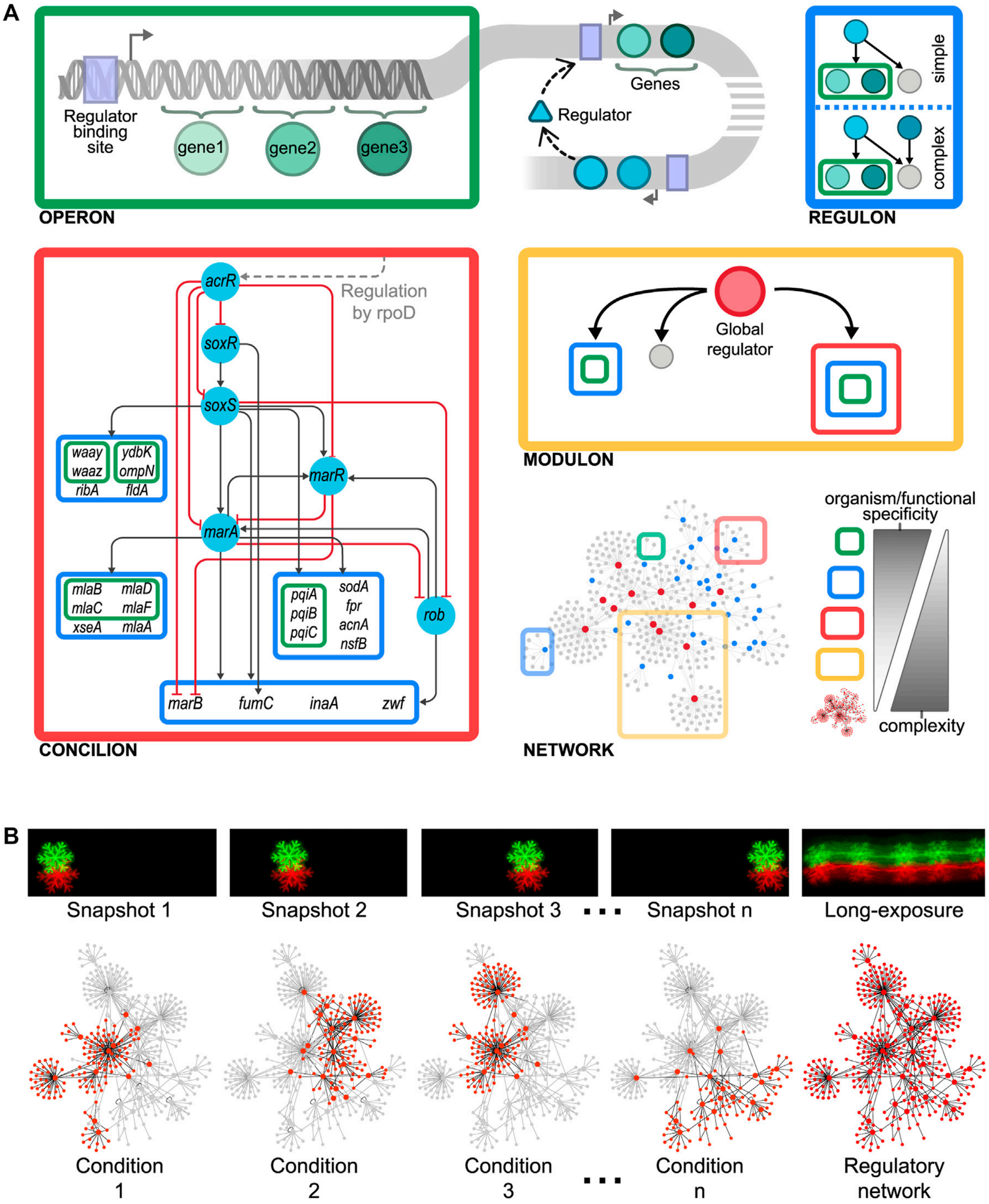


FIGURE 1 | (A) Organizational layers shaping the modular hierarchy of the gene regulatory organization as gene < operon < regulon < concilion < modulon. A biological example of the here-proposed concilion is the "response to multiple stresses" module found in *E. coli* (Escorcia-Rodriguez et al., 2020). The grey dashed line shows that *acrR* is globally controlled by *rpoD*, which also controls other concilions and regulons (**Figure 2A**). The master regulators in this hierarchy are SoxR and SoxS, which respond to oxidative stress through sensing superoxide and nitric oxide. SoxS, MarA, and Rob bind as monomers to the same DNA site, a 20-bp degenerated sequence known as Mar/Sox/Rob box. The differential regulation of these genes could be archived by the degeneracy of their DNA binding sites or by the regulators' concentration and the different affinities for the Mar/Sox/Rob box (Martin et al., 1999; Chubiz et al., 2012). The presence of several paralogous regulators (members of the AraC/XylS family) recognizing the same DNA binding site allows to archive a differential response by activating the same genes in response to different environmental cues (Martin et al., 2008). This phenomenon, known as commensurate regulon activation, enables bacteria to mount a proportionate response of the *marA/soxS/rob* regulon to the stress signal, keeping the number of activated genes to the minimum necessary to cope with prolonged stress (Martin et al., 2008; Wall et al., 2009). This balances the energetic cost of gene expression against the intensity of the stress. **(B)** Curated reconstructed regulatory networks merge many individual condition-specific subnetworks (such as picture snapshots) into a single network model thus capturing all the possible dynamic trajectories (such as a long-exposure photo does). Consequently, curated regulatory networks are not static representations of regulation, as they embed all the potential regulations that can occur thus constraining the large number of organizations a regulatory network could potentially have.

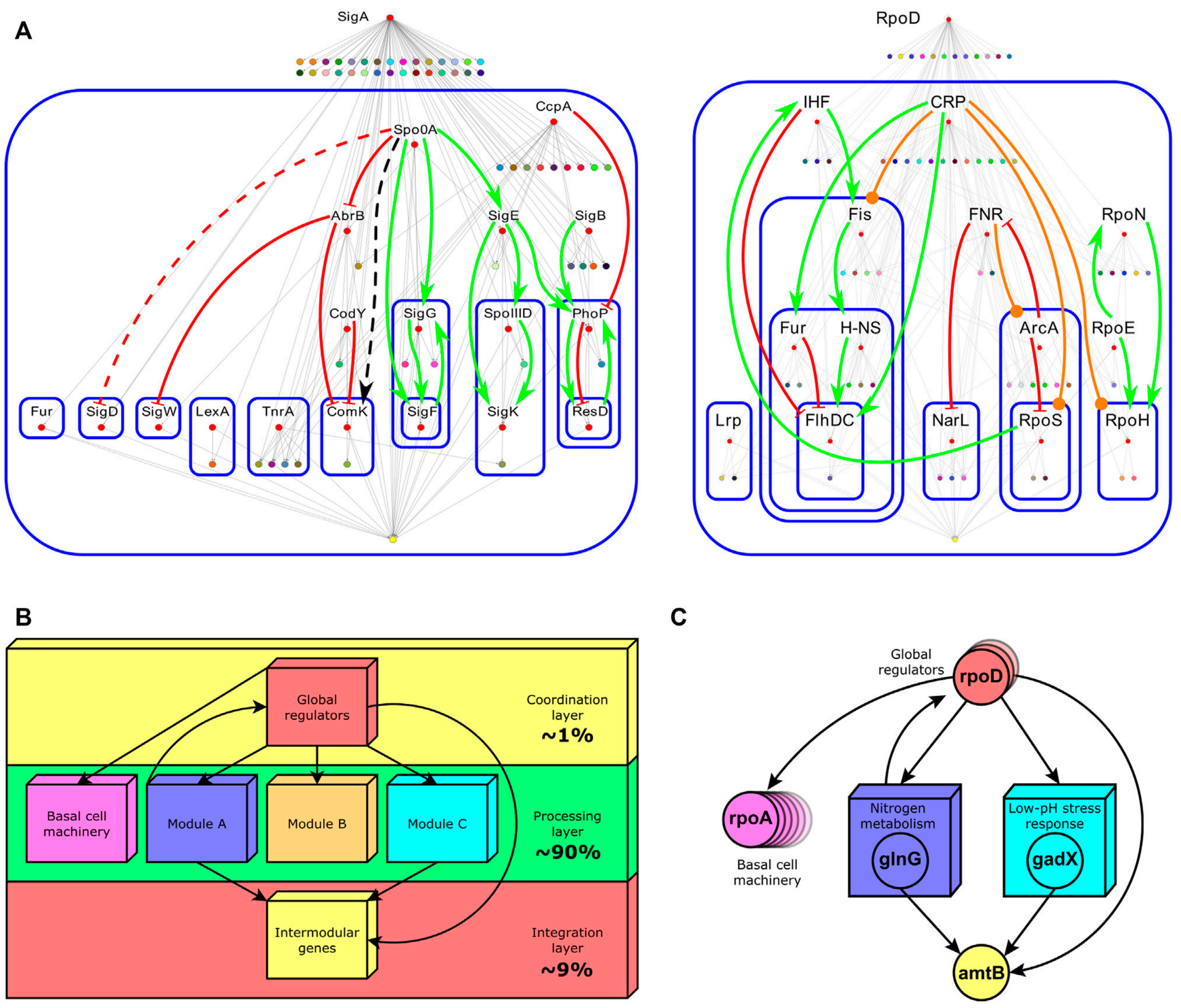


FIGURE 2 | (A) Hierarchies identified by the theoretical pleiotropy approach for *B. subtilis* (left) and *E. coli* (right). Labeled red nodes are global regulators. Nodes composing modules were shrunk into a single colored node. At the bottom of each figure, the yellow node contains the set of intermodular genes. Continuous arrows (red for negative interactions, green for positive ones, orange for duals, and black for interactions whose sign is unknown) indicate regulatory interactions between global regulators. Blue rounded-corner rectangles bound hierarchical layers. For a detailed description of this figure, the reader is referred to the original caption (Freyre-Gonzalez et al., 2012)[1]. **(B)** The common functional architecture found across bacteria by the NDA. Percentages indicate the fraction of genes in the GRN composing that layer. **(C)** A biological example of each layer composing the functional architecture of the *E. coli* GRN. The global regulator *rpoD* is one of the several global regulators controlling genes in many modules (concilions and regulons). Global regulators also control many single genes or operons not regulated by local regulators (basal machinery). Two examples of modules, ‘Nitrogen metabolism’ and ‘Low-pH stress response’, are shown. They jointly control the intermodular gene *amtB via* the local regulators *glnG* (NtrC) and *gadX* (GadX). NtrC is the general regulator of the nitrogen assimilation pathway. GadX is one of the central regulators of the glutamate-dependent acid resistance system (GAD system). The *amtB* gene encodes an NH_4^+ antiporter. Disruption of this gene impaired the growth on ammonium only under acidic conditions. Ammonium is also a precursor of glutamate, which plays a central role in the GAD system. This shows that intermodular genes integrate disparate physiological responses coming from different modules.

to the term module—by a common pleiotropic regulatory protein (Iuchi and Lin, 1988) (**Figure 1A**, center right). Here, pleiotropy implies that operons and regulons under control are no longer functionally related. Therefore, mutations in the pleiotropic regulatory protein controlling the modulon give rise to alterations in multiple phenotypic traits in a cell, confirming that global regulators are involved in disparate physiological functions. A pleiotropic or global regulator is responsible for sensing and responding to signals of general interest for the cell such as DNA[1] damage, stresses, or energy levels (Freyre-Gonzalez et al., 2008). Each global regulator shapes only one modulon and these could overlap by the co-regulation of some genes. Global regulators also shape a hierarchy comprising chains of command having each a specific physiology as has been previously reported for *E. coli* (Freyre-Gonzalez et al., 2008) and *Bacillus subtilis* (Freyre-Gonzalez et al., 2012) (**Figure 2A**). These chains of command modulate the local responses carried on, at the regulon level, by local regulators according to general interest environmental cues (e.g., low glucose, heat, high oxidizing power). Hence, modulons mostly shape a top-down hierarchy that could be seen as the global control device of the cell responsible for the coordination of lower functionally related structures. An interesting biological example of this global control device and its chains of command was outlined for the global regulator CtrA of Caulobacter crescentus (Laub et al., 2000).

THE MISSING PIECE: COORDINATING A SINGLE FUNCTION USING A HIERARCHY OF LOCAL REGULATORS, THE CONCILION

Modularity is an organizing principle in the cell (Hartwell et al., 1999). Genomic islands (e.g., pathogenicity islands, secretion islands, antimicrobial resistance islands, and metabolic islands) and compartmentalization are clear examples of this. As we previously discussed, genes are grouped into operons, regulons, and modulons. All these are kinds of modules shaping the levels of the genetic organization. Indeed, regulons have been considered by far the ultimate level of genetic organization for functionally-related genes (Gutierrez-Rios et al., 2003). However, some complex processes, involving operons, regulons, or both devoted to closely related functions, require the coordinated expression, controlled in both time and quantity, of different regulons. Besides, processing genetic and environmental information may require both 1) dividing tasks into specialized processing units and 2) integrating the resulting information. For example, an antibiotics resistance module may comprise operons or regulons each responsible for providing resistance to different antibiotics. Hence, operons and regulons must be embedded into a complex structure that cannot be reduced into a simple regulon of regulons but that still is responsible for a unique, well-defined physiological function.

We defined this novel structure, previously only loosely named module, as the concilion [kon'si.li.on]. The term is derived from the Latin noun *concilium*, council or meeting, and the verb *conciliō*, to unite, to bring together. This refers to the group of structural genes and their local regulators responsible for a single function that, organized hierarchically, coordinates a response in a way reminiscent of the deliberation and negotiation that take place in a council (**Figure 1A**, bottom left). Concilions may be differentiated from regulons because the former exhibits interactions between their regulators resembling a hierarchical circuit that could even include some feedback and cross-regulation. Moreover, concilions do not contain any global regulator, they are local regulation devices devoted to a unique, well-defined function, contrary to modulons that include a global regulator by definition and control a diversity of functions. By analyzing a non-redundant set containing the most recent GRN for each of the 42 bacteria in Abasy Atlas, we found that, on average, roughly 17% of the modules identified by the natural decomposition approach (NDA, see next section) in a GRN are concilions. Furthermore, in the most recent reconstruction of the *E. coli* GRN (Abasy Atlas regnetid: 511145_v2020_sRDB18-13), we found that about 25% of the modules are concilions whereas the remainder modules are simple or complex regulons, highlighting the important role of the concilion in the functional architecture.

A biological example of this novel genetic organizational level is provided by the "response to multiple stresses" module found in *E. coli* (Escorcia-Rodriguez et al., 2020). This concilion comprises several regulons organized into a regulatory cascade mainly controlled by SoxR, SoxS, Rob, MarR, and MarA, which shapes a hierarchy regulating 22 structural genes, many of them regulated by two or more regulators, involved in the response of *E. coli* to stress from antibiotics, organic compounds, mechanical, oxidative, and xenobiotics (https://abasy.ccg.unam.mx/module?regnetid=511145_v2020_sRDB18-13_eStrong&class=39.2). Therefore, the different organizational layers shape the modular hierarchy of the gene regulatory organization as gene < operon < regulon < concilion < modulon. As we ascend in this hierarchy network complexity increases whereas functional and organism specificity decrease (**Figure 1A**, bottom right). This introduces at least two new problems for the study of genetic organization: 1) how a concilion can be identified, and 2) how the hierarchy governing these different genetic levels can be inferred.

UNRAVELING THE COMMON FUNCTIONAL ARCHITECTURE AND SYSTEM ELEMENTS OF GLOBAL GRNS

Studying the system dynamics of large-scale regulatory networks is challenging. Using a standard differential equations model to study the evolution in time of a system having thousands of interactions renders the model prohibitively complex. Moreover, despite the large availability of genomic data, incomplete knowledge of the system also hinders this goal (e.g., the poor availability of kinetic parameters). Therefore, as system complexity increases less detail must be included in the model (Bornholdt, 2005). On the other hand, the study of the system organization is fundamental as it constrains the possible dynamic outcomes (Ruklisa et al., 2019). Traditionally, one is interested in those genes responding to a particular condition, while this is interesting to study a specific response it is just an instantaneous

[1] Reprinted from Journal of Biotechnology, 161:3, Julio A. Freyre-González, Luis G. Treviño-Quintanilla, Ilse A. Valtierra-Gutiérrez, Rosa María Gutiérrez-Ríos, José A. Alonso-Pavón, Prokaryotic regulatory systems biology: Common principles governing the functional architectures of *Bacillus subtilis* and *Escherichia coli* unveiled by the natural decomposition approach, 278-286, Copyright (2012), with permission from Elsevier.

snapshot of the system. The combinatorial nature of gene expression requires many individual condition-specific subnetworks (akin to picture snapshots) merged into a single network model to capture all the possible dynamic trajectories, in the way a long-exposure photo does (**Figure 1B**). This global network model is not a static representation of regulation, contrary to the specific-condition network. Instead, it embeds all the potential regulations that can occur, forming a regulatory landscape by constraining the large number of organizations a regulatory network could potentially have.

Curation efforts have allowed the reconstruction of updated regulatory networks for many organisms, alleviating the large-scale study of the architecture of regulatory networks. Further curation can help to improve the current reconstructions and even increase the number of organisms with an available reconstructed regulatory network. However, the massive curation of regulatory networks is limited by competitive funding with short grant cycles, which renders long-term funding, if available, uncertain, although alternative subscription-based funding models have been proposed (Reiser et al., 2016). Recently, Abasy Atlas (https://abasy.ccg.unam.mx) has gathered the largest collection of disambiguated and homogenized regulatory networks with experimentally validated interactions (Ibarra-Arellano et al., 2016). Such networks cover 42 bacteria distributed in nine species, including historical snapshots of the regulatory network reconstruction of some organisms at different stages of curation (Escorcia-Rodriguez et al., 2020). The construction of Abasy Atlas has exposed the poor knowledge we have about regulation in bacteria as only roughly 10% of the organisms in Abasy Atlas have a reconstructed regulatory network with interaction completeness > 65%. This statistic is based on a recent model developed to quantify the total number of interactions that the regulatory network of an organism will have according to its genome size (Campos and Freyre-Gonzalez, 2019; Escorcia-Rodriguez et al., 2020). This interaction completeness model is implemented and available in Abasy Atlas since version 2.2. Regulatory networks deposited in Abasy Atlas include different types of regulations (e.g., protein-DNA, small RNAs, and protein-protein interactions). Abasy Atlas also provides the system elements identified by the natural decomposition approach (NDA) that compose the functional architecture of a regulatory network.

The NDA leverages the global structure of a regulatory network to define mathematical diagnostic criteria and an algorithm to identify these system elements by the controlled decomposition of a network (Freyre-Gonzalez et al., 2008; Freyre-Gonzalez and Trevino-Quintanilla, 2010; Freyre-Gonzalez et al., 2012; Ibarra-Arellano et al., 2016; Freyre-Gonzalez and Tauch, 2017; Escorcia-Rodriguez et al., 2020; 2021). First, the κ-value is computed as the solution ($\sqrt[\alpha+1]{\alpha\gamma} \cdot k_{\mathrm{out}_{\max}}$) to the equation $\mathrm{d}C(k_{\mathrm{out}})/\mathrm{d}k_{\mathrm{out}} = -1$, where $C(k_{\mathrm{out}}) = \gamma k_{\mathrm{out}}^{-\alpha}$ is the clustering coefficient distribution of a GRN as a function of the out-connectivity (k_{out}) and is obtained by robust least-squares fitting. The global regulators are identified as those having out-connectivity > κ-value. The global regulators and their interactions are removed from the network to naturally reveal the modules (remaining connected subgraphs) and the basal machinery (disconnected nodes). Intermodular genes are identified as structural genes (nodes having zero out-connectivity ($k_{\mathrm{out}} = 0$) and therefore no coding for regulators) being controlled by different modules and then integrating disparate physiological responses. For further details on the NDA methodology, please see **Figures 1**, **2** in both (Ibarra-Arellano et al., 2016; Freyre-Gonzalez and Tauch, 2017). Sensitivity analyses have shown that the global regulators are the most robust to network incompleteness, whereas the intermodular genes are the most labile. By focusing on the modular and basal machinery genes, it has been observed that the NDA is highly robust to incompleteness in the set of interactions and more labile to incompleteness in the set of genes. This suggests that NDA predictions from GRNs having high network genomic coverage are quite reliable (Freyre-Gonzalez and Tauch, 2017). These observations have been supported by analyzing historical snapshots of the *E. coli* GRN (Escorcia-Rodriguez et al., 2020). Additionally, an assessment of the NDA predictions obtained by using three network models of the *C. glutamicum* GRN with different confidence degrees, including the addition of small RNAs, and an analysis of historical snapshots, have also confirmed these observations (Escorcia-Rodriguez et al., 2021).

All together global regulators, modules comprising modular genes, basal machinery genes, and intermodular genes compose a non-pyramidal, three-tier, diamond-like hierarchy common to all the organisms in Abasy Atlas (**Figure 2B**). The diamond-like nature of the functional architecture follows from the asymmetry in the number of genes composing each layer. The coordination layer comprises roughly 1% of the genes in the network, whereas the processing layer, composed of modular and basal machinery genes, accounts for about 90%, and the integration layer comprises roughly 9%. The global regulators (coordination layer) modulate the expression of genes belonging to the two lower layers (processing and integration), whereas some feedback could occasionally occur between the processing and coordination layers. Modules identified by the NDA can be concilions or regulons, but neither modulons nor single operons. Nevertheless, modulons globally coordinate modules. Basal machinery genes account for the cell's housekeeping functions and are controlled only by global regulators (Freyre-Gonzalez et al., 2012). Each module is responsible for a specific different function, whose combinatorial expression allows the cell to cope with a variety of environments. Remarkably, the NDA revealed that modules are locally independent meaning that there is no cross-regulation between them (Freyre-Gonzalez et al., 2012). Global regulators only coordinate the modules, and intermodular genes integrate some of their responses. Intermodular genes compose the integration layer. They were first identified by the NDA, they integrate, at the promoter level, the response of functionally disparate modules, and they thus enable the cell to cope with complex environments such as the assimilation of nitrogen under acidic conditions (**Figure 2C**) (Freyre-Gonzalez and Trevino-Quintanilla, 2010).

An alternative approach is to study expression data to elucidate the underlying network structures governing gene expression (Saelens et al., 2018). Recent works have applied independent component analysis (ICA) to transcription datasets to unravel the signals that govern gene expression in *E. coli*, *S. aureus*, and *B. subtilis* (Sastry et al., 2019; Rychel et al., 2021). The analysis produces a series of so-called iModulons (unrelated to the traditional term modulon, see above), a group of genes that are governed by a certain signal. This signal in many cases can be assigned to a certain regulator, based on biological knowledge. A gene can be included in more than one iModulon and some iModulons are assigned to more than one regulator, which is consistent with the existence of complex regulons. This analysis partially reconstructs some of the known regulons of the network and even aids in predicting new regulatory interactions.

DEALING WITH GRNS INCOMPLETENESS

From the perspective of rational synthetic biology, the top-down approach can be applied to identify disposable components in an organism using a global GRN (Lastiri-Pancardo et al., 2020). So far, not even the model organisms in gene regulation have a complete experimentally supported GRN (Escorcia-Rodriguez et al., 2020) because of the time and resource consumption needed for experimental validation and curation. Therefore, network inference is currently one of the best alternatives to reconstructing complete GRNs. However, it is a still-going challenge that, on one hand, has been approached through a plethora of transcriptomics-based strategies ranging from mechanistic models to machine learning, all of them with modest to poor results (Marbach et al., 2012). Network inference based on the identification of regulatory binding sites has performed significantly better (Zorro-Aranda et al., 2022), but it requires a prior network for its application. One way to deal with this limitation is to transfer regulatory information from one organism to another (Alkema et al., 2004). Nevertheless, this approach requires the organisms to be similar enough so the interactions are conserved (McCue et al., 2002; Escorcia-Rodriguez et al., 2021), and prior regulatory information for the source organism is still required. Inferences based on gene expression data have also benefited from the integration of biological information. For instance, the pre-selection of transcription factors (TFs) from experimental data constrains the number of potential inferences, and the application of structural properties of biological GRNs improves the assessment of the predictions (Zorro-Aranda et al., 2022). Other works have also shown improvements in the inference of regulatory networks integrating multiple omics data (Cheng et al., 2011; Banf and Rhee, 2017) and network structure (Castro et al., 2019). This suggests that the integration of biological data and network structure might be the approach to pursue in the inference of GRNs.

There is no straightforward nor standard way to infer a global regulatory network. A few precomputed inferences based on sequence or transcriptomics are scattered across the literature and organism-specific databases (Galan-Vasquez et al., 2020; Parise et al., 2020). Most of these inferences come from different approaches making it difficult to assess them. Besides, for the organisms with transcriptomic data, we need to gather and normalize the data to apply one of the top-ranking tools in previous assessments (Marbach et al., 2012). There exist databases hosting inferences of regulatory networks based on regulatory binding sites [e.g., PRECISE (Novichkov et al., 2013)]. However, these predictions have not been systematically assessed. We need to standardize the benchmarking of network inference tools with biological datasets and GRN gold standards used as reference. This way, we could keep pace with the rate of emerging methodologies. Moreover, the incompleteness of the GRN gold standards hinders proper assessment of inferred networks as actual true positive interactions are incorrectly labeled as false positive if they are not part of the current gold standard. We can leverage the constrained space for structural properties found in biological GRNs (Campos and Freyre-Gonzalez, 2019; Escorcia-Rodriguez et al., 2021) to verify if the inferred networks have similar properties.

Once we know the inferred networks behave as the biological ones, we can study their functional architecture and system-level components (Freyre-Gonzalez et al., 2012). Although the diamond-like structure has been found across all the organisms in Abasy Atlas, the system-level conservation has been quantitatively evaluated only between *E. coli* and *B. subtilis* (Freyre-Gonzalez et al., 2012), and *Corynebacterium glutamicum* and *Streptomyces coelicolor* (Zorro-Aranda et al., 2022). Future work assessing the conservation across all the available organisms and the robustness of the node classification to network incompleteness would shed light on the missing interactions for incomplete networks and their hierarchical role in the global network.

CONSISTENCY OF GRNS: CORRELATION DOES NOT IMPLY CAUSATION

The consistency between GRNs and expression data has been previously studied by assuming a causal effect between the expression of the TFs and their target genes (TGs). Recent studies using expression data in *E. coli* and *C. glutamicum* have assessed this causal effect by using correlations to show a weak correlation of the known regulatory TF-TG pairs compared to all the possible random pairs as background (Larsen et al., 2019; Parise et al., 2021). Moreover, repressor interactions were associated with a positive correlation, rather than the expected negative correlation. This apparent inconsistency between GRNs and expression data may be explained by some molecular factors that cause known TF-TG pairs not to correlate well, e.g., the time delay between the stimulus and the regulatory response or TFs not being in their allosteric active configuration (Yu et al., 2003; Maier et al., 2009; Ghazalpour et al., 2011). Thus, we should not attempt to invalidate, through correlations of high-throughput expression data, reconstructed GRNs that are the result of experiments showing the physical binding of a TF to a DNA binding site. Further, expression data might capture false positive interactions and lead to an inherently noisy reconstruction of GRNs because found interactions are based on correlations and not necessarily causal.

Instead, an alternative approach, not yet reported, is to assess consistency within expression data considering the GRN architecture and organization. The functional architecture found in bacterial regulatory networks by the NDA (Escorcia-Rodriguez et al., 2020) (**Figure 2B**) proposes a robust partitioning of the network into physiologically correlated gene clusters and specific interaction roles for each regulatory interaction. This partitioning of the network may allow finding pairs of expressed genes that are significantly co-expressed across conditions by removing the noise in the previously unstructured set of interactions using a properly structured background. As mentioned above, expression data have been analyzed using ICA yielding significant biological results and partially reconstructing known regulons (Sastry et al., 2019; Rychel et al., 2021). This would be entirely impossible if expression data were completely inconsistent with the known structure of GRNs.

INTEGRATING QUANTITATIVE INFORMATION INTO NETWORK REPRESENTATIONS OF GENE REGULATION

Weighted gene co-expression networks have been widely and successfully used to identify biologically relevant subgraphs, outperforming approaches based solely on network structure (Li et al., 2011; Niu et al., 2019; Farhadian et al., 2021). Perhaps including quantitative information in the network could aid structure-based approaches, such as the NDA, in discovering relevant modules. Optimally partitioning the network into subgraphs comprising strong interactions could also help identify sections of the network that can be modeled independently.

Research on GRNs has focused mainly on structural aspects, leaving out any quantitative information about how a certain regulator affects the expression of its targets. Although the modeling of gene expression dynamics based on Hill kinetics and differential equations becomes prohibitively complex as network size increases, simpler models could perhaps yield interesting information about how GRNs are globally organized. A first approach could be representing the network as a weighted graph, i.e., having each edge on the network assigned a certain weight that represents the strength of the interaction. The sole definition of what this strength would be (the affinity of the TF to its binding site, the TF-TG correlation, or some other measure) is itself a challenge as it is inherently related to the data used to quantify this information.

Integrating quantitative information into the network could yield valuable insights into the dynamical stability of the system as a whole or provide parameters with which to model small circuits within the network. Gene regulatory networks seem to be constrained in their density, tending towards lower values as network nodes increase, following a power law (Campos and Freyre-Gonzalez, 2019). In that study, the authors discuss that this restriction may stem from the necessity of dynamic stability, as predicted by the May-Wigner theorem (Gardner and Ashby, 1970; May, 1972). A 2018 study on the dynamics of phage λ demonstrated that, although some of its behavior can be solely explained by the structure of its network, the relative ordering of transcription factor binding site affinities determined modified behaviors of the attractors of the system (i.e., the set of the stable states the system arrives after perturbation) (Ruklisa et al., 2019). Advancing global network models from the purely qualitative to the quantitative are surely one of the ongoing challenges of biological network science, and essential to furthering our understanding of dynamic living systems.

EVOLUTION OF GRNS FROM A SYSTEM-LEVEL PERSPECTIVE

In a seminal paper in 1962, Herbert Simon proposed the idea that the evolution of a complex structure from simple elements must proceed through a hierarchy of potential stable subassemblies (Simon, 1962). In his parable of the two watchmakers, Simon argues that these hierarchical structures will evolve faster than non-hierarchical counterparts of similar size. Consequently, in the study of the evolution of complex structures such as complex biological networks, it is imperative to adopt an approach that considers how these potential stable subassemblies have played a role in their evolution. These subassemblies could be operons, regulons, concilions, or modulons in GRNs, all of them collectively referred to as systems hereafter. Therefore, for the study of the evolution of GRNs, we need a system-level approach.

Previous evolutionary studies have focused on the effect of gene duplication and horizontal gene transfer in the evolution of GRNs but without taking into account the network organization and how these mechanisms have given rise to its functional architecture (Madan Babu and Teichmann, 2003; Teichmann and Babu, 2004; Price et al., 2008). Further studies have assessed the conservation and evolution of GRNs by using networks inferred through orthology (Madan Babu et al., 2006) or biding sites prediction (Gonzalez Perez et al., 2008). All these advances have been properly summarized (Janga and Collado-Vides, 2007; Babu, 2010). Recently, some studies on eukaryotes have aimed to study how modularity evolves in developmental GRNs by using gene co-expression data (Peter and Davidson, 2011; Verd et al., 2019) or completely theoretical approaches (Espinosa-Soto and Wagner, 2010; Espinosa-Soto, 2018). An interesting study focuses on exploring the evolution of non-developmental GRNs (Defoort et al., 2018). By using genomic phylostratigraphy (Domazet-Loso et al., 2007), the authors explore the evolution at the level of small regulatory subgraphs (i.e., network motifs) of a mix of different types of GRNs in two eukaryotic organisms. Whereas this is an interesting study, the question of how evolution shapes the systems composing a GRN and its functional architecture is still an open question.

The lack of reliable methodologies to identify the system components integrating a GRN and the low completeness and standardization of GRNs have limited the study of its evolution from a systems perspective. Previous analyses have shown that the system elements proposed by the NDA are poorly conserved and that their evolution is possibly driven by evolutionary convergence (Freyre-Gonzalez et al., 2012). The recent availability of databases providing homogenized and standardized GRNs (Escorcia-Rodriguez et al., 2020), including the modules and system-level elements composing each GRN, provides the basis to explore how these systems have been shaped by evolution and whether stable subassemblies have arisen during the evolution of the currently known systems.

DISCUSSION

Without the study of the basic principles governing cell systems, it will be impossible for synthetic biology to become a true biological engineering discipline as has been defined by a European NEST (New and Emerging Science and Technology) High-Level Expert Group (European Commission, 2005; Pei et al., 2012) and repeatedly elsewhere (Serrano, 2007; Cheng and Lu, 2012; Bartley et al., 2017; Hanczyc, 2020). Even if the aim of being a true biological engineering discipline becomes elusive, the study of these fundamental principles is necessary to improve our basic understanding of biological complex systems (Schwille, 2011). All the themes presented in this paper are interconnected. Therefore, advance in one area affects the others. For example, having a model that describes the global organization of GRNs helps to delimit and guide their study in dynamics and evolution, as well as improve the understanding of their consistency with expression data. In turn, improvements in these subjects help to refine this model of the global network organization. Furthermore, all these topics together help to infer more and better GRNs incrementally improving our understanding of genomic regulation. Overall, during the last decade, some basic principles governing the still incomplete GRNs of a few organisms have been found. It is time to continue the research of these basic principles of biological complex networks to contribute to achieving rational synthetic biology.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in https://abasy.ccg.unam.mx and the article/Supplementary Materials, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JF-G Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Visualization, Writing—original draft, Writing—review and editing. JE-R, LG-M, JM-V Investigation, Visualization, Writing—original draft, Writing—review and editing. CT-F, AZ-A Investigation, Writing—original draft.

FUNDING

This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) [IN202421 to JF-G].

ACKNOWLEDGMENTS

We thank Felipe B. and Victor M. Barrera Rivera for their support with long-exposure photography used in **Figure 1B**. JE-R is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). He received fellowship 959406 from CONACYT. We apologize to all those colleagues whose research was not cited due to space constraints.

REFERENCES

Abil, Z., and Danelon, C. (2020). Roadmap to Building a Cell: An Evolutionary Approach. *Front. Bioeng. Biotechnol.* 8, 927. doi:10.3389/fbioe.2020.00927

Alkema, W. B. L., Lenhard, B., and Wasserman, W. W. (2004). Regulog Analysis: Detection of Conserved Regulatory Networks across Bacteria: Application to *Staphylococcus aureus*. *Genome Res.* 14 (7), 1362–1373. doi:10.1101/gr.2242604

Babu, M. M. (2010). Structure, Evolution and Dynamics of Transcriptional Regulatory Networks. *Biochem. Soc. Trans.* 38 (5), 1155–1178. doi:10.1042/BST0381155

Banf, M., and Rhee, S. Y. (2017). Enhancing Gene Regulatory Network Inference through Data Integration with Markov Random Fields. *Sci. Rep.* 7, 41174. doi:10.1038/srep41174

Bartley, B. A., Kim, K., Medley, J. K., and Sauro, H. M. (2017). Synthetic Biology: Engineering Living Systems from Biophysical Principles. *Biophysical J.* 112 (6), 1050–1058. doi:10.1016/j.bpj.2017.02.013

Bornholdt, S. (2005). Less Is More in Modeling Large Genetic Networks. *Science* 310 (5747), 449–451. doi:10.1126/science.1119959

Campos, A. I., and Freyre-González, J. A. (2019). Evolutionary Constraints on the Complexity of Genetic Regulatory Networks Allow Predictions of the Total Number of Genetic Interactions. *Sci. Rep.* 9 (1), 3618. doi:10.1038/s41598-019-39866-z

Castro, D. M., de Veaux, N. R., Miraldi, E. R., and Bonneau, R. (2019). Multi-study Inference of Regulatory Networks for More Accurate Models of Gene Regulation. *PLoS Comput. Biol.* 15 (1), e1006591. doi:10.1371/journal.pcbi.1006591

Cheng, A. A., and Lu, T. K. (2012). Synthetic Biology: an Emerging Engineering Discipline. *Annu. Rev. Biomed. Eng.* 14, 155–178. doi:10.1146/annurev-bioeng-071811-150118

Cheng, C., Yan, K.-K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., et al. (2011). Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data. *PLoS Comput. Biol.* 7 (11), e1002190. doi:10.1371/journal.pcbi.1002190

Chubiz, L. M., Glekas, G. D., and Rao, C. V. (2012). Transcriptional Cross Talk within Themar-Sox-robRegulon in *Escherichia coli* Is Limited to therobandmarRABOperons. *J. Bacteriol.* 194 (18), 4867–4875. doi:10.1128/JB.00680-12

de Daruvar, A., Collado-Vides, J., and Valencia, A. (2002). Analysis of the Cellular Functions of *Escherichia coli* Operons and Their Conservation in Bacillus Subtilis. *J. Mol. Evol.* 55 (2), 211–221. doi:10.1007/s00239-002-2317-1

Defoort, J., Van de Peer, Y., and Vermeirssen, V. (2018). Function, Dynamics and Evolution of Network Motif Modules in Integrated Gene Regulatory Networks of Worm and Plant. *Nucleic Acids Res.* 46 (13), 6480–6503. doi:10.1093/nar/gky468

Domazet-Lošo, T., Brajković, J., and Tautz, D. (2007). A Phylostratigraphy Approach to Uncover the Genomic History of Major Adaptations in Metazoan Lineages. *Trends Genet.* 23 (11), 533–539. doi:10.1016/j.tig.2007.08.014

Escorcia-Rodríguez, J. M., Tauch, A., and Freyre-González, J. A. (2020). Abasy Atlas v2.2: The Most Comprehensive and Up-To-Date Inventory of Meta-

Curated, Historical, Bacterial Regulatory Networks, Their Completeness and System-Level Characterization. *Comput. Struct. Biotechnol. J.* 18, 1228–1237. doi:10.1016/j.csbj.2020.05.015

Escorcia-Rodríguez, J. M., Tauch, A., and Freyre-González, J. A. (2021). Corynebacterium Glutamicum Regulation beyond Transcription: Organizing Principles and Reconstruction of an Extended Regulatory Network Incorporating Regulations Mediated by Small RNA and Protein-Protein Interactions. *Microorganisms* 9 (7), 1395. doi:10.3390/microorganisms9071395

Espinosa-Soto, C. (2018). On the Role of Sparseness in the Evolution of Modularity in Gene Regulatory Networks. *PLoS Comput. Biol.* 14 (5), e1006172. doi:10.1371/journal.pcbi.1006172

Espinosa-Soto, C., and Wagner, A. (2010). Specialization Can Drive the Evolution of Modularity. *PLoS Comput. Biol.* 6 (3), e1000719. doi:10.1371/journal.pcbi.1000719

European Commission (2005). *Synthetic Biology: Applying Engineering to Biology : Report of a NEST High-Level Expert Group*. Office for Official Publications of the European Communities. London, UK: Luxembourg.

Farhadian, M., Rafat, S. A., Panahi, B., and Mayack, C. (2021). Weighted Gene Co-expression Network Analysis Identifies Modules and Functionally Enriched Pathways in the Lactation Process. *Sci. Rep.* 11 (1), 2367. doi:10.1038/s41598-021-81888-z

Freyre-Gonzalez, J. A., Alonso-Pavon, J. A., Trevino-Quintanilla, L. G., and Collado-Vides, J. (2008). Functional Architecture of *Escherichia coli*: New Insights provided by a Natural Decomposition Approach. *Genome Biol.* 9 (10), R154. doi:10.1186/gb-2008-9-10-r154

Freyre-González, J. A., and Tauch, A. (2017). Functional Architecture and Global Properties of the Corynebacterium Glutamicum Regulatory Network: Novel Insights from a Dataset with a High Genomic Coverage. *J. Biotechnol.* 257, 199–210. doi:10.1016/j.jbiotec.2016.10.025

Freyre-Gonzalez, J. A., and Trevino-Quintanilla, L. G. (2010). Analyzing Regulatory Networks in Bacteria. *Nat. Educ.* 3 (9). 1.

Freyre-González, J. A., Treviño-Quintanilla, L. G., Valtierra-Gutiérrez, I. A., Gutiérrez-Ríos, R. M., and Alonso-Pavón, J. A. (2012). Prokaryotic Regulatory Systems Biology: Common Principles Governing the Functional Architectures of Bacillus Subtilis and *Escherichia coli* Unveiled by the Natural Decomposition Approach. *J. Biotechnol.* 161 (3), 278–286. doi:10.1016/j.jbiotec.2012.03.028

Galán-Vásquez, E., Luna-Olivera, B. C., Ramírez-Ibáñez, M., and Martínez-Antonio, A. (2020). RegulomePA: A Database of Transcriptional Regulatory Interactions in Pseudomonas aeruginosa PAO1, *Database (Oxford)*. 2020: baaa106. doi:10.1093/database/baaa106

Gardner, M. R., and Ashby, W. R. (1970). Connectance of Large Dynamic (Cybernetic) Systems: Critical Values for Stability. *Nature* 228 (5273), 784. doi:10.1038/228784a0

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., et al. (2011). Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *PLoS Genet.* 7 (6), e1001393. doi:10.1371/journal.pgen.1001393

González Pérez, A. D., González González, E., Espinosa Angarica, V., Vasconcelos, A. T. R., and Collado-Vides, J. (2008). Impact of Transcription Units Rearrangement on the Evolution of the Regulatory Network of Gamma-Proteobacteria. *BMC Genomics* 9, 128. doi:10.1186/1471-2164-9-128

Gottesman, S. (1984). Bacterial Regulation: Global Regulatory Networks. *Annu. Rev. Genet.* 18, 415–441. doi:10.1146/annurev.ge.18.120184.002215

Gutiérrez-Ríos, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R., et al. (2003). Regulatory Network of *Escherichia coli*: Consistency between Literature Knowledge and Microarray Profiles. *Genome Res.* 13 (11), 2435–2443. doi:10.1101/gr.1387003

Hanczyc, M. M. (2020). Engineering Life: A Review of Synthetic Biology. *Artif. Life* 26 (2), 260–273. doi:10.1162/artl_a_00318

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From Molecular to Modular Cell Biology. *Nature* 402 (6761 Suppl. l), C47–C52. doi:10.1038/35011540

Ibarra-Arellano, M. A., Campos-González, A. I., Treviño-Quintanilla, L. G., Tauch, A., and Freyre-González, J. A. (2016). Abasy Atlas: A Comprehensive Inventory of Systems, Global Network Properties and Systems-Level Elements across Bacteria,*Database* 2016. baw089. doi:10.1093/database/baw089

Iuchi, S., and Lin, E. C. (1988). arcA (Dye), a Global Regulatory Gene in *Escherichia coli* Mediating Repression of Enzymes in Aerobic Pathways. *Proc. Natl. Acad. Sci. U.S.A.* 85 (6), 1888–1892. doi:10.1073/pnas.85.6.1888

Jacob, F., Perrin, D., Sanchez, C., and Monod, J. (1960). Operon: a Group of Genes with the Expression Coordinated by an Operator. *C R. Hebd. Seances Acad. Sci.* 250, 1727–1729.

Jacob, F., and Monod, J. (1961). Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.* 3, 318–356. doi:10.1016/s0022-2836(61)80072-7

Janga, S. C., and Collado-Vides, J. (2007). Structure and Evolution of Gene Regulatory Networks in Microbial Genomes. *Res. Microbiol.* 158 (10), 787–794. doi:10.1016/j.resmic.2007.09.001

Larsen, S. J., Röttger, R., Schmidt, H. H. H. W., and Baumbach, J. (2019). E. Coligene Regulatory Networks Are Inconsistent with Gene Expression Data. *Nucleic Acids Res.* 47 (1), 85–92. doi:10.1093/nar/gky1176

Lastiri-Pancardo, G., Mercado-Hernández, J. S., Kim, J., Jiménez, J. I., and Utrilla, J. (2020). A Quantitative Method for Proteome Reallocation Using Minimal Regulatory Interventions. *Nat. Chem. Biol.* 16 (9), 1026–1033. doi:10.1038/s41589-020-0593-y

Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M., and Shapiro, L. (2000). Global Analysis of the Genetic Network Controlling a Bacterial Cell Cycle. *Science* 290 (5499), 2144–2148. doi:10.1126/science.290.5499.2144

Li, W., Liu, C.-C., Zhang, T., Li, H., Waterman, M. S., and Zhou, X. J. (2011). Integrative Analysis of Many Weighted Co-expression Networks Using Tensor Computation. *PLoS Comput. Biol.* 7 (6), e1001106. doi:10.1371/journal.pcbi.1001106

Maas, W. K., and Clark, A. J. (1964). Studies on the Mechanism of Repression of Arginine Biosynthesis in *Escherichia coli*. *J. Mol. Biol.* 8, 365–370. doi:10.1016/s0022-2836(64)80200-x

Madan Babu, M., Teichmann, S. A., and Aravind, L. (2006). Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks. *J. Mol. Biol.* 358 (2), 614–633. doi:10.1016/j.jmb.2006.02.019

Madan Babu, M., and Teichmann, S. A. (2003). Evolution of Transcription Factors and the Gene Regulatory Network in *Escherichia coli*. *Nucleic Acids Res.* 31 (4), 1234–1244. doi:10.1093/nar/gkg210

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and Protein in Complex Biological Samples. *FEBS Lett.* 583 (24), 3966–3973. doi:10.1016/j.febslet.2009.10.036

Marbach, D., Costello, J. C., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., et al. (2012). Wisdom of Crowds for Robust Gene Network Inference. *Nat. Methods* 9 (8), 796–804. doi:10.1038/nmeth.2016

Martin, R. G., Bartlett, E. S., Rosner, J. L., and Wall, M. E. (2008). Activation of the *Escherichia coli* marA/soxS/rob Regulon in Response to Transcriptional Activator Concentration. *J. Mol. Biol.* 380 (2), 278–284. doi:10.1016/j.jmb.2008.05.015

Martin, R. G., Gillette, W. K., Rhee, S., and Rosner, J. L. (1999). Structural Requirements for Marbox Function in Transcriptional Activation of Mar/sox/rob Regulon Promoters in *Escherichia coli*: Sequence, Orientation and Spatial Relationship to the Core Promoter. *Mol. Microbiol.* 34 (3), 431–441. doi:10.1046/j.1365-2958.1999.01599.x

May, R. M. (1972). Will a Large Complex System Be Stable? *Nature* 238 (5364), 413–414. doi:10.1038/238413a0

McCue, L. A., Thompson, W., Carmack, C. S., and Lawrence, C. E. (2002). Factors Influencing the Identification of Transcription Factor Binding Sites by Cross-Species Comparison. *Genome Res.* 12 (10), 1523–1532. doi:10.1101/gr.323602

Niu, X., Zhang, J., Zhang, L., Hou, Y., Pu, S., Chu, A., et al. (2019). Weighted Gene Co-expression Network Analysis Identifies Critical Genes in the Development of Heart Failure after Acute Myocardial Infarction. *Front. Genet.* 10, 1214. doi:10.3389/fgene.2019.01214

Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., et al. (2013). RegPrecise 3.0 - A Resource for Genome-Scale Exploration of Transcriptional Regulation in Bacteria. *BMC Genomics* 14, 745. doi:10.1186/1471-2164-14-745

Osbourn, A. E., and Field, B. (2009). Operons. *Cell. Mol. Life Sci.* 66 (23), 3755–3775. doi:10.1007/s00018-009-0114-3

Parise, D., Parise, M. T. D., Kataka, E., Kato, R. B., List, M., Tauch, A., et al. (2021). On the Consistency between Gene Expression and the Gene Regulatory

Network of Corynebacterium Glutamicum. *Netw. Syst. Med.* 4 (1), 51–59. doi:10.1089/nsm.2020.0014

Parise, M. T. D., Parise, D., Kato, R. B., Pauling, J. K., Tauch, A., Azevedo, V. A. d. C., et al. (2020). CoryneRegNet 7, the Reference Database and Analysis Platform for Corynebacterial Gene Regulatory Networks. *Sci. Data* 7 (1), 142. doi:10.1038/s41597-020-0484-9

Pei, L., Gaisser, S., and Schmidt, M. (2012). Synthetic Biology in the View of European Public Funding Organisations. *Public Underst. Sci.* 21 (2), 149–162. doi:10.1177/0963662510393624

Peter, I. S., and Davidson, E. H. (2011). Evolution of Gene Regulatory Networks Controlling Body Plan Development. *Cell* 144 (6), 970–985. doi:10.1016/j.cell.2011.02.017

Price, M. N., Dehal, P. S., and Arkin, A. P. (2008). Horizontal Gene Transfer and the Evolution of Transcriptional Regulation in *Escherichia coli*. *Genome Biol.* 9 (1), R4. doi:10.1186/gb-2008-9-1-r4

Reiser, L., Berardini, T. Z., Li, D., Muller, R., Strait, E. M., Li, Q., et al. (2016). Sustainable Funding for Biocuration: The Arabidopsis Information Resource (TAIR) as a Case Study of a Subscription-Based Funding Model,*Database* 2016. baw018. doi:10.1093/database/baw018

Ruklisa, D., Brazma, A., Cerans, K., Schlitt, T., and Viksna, J. (2019). Dynamics of Gene Regulatory Networks and Their Dependence on Network Topology and Quantitative Parameters - the Case of Phage λ. *BMC Bioinforma.* 20 (1), 296. doi:10.1186/s12859-019-2909-z

Rychel, K., Decker, K., Sastry, A. V., Phaneuf, P. V., Poudel, S., and Palsson, B. O. (2021). iModulonDB: a Knowledgebase of Microbial Transcriptional Regulation Derived from Machine Learning. *Nucleic Acids Res.* 49 (D1), D112–D120. doi:10.1093/nar/gkaa810

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A Comprehensive Evaluation of Module Detection Methods for Gene Expression Data. *Nat. Commun.* 9 (1), 1090. doi:10.1038/s41467-018-03424-4

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., et al. (2019). The *Escherichia coli* Transcriptome Mostly Consists of Independently Regulated Modules. *Nat. Commun.* 10 (1), 5536. doi:10.1038/s41467-019-13483-w

Schwille, P. (2011). Bottom-up Synthetic Biology: Engineering in a Tinkerer's World. *Science* 333 (6047), 1252–1254. doi:10.1126/science.1211701

Serrano, L. (2007). Synthetic Biology: Promises and Challenges. *Mol. Syst. Biol.* 3, 158. doi:10.1038/msb4100202

Simon, H. A. (1962). The Architecture of Complexity. *Proc. Am. Philosophical Soc.* 106 (6), 467–482.

Teichmann, S. A., and Babu, M. M. (2004). Gene Regulatory Network Growth by Duplication. *Nat. Genet.* 36 (5), 492–496. doi:10.1038/ng1340

Verd, B., Monk, N. A., and Jaeger, J. (2019). Modularity, Criticality, and Evolvability of a Developmental Gene Regulatory Network. *Elife* 8. e42832. doi:10.7554/eLife.42832

Wall, M. E., Markowitz, D. A., Rosner, J. L., and Martin, R. G. (2009). Model of Transcriptional Activation by MarA in *Escherichia coli*. *PLoS Comput. Biol.* 5 (12), e1000614. doi:10.1371/journal.pcbi.1000614

Yu, H., Luscombe, N. M., Qian, J., and Gerstein, M. (2003). Genomic Analysis of Gene Expression Relationships in Transcriptional Regulatory Networks. *Trends Genet.* 19 (8), 422–427. doi:10.1016/S0168-9525(03)00175-6

Zorro-Aranda, A., Escorcia-Rodríguez, J. M., González-Kise, J. K., and Freyre-González, J. A. (2022). Curation, Inference, and Assessment of a Globally Reconstructed Gene Regulatory Network for Streptomyces Coelicolor. *Sci. Rep.* 12 (1), 2840. doi:10.1038/s41598-022-06658-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

II. Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization

(Escorcia-Rodríguez et al., 2020)

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization

Juan M. Escorcia-Rodríguez [a], Andreas Tauch [b], Julio A. Freyre-González [a,*]

[a] Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n, Col. Chamilpa, 62210 Cuernavaca, Morelos, Mexico
[b] Centrum für Biotechnologie (CeBiTec). Universität Bielefeld, Universitätsstraße 27, 33615 Bielefeld, Germany

ARTICLE INFO

Article history:
Received 13 November 2019
Received in revised form 4 May 2020
Accepted 9 May 2020
Available online 16 May 2020

Keywords:
Systems biology
Regulatory networks
Historical snapshots
Completeness
Modules
Global regulators
Intermodular genes
Meta-curation

ABSTRACT

Some organism-specific databases about regulation in bacteria have become larger, accelerated by high-throughput methodologies, while others are no longer updated or accessible. Each database homogenize its datasets, giving rise to heterogeneity across databases. Such heterogeneity mainly encompasses different names for a gene and different network representations, generating duplicated interactions that could bias network analyses. Abasy (**A**cross-**ba**cteria **sy**stems) Atlas consolidates information from different sources into meta-curated regulatory networks in bacteria. The high-quality networks in Abasy Atlas enable cross-organisms analyses, such as benchmarking studies where gold standards are required. Nevertheless, network incompleteness still casts doubts on the conclusions of network analyses, and available sampling methods cannot reflect the curation process. To tackle this problem, the updated version of Abasy Atlas presented in this work provides historical snapshots of regulatory networks. Thus, network analyses can be performed at different completeness levels, making possible to identify potential bias and to predict future results. We leverage the recently found constraint in the complexity of regulatory networks to develop a novel model to quantify the total number of regulatory interactions as a function of the genome size. This completeness estimation is a valuable insight that may aid in the daunting task of network curation, prediction, and validation. The new version of Abasy Atlas provides 76 networks (204,282 regulatory interactions) covering 42 bacteria (64% Gram-positive and 36% Gram-negative) distributed in 9 species (*Mycobacterium tuberculosis, Bacillus subtilis, Escherichia coli, Corynebacterium glutamicum, Staphylococcus aureus, Pseudomonas aeruginosa, Streptococcus pyogenes, Streptococcus pneumoniae*, and *Streptomyces coelicolor*), containing 8459 regulons and 4335 modules.

Database URL: https://abasy.ccg.unam.mx/.

1. Background

Regulation at the gene transcription level is a fundamental process for bacteria to adapt to different media conditions and to cope with adverse environments. Transcription factors (TFs) mainly mediate this process. They are proteins capable to promote or hinder the transcription of their target genes (TGs). A TF-coding gene and its TGs conform a regulon, multiple regulons can be assembled to construct a gene regulatory network (GRN) where nodes and edges depict genes and interactions, respectively. Given the different specificity across TFs, they can contribute to organism adaptation in different levels which provides hierarchical and modular properties to GRNs in bacteria [1].

The increasing number of experimental strategies to study the transcriptional machinery [2] has allowed the community to unveil novel regulatory interactions. Despite curation efforts, many interactions remain buried in publications and are not integrated into a GRN yet. Organism-specific databases offer expertise and often are the primary resource for further research on the organism of interest. Such databases include RegulonDB [3] for *Escherichia coli*, DBTBS [4] and SubtiWiki [5] for *Bacillus subtilis*, CoryRegNet [6]

Abbreviations: GRN, Gene regulatory network; TFs, Transcription factors; TGs, Target genes; NDA, Natural decomposition approach.

[*] Corresponding author.

E-mail address: jfreyre@ccg.unam.mx (J.A. Freyre-González).

https://doi.org/10.1016/j.csbj.2020.05.015
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

for *Corynebacterium glutamicum* and MtbRegList [7] for *Mycobacterium tuberculosis*. Nonetheless, many of those databases are no longer updated or accessible [8]. Besides, the availability of multiple organism-specific databases gives rise to heterogeneity, which could bias results when cross-organisms analyses are performed. Such heterogeneity encompasses different names for the same gene and different network representations. This is even a problem for a single organism when complementary databases are integrated.

The analysis of global properties through multiple bacteria have revealed similarities among them [9–14]. Nonetheless, those studies have been limited to only a few organisms and results need to be validated with the most complete GRNs [15]. Besides, the study of the effect of network incompleteness on network structural analyses has been hindered by the limitations in databases to identify when a set of novel interactions is reported, and the experimental evidence supporting those interactions. Since no GRN curation model has been developed, works to study this phenomenon have been limited to simulate the curation process by decomposition or reconstruction of the GRNs by different random models [16,17].

Diverse databases cope with information inconsistency, such as CollecTF [18] for experimentally-validated TF binding sites in bacteria, and GSDB [19] for 3D chromosome and genome topological structures. Other resources integrating and homogenizing experimentally-validated data with computational predictions include STRING [20] for protein-protein interaction networks, SwissRegulon [21] for regulatory sites in prokaryotes and eukaryotes organisms, PRODORIC [22] for DNA binding sites for prokaryotic TFs, RegNetwork [23] for transcriptional and posttranscriptional regulatory relationships for human and mouse, and Network Portal (http://networks.systemsbiology.net/) for coregulation networks. But, poor efforts have been carried out to provide consolidated, disambiguated, homogenized high-quality GRNs on a global scale, their structural properties, system-level components, and their historical snapshots to trace their curation process.

Abasy Atlas v1.0 was originally conceived to fill this gap by making a cartography of the functional architectures of a wide range of bacteria [12]. Our database provides a comprehensive atlas of annotated functional systems (hereinafter also referred to as modules), statistical and structural network properties, and system-level elements for reconstructed and meta-curated (homogeneous and disambiguated) GRNs across 42 bacteria, including pathogenically and biotechnologically relevant organisms. Abasy Atlas is the first database in providing predictions of global regulators, basal machinery genes, members of functional modules, and intermodular genes based on the system-level elements predicted for the natural decomposition approach (NDA) in several bacteria [9,11–13]. The NDA is a biologically motivated mathematical approach leveraging the global structural properties of a GRN to derive its architecture and classify its genes into one of the four above-mentioned categories of system-level elements. Abasy Atlas was also designed to provide statistical and structural properties characterizing the GRNs, such as their associated power laws, percentage of regulators, network density and giant component size, and the number of feedforward and feedback motifs among others.

In this work, we present the expanded version of Abasy (**A**cross-**ba**cteria **sy**stems) Atlas, which consolidates information from different sources into historical snapshots of meta-curated GRNs in bacteria. Each historical snapshot represents the integrated knowledge we had about a GRN at a given time point. The new Abasy Atlas v2.2 makes possible to study the effect of network incompleteness across bacteria on diverse GRNs analyses, to identify potential bias and improvements, and to predict future results with more complete GRNs. Besides, Abasy Atlas GRNs integrates regulation mediated by regulatory proteins, small RNAs, sigma factors and regulatory complexes to better understand the biological systems [24]. This global representation of the GRNs eases their use because the organism-specific databases usually represent each network in a different file and different format, which can convolute the parsing of the network flat files and the integration of information.

While most proteins regulate gene transcription as homodimeric complexes, the regulation of gene expression can also be achieved by heteromeric complexes, whose subunits are encoded by different genes. Despite previous integrative approaches merging different level components [25–27], heterodimeric complexes have not been properly represented in most of them nor databases. One of the most common representations is to assign the regulations to each subunit, leading to a duplicated representation of the interaction in the GRNs. The new Abasy Atlas v2.2 provides a homogeneous representation for heteromeric complexes, when information is available, preserving the regulatory information and avoiding duplicated, misleading interactions.

In summary, Abasy Atlas v2.2 provides historical snapshots of reconstructed and meta-curated GRNs across bacteria, their completeness level, topological properties, and system-level components, enabling network completeness-dependent analyses for multiple organisms. Besides, the homogeneity of gene symbols, interactions confidence level, and network representation allow Abasy Atlas GRNs to be used as gold standards for benchmarking purposes, such as those to assess GRN predictions and theoretical models. In the section "Functionality", we describe studies that would be benefited from the functionality of Abasy Atlas v2.2 [28–35].

Abasy Atlas does not intend to replace organism-specific databases containing regulatory interactions with biological information such as regulatory sites. Conversely, it fills a gap by providing a consolidated version of bacterial GRNs on a global scale, their structural properties, system-level components, and their historical snapshots to trace their curation process. Abasy Atlas is cross-linked to diverse external databases providing biological, genomic, and molecular details. Cross-links to organism-specific databases included as a source for each GRN are also provided. From there, the user can further inquire about biological considerations such as binding sites annotation, TF conformation, genome annotation, and chromosomal conformation. All essential data when studying the molecular mechanisms and evolution of GRNs in bacteria. In this way, Abasy Atlas serves as an across-organisms database coping with information inconsistency and providing high-quality GRNs on a global scale.

Remarkable uses of previous versions of Abasy Atlas [12] comprise the characterization of *C. glutamicum* GRN [13], the integration of gene regulatory interactions to metabolism to identify the relevant TGs suitable for strain improvement [36], and comparative genomic analyses to characterize the transcriptome profile of *Corynebacterium pseudotuberculosis* in response to iron limitation [37]. Abasy Atlas v2.0 was used to identify evolutionary constraints on the complexity of GRNs enabling the study of three models to predict the total number of genetic interactions [14]. The latter allowed to compute an interaction coverage as a proxy of network completeness, which improves the biased network genomic coverage (fraction of the genome in the network). Abasy Atlas V2.2 could be useful to improve these works since more complete GRNs provide more information regarding transcriptional regulation in medically and biotechnologically relevant organisms such as *M. tuberculosis* and *C. glutamicum*. Also, to improve models developed with the previous version of Abasy, such as the novel network completeness model presented in the section “Estimating GRNs completeness by leveraging their constrained complexity”.

2. A primer on the natural decomposition approach: Predicting global regulators, modular genes shaping functional systems, basal machinery genes, and intermodular genes

Abasy Atlas was designed to provide annotations of the modules and system-level elements integrating each GRN. These predictions are computed by using the NDA. The NDA is a large-scale modeling approach characterizing the circuit wiring and its global architecture. It defines a mathematical-biological framework providing criteria to identify the four classes of system-level elements shaping GRNs: global regulators, modular genes shaping functional systems, basal machinery genes, and intermodular genes. Studies have shown that regulatory networks are highly plastic [38]. Despite this plasticity, by applying the NDA our group has found that there are organizational principles conserved by convergent evolution in the GRNs of phylogenetically distant bacteria [11]. The high predictive power of the NDA has been proven in previous studies by applying it to the phylogenetically distant *E. coli* [9], *B. subtilis* [11], and *C. glutamicum* [13], and by comparing it with other methods to identify modules [39].

The NDA defines objective criteria (e.g., the κ-value to identify global regulators) to expose functional systems and system-level elements in a GRN, and rules to reveal its functional architecture by controlled decomposition (Supplementary Fig. 1). It is based on two biological premises [10,11]: (1) a module is a set of genes cooperating to carry out a particular physiological function, thus conferring different phenotypic traits to the cell. (2) Given the pleiotropic effect of global regulators, they must not belong to modules but rather coordinate them in response to general-interest environmental cues.

According to the NDA, every gene in a GRN is predicted to belong to one out of four possible classes of system-level elements, which interrelate in a non-pyramidal, three-tier, hierarchy shaping the functional architecture [10–13] as follows (Supplementary Fig. 2): (1) Global regulators are responsible for coordinating both the (2) basal cell machinery, composed of strictly globally regulated genes and (3) locally autonomous modules (shaped by modular genes), whereas (4) intermodular genes integrate, at the promoter level, physiologically disparate module responses eliciting combinatorial processing of environmental cues.

3. Construction and content

3.1. Abasy Atlas current content

Abasy Atlas v2.2 provides the most complete set of experimentally curated GRNs across bacteria. Abasy Atlas represents regulatory interactions by using network models where nodes represent genes or regulatory protein complexes, and directed links depict regulatory interactions. Since the release of Abasy Atlas v1.0 in 2016 [12], the number of GRNs has increased from 50 to 76 (+52%) covering 42 bacteria (64% Gram-positive and 36% Gram-negative) distributed in 9 species (*Mycobacterium tuberculosis*, *Bacillus subtilis*, *Escherichia coli*, *Corynebacterium glutamicum*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, *Streptococcus pyogenes*, *Streptococcus pneumoniae*, and *Streptomyces coelicolor*) and 41 strains (Fig. 1A and Supplementary Fig. 3).

These 76 GRNs comprise 204282 regulatory interactions (+160%) organized into 8459 (+128%) regulons and 4335 modules (+144%). We homogenized the representation of heteromeric TFs and their subunits and obtained a total of 12 heteromeric TFs, all of them in the GRN of *E. coli* K-12. However, this paves the way for a homogeneous representation of GRNs that will be propagated to more organisms in a future version of Abasy Atlas, when information regarding heteromeric TFs for these organisms is available. A total of 20 historical snapshots for the model organisms *M. tuberculosis*, *B. subtilis*, *E. coli*, and *C. glutamicum* were also included in the Abasy Atlas v2.2.

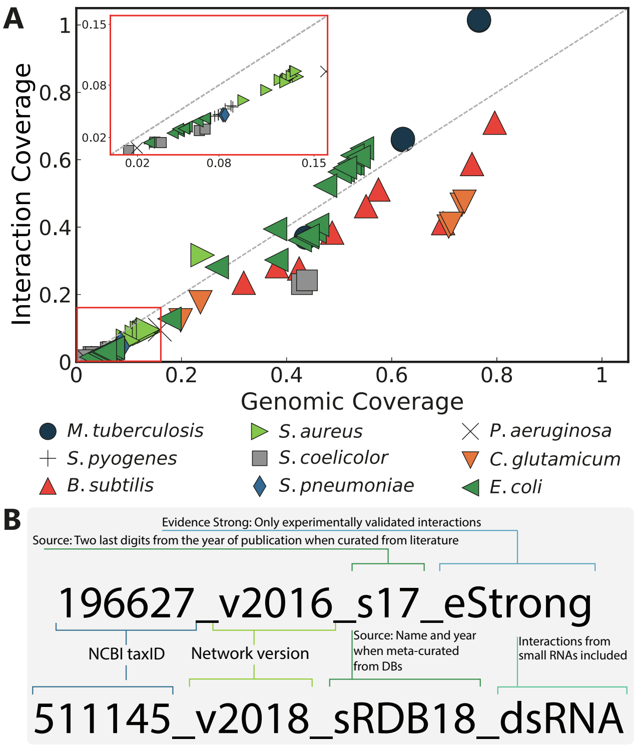


Fig. 1. Abasy Atlas content. (A) Completeness measured as genomic and interaction coverage for the GRNs in Abasy, 76 networks covering 42 bacteria distributed in 9 species. (B) Examples describing the format of the Abasy identifiers. The most complete *C. glutamicum* GRN (upper) filtered to contain only “strong” interactions, and the most recent, meta-curated *E. coli* GRN (lower).

3.2. Unique machine-readable, user-friendly identifiers for each GRN reconstruction

Studies using GRNs from organism-specific databases usually cite the source database. However, while some articles specify the GRNs used [28,39], others do not [9,40]. This drives to a reproducibility problem when the database updates the GRN and does not provide the historical snapshots. To cope with this problem, a machine-readable and user-friendly identifier was assigned to each network to ease reporting and identification when using the database.

Network identifiers are constructed as follows: Five fields are separated by an underscore, three are mandatory and two are optional. The first field represents the NCBI taxonomy ID of the organism (mandatory). The second field, preceded by a “v”, which stands for version, is the year when the network was reconstructed (mandatory). The field starting with an “s” provides information about the sources from which the network was reconstructed (mandatory). The confidence level of the evidence supporting the regulatory interactions is described by an optional field starting with an “e”. When this field is omitted means that the reconstruction contains all the available interactions disregarding the confidence level of evidence, whereas “strong” is used for those GRNs reconstructed only with interactions validated by direct experimental evidence. An optional description field, preceded by a “d”, enables to include keywords such as ‘‘sRNA” for GRNs containing sRNAs-controlled regulons (Fig. 1B).

The source field, that starting with an ‘‘s”, is composed by a database name abbreviation and year when meta-curated from databases, and the last two digits of the publication year when curated from literature (see Supplementary Table 1 for a complete list of data sources abbreviations and references). On the ‘‘Browse” page of Abasy Atlas, the user can identify the source for each GRN, as well as for the subnetworks when the GRN is a meta-curation from different sources.

3.3. Historical snapshots of the GRNs

Network theory-based approaches to study the organizing principles governing GRNs have been pointed to be biased by the curation process and incompleteness [16,41]. Nevertheless, those studies have been mainly applied to subnetworks sampled by different random computational algorithms that cannot reproduce faithfully the curation process by the scientific community. To bring an alternative solution to this problem, we have been curating organism-specific databases and literature during the construction of Abasy Atlas in different time points for several organisms (hereinafter referred to as historical snapshots). Namely, nine historical snapshots for *E. coli*, four for *C. glutamicum*, four for *B. subtilis*, and three for *M. tuberculosis* (Fig. 2).

Each historical snapshot represented in Fig. 2 is the most complete version of the GRNs at that time point. However, individual GRNs are also available. For example, the historical snapshot of the GRN of *B. subtilis* in 2017 (224308_v2017_sDBTBS08-15-SW18, Fig. 2) integrates regulatory interactions from two organism-specific databases (DBTBS [42] and SubtiWiki [5]) and one article [43] (Fig. 3). The individual GRNs are available with their own network ID (224308_v2008_sDBTBS08_eStrong, 224308_v2015_s15, and 224308_v2017_sSW18, respectively). Note that the GRN from DBTBS is also the first historical snapshot for *B. subtilis* (Fig. 2), and GRNs from different sources do not need to be from the same year since a new historical snapshot integrates every previous GRNs. The network integration and homogenization from different sources enables cross-bacteria analyses with the historical snapshots.

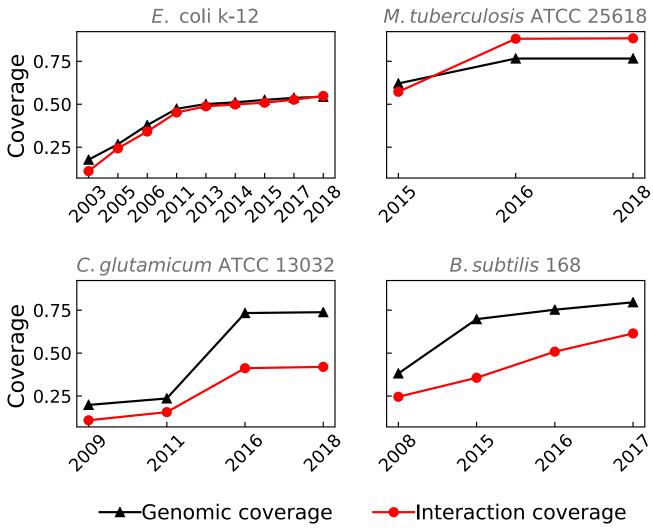


Fig. 2. Historical snapshots for GRNs of model organisms. The completeness of the network can be measured as genomic coverage (fraction of the genome included in the GRN, black triangles) and interaction coverage (fraction of the known interactions relative to the complete network, red circles). It is evident that for some networks genomic coverage overestimates completeness as some networks may be classified as almost completed in terms of genomic coverage whereas many interactions are still missing. For instance, the GRN for *C. glutamicum* in 2016 is a meta-curation of the network from 2011 and a set of interactions curated in [13] including the *sigA* housekeeping sigmulon. On the other hand, the GRN for *M. tuberculosis* in 2016 is the most complete in terms of interaction coverage (97.7%) since it integrates the network from 2015 with novel interactions curated from the literature. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

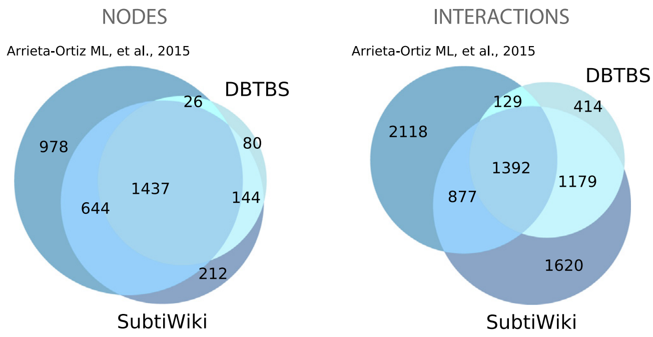


Fig. 3. Complementary sources to reconstruct the meta-curated GRN for *B. subtilis*. A poor overlap is observed between the different sources used to reconstruct the meta-curated GRN for *B. subtilis*, mainly for interactions. This highlights the need for the meta-curation since the organism-specific databases do not fully cover each other nor the dataset not previously hosted in any database. Abasy provides homogeneous meta-curations integrating all the available information.

We will continue querying organism-specific databases and curating literature periodically to obtain more complete versions of each GRNs. Also, we will extend the historical snapshots to other organisms as information will be available.

3.4. Meta-curation of GRNs: Quality control coping with inconsistency and preserving information from the different sources

The heterogeneity in gene symbols and network representations often conduces to redundancy and loss of information. Consequently, this heterogeneity can result in misleading network reconstructions. The meta-curation process mainly consists of homogenizing gene symbols and network representation before merging interactions from different sources. To cope with gene symbols disagreement among regulatory datasets from different sources, we gathered gene name, locus tag, and synonyms for each gene in the GRNs. Then, we developed an algorithm to map gene symbols onto unambiguous canonical gene names and locus tags. This allowed us to remove a total of 223 redundant nodes and 412 redundant interactions from the current set of GRNs (Supplementary Fig. 4). We refer the reader to version 1.0 of Abasy Atlas for further information about the gene symbols disambiguation algorithm [12]. For the graphical network representation, we use the unambiguous canonical gene name when available or locus tag. This eases to identify genes of interest. However, the mapping of gene identifiers allows the user to use the search box with different gene symbols and synonyms mapping to the same gene and navigate through the neighborhood of the gene of interest.

Abasy Atlas also provides the confidence level supporting each interaction since GRNs composed with different confidence-levels may bias their structural properties [14]. Therefore, a ‘‘strong” or ‘‘weak” confidence level is assigned to each interaction according to an expanded scheme based on the one proposed by RegulonDB [44,45]. The basic idea of the confidence level scheme is to label as ‘‘strong” only those interactions with direct, non-ambiguous experimental support such as DNA binding of purified TF [45]. Besides, the meta-curated networks that merge regulons from different sources also integrate the effect and the evidence level. This makes the GRNs from Abasy Atlas the most complete collection of homogenous versions in contrast to those individual GRNs available in organism-specific databases.

One of the main caveats of consolidating networks is the non-machine readable, heterogeneous way to represent the information about the way a TF regulates a specific TG and the evidence supporting such interaction, mainly for community-updated databases. To tackle this problem, we manually curate those attributes from different sources when available. Thus, Abasy Atlas makes possible to know in a homogenous fashion whether a TF promotes or hinders its TGs transcription even for interactions from a community-updated database such as SubtiWiki. Therefore, if the same interaction from a different source share effect but diverge on evidence, the interaction and the "strong" evidence is conserved since one directly experimentally validated interaction is enough to classify the edge as "strong" [45]. On the other hand, in case of different effects and the same evidence level, both effects are conserved in a single dual interaction to avoid redundancy. In the case that both attributes are different, only the "strong" interaction is conserved (Supplementary Fig. 5). This meta-curation process allows us to reconstruct the most complete GRNs available preserving information from the different complementary sources (Fig. 3).

3.5. Meta-curation of GRNs: Quality control filtering spurious interactions by reassessing the confidence level of each interaction

We perform a meta-curation process to reduce the number of spurious interactions, thereby reassessing the confidence level of the interactions. Although networks with "weak" evidence are a valuable resource to study the transcriptional regulation, only directly experimentally validated interactions offer the reliability needed to use GRNs as gold standards. Abasy Atlas eases the selection of gold standards for benchmarking purposes through ready-to-download filtered "strong" GRNs (Supplementary Fig. 6).

Using the historical snapshots of the *E. coli* GRNs, we analyzed how often a regulatory interaction identified by a "weak" methodology was validated as "strong" evidence. We found that the number of interactions identified for each methodology varies in a wide range, as well as its fraction of predictions validated as "strong" (Fig. 4A). Namely, "inferred computationally without human oversight" (ICWHO) is the evidence with the lowest fraction of validated interactions (Fig. 4A and Supplementary Fig. 7). On the other hand, "RNA-polymerase footprinting" (RPF) is the only methodology having a 100% of interactions validated as "strong" evidence, and >50% of "gene expression analysis" (GEA) predictions have been validated despite being the "weak" evidence with the highest number of predictions.

We further analyzed the effect of the interactions with ICWHO as its unique evidence, and found that most of these interactions were present in the 2013 and 2014 time points but no longer in 2015 or later. Being this the reason for the outstanding completeness of these network reconstructions and its unusual system-level elements proportions (Fig. 4B). For this reason, we decided to exclude predictions being supported only by the ICWHO evidence in Abasy Atlas. This analysis highlights the capability of the system-level properties to assess GRNs quality. It is important to note that despite the small fraction of validated interactions inferred by "non-traceable author statement" (NTAS) (Supplementary Fig. 7), we did not remove interactions supported only by this evidence since the number of predicted interactions is very small (Fig. 4A).

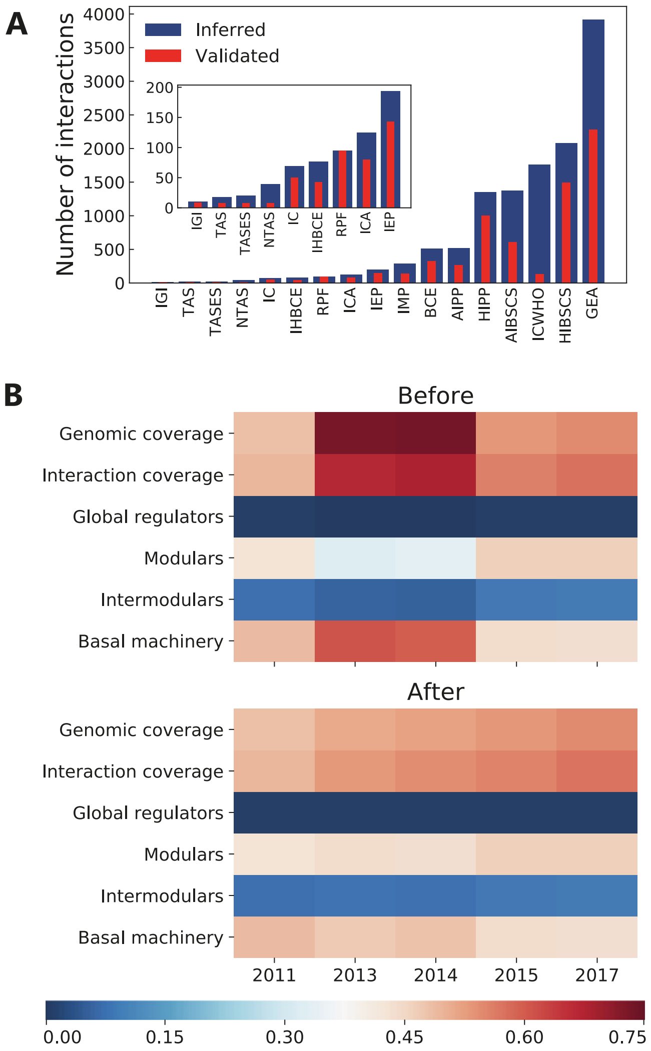


Fig. 4. (A) Number of interactions identified by methods described as "weak" in [3] and how many of these interactions have been validated by "strong" evidence. IGI (inferred from genetic interaction), TAS (traceable author statement), TASES (traceable author statement to experimental support), NTAS (non-traceable author statement), IC (inferred by curator), IHBCE (inferred by a human based on computational evidence), RFP (RNA-polymerase footprinting), ICA (inferred by computational analysis), IEP (inferred from expression pattern), IMP (inferred from mutant phenotype), BCE (binding of cellular extracts), AIPP (automated inference of promoter position), HIPP (human inference of promoter position), AIBSCS (automated inference based on similarity to consensus sequences), ICWHO (inferred computationally without human oversight), HIBSCS (human inference based on similarity to consensus sequences), GEA (gene expression analysis) [59]. (B) Effect of removing spurious interactions through the meta-curation process. System-level elements (global regulators, modular, intermodular, and basal-machinery genes) values represent its fraction from the total genes in the *E. coli* GRN historical snapshots before and after removal of interactions supported only by the ICWHO evidence.

3.6. Estimating GRNs completeness by leveraging their constrained complexity

The ability to quantify the total number of interactions in the complete GRN of an organism is a valuable insight that will leverage the daunting task of curation, prediction, and validation by enabling the inclusion of prior information about the network structure. Besides, the ability to track the completeness, quantified as the fraction of the known interactions from the total number in the complete network (interaction coverage), through different historical snapshots could allow to develop models on how new regulatory interactions are discovered and to provide a framework to assess network analysis and network inference tools. But, poor efforts have been directed towards the longstanding problem of how to assess the completeness of these networks. Traditionally, network genomic coverage has been used as a proxy of completeness. The genomic coverage of a regulatory network is the fraction of genes in the network relative to the genome size. Nevertheless, this measure poses potential biases as it neglects regulatory redundancy and the combinatorial nature of gene regulation, thus potentially overestimating network completeness.

For example, the addition of a global regulon or sigmulon (perhaps discovered by high-throughput methodologies) to a quite incomplete regulatory network could bias the genomic coverage. Assume you have a regulatory network with a genomic coverage of 15% (600/4000) and 700 interactions. You then found a paper reporting the promoter mapping for the corresponding housekeeping sigma factor, whose sigmulon has 3000 genes (400 of which were already in the original network). Next, you found that 100 out of the 3000 interactions in the global sigmulon already exist in our original network. You then integrate all the remaining 2900 new interactions to your original network to found that your resulting network has a new genomic coverage of 80% (3200/4000) and 3600 interactions. This new high genomic coverage may suggest a highly complete network but it is indeed the same quite incomplete original network plus a single global sigmulon. To clarify this, assume that the total number of interactions in the complete network is 10000, then the completeness of this new network is 36% (3600/10000). Whereas the curation of a single housekeeping sigmulon increased the completeness ~30% (3600/10000 – 700/10000), the new completeness is still low, and the genomic coverage is highly overestimating when is used as a proxy for completeness. Therefore, to state the completeness of a regulatory network correctly, it is fundamental to estimate the total number of interactions. Two recent works have simultaneously provided estimations on the size of GRNs [14,46].

On one hand, the RegulonDB team carried out an exploratory analysis [46]. They used a single version of the *E. coli* regulatory network and high-throughput datasets of binding experiments for around 15 TFs. By assuming a linear model, they found an upper-bound estimate of 45759 regulatory interactions. They claimed that only one-third of the ~46000 would affect gene expression, concluding that the complete network comprises only around 13000 interactions.

Alternatively, our group recently explored the constraints on several structural properties of the 71 regulatory networks deposited in Abasy Atlas v2.0 [14]. We found that the network density (d) as a function of the number of genes (n) follows a power law as $d \sim n^{-\gamma}$ with $\gamma \approx 1$. Since 1972, a seminal paper by Robert May showed that the frontier between dynamical stability and instability for a complex system follows a power law as $d \sim n^{-1}$, relating complexity quantified via the density of interactions and the number of variables (the size of the system) [47]. The density of interactions (network density) is the fraction of potential interactions that are real interactions, thus a constraint in network density implies a constraint in the total number of interactions in the complete network. As we found that density is constrained in GRNs, we explored three possible models to predict the total number of interactions as a function of the number of genes (see Fig. 4 in [14]): edge regression (assuming linearity, R^2 = 0.90), density invariance (assuming an invariant density, R^2 = 0.86) and density proportionality (assuming an exponential decay, R^2 = 0.91). All the models had a good fit to the data ($0.86 \leq R^2 \leq 0.91$), with small differences between them. These models predicted that the total number of interactions in the complete *E. coli* regulatory network is ~10000, ~14000, and ~11000, respectively.

After publication, we reformulated the problem. As regulatory networks are directed and self-regulations are allowed, the maximum number of possible interactions ($I_{\max}$) is n^2 as each of the n genes could regulate to other n genes including itself (self-regulation). The density of a regulatory network must be then computed as

$$d = \frac{I}{I_{\max}} = \frac{I}{n^2}$$

By introducing this equation into the power law found for the density of the Abasy Atlas networks ($d \sim n^{-\gamma}$), we derived another power law modeling the total number of interactions in the regulatory network as a function of the number of genes as

$$I = dn^2 \sim n^{-\gamma}n^2 \sim n^{2-\gamma}$$

This model has a better fit to data (Fig. 5, R^2 = 0.98) than the previous three models, and allows us to compute the total number of interactions in the regulatory network of an organism as $I_{\text{total}} \sim (\text{genome size})^{2-\gamma}$. We implemented this model in Abasy Atlas v2.2 to provide estimations on the completeness of each regulatory network, including confidence intervals. The power-law model predicts that the complete *E. coli* regulatory network will have 11656 total regulatory interactions. This model can learn the tendency in the number of interactions, and it improves as more regulatory networks are included in Abasy Atlas. That is one of the reasons motivating us to continue expanding Abasy Atlas by adding new organisms and historical snapshots.

3.7. Homogeneous representation for heteromeric transcription factor complexes

Even though heterodimeric regulatory complexes are not overrepresented in regulatory networks, some of them are global regulators and their interactions control up to ~10% of the genome and represent a valuable percent of the whole network (~6% in *E. coli* GRNs). IHF is a global regulator histone-like protein of *E. coli* that regulates transcription as a heterodimeric complex that is shaped by two different proteins: IhfA and IhfB. Although both subunits can form homodimeric complexes, the affinity for DNA is much lower [48], and no regulation in such fashion has been reported. For this reason, assigning the regulatory activity to each subunit (a gene-gene representation, Fig. 6B) is a misleading representation. Additionally, the RpoS sigma factor allows the transcription of both subunits conforming IHF, which in turn also regulates its subunits (Fig. 6A). Such interesting autoregulation cannot be properly represented in a gene-gene based representation (Fig. 6B). Conversely, a representation of the IHF heteromeric complex regulating *ihfA* and *ihfB* is better as it depicts the IHF conformation and links them to the TFs regulating their transcription.

This representation is also useful for subunits of heteromeric regulatory complexes that can exhibit regulation in a homodimeric fashion, such as the *relB* product regulating *relE*, *hokD*, and its transcription both as a homodimer and as part of the RelBE complex with *relE* (Fig. 6C). This RelE-RelB toxin-antitoxin system in *E. coli* [49] is not properly represented in a gene-gene network (Fig. 6D) as it shows regulatory activity by the *relE* product on its own. This representation eases the application of the networks as gold standards for inference methods such as those based on the DNA sequence and TF binding sites prediction. For analysis requiring GRNs composed only by genes, Abasy Atlas provides the required information to identify the classification of each biological entity (Supplementary Fig. 8). Currently, Abasy Atlas comprises 12 heteromeric TFs, all of them in the meta-curated GRN of *E coli* K-12 obtained from RegulonDB [46]. Future development includes the addition of heteromeric TFs in those organisms where this information is available.

4. Updates for model organisms

4.1. Corynebacterium glutamicum ATCC 13,032

The PubMed database was screened to find papers published between January 2017 and August 2018 and describing new tran-

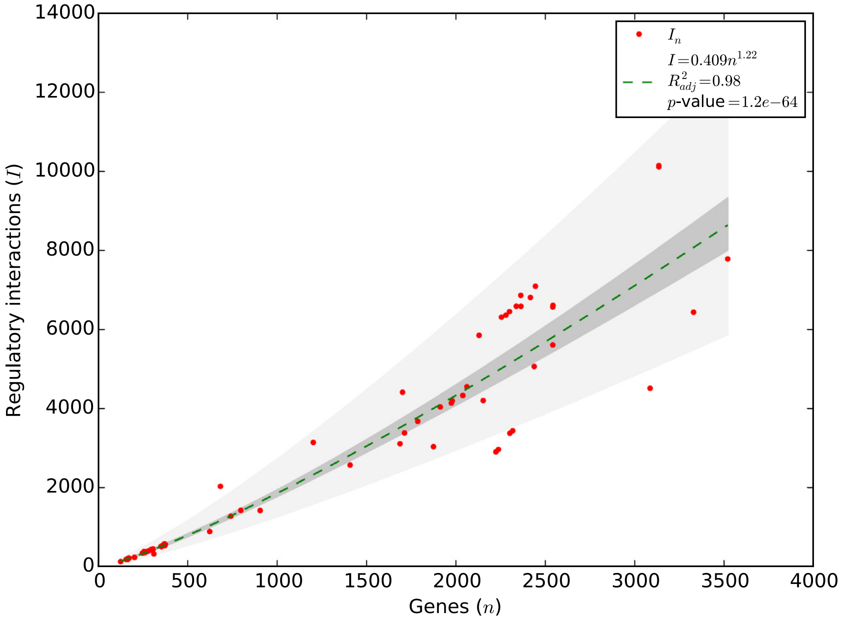


Fig. 5. The constrained complexity of regulatory networks allows computing their total number of interactions. The number of interactions in the Abasy GRNs follows a power law with the number of genes as $I \sim n^{2-\gamma}$ ($R^2 = 0.98$), where γ is the exponent of the power law found for the density of these networks. This power law may be used to compute the total number of interactions (I_{total}) in the regulatory network of an organism as $I_{\text{total}} \sim (\text{genome size})^{2-\gamma}$.

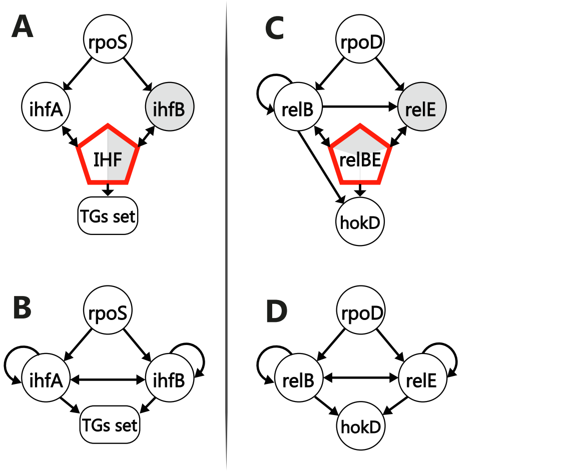


Fig. 6. Homogeneous network representation. Heteromeric-complex-base gene representation for IHF (A) and RelBE (C). Misrepresentation of gene expression regulation where heteromeric protein complexes are involved for IHF (B) and RelBE (D) systems. RelB can regulate itself as a homomeric-complex, and as a heteromeric-complex with *relE* (C). Besides, *relE* can regulate neither its transcription nor RelB transcription on its own, as could be misinterpreted from (D). This same misrepresentation is observed for the IHF complex where neither of the subunits has regulatory activity as a homomeric complex.

scriptional regulatory interactions of *C. glutamicum*, in addition to the comprehensive data set previously deposited in Abasy Atlas [13]. Four new regulators of different types have been examined in detail, exerting in total 63 new direct transcriptional interactions. Moreover, the predicted regulatory role of the AraC/XylR-type protein Cg2965 (PheR) has been confirmed by experimental data [50,51]. PheR activates the expression of the *phe* gene (*cg2966*) encoding phenol hydroxylase, allowing *C. glutamicum* to degrade phenol by a meta-cleavage pathway. Electrophoretic mobility shift assays (EMSAs) demonstrated a direct interaction of the purified PheR protein with the *phe* promoter region [51]. The MarR-type regulator CrtR (Cg0725) is encoded upstream and in divergent orientation of the carotenoid biosynthesis operon *crtEcg0722crtBIYEb* in *C. glutamicum*. DNA microarray experiments revealed that CrtR acts as a repressor of the *crt* operon. Additional EMSAs with purified CrtR showed that CrtR binds to a region overlapping the −10 and −35 promoter sequences of the *crt* operon [52].

The two-component system EsrSR (Cg0707/Cg0709) controls a regulon involved in the cell envelope stress response of *C. glutamicum* [53]. Interestingly, the integral membrane protein EsrI (Cg0706) acts as an inhibitor of EsrSR under non-stress conditions. The resulting three-component system EsrISR directly regulates a broad set of genes, including the *esrI-esrSR* locus itself, and genes encoding heat shock proteins (*clpB*, *dnaK*, *grpE*, *dnaJ*), ABC transporters and putative membrane-associated or secreted proteins of unknown function. Among the target genes of EsrSR is moreover *rosR* (*cg1324*) encoding a hydrogen peroxide-sensitive transcriptional regulator of the MarR family and playing a role in the oxidative stress response of *C. glutamicum* [53,54].

The extracytoplasmic function sigma factor SigD (Cg0696) is a key regulator of mycolate biosynthesis genes in *C. glutamicum* [55]. Chromatin immunoprecipitation coupled with DNA microarray (ChIP-chip) analysis detected SigD-binding regions in the genome sequence, thus establishing a consensus promoter sequence for this sigma factor. The conserved DNA sequence motif 5′-GTAAC-$N_{17(16)}$-CGAT-3′ was found in all ChIP-chip peak regions and presumably corresponds to the −35 and −10 promoter regions recognized by SigD. The *rsdA* (*cg0697*) gene, located immediately downstream of *sigD*, is under direct control of a SigD-dependent promoter and encodes the corresponding SigD anti-sigma factor [55].

The WhcD protein (Cg0850) interacts with WhiA (Cg1792) to exert jointly an important regulatory effect on cell division genes of *C. glutamicum* [56]. WhiA is an exceptional transcriptional regulator as it has been classified as a distant homolog of homing endonucleases that retained only DNA binding activity [57]. Binding of the WhcD-WhiA complex to the promoter region of the cell division gene *ftsZ* was observed by EMSAs using purified fusion proteins, although WhcD alone did not bind to the genomic DNA.

The sequence motif 5′-GACAC-3′ was found to be important for binding of the WhcD-WhiA complex to the DNA. Additionally, loss of the DNA-binding activity of WhiA in the presence of an oxidant indicated a regulatory role for this protein to control cell division of *C. glutamicum* under oxidative stress conditions [56].

We merge these interactions with the previous version of the GRN for *C. glutamicum* and included as a new historical snapshot (196627_v2018_s17) with 2317 genes (73.8% of genomic coverage) and 3444 interactions (45.8% of interaction coverage) (Fig. 2). The "strong" version of the network was also included, containing a total of 2237 genes (71.3% of genomic coverage) and 2969 interactions (39.5% of interaction coverage).

4.2. *Mycobacterium tuberculosis H37Rv*

Chauhan et al. [58] reported 41 experimentally validated interactions among sigma factors and transcribed genes in the human pathogen *M. tuberculosis*. These interactions were added to the most recent *M. tuberculosis* GRNs and deposited in Abasy Atlas. The regulations among the sigma factors and TGs constitute a valuable contribution to the understanding of how *M. tuberculosis* sigma factors regulate their expression and therefore, their cellular concentrations to compete for the available RNA polymerases. Historical snapshots for the years 2015, 2016, and 2018 are available so far (Fig. 2).

4.3. *Bacillus subtilis subtilis 168*

Interactions from the most recent big update of SubtiWiki [5] were merged with the last version of Abasy Atlas including interactions from DBTBS [4] and a non-database hosted publication [43]. The result represents a new time point in the *B. subtilis* GRN history. Until now, four historical snapshots are available for this representative Gram-positive organism (Fig. 2), being the last one the GRN with the highest genomic coverage in Abasy Atlas.

4.4. *Escherichia coli K-12 MG1655*

RegulonDB [46] is one of the first organism-specific databases for transcriptional regulation data and it continues being updated. This makes *E. coli* the organism with a higher number of historical snapshots. Meta-curated GRNs from 2003 to 2018 depict the effect of the curation process in this Gram-negative model organism (Fig. 2). The meta-curation of the GRNs in Abasy Atlas reassesses the confidence level of the interactions (see "Construction and content"), and integrates the regulations by TFs, sRNAs, and sigma factors from RegulonDB into a global regulatory network.

5. Utility and discussion

5.1. User interface

From the "Home" page, you can find the description and statistics of Abasy Atlas, as well as links of interest. In the "Browse", page you can find the species for which a global GRN is deposited in Abasy Atlas, along with the number of items (networks) for such species. Further, you can click on the species to identify the strains available and even the confidence level you need. After the selection of the strain and the confidence level, you will find the historical snapshots available for the GRN of interest, as well as additional information such as the genomic and interaction coverage, data sources, and fraction of the system-level components predicted by the NDA (Supplementary Fig. 9). By clicking on "Global properties", you will find statistical and structural properties characterizing the GRN of interest. Such properties include the number of transcription factors, network density, size of the giant component, number of feedforward and feedback motifs, among others. On the same page, you can find the plots for degree, out-degree and clustering coefficient distributions (Supplementary Fig. 10). We fitted these distributions to a power-law using robust linear regression of log–log-transformed data with Huber's T for M-estimation. This overcomes the negative effect of outliers, in contrast to ordinary least squares, which is highly sensitive to outliers in data.

You can directly search for a specific gene in the upper-right box from any page. Once you are visualizing the subnetwork of interest, using the interactive panel (Supplementary Fig. 11) you can customize the visualization with several buttons and download the subnetwork as a high-definition PNG image, as well as the JSON file. Every global network can be downloaded from the "Downloads" page (Supplementary Fig. 6). Regulatory networks are provided in JSON data-interchange format, including NDA predictions and, when available, effect and evidence supporting regulatory interactions. JSON is an open standard file format, which is a lightweight, language-independent, widely used, data-interchange format supported by >50 programming languages (e.g., Python, R, Matlab, Perl, Julia, JavaScript, PHP) through a variety of readily available libraries. JSON uses human-readable text to store and transmit data objects consisting of attribute–value pairs and array data types. The JSON data files downloadable from Abasy Atlas are readily importable into Cytoscape for further analyses. Gene information and module annotation flat files in tab-separated-value file formats are also available for download. Information on how to parse the JSON files is available in the "Downloads" page. The citation policy, and the methodology to identify the system-level elements and to predict the interaction coverage is available in the "About" page. You can find additional help on the "Help" page, and contact us on the "Contact" page for any subject; we will appreciate your feedback.

5.2. Functionality

Following, we describe some remarkable cases where this new version of Abasy Atlas could have been applied to improve the studies:

The DREAM5 consortium assessed to identify the best methodology to predict GRNs from gene expression data [28] using *E. coli* and *Staphylococcus aureus* as prokaryotic models. However, they did not study how its assessment was affected by network incompleteness. This analysis can be carried out by using the set of the historical snapshots for model organisms as gold standards. The same could be applied for other assessments such as identifying the best tools to predict TF binding sites [29], DNA motifs [29,30,59], and functional modules [31].

Further, Abasy Atlas could be used to extend those benchmarking studies to include more organisms. For example, DREAM5 considered only *E. coli* as a prokaryotic model to compute the overall score because a sufficiently large set of experimentally validated interactions for *S. aureus* did not exist at that time [28]. Currently, Abasy Atlas provides GRNs for 13 *S. aureus* strains, being USA300/TCH1516 the most complete one with 25 and 30.6% of genomic and interaction coverage, respectively.

In addition to benchmarking improvements, the comprehensive atlas of GRNs that Abasy Atlas provides could be applied to study the communication that exists between the regulation of gene transcription with other mechanisms such as protein–protein interactions and metabolism [32–34]. Even when only the regulation of gene transcription is studied, across-organisms information provided by Abasy Atlas can be used to trace the evolution of the GRN in bacteria, and compare them using gene orthology and network alignment [35]. Future development of Abasy Atlas includes GRNs comparative analyses based on their structural properties.

5.3. Future development

Despite high-throughput strategies to study transcriptional regulation, there is a lack of novel interactions reported in contrast with earlier years (Fig. 2). Besides, only a handful of organisms have been experimentally studied. Computational approaches have been a hopeful option for non-model organisms and a plethora of algorithms to infer GRNs have emerged. Nonetheless, many of them are based solely on statistical approaches lacking biological constraints to filter spurious interactions. Previous assessments of tools to infer GRNs have unveiled their poor performance but also have shed light on the possibility to increase precision by consensus approaches and biological constraints [28].

Future development of Abasy Atlas aims to include inferred non-model organisms GRNs in a conservative fashion by different consensus-based approaches and the application of currently available data to validate predicted networks by using GRN organizing constraints, such as the composition of system-level elements (Fig. 4B) and network structural properties. The addition of heteromeric TFs for more organisms is also considered in the short-term future development. Mainly for the model organisms *C. glutamicum* and *B. subtilis* for which more information regarding regulation by heteromeric TFs is available. Besides, historical snapshots for non-model organisms already available in Abasy Atlas, such *Streptomyces coelicolor* will be included, while continuing including additional historical snapshots for model organisms curated from the literature and organism-specific databases. Finally, a python library providing an API to allow programmatic access to Abasy Atlas, and a REST API are under development.

6. Conclusions

Beyond the regulon level, Abasy Atlas provides the most complete and reliable set of GRNs for many bacterial organisms, which can be used as the gold standard for benchmarking purposes and training data for modeling and network prediction. Besides, Abasy Atlas provides historical snapshots of regulatory networks. Therefore, network analyses can be performed with GRNs having different completeness levels, making it possible to identify how a methodology is affected by the incompleteness, to pinpoint potential bias and improvements, and to predict future results. Additionally, Abasy Atlas is the first database providing estimations on the completeness of GRNs, their global regulators, modules, and other system-level components. The estimation of the total number of regulatory interactions a GRN could have is a valuable insight that may aid in the daunting task of network curation, prediction, and validation. Furthermore, the prediction of the system-level elements in GRNs has allowed unraveling the complexity of these networks and provides new insights into the organizing principles governing them, such as the diamond-shaped, three-tier, hierarchy unveiled by the NDA. The GRNs in Abasy Atlas have been meta-curated to avoid heterogeneity such as inconsistencies in gene symbols and heteromeric regulatory complexes representation. This enables large-scale comparative systems biology studies aimed to understand the common organizing principles and particular lifestyle adaptations of regulatory systems across bacteria and to implement those principles into future work such as the reverse engineering of GRNs.

Availability and requirements

Abasy Atlas is available for web access at https://abasy.ccg.unam.mx. If you use any material from Abasy Atlas please cite properly. Use of Abasy Atlas and each downloaded material is licensed under a Creative Commons Attribution 4.0 International License. Permissions beyond the scope of this license may be available at jfreyre@ccg.unam.mx. **Disclaimer:** Please note that original data contained in Abasy Atlas may be subject to rights claimed by third parties. It is the responsibility of users of Abasy Atlas to ensure that their exploitation of the data does not infringe any of the rights of such third parties.

Funding

This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM), México [IN205918 to JAFG].

CRediT authorship contribution statement

Juan M. Escorcia-Rodríguez: Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Andreas Tauch:** Validation, Investigation, Data curation, Writing - original draft. **Julio A. Freyre-González:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all anonymous reviewers for their positive and encouraging suggestions that helped to improve the manuscript. JMER was supported by an undergraduate fellowship from DGAPA-UNAM. He is also a PhD student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 959406 from CONACYT.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.05.015.

References

[1] Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet 2004;5:101–13.

[2] Geertz M, Maerkl SJ. Experimental strategies for studying transcription factor-DNA binding specificities. Brief Funct Genomics 2010;9:362–73.

[3] Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res 2016;44:D133–143.

[4] Makita Y, Nakao M, Ogasawara N, Nakai K. DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. Nucleic Acids Res 2004;32:D75–77.

[5] Zhu B, Stulke J. SubtiWiki in 2018: from genes and proteins to functional network annotation of the model organism Bacillus subtilis. Nucleic Acids Res 2018;46:D743–8.

[6] Pauling J, Rottger R, Tauch A, Azevedo V, Baumbach J. CoryneRegNet 6.0–Updated database content, new analysis methods and novel features focusing on community demands. Nucleic Acids Res 2012;40:D610–614.

[7] Jacques PE, Gervais AL, Cantin M, Lucier JF, Dallaire G, et al. MtbRegList, a database dedicated to the analysis of transcriptional regulation in Mycobacterium tuberculosis. Bioinformatics 2005;21:2563–5.

[8] Wren JD, Bateman A. Databases, data tombs and dust in the wind. Bioinformatics 2008;24:2127–8.

[9] 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.

[10] Freyre-Gonzalez, JA, Trevino-Quintanilla, LG (2010) Analyzing Regulatory Networks in Bacteria. Nature Education 3: 24.

[11] Freyre-Gonzalez JA, Trevino-Quintanilla LG, Valtierra-Gutierrez IA, Gutierrez-Rios RM, Alonso-Pavon JA. Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and Escherichia coli unveiled by the natural decomposition approach. J Biotechnol 2012;161:278–86.

[12] Ibarra-Arellano, MA, Campos-Gonzalez, AI, Trevino-Quintanilla, LG, Tauch, A, Freyre-Gonzalez, JA (2016) Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. Database (Oxford) 2016

[13] Freyre-Gonzalez JA, Tauch A. Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: Novel insights from a dataset with a high genomic coverage. J Biotechnol 2017;257:199–210.

[14] Campos AI, Freyre-Gonzalez JA. Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. Sci Rep 2019;9:3618.

[15] Beber, ME, Muskhelishvili, G, Hutt, MT (2016) Effect of database drift on network topology and enrichment analyses: a case study for RegulonDB. Database (Oxford) 2016.

[16] Lima-Mendez G, van Helden J. The powerful law of the power law and other myths in network biology. Mol Biosyst 2009;5:1482–93.

[17] Sanz J, Cozzo E, Borge-Holthoefer J, Moreno Y. Topological effects of data incompleteness of gene regulatory networks. BMC Syst Biol 2012;6:110.

[18] Kilic S, White ER, Sagitova DM, Cornish JP, Erill I. CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Res 2014;42:D156–160.

[19] Oluwadare, O, Highsmith, M, Cheng, J (2019) GSDB: a database of 3D chromosome and genome structures reconstructed from Hi-C data. bioRxiv: 692731.

[20] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13.

[21] Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res 2013;41:D214–220.

[22] Eckweiler D, Dudek CA, Hartlich J, Brotje D, Jahn D. PRODORIC2: the bacterial gene regulation database in 2018. Nucleic Acids Res 2018;46:D320–6.

[23] Liu, ZP, Wu, C, Miao, H, Wu, H (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database (Oxford) 2015

[24] Greene CS, Troyanskaya OG. Integrative systems biology for data-driven knowledge discovery. Semin Nephrol 2010;30:443–54.

[25] Antiqueira L, Janga SC, Costa Lda F. Extensive cross-talk and global regulators identified from an analysis of the integrated transcriptional and signaling network in Escherichia coli. Mol Biosyst 2012;8:3028–35.

[26] Covert MW, Xiao N, Chen TJ, Karr JR. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. Bioinformatics 2008;24:2044–50.

[27] Wang YC, Chen BS. Integrated cellular network of transcription regulations and protein-protein interactions. BMC Syst Biol 2010;4:20.

[28] Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. Wisdom of crowds for robust gene network inference. Nat Methods 2012;9:796–804.

[29] Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 2005;23:137–44.

[30] Jayaram, N, Usvyat, D, AC, RM (2016) Evaluating tools for transcription factor binding site prediction. BMC Bioinformatics.

[31] Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun 2018;9:1090.

[32] Simeonidis E, Chandrasekaran S, Price ND. A guide to integrating transcriptional regulatory and metabolic networks using PROM (probabilistic regulation of metabolism). Methods Mol Biol 2013;985:103–12.

[33] Chandrasekaran S, Price ND. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. Proc Natl Acad Sci U S A 2010;107:17845–50.

[34] Banos DT, Trebulle P, Elati M. Integrating transcriptional activity in genome-scale models of metabolism. BMC Syst Biol 2017;11:134.

[35] Zepeda H, Considine RV, Smith HL, Sherwin JR, Ohishi I, et al. Actions of the Clostridium botulinum binary toxin on the structure and function of Y-1 adrenal cells. J Pharmacol Exp Ther 1988;246:1183–9.

[36] Koduru L, Lakshmanan M, Lee DY. In silico model-guided identification of transcriptional regulator targets for efficient strain design. Microb Cell Fact 2018;17:167.

[37] Ibraim IC, Parise MTD, Parise D, Sfeir MZT, de Paula Castro TL, et al. Transcriptome profile of Corynebacterium pseudotuberculosis in response to iron limitation. BMC Genomics 2019;20:663.

[38] Price MN, Dehal PS, Arkin AP. Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol 2007;3:1739–50.

[39] Freyre-Gonzalez JA, Manjarrez-Casas AM, Merino E, Martinez-Nunez M, Perez-Rueda E, et al. Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis. BMC Syst Biol 2013;7:127.

[40] Morrison MD, Fajardo-Cavazos P, Nicholson WL. Comparison of Bacillus subtilis transcriptome profiles from two separate missions to the International Space Station. npj Microgravity 2019;5:1.

[41] Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M. Effect of sampling on topology predictions of protein-protein interaction networks. Nat Biotechnol 2005;23:839–44.

[42] Sierro N, Makita Y, de Hoon M, Nakai K. DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res 2008;36:D93–96.

[43] Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, et al. An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. Mol Syst Biol 2015;11:839.

[44] Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M, Santos-Zavaleta A, Penaloza-Spinola MI, et al. RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res 2008;36:D120–124.

[45] Weiss V, Medina-Rivera A, Huerta AM, Santos-Zavaleta A, Salgado H, et al. Evidence classification of high-throughput protocols and confidence integration in RegulonDB. Database (Oxford) 2013;2013:bas059.

[46] Santos-Zavaleta A, Salgado H, Gama-Castro S, Sanchez-Perez M, Gomez-Romero L, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res 2019;47:D212–20.

[47] May RM. Will a large complex system be stable?. Nature 1972;238:413–4.

[48] Zulianello L, de la Gorgue de Rosny E, van Ulsen P, van de Putte P, Goosen N. The HimA and HimD subunits of integration host factor can specifically bind to DNA as homodimers. EMBO J 1994;13:1534–40.

[49] Gotfredsen M, Gerdes K. The Escherichia coli relBE genes belong to a new toxin-antitoxin gene family. Mol Microbiol 1998;29:1065–76.

[50] Brinkrolf K, Brune I, Tauch A. Transcriptional regulation of catabolic pathways for aromatic compounds in Corynebacterium glutamicum. Genet Mol Res 2006;5:773–89.

[51] Chen C, Zhang Y, Xu L, Zhu K, Feng Y, et al. Transcriptional control of the phenol hydroxylase gene phe of Corynebacterium glutamicum by the AraC-type regulator PheR. Microbiol Res 2018;209:14–20.

[52] Henke NA, Heider SAE, Hannibal S, Wendisch VF, Peters-Wendisch P. Isoprenoid Pyrophosphate-Dependent Transcriptional Regulation of Carotenogenesis in Corynebacterium glutamicum. Front Microbiol 2017;8:633.

[53] Kleine B, Chattopadhyay A, Polen T, Pinto D, Mascher T, et al. The three-component system EsrISR regulates a cell envelope stress response in Corynebacterium glutamicum. Mol Microbiol 2017;106:719–41.

[54] Bussmann M, Baumgart M, Bott M. RosR (Cg1324), a hydrogen peroxide-sensitive MarR-type transcriptional regulator of Corynebacterium glutamicum. J Biol Chem 2010;285:29305–18.

[55] Toyoda K, Inui M. Extracytoplasmic function sigma factor sigma(D) confers resistance to environmental stress by enhancing mycolate synthesis and modifying peptidoglycan structures in Corynebacterium glutamicum. Mol Microbiol 2018;107:312–29.

[56] Lee DS, Kim P, Kim ES, Kim Y, Lee HS. Corynebacterium glutamicum WhcD interacts with WhiA to exert a regulatory effect on cell division genes. Antonie Van Leeuwenhoek 2018;111:641–8.

[57] Knizewski L, Ginalski K. Bacterial DUF199/COG1481 proteins including sporulation regulator WhiA are distant homologs of LAGLIDADG homing endonucleases that retained only DNA binding. Cell Cycle 2007;6:1666–70.

[58] Chauhan R, Ravi J, Datta P, Chen T, Schnappinger D, et al. Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis. Nat Commun 2016;7:11062.

[59] Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res 2013;41:D203–213.

III. *Corynebacterium glutamicum* Regulation beyond Transcription: Organizing Principles and Reconstruction of an Extended Regulatory Network Incorporating Regulations Mediated by Small RNA and Protein-Protein Interactions

(Escorcia-Rodríguez et al., 2021)

microorganisms

MDPI

Article

Corynebacterium glutamicum Regulation beyond Transcription: Organizing Principles and Reconstruction of an Extended Regulatory Network Incorporating Regulations Mediated by Small RNA and Protein–Protein Interactions

Juan M. Escorcia-Rodríguez [1], Andreas Tauch [2] and Julio A. Freyre-González [1,*]

[1] Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n, Col. Chamilpa, Cuernavaca 62210, Morelos, Mexico; escorcia@ccg.unam.mx

[2] Centrum für Biotechnologie (CeBiTec), Universität Bielefeld, Universitätsstraße 27, 33615 Bielefeld, Germany; tauch@cebitec.uni-bielefeld.de

* Correspondence: jfreyre@ccg.unam.mx

check for updates

Citation: Escorcia-Rodríguez, J.M.; Tauch, A.; Freyre-González, J.A. *Corynebacterium glutamicum* Regulation beyond Transcription: Organizing Principles and Reconstruction of an Extended Regulatory Network Incorporating Regulations Mediated by Small RNA and Protein–Protein Interactions. *Microorganisms* **2021**, *9*, 1395. https://doi.org/10.3390/microorganisms9071395

Academic Editor: Hiroya Yurimoto

Received: 20 December 2020
Accepted: 12 January 2021
Published: 28 June 2021

Abstract: *Corynebacterium glutamicum* is a Gram-positive bacterium found in soil where the condition changes demand plasticity of the regulatory machinery. The study of such machinery at the global scale has been challenged by the lack of data integration. Here, we report three regulatory network models for *C. glutamicum*: *strong* (3040 interactions) constructed solely with regulations previously supported by directed experiments; *all evidence* (4665 interactions) containing the *strong* network, regulations previously supported by nondirected experiments, and protein–protein interactions with a direct effect on gene transcription; *sRNA* (5222 interactions) containing the *all evidence* network and sRNA-mediated regulations. Compared to the previous version (2018), the *strong* and *all evidence* networks increased by 75 and 1225 interactions, respectively. We analyzed the system-level components of the three networks to identify how they differ and compared their structures against those for the networks of more than 40 species. The inclusion of the sRNA-mediated regulations changed the proportions of the system-level components and increased the number of modules but decreased their size. The *C. glutamicum* regulatory structure contrasted with other bacterial regulatory networks. Finally, we used the *strong* networks of three model organisms to provide insights and future directions of the *C. glutamicum* regulatory network characterization.

Keywords: *Corynebacterium glutamicum*; regulatory interactions; regulatory network; curation; network inference; systems; modules; NDA; regulogs

1. Introduction

Corynebacterium glutamicum is a Gram-positive soil bacterium, industrially relevant due to its amino acid production proficiency. It is also a model organism for the study of regulatory networks [1], along with other bacteria such as *Escherichia coli*, *Bacillus subtilis*, and *Streptomyces coelicolor*. These model organisms are usually compared, and diverse differences have been found (e.g., while *C. glutamicum* grows by apical elongation, *B. subtilis* and *E. coli* grow by lateral elongation [2]). Some aspects of the transcriptional regulatory mechanism of *C. glutamicum* have also been found to be different from those in other model organisms [3]. In contrast to *E. coli*, repression is the most common regulatory mechanism in *C. glutamicum* [4], and unlike *B. subtilis* and *E. coli*, which have diauxic growth due to the preferential consumption of one carbon source over others, *C. glutamicum* cometabolizes glucose with several other carbon sources [3]. In terms of σ factors, *E. coli* and *C. glutamicum* have seven, while *B. subtilis* has 17, and over 60 σ factors have been found in the *Streptomyces* species [5].

One of the challenges for the study of their transcriptional machinery at the global scale is the lack of data integration and the incompleteness of their global regulatory networks despite being model organisms [1]. The network incompleteness situation is worst for nonmodel organisms for which little or none is known about their transcriptional machinery. Even though high-throughput technologies speed up the reconstruction of regulatory networks, network models reconstructed solely with high throughput experiments present unusual structural properties when compared with other reconstructions performed mainly by conventional experiments (e.g., lower clustering coefficient [6]). Moreover, the number of sequenced genomes scales rapidly, especially for bacteria, so that even with high throughput experiments, we cannot cope with all of them. Computational approaches for the inference of regulatory networks based on gene expression data are still emerging. Proof of that is their modest performance for model organisms in the DREAM5 challenge [7] and the inconsistency between gene expression data and the model used for regulatory networks [8], although a reassessment with more complete networks and a larger number of model organisms is required [1]. An integrative approach of expression data and regulatory binding sites have shown to improve the prediction, but most of that improvement is by the binding sites approach, which provides more biological information (e.g., [8]).

When inferring regulatory interactions with transcription factor (TF) binding sites data, the approaches can be classified into three major groups: phylogenetic footprinting, regulon expansion, and regulatory interaction transfer. The latter two approaches require previous regulatory information to increase the target genes (TGs) for the TFs in a network or transfer the regulatory information between organisms, respectively. On the other hand, phylogenetic footprinting does not require previous regulatory information but is limited to the identification of coregulated genes by a common TF. However, when the cognate regulator is unknown, its identification is not trivial [9,10] due to the small size of the regulatory sequences and their overlap for some close homologous proteins. The transfer of regulatory interactions can be directly through the orthology of both TF and TG conservation or by filtering for TF binding sites in the promoter region of the TG (also known as a regulog analysis [11]). The latter provides the best results, helping us to reduce spurious interactions that are not conserved in the organism of interest [12].

Previously, we studied the functional architecture of the *C. glutamicum* regulatory network with regulations by TFs binding to DNA acting at the level of transcription initiation (transcriptional regulatory network) and compared its connectivity distribution to those in *E. coli* and *B. subtilis* regulatory networks [13]. Since then, a plethora of studies has continued unveiling novel transcriptional regulatory mechanisms in *C. glutamicum*. However, the study of the regulatory mechanisms has not been restricted to TF–DNA interactions. Some protein–protein interactions (PPis) are directly involved in transcription regulation (e.g., adenylated GlnK binding to AmtR (repressor) to release it from the DNA). Additionally, the inclusion of post-transcriptional regulations mediated by sRNAs into global regulatory networks has been performed in other organisms (e.g., [14] in *E. coli* as an undirected network). Previous versions of the *C. glutamicum* transcriptional regulatory network have been used for the transfer of regulatory interactions to other corynebacterial strains hosted in the CoryneRegNet database [15], the construction of a model for the inference of the number of interactions once the regulatory networks are complete [6], for an assessment of the NDA robustness to random remotion of nodes and interactions [13], as the gold standard for the benchmarking of a network inference approach based on sequence data (unpublished results), and as a reference for the identification of global regulators [16].

Here, we update the two previous transcriptional regulatory network models for *C. glutamicum* (Abasy IDs: 196627_v2018_s17 and 196627_v2018_s17_eStrong; hereinafter referred to as *all evidence* and *strong*, respectively) with hundreds of curated TF–DNA interactions, their effect, and their corresponding confidence level. In the *all evidence* network, we also included curated PPi that have a direct effect on gene transcription, such as anti-σ–σ factor interactions and the formation of heteromeric regulatory complexes. We incorporated interactions mediated by regulatory small RNAs acting at the post-transcriptional level in a third network model (hereinafter referred to as *sRNA)*. We deposited all three network models in the new v2.4 of Abasy Atlas. Our continuous curation of the *C. glutamicum* regulatory network has produced a set of five historical snapshots that, together, recount the curation process that has spanned 11 years. These historical snapshots are also available in Abasy Atlas.

After this update, *C. glutamicum* moves from the fourth to the second position among the organisms with the most complete regulatory network in Abasy Atlas, according to our recently published model of the total number of interactions a complete regulatory network has [6]. We discuss the global structural properties of the three network models in the context of the previous versions of the transcriptional regulatory models and more than 40 other bacterial networks from Abasy Atlas, the most complete collection of experimentally validated regulatory networks [1]. We analyzed the organizing principles and the system-level components of the three networks to identify the effects of the inclusion of interactions supported by nonstrong experiments, protein–protein interactions, and post-transcriptional layer regulation by sRNAs. Finally, we use strongly supported regulatory networks from *S. coelicolor*, *B. subtilis*, and *E. coli* to gain knowledge of the DNA-binding TFs for which no TGs have been characterized in *C. glutamicum*, and we provide a list of potential interactions retrieved through a strict and conservative computational pipeline using the most precise tools to identify regulations.

A Primer on Analyzing Regulatory Networks

The concepts and procedures used in the field of network biology have been summarized and explained in-depth and with great clarity in previous works [17–21]. Nevertheless, in this section, we summarize the state of the knowledge and main concepts required to analyze the relationship between the structure and function of regulatory networks.

The abstraction of a regulatory network can be represented as a group of nodes and directed arcs. The nodes represent the entities of the network (commonly genes or sRNAs), and the arcs represent the direction of the interaction between two nodes. For example, the requirement of GlxR for the transcription of *ramA* can be represented as *glxR* → *ramA*, while a negative effect (such as GlxR on *acnR* transcription) is usually represented as *glxR* ⊣ *acnR*. We use the gene symbol (or locus tag in the case no name has been assigned yet) to consistently represent the sequence of the interactions, for example, *sigA* → *glxR* ⊣ *acnR*, and so on (the housekeeping σ factor is required for the transcription of *glxR*, and GlxR hinders the transcription of *acnR*). Nodes representing other biological entities can also be included in the network. The *C. glutamicum* networks herein reported contain three types of nodes: genes, heteromeric protein complexes, and sRNAs. Heteromeric protein complexes are conformed by two or more regulatory proteins transcribed by different genes and are included in the network to reduce redundancy and improve representation accuracy [1]. The effect of the sRNA regulatory interactions is carried out at the post-transcriptional level. These interactions are included in the networks with an sRNA label in their corresponding Abasy ID [1]. The importance of the inclusion of sRNAs in bacterial regulatory networks is relatively recent [14], and there is little information regarding these types of interactions in bacterial regulatory networks.

Once the interactions are merged to form a global regulatory network, we can compute the **connectivity degree** of the nodes (k), which represents the number of interactions of a node with the rest of the network, regardless of the direction. In some scenarios, the connectivity degree can be more informative if the direction of the interactions is considered. The out-degree ($\mathbf{k_{out}}$) of a node is the number of nodes it regulates. The nodes with a k_{out} greater than zero are defined as regulators. The k_{out} is the most applied connectivity in regulatory networks (e.g., for the identification of proteins required for the transcription of a large fraction of the network: global regulators). The in-degree ($\mathbf{k_{in}}$) is the number of regulators involved in the transcription of a given gene/sRNA. An exception is the incoming interactions in heteromeric complexes that represent the formation of the complex instead of their regulation, despite that the relationship is causal, as the presence of the subunits is required to produce the heteromeric complex. Hence, the heteromeric complexes have incoming interactions only from the subunits required for its conformation, while the subunits have outgoing interactions only to the heteromeric complexes they are part of [1]. These types of interactions are underrepresented in the network and, therefore, not specified in most cases. $\mathbf{k_{max}}$ is defined as the largest connectivity value of the network and equals the k_{out} of the global regulator with the largest set of TGs. The **auto-regulations** represent a direct transcriptional effect of the regulator onto its own coding sequence.

The average clustering coefficient quantifies the modularity of a network. This structural property is an example where the direction of the interactions is disregarded, as modularity is defined as the degree to which the components of a system are separated or combined. The **clustering coefficient** of a node is defined as the fraction of its neighboring nodes that are connected to each other, relative to the potential interactions that could exist among them. For example, node A, having as neighbors only the nodes B and C, will have a clustering coefficient of one if an interaction exists between B and C (regardless of the direction of the interaction) because the potential number of interactions between the neighbors of A is only one. The clustering coefficient of A is zero if there is no interaction between B and C. Once the clustering coefficient is calculated for every node having at least two neighbors in the network, the values are averaged. For an illustrated example, please see Box 1 in reference [19]. **C(k)** shows a distribution of the average clustering coefficient for the nodes with connectivity k. Similarly, the distribution of the connectivity of the nodes is denoted as **P(k)**, provided by the probability of a node having k interactions. It has been previously debated whether the P(k) of real networks is truly governed by a power-law distribution, where a few nodes have most of the interactions [22]. Recently, using several statistic methods, we demonstrated that regulatory networks truly follow a power-law distribution—they fit other power-law-like distributions better than a Poisson distribution, regardless of the completeness of the network—and that the sole coefficient of determination (R^2) is a good proxy to assess the goodness-of-fit of the model [6].

A network component is a group of nodes in which every pair is connected by at least one path. Regulatory networks do not always comprise a single component. Commonly, small groups of nodes can be isolated from the rest of the network. This is frequently observed in nonmodel organisms for which only some groups of nodes have been studied. Whether regulatory networks are truly multicomponent, or this is only a consequence of network incompleteness, is still an open question. The **giant component** is the largest component of the network, and its size is determined by the number of nodes it covers. In regulatory networks, the global TFs, such *sigA*, increase the fraction of nodes in the giant component. The higher the fraction of nodes in the giant component, the more cohesive the network is. The giant component of a network is the representative part of the network for most structural properties such as density. **Network density** is the fraction of interactions from the fully connected network (where every node would have a directed interaction to itself and every other node in the network) that exists in the actual network. The detection of a constrained space for density values in bacterial regulatory networks [6] allowed us to infer the number of interactions expected once the curation of the network is completed [1] in order to identify some differences in the curation state of the regulatory networks.

Most of the definitions mentioned before are applied to the **κ-value** (Kappa value), which is defined as the point of the $C(k_{out})$ distribution where the change in normalized k_{out} connectivity equals the change in the clustering but with the opposite sign. The κ-value is used as a threshold for the identification of global regulators and has shown high precision and sensitivity to different bacterial regulatory networks such as *E. coli* [23], *B. subtilis* [24], and *S. coelicolor* (unpublished results) while being conservative (high precision, low sensitivity) on an earlier version of the *C. glutamicum* regulatory network [13].

The global regulators shape the highest hierarchy in the diamond-shaped structure unveiled by the natural decomposition approach (NDA). The **NDA** is an in-silico technique that deconstructs a regulatory network to naturally identify its structure and reconstructs it with the nodes classified into one of four classes: global regulators (**GRs**), modular nodes (**Mds**), intermodular nodes (**IMs**), and basal machinery (**BM**). Global regulators are the TFs with a low clustering coefficient and a k_{out} greater than the κ-value. Once the GRs have been identified, the BM is also unveiled as the TGs that are regulated only by GRs. The direct GR–BM regulation is required for fast responses without previous modulation of intermediates. GR and BM nodes and their interactions are removed from the network as well as the rest of the nodes with $k_{out} = 0$ (putative structural genes). The remotion of these structural genes will lead to isolated groups of Mds (modules) that work together for a common purpose. Finally, the structural genes are reinserted into the network, preserving their original interactions, and they are included into the module of their regulators only if all their regulators are from the same module. Otherwise, they are included as IMs, integrating the signals from different modules. For further details about the NDA methodology, please see Figures 1 and 2 in [13], where the NDA is described and applied to an earlier version of the *C. glutamicum* transcriptional regulatory network. Noteworthy, this diamond-shaped hierarchy has been found to be structurally conserved even between phylogenetically distant organisms [24]. The NDA classification is robust to random remotion of interactions and nodes [13], but the curation state of the network can alter the class of some nodes. This applies mainly to the IM and BM nodes that can be included in the Md class in a later (more complete) version of the network.

2. Materials and Methods

2.1. Curation and Network Definition

Four types of interactions were defined for consideration in this new version of the *C. glutamicum* networks: (1) homomeric-TF–DNA comprehending interactions between DNA-binding TFs (including σ factors) and the DNA, altering the gene expression; (2) sRNA–RNA interactions, occurring at the post-transcriptional level, modulating the concentration of the proteins; (3) protein–protein interactions class 1 (PPi-cI), defined as PPis with a causal regulatory effect, such as anti-σ–σ interactions; (4) PPi class 2 (PPi-cII), a form of TF–DNA interaction where the TF is a heteromeric protein complex its with cognate subunits—complex interactions. Two levels of confidence are defined for the interactions: strong, if the interaction is supported by a TF–DNA direct binding experiment (e.g., footprinting with purified protein), and weak, otherwise. Even though other types of interactions considered for this version might be supported by a direct experiment (e.g., yeast two-hybrid assay for PPi-cI), we only included homomeric-TF–DNA and heteromeric-TF–DNA interactions (PPi-cII) in the *strong* network. The *all evidence* network includes interactions supported by any experimental evidence, keeping the label "strong" only for those interactions taken from the *strong* network. For the *all evidence* network, all but the sRNA-mediated regulations are considered, while the *sRNA* network includes every type of interaction regardless of the experiment supporting it. The three networks reconstructed in this work have been deposited in the new v2.4 of Abasy Atlas.

The curation of strong interactions was carried out manually by screening the PubMed library for publications describing regulatory interactions of *C. glutamicum*. Interactions are classified as strong when the respective paper contains experimental evidence of a TF–DNA interaction. In most cases, the TF of interest is purified and its direct interaction with DNA is demonstrated in vitro. Approaches like this also lead to the experimental identification of the DNA binding site sequence. For the recovery of weakly supported interactions, we reviewed the literature to identify TGs for the TFs already present in the *all evidence* network. We used as keywords "glutamicum", "regulon", ″target genes", and the name symbol of the gene or its locus tag. Then, we followed a set of rules to include the interactions for every TF–TG pair of nodes: (1) an interaction does not exist in the network unless it is already in the previous version; (2) an interaction that is not part of the previous version does not exist unless there is experimental evidence to support the interaction; (3) an interaction supported solely by computational predictions is not included in any of the networks; (4) an interaction weakly supported by an experiment is part of the network until contradictory evidence is found (e.g., gene overexpression supported by microarrays data but invalidated by RT-PCR).

We included in the *sRNA* network the regulatory interactions by anti-sense sRNAs from reference [25]. The authors included as anti-sense sRNA every sRNA that is transcribed in the opposite strand of a gene, starting within 100 nt of the 5′-end of an opposite CDS or within 60 nt from the 3′-end of an opposite CDS [25]. The authors identified two other types of sRNAs, but regulatory interactions were only assigned to anti-sense sRNAs. For the name of the sRNAs, we used the nomenclature suggested by the authors—*cgb_xxxxx*—to ease the identification of the nodes representing sRNAs in the *sRNA* network. The effect of the interactions was set to unknown—"?"—and most of the sRNAs regulate the gene transcribed in the opposite DNA strand. We included the sRNAs as independent nodes. We acknowledge that this artificially increases the genomic coverage for the *sRNA* network (counting twice the genes with an asRNA). However, assigning the interaction to the coding gene would be misleading and would inflate the number of self-loops in the network even when the sRNAs might be transcribed through its own promoter. As previously discussed, interaction coverage is a better proxy for network completeness than genomic coverage [6]. Although the authors provide the σ factors required for the transcription of the sRNAs, we did not include these σ-DNA interactions as they were solely supported by DNA-binding motif computational predictions and we have identified a high number of false-positives in the search for binding sites for σ factors. Moreover, interactions supported solely by computational predictions are not considered for Abasy Atlas networks [1]. Interactions involving a protein-coding gene not mapping to a cgl-number or from another strain are not included in the networks but collected in a separated file (Table S1).

2.2. Genome Annotation and Upstream Sequences

Genome annotations used in this work were retrieved from NCBI [26] for the following organisms (accession code and version): *Corynebacterium glutamicum* ATCC 13032 (NC_006958.1), *Streptomyces coelicolor* A3(2) (NC_003888.3), *Bacillus subtilis* subsp. subtilis str. 168 (NC_000964.3), and *Escherichia coli* str. K-12 substr. MG1655 (NC_000913.3). Upstream (up to −300 to +50) sequences with reference to the translation-start codon, for the four genomes, were retrieved from the RSAT suite [27] with the retrieve-seq tool, preventing overlap with neighboring genes.

2.3. Regulatory Networks for Other Organisms

All the regulatory networks used in this work were downloaded from Abasy Atlas, a large collection of manually curated transcriptional regulatory networks [1]. The set of nonredundant networks is defined as the most recent regulatory networks for each organism available in Abasy Atlas, resulting in a dataset of 42 regulatory networks for 42 bacterial strains. When using the nonredundant set as a background for the herein reported regulatory networks of *C. glutamicum*, the set includes the regulatory networks of all other organisms (41) plus the three herein reported networks.

2.4. System-Level Components

Nodes were classified into one of the four system-level component classes: GRs, BM, Mds, and IMs were retrieved from Abasy Atlas. The classification of the nodes has been previously described [13]. In the following paragraph, we briefly describe the NDA, the approach used for the classification of the nodes and module identification: The κ-value is computed for the identification of GRs. Every node with a number of directly regulated TGs greater than the κ-value is classified as a GR and removed from the network, along with their interactions. The remotion of the global regulator nodes leaves some nodes isolated. The isolated nodes that are solely regulated by global regulators are classified as BM, representing structural components required for elemental functions such as the subunits for RNA core polymerase. The nodes with no regulated genes in the remaining network are labeled as structural nodes and removed in order to identify an isolated group of nodes (modules) to be classified as Mds. The nodes labeled as structural are reintegrated to the network as part of a module if all of their regulators belong to the same module; otherwise, they are labeled as IM components, which integrate the signals from two or more modules responding to different conditions.

2.5. Comparison of Nodes and Interactions of C. glutamicum with Other Bacterial Regulatory Networks

To quantify the fraction of *strong* interactions in each network, we computed the ratio of regulatory interactions classified as *strong* in each of the *all evidence* regulatory networks deposited in Abasy Atlas, including the *all evidence C. glutamicum* network herein reported, and plotted the distribution. We reconstructed the previously reported model, developed to predict the size of regulatory networks [1], by using an expanded dataset including the herein reported *C. glutamicum* regulatory networks and robust linear regression. We then reassessed the goodness-of-fit of the model by recomputing the adjusted coefficient of determination. Regulatory networks of *C. glutamicum* were highlighted in the distributions to ease identification and comparison with previous versions.

2.6. Global Structural Properties

All the structural properties reported in this work were retrieved from Abasy Atlas [1] version 2.4. For comparison with other bacteria, the values reported were normalized as follows: The number of autoregulations was normalized by the number of regulatory nodes (those with the potential to have an autoregulation). To ease the comparison of density values in a plot, each of them was multiplied by 10. Please note that this modification is used only to compare the properties. The k_{max} was normalized by the number of nodes in the network (potential targets). The κ-value was normalized by the k_{max}. The size of the giant component was normalized by the number of nodes in the network. No normalization was applied to compare the *C. glutamicum* network across versions and evidence levels. Instead, we used a log2-fold change ratio of the properties' value relative to the corresponding value for the earliest network in the case of different versions and the smallest network in the case of comparing different evidence levels.

2.7. System-Level Components

Node classification, module identification, and their annotation were retrieved from Abasy Atlas [1] version 2.4. For the graphic representation of node classification, the values were computed using a log10 scale. For the representation of module size, actual values were used for the treemapping plot. For distribution of the number of modules, the nonredundant set of regulatory networks from Abasy Atlas version 2.4 was used, and the herein reported networks were highlighted and labeled to ease identification. For the comparison of the nodes in each NDA class for the three networks reported here, we used the Simpson similarity index, defined as the number of common elements between two sets divided by the minimum of the two numbers. Hence, the similarity index can take values from zero (no overlap at all between the two sets) to one (one set is a subset of the other). For the interactions from GRs and Mds to the four classes, we computed the fraction of interactions between each class, ignoring interactions from BM and IM classes (less than 1% of the network), which are attributed to missing interactions that will be included in the future curation of the network (e.g., *cgb_20925* regulating *sigA*). Matplotlib, Seaborn, Numpy, and Squarify libraries from Python were used to compute and plot the results.

2.8. Regulog Analysis

For the selection of source organisms, we used the last *strong* version of those organisms having strong regulatory networks, namely, *Escherichia coli* K-12 MG1655 (Abasy ID: 511145_v2020_sRDB18-13_eStrong), *Bacillus subtilis* strain 168 (Abasy ID: 224308_v2008_sDBTBS08_eStrong), and a curated *Streptomyces coelicolor* network, with curated strong interactions until 2019 (unreported network). Regulog analysis is based on the premise that regulatory sites are more conserved than the rest of the noncoding sequences because they are required for the cell to survive. Given the basis of the approach, the best strategy is to use phylogenetically closely related organisms [11,28]. Unfortunately, model organisms for which a *strong* regulatory network is available are phylogenetically far from each other, but we still can use them to study essential, conserved interactions [24]. The closest model organism with a highly complete regulatory network is *Mycobacterium tuberculosis* (Abasy ID: 83332_v2018_s11-12-15-16), but its regulog analysis has been previously used to transfer interactions in the opposite direction (from *C. glutamicum* to *M. tuberculosis*) [29], and the remaining interactions are mostly supported by weak evidence.

For the identification of orthologous genes, we used the OMA standalone [30] with genome sequences from NCBI (see above). We used the OMA classification of orthology relationship type and kept only the one-to-one orthology relationships. To construct the position weight matrices, we used MEME [31], Bioprospector [32], and MDscan [33] with the upstream sequence of TGs for each TF with at least one *strong* evidence supporting the interaction. Upstream sequences were defined as up-to -300 to +50 bp, relative to the translation-start codon. Then, we used FIMO [34] to find individual matches of the matrices in the upstream sequences of the complete set of *C. glutamicum* one-to-one orthologous genes using a p-value of 1×10^{-4} as a threshold to form TF–TG putative interactions. Gene identifiers for the TFs and TGs were mapped to the *C. glutamicum* genome annotation, and the interactions obtained with each of the three motif-finding tools were integrated by a vote-counting approach, which has been found to improve predictions [7], prioritizing the interactions considered as "more reliable" by the three motif-finding tools.

3. Results and Discussion

3.1. The Regulatory Networks of C. glutamicum and Potential Applications

In this section, we report the new regulatory network models of *C. glutamicum*, their differences, and the statistics comparing them with the previous version and discuss some potential applications of our network models. We reconstructed three regulatory network models: (1) The *strong* network (Abasy ID: 196627_v2020_s21_eStrong), conformed solely by DNA-binding TFs—mediated interactions that are supported by a direct experiment (e.g., footprinting with purified protein); (2) The *all evidence* network (Abasy ID: 196627_v2020_s21), conformed by every type of interaction at the transcriptional level that is supported by any experimental evidence and not discarded by any other; (3) The *sRNA* network (Abasy ID: 196627_v2020_s21_dsRNA), containing the *all evidence* network plus 545 post-transcriptional interactions mediated by regulatory sRNAs (Figure 1). The *strong* network is a subset of the *all evidence* network, while the *all evidence* network is a subset of the *sRNA* network (Figure S1). We deposited the three reconstructed networks in the new v2.4 of Abasy Atlas, each of them providing a different level of completeness (Figure S2) that is useful in different scenarios. For example, even though the *strong* network is the smallest one, the confidence level of its interactions makes this network the best alternative to be used as the gold standard for benchmarking approaches for the inference of directed regulatory networks (such as those based on regulatory binding sites). On the other hand, benchmarking of network inference tools based on transcriptomic data might tend to be penalized when using only the *strong* network, as it only contains direct TF–DNA interactions that cannot accurately be predicted based solely on transcriptomic data [8]. In that case, the *all evidence* network can be used as the gold standard, as it includes a broader scope of experimentally supported interactions that have not been reported as spurious. The *sRNA* network is the most comprehensive and, therefore, the best suited to study the biological regulatory mechanisms of *C. glutamicum*. Having reliable regulatory network models has proven to be important even for synthetic biology, for example, to engineer resource allocation by rationally modifying the transcriptional regulatory network [35].

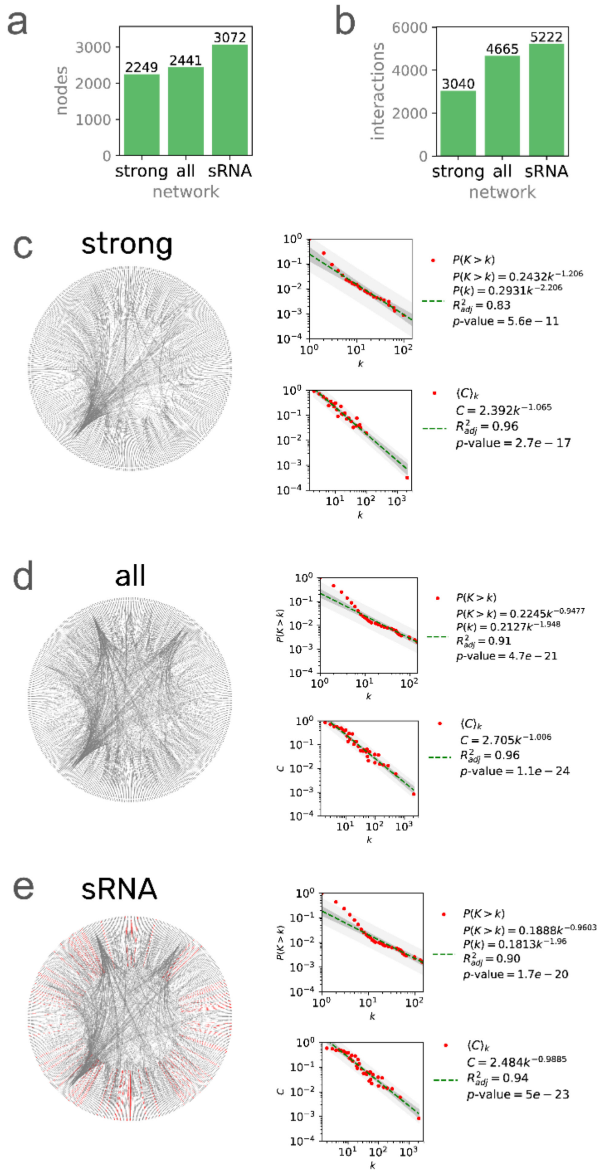


Figure 1. Three network models of the *C. glutamicum* regulatory network. The number of nodes (**a**) and interactions (**b**) for the three networks. Network, P(k), and C(k) distributions for (**c**) 196627_v2020_s21_eStrong (*strong*), (**d**) 196627_v2020_s21 (*all evidence*), and (**e**) 196627_v2020_s21_dsRNA (*sRNA*) networks. Network plots were generated with Circos [36] using the leftmost gene/sRNA coordinates to sort the nodes clockwise. Nodes with no coordinates in the genome annotation were disregarded.

3.2. Global Networks of C. glutamicum Are Quite Different from other Bacterial Networks in Terms of Their Structural Properties

In this section, we analyze the global structural properties of the *C. glutamicum* regulatory networks in the context of the whole Abasy Atlas dataset. Previously, our group found a constrained complexity in the regulatory networks [6] and leveraged it to create a model for the inference of the size of regulatory interactions expected once network curation is complete [1]. We identified a few networks falling outside of the prediction area (see Figure 5 in reference [1]), *C. glutamicum* being one of those organisms, namely, for the later versions containing the sigmulons of the housekeeping σ factor *sigA*. We found that this was a result of a low number of weakly supported interactions in contrast with other bacterial regulatory networks (Figure 2a), mainly because the *C. glutamicum* regulatory network has been highly curated in-house, giving preference to strongly supported interactions and resulting in an overrepresentation of these interactions in contrast to other bacterial regulatory networks. The inclusion of weakly supported interactions better fits the *C. glutamicum* network into the model (Figure 2b). Note that the *strong* version of the network follows the model poorly as *sigA* directly regulates 85% of the network nodes. Moreover, the fit of the *all evidence* network is affected by the inclusion of sRNA-mediated interactions (RNA in Figure 2b). This is a result of many sRNAs regulating only one gene in most cases.

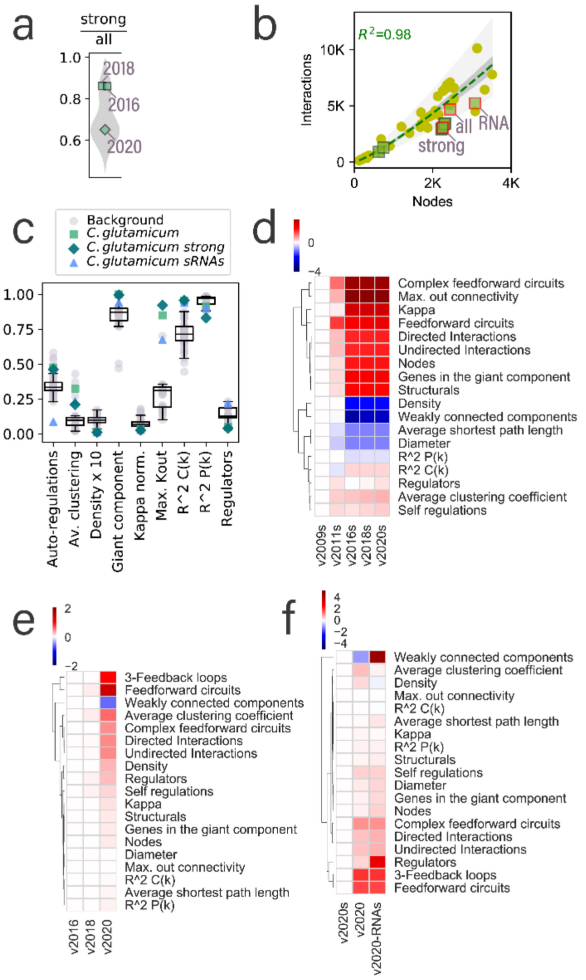


Figure 2. Structural properties of the *C. glutamicum* networks. (**a**) Distribution of the fraction of the strong interactions in the *all evidence* networks, including at least one strong interaction. *C glutamicum* networks are highlighted and labeled. 2009 and 2011 versions of the *C. glutamicum* network are not included as they do not have a cognate *all evidence* network. (**b**) Inclusion of the three networks presented in this work into the previous model reported in [1] for the inference of the number of interactions for the regulatory networks. *C. glutamicum* networks are marked with green squares, and the three networks reported in this work are highlighted with a red outline and labeled. The rest of the data points (yellow dots) are the rest of the Abasy Atlas database used for reference. (**c**) Comparison of *C. glutamicum* structural properties with the nonredundant set of bacterial networks used as background. Boxplots were drawn, including the nonredundant data set and the *C. glutamicum* networks reported in this work. (**d**) Heatmap values are the log2-fold change of the *C. glutamicum* regulatory networks for *strong* networks of versions 2011, 2016, 2018, and 2020, relative to the earliest *strong* version (2009). The v2009 column is included for clarity. Properties are clustered to ease the identification of those that have increased, decreased, or remained virtually unchanged. Heatmaps (**e**,**f**) also represent the log2-fold change values relative to the leftmost column of (**e**) for versions of the *all evidence* network and (**f**) for the three different network models presented in this work to highlight the impact of the inclusion of sRNA-mediated interactions into the structural properties of the network.

Related to this, we expect the average clustering coefficient to decrease as the node/interaction ratio increases. The clustering coefficient of a node in the network is determined by the fraction of neighbors connected to each other. As expected, the average clustering coefficient of the *all evidence* network is higher than the other two networks of the same time frame (network version) (Figure 2c) as it exhibits a better equilibrium (closest to 1) of the genomic/interaction coverage ratio (Figure S2). Interestingly, despite the *C. glutamicum* networks exhibiting a higher node/interaction ratio, they have a higher clustering coefficient than most of the bacterial regulatory networks (Figure 2c), perhaps because of a higher level of curation of the organism due to its biotechnological relevance. The density of the *C. glutamicum* networks is slightly lower than the rest of the bacterial regulatory networks. However, note that this difference is so small that even a 10-time magnification of the variance of density values is very small (Figure 2c). This is expected due to the constraint governing the complexity of regulatory networks [6].

The fraction of nodes acting as transcriptional regulators is constrained in bacteria, beyond considering only the DNA-binding TFs (Figure 2c). The *C. glutamicum* regulatory network models show a different behavior; while the network including sRNA-mediated interactions falls on the upper boundary (~25%), the other two networks fall on the lower boundary of the distribution (5%), even when the latter includes most of the DNA-binding TFs of *C. glutamicum*. For most organisms, the k_{max} is below 50% of the nodes in the network. However, the regulatory networks for *C. glutamicum* are outliers in the distribution (Figure 2c) due to the *sigA* interactions. The size of the giant component can be represented by the fraction of the network it comprehends. For most regulatory networks, this fraction is close to one (Figure 2c), especially in the case of *C. glutamicum*, whose networks with no sRNA regulation are practically a single component, showing the cohesiveness of these networks.

The κ-value is the threshold to identify global regulators. Every network has a different κ-value that relies on its hubness and modularity, but larger k_{max} values result in larger κ-values. To make the κ -values comparable, we normalized them by the k_{max} of the cognate network, allowing κ to take values between 0 and 1. Interestingly, the normalized κ-value seems to be also constrained to values lower than 0.25, and the values for the three networks of *C. glutamicum* are overlapped. This suggests that the κ-value is robust to the inclusion of weakly supported interactions and sRNAs. Moreover, this agrees with previous analysis on the robustness of the inference of global regulators to random removal of nodes and interactions [13]. However, in-depth studies with other sampling approaches and other organisms are required. Autoregulations in a regulatory network allow mechanisms to modulate themselves. A higher number of autoregulations in the networks provide a faster response of the organism to the changing conditions [37]. *C. glutamicum* requires the adaptation to different media conditions in the soil; therefore, a high number of autoregulations is expected (Figure 2c–f), where the *strong* and *all evidence* networks are above most regulatory networks. However, the fraction of autoregulations in the network containing sRNA-mediated interactions is much lower because of the large number of regulatory sRNAs that bind to other RNA but not to themselves.

3.3. System-Level Components of the C. glutamicum Regulatory Networks

The regulation of gene transcription is organized into different hierarchical layers. Previously, we have described a large-scale modeling approach to characterize the nodes of a regulatory network: the NDA (natural decomposition approach). The NDA classifies each node of the network into one of four system-level components: GRs, BM, Mds, and IMs. Regulatory networks having a diamond-shaped hierarchy have been found in different bacteria such as *E. coli* [23,24], *B. subtilis* [24], and a previous version of the *C. glutamicum* transcriptional regulatory network [13]. The hierarchy is divided into three layers (Figure 3a): the top layer, composed solely of global regulators (coordination layer), is the smallest one and can directly regulate the four NDA classes; the middle layer (processing) is composed of Mds and BM, the two largest NDA components, both regulated by the coordination layer, but with only the Md class providing feedback to the top layer (i.e., some Md TFs regulate GRs); the last layer (integration) assimilates the combinatorial disparate signals provided by GRs and Md TFs belonging to different modules into a single coordinated response, essential to adapting to environmental changes.

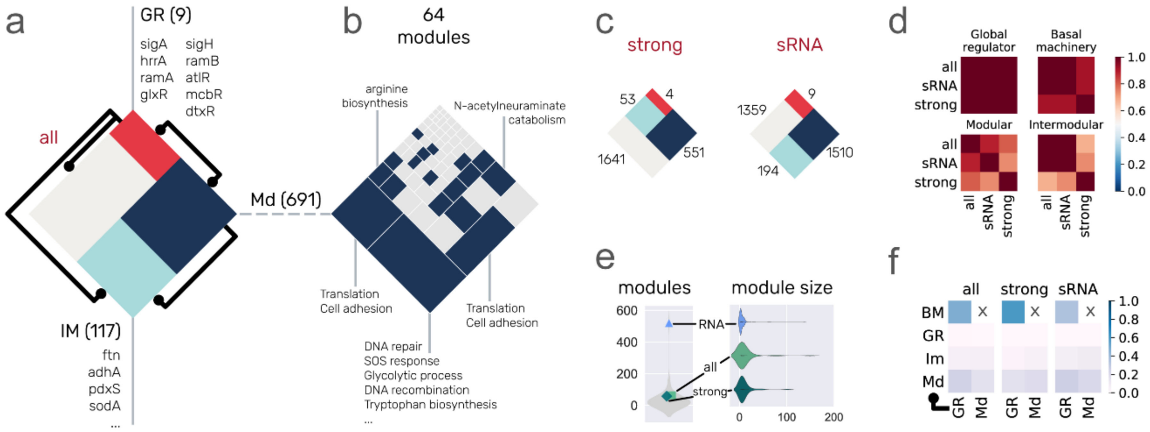


Figure 3. System-level classification of the networks. (**a**) The diamond represents the complete set of nodes in the network, which are classified in one of the four classes: global regulators (red), modular (dark blue), intermodular (light blue), and basal machinery (gray, 1624 nodes). The size of the classes is proportional to the size of the *all evidence* network on a logarithmic scale. Black lines represent the interactions between the two classes. We listed the global regulators and some examples of intermodular genes. The modular class is further divided into 64 locally independent modules in the *all evidence* network (**b**). Modules enriched with a biological function are colored in blue. The size of the sections is proportional to the size of the modules. Similar to the *all evidence* network in panel (**a**), panel (**c**) shows the proportion of the NDA classes for the *strong* and *sRNA* networks. (**d**) Heatmaps of similarity index between the three *C. glutamicum* networks for each one of the four NDA classes. The color bar shows that more than half of the nodes in the class are conserved for each class among the three networks, showing the precision of network node classification. (**e**) Distribution of the number of modules and their size. Light gray distribution of the number of modules was drawn using the nonredundant set of networks, including the three *C. glutamicum* networks. (**f**) The fraction of network interactions between the four classes for each one of the *C. glutamicum* networks.

Using the *all evidence* network as an example, the coordination layer is composed of nine GRs (Figure 3a). As expected, the first GR, when sorted by their K_{out}, is the housekeeping σ factor (*sigA*), required for the transcription of 85% of the nodes in the network. It is followed by the dual regulator *hrrA*, involved in the transcription of 21% of the network. The rest of the global regulators (and their corresponding rounded regulated network percentage) are *ramA* (11%), *glxR* (8%), *sigH* (6%), *ramB* (5%), *atlR* (4%), *mcbR* (4%), and *dtxR* (3%). The difference in regulated genes by the first and second global regulators is enormous, and this gap becomes smaller for the rest of the TFs. This is what provides the hierarchical structure to the network fitting a power-law distribution (a small fraction of nodes has most of the interactions). More than 66% of the *all evidence* network nodes are classified as BM. Examples of BM are the *rpoA*, *rpoB*, *rpoC*, and *rpoZ*, genes coding for RNA polymerase subunits.

Please note that the BM class is composed of nonregulators and is inferred based on their regulation solely by GRs. Therefore, some of its members can be transferred to the Md or IM class if they are found to be regulated by a TF from the Md class. However, it is very unlikely for a structural gene belonging to the Md class to become part of the BM (because it requires losing regulations mediated by an Md TF) and even less likely for IMs because it would require the loss of at least two Md-mediated interactions. For these reasons, a regulatory network with high genomic coverage tends only to reduce the BM as more interactions are included. On the other hand, regulatory networks with low genomic coverage are highly likely to be lacking interactions by GRs and their BM will increase with genomic coverage. It was the case for the large increase in genomic coverage in a previous update of the *C. glutamicum* transcriptional regulatory network from 2011 (genomic coverage: ~24%) to 2016 (genomic coverage: ~71%), which was mainly due to the inclusion of the *sigA* sigmulon, causing an increment of BM from 60% to 77% of the network. The Md class is composed of ~28% (691/2441) of the network, divided into locally independent modules (see below). Finally, the IM class is composed of ~5% (117/2441) of the genes in the network, all of them being structural genes (nonregulators with $k_{out} = 0$).

The Md class is further divided into locally independent modules, groups of genes that are combinatorially expressed in response to specific media conditions. In the case of the *all evidence* network, the Md class is divided into 64 modules, 18 of them (28%) enriched with one or more biological functions (Figure 3b). We used a "guild-by-association" approach to assign a biological function to nodes that have no previous annotation due to poorly annotated orthologs but belong to enriched functional modules (e.g., a module where all but one node has a GO annotation for DNA repair) [38]. The proportions for each NDA class are conserved in the network containing only strongly supported interactions, BM being the largest class, followed by Mds, IMs, and lastly, GRs. On the other hand, when regulations mediated by sRNAs are integrated (*sRNA* network) to the *all evidence* network, the proportions change for the BM and Md classes, the Md class being the largest one (Figure 3c). The number of modules is largely increased with the inclusion of sRNA regulations (Figure 3e), being an outlier in the distribution of the number of modules of bacterial regulatory networks, while the *strong* and *all evidence* networks have similar values. Even though the *sRNA* regulatory network is larger (Figure 1) and every sRNA but *cgb_20925* is included in the Md class, this does not compensate for the number of modules in the network. This is observed when we compare the distribution of the size of the modules in the networks (Figure 3e). This is also a result of the sRNAs regulating many of the nodes that are solely regulated by *sigA* in the *all evidence* network, transferring them from the BM class to the Md class and decreasing the BM class from 66.5% to 44.2% of the network.

Comparison of the size of the classes provides insights into their differences and similarities; contrasting the elements of each class contributes more to the comparative purpose. We used the Simpson similarity index to identify the overlap of two classes, taking as reference the smallest one in each comparison. Thus, the Simpson similarity index for two sets, one being a subset of the other, is 1. On the other hand, two sets having no overlap at all have an index of 0, and two sets where half of the smallest one is a subset of the largest one will have 0.5 as an index. For each NDA class, we computed the Simpson similarity index for every pair of networks and found that the *all evidence* and *sRNA* networks are more similar to each other than to the *strong* network (Figure 3d). This is expected since the *all evidence* network is a subset of the *sRNA* network (Figure S1). Please note that even though one network is a subset of the other, NDA classification is performed independently for each network; therefore, the class of a node can change from one network to another. Previous analysis of the robustness of the NDA classifications to random remotion of nodes and interactions showed the IM class is the least conserved class [13]. Surprisingly, this was not the case in the class conservation across network models, where the Md class was the least conserved (Figure 3d). This was caused by the inclusion of sRNAs in the Md class. On the other hand, the similarity index of the IM class between the *all evidence* and *sRNA* networks was not affected because even though the number of intermodular nodes increased (from 117 to 194), one is a subset of the other. Consistent with the previous robustness analysis of the *C. glutamicum* network to random interactions remotion [13], the basal machinery is well conserved, while the GR class is the most conserved class, with a similarity index of 1 for the three comparisons between the networks. This is because the *all evidence* and *sRNA* networks have the same global regulators (listed in Figure 3a), and the *strong* network has four of these nine global regulators (*sigA*, *sigH*, *dtxR*, and *glxR*).

When analyzing the communication between classes (Figure 3f), most of the interactions in the network occur from GR → BM, followed by GR → Md and Md → Md (regulations between modular TFs). For the *sRNA* network, GR → BM is decreased, while the Md → Md interactions are increased due to the inclusion of sRNAs in the Md class, regulating nodes that used to be part of the BM but are now included in the Md class. The GR and IM classes have virtually the same fraction of regulations coming from GRs and Md TFs in the *C. glutamicum* network, but further investigation in other organisms is required to assess the conservation of the proportions.

3.4. Recovering Conserved Interactions from Other Model Organisms

Regulog analysis is based on the premise that a TF–TG interaction from organism A is conserved in organism B if B has an ortholog of the TF, an ortholog of the TG, and a binding site for the TF in the promoter region of TGs [11]. As regulatory networks are highly plastic, a caveat of the regulog analysis is the functional divergence of one of the components involved in the interaction, especially for the TF [39]. Therefore, this analysis is usually applied between phylogenetically closely related organisms and is useful to transfer interactions from one model organism to others, for example, from *C. glutamicum* to other Corynebacteriales [15]. However, model organisms for the study of regulatory networks are phylogenetically far from each other, which allows the transfer of interactions from model organisms across several bacterial genera [10,15,40]. We restricted our source organisms to purely *strong* networks as they only contain directed TF–DNA interactions supported by at least *strong* evidence, namely, *E. coli*, *B. subtilis*, and *S. coelicolor*. Please note that despite the high completeness of the network for *M. tuberculosis* [1] and the closeness to *C. glutamicum* in the phylogeny, we did not use this network as a source since it was mainly constructed using only high-throughput technologies without further confirmation with directed experiments. This causes an unusually lower clustering coefficient for the network (see Figure 5 in reference [6]). Moreover, *C. glutamicum* has been used as a source organism for the inference of regulatory interactions in *M. tuberculosis* [41]. We acknowledge the caveats of using distant organisms for regulog analysis; for this reason, we applied strict conditions during the entire workflow, prioritizing precision at the expense of losing many potential interactions.

Using *S. coelicolor, B. subtilis*, and *E. coli* as source organisms (Figure 4a), we aimed to identify conserved interactions despite their phylogenetic distance (especially for *B. subtilis* and *E. coli*). To do so, first, we identified the pair-wise genome-wide orthologs between the source organisms and *E. coli* with the OMA standalone package [30], and we kept only the one-to-one orthology relationships as they have a higher probability of being *bona fide* orthologs, more likely to conserve their functions [42]. We kept 1117 one-to-one orthology relationships for *S. coelicolor* out of the total 2480 (45%), 661 out of 1480 (45%) for *B. subtilis*, and 641 out of 1488 (43%) for *E. coli* (Figure 4b). As expected, there was a greater number of one-to-one orthologous genes with *S. coelicolor* due to its phylogenetic closeness compared with the other two source organisms. Just by filtering orthologs, we restrained more than 50% of nodes to be included in the transferred interactions. The next filter is due to the completeness of the source networks since we can only transfer interactions between nodes already present in the source networks (Figure 4c). From there, we were primarily interested in TFs (white inner circles in Figure 4c), but we only considered those with at least one TG with a one-to-one ortholog in *C. glutamicum*, resulting in a total of 8, 7, and 13 potential TFs/regulons to be transferred from *S. coelicolor, B. subtilis*, and *E. coli*, respectively (colored inner circles in Figure 4c). However, the number of potential interactions to be transferred was reduced when we searched for a TF binding site in the promoter sequences of the orthologous TG in *C. glutamicum*; 24 out of the 479 interactions from the *S. coelicolor* network were conserved, along with the TF binding site, 17 out of 2576 from the *B. subtilis* network, and 70 out of 4653 from the *E. coli* network. We recovered more regulogs from *E. coli* due to the completeness of the source network. We lost many interactions through the stringent filters we applied, but we expect these conserved interactions to be true-positives. As mentioned above, the main goal of this interactions transfer is to detect interactions for the *C. glutamicum* TFs that are still missing in the network (Figure S3) despite the exhaustive work of the community to model the network. We retrieved interactions for a total of five DNA-binding TFs not considered in the current curation state of the network (Figure 4e).

Given that the *C. glutamicum* regulatory mechanism is already one of the most studied and curated (Figure S2), most of the TFs that were retrieved from regulogs were already present in the *all evidence* network (Figure 4e). However, in terms of interactions, 82 out of the 111 interactions were not present in any of the *C. glutamicum* curated networks (Figure 4f). There was poor overlap between the regulogs obtained from each organism. There was one common TF between *E. coli* and *S. coelicolor* (Zur) and another one between *E. coli* and *B. subtilis* (LexA) (Figure 4g,h).

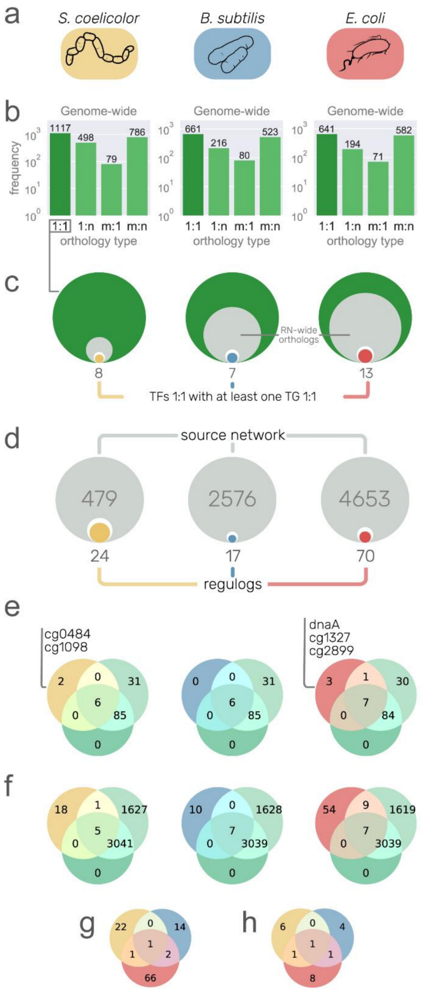


Figure 4. Putative regulons from other model organisms. (**a**) Networks with strong interactions of *S. coelicolor, B. subtilis*, and *E. coli*. used as a source of information. Rounded rectangles color is used to relate the organism to the rest of the figure. (**b**) Orthology relationship type between source organisms and *C. glutamicum*. Only one-to-one relationships were used for downstream analysis. (**c**) Size comparison between the one-to-one orthology genes (green circles), the orthologs with at least one interaction in the source network (inner gray circle), transcription factor (TF) orthologs (inner white circle), and TF orthologs with at least one target gene (TG) with one-to-one orthology relationship (inner colored circles and numbers). (**d**) Size comparison between the source networks (gray circles with large gray numbers), TF–TG pairs conserved as orthologs one-to-one in *C. glutamicum* (inner white circles), and the interactions conserved with a TF binding site in the promoter region of the TG (colored inner circles and numbers of regulogs). (**e**) Venn diagrams showing the overlap of TFs between three sets: the *strong* network (green circle), the *all evidence* network (light green circle), and the interactions from the source organisms with the unique TFs listed. (**f**) Venn diagrams showing the overlap of interactions between the *strong* network, the *all evidence* network, and the regulogs network from source organisms. (**g**,**h**) Euler diagrams showing poor overlap between the regulogs (**g**), and their TFs (**h**).

In the following section, we describe some of the conserved regulations in *C. glutamicum*. From *S. coelicolor*, 24 interactions were conserved. The interaction of Zur (*cg2502*) regulating *cg0042* is already part of the *strong* network (Figure S4). Another interaction is by RegX3 (*cg0484*), an essential response regulator of the SenX3–RegX3 two-component system [43]. RegX3 has a one-to-one orthology relationship with PhoP (*SCO4230*) from *S. coelicolor,* as does the gene *amtB* (*cg2261*) with *SCO5583*, for which a regulatory site for the PhoP ortholog in their upstream region is conserved. However, the interaction could not be transferred from *E. coli* or *B. subtilis* because a many-to-many orthology relationship was found for RegX3 in both organisms and, therefore, discarded. RegX3 has been characterized as a gene coding a regulator of phosphate-dependent gene expression in *Mycobacterium smegmatis* [44], required for virulence in *M. tuberculosis* [45], but its regulon has not been characterized in *C. glutamicum*. PhoP represses *amtB* and other nitrogen genes in *S. coelicolor* [46]. Previous work showed that *amtB* is required for ammonium uptake in *C. glutamicum* [47]. A binding site for PhoP was found 87–69 bp upstream of the *cg2261* translation start codon. This agrees with the mechanism of *amtB* regulation in *S. coelicolor*, binding upstream of the CDS and repressing its transcription by regulating a promoter in the upstream sequence from the binding site [46]. From *B. subtilis*, 17 interactions were fully conserved. For example, an autoregulation for LexA that was already part of the *strong* network (Figure S5). Cg1098 is an ortholog of SCO3129, a TetR family regulator involved in *S. coelicolor* osmotic stress [48]. In *S. coelicolor*, it regulates the transcription of two (*SCO3128* and *SCO3130*) genes and its own. However, only the autoregulation was fully conserved in *C. glutamicum*. Most of the characterized TetR family regulators regulate their own transcription [49].

From *E. coli*, we recovered a total of 70 interactions, for example, ArgR regulating *argC* (Figure S6), LexA (cg2114) regulating *recA* (Figure S7), and NrdR regulating *nrdI* (Figure S8). While the first two interactions are already included in the *strong* network, the latter is only included in the *all evidence* network. The gene *cg1327* has *b1334* as an ortholog, coding for the FNR global regulator in *E. coli*. For this protein, the regulation of *hmp* (*cg3141*) and the autoregulation were fully conserved. However, the *cg1327* gene is currently part of the basal machinery in the *C. glutamicum* network due to unreported characterization of its regulon. The gene *cg2899* codes for a regulator of the LysR family and is an ortholog of *b2537* (HcaR) in *E. coli*, regulating *hcaE*, which is an ortholog of *cg2637* (*benA*) in *C. glutamicum*, only regulated by GlxR and BenR in the *all evidence* network. In contrast with *C. glutamicum*, in *E. coli*, *hcaR* and *hcaE* are divergently transcribed, sharing the same promoter recognized by HcaR. The gene *cg0350* encodes for the GlxR ortholog to CRP in *E. coli*, both being global regulators in their corresponding networks. The regulation of CRP to *dadA* (*b1189*) is fully conserved in *C. glutamicum* for their orthologs (GlxR and its target *cg3340*, repectively). The gene *cg3340* is currently regulated only by SigA. The other TG conserved is *cg2175* (with *b3167* as its ortholog in *E. coli*), which codes for a ribosome binding protein. However, none of the two targets were identified in a previous *in silico* analysis of the GlxR regulon in *C. glutamicum* [50]. The gene *cg1425*, coding for LysG (ArgP encoded by *b2916* in *E. coli*), regulates *dnaA* that is not part of the current *C. glutamicum* network. However, none of the three interactions were conserved in *C. glutamicum*. DnaA, besides being the protein for DNA replication initiation, is a transcriptional regulator that controls the transcription of its own coding gene and at least 10 others in *E. coli*. The autoregulation and the regulation of the other four genes (*cg0004*, *cg0005*, *cg1525*, and *cg1550*) were fully conserved in *C. glutamicum* (Table S2). Zur is encoded by *cg2502*, ortholog to *b0683* in *E. coli*. A regulation from Zur to *cg2183* was recovered from the *oppC* gene in *E. coli*. The interactions are not part of the current networks for *C. glutamicum*. LldR is encoded by *cg3224*, ortholog to *b2980* (*glcC* in *E. coli*), which regulates *glcB*. The interaction was conserved in *C. glutamicum* but not present in the current networks, although the LldR regulon has 12 TGs already.

These results show that even though some interactions that are already known in *C. glutamicum* are recovered, the rate of recovered interactions is low. Therefore, for long phylogenetic distances, it might be better to discriminate false-positives after a mildly lax prediction. We noticed that most of the interactions are lost due to the conservative approach of using only one-to-one orthologs. A potential solution for this is the use of other orthology relationships, with subsequent discrimination of false-positives through the conservation of regulogs not only in *C. glutamicum* but also in other closely related organisms, conferring greater confidence values to those interactions highly conserved.

4. Conclusions

In this work, we update the *C. glutamicum* regulatory network by manual curation of the literature. We also went beyond the regulation of transcription initiation to incorporate regulations mediated by protein–protein interactions and small RNAs. Three network models with different confidence levels were reconstructed and deposited in the new v2.4 of Abasy Atlas (https://abasy.ccg.unam.mx (accessed on 1 January 2021)). Poor efforts have been carried out to provide consolidated, disambiguated, homogenized high-quality regulatory networks on a global scale, with their structural properties, system-level components, and historical snapshots to trace their curation process. We originally conceived Abasy Atlas to fill this gap by making a cartography of the functional architectures of regulatory networks for a wide range of bacteria.

This work provides the most complete and reliable set of *C. glutamicum* regulatory networks, which can be used as the gold standard for benchmarking purposes and training data for modeling. The *C. glutamicum* regulatory networks have been metacurated to avoid heterogeneity such as inconsistencies in gene symbols and heteromeric regulatory complexes representation. This enables large-scale comparative systems biology studies to understand the common principles and particular lifestyle adaptations of regulatory systems across bacteria and to implement those principles into future work such as the reverse engineering of regulatory networks. The historical snapshots deposited in Abasy Atlas allow us to carry out network analyses at different incompleteness levels, making it possible to identify how a methodology is affected, to pinpoint potential bias and improvements, and to predict future results. Regulatory network models, gene information, and module annotations can be downloaded from the "Downloads" section in Abasy Atlas (https://abasy.ccg.unam.mx/downloads (accessed on 1 January 2021)). The same web page provides useful information about the downloadable files.

Supplementary Materials: The following are available online at https://www.mdpi.com/2076-2607/9/7/1395/s1. The supplementary file contains images to support the discussion in the main text and a table with several interactions in strains of *C. glutamicum* other than ATCC 13032.

Author Contributions: Conceptualization, J.A.F.-G.; methodology, J.M.E.-R., A.T., and J.A.F.-G.; software, J.M.E.-R. and J.A.F.-G.; validation, J.M.E.-R., A.T., and J.A.F.-G.; formal analysis, J.M.E.-R. and J.A.F.-G.; investigation, J.M.E.-R., A.T., and J.A.F.-G.; resources, J.A.F.-G.; data curation, J.M.E.-R., A.T., and J.A.F.-G.; writing—original draft preparation, J.M.E.-R.; writing—review and editing, J.M.E.-R., A.T., and J.A.F.-G.; visualization, J.M.E.-R.; supervision, J.A.F.-G.; project administration, J.A.F.-G.; funding acquisition, J.A.F.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM; IN205918 and IN202421 to J.A.F.-G.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data sets supporting the results of this work are publicly accessible from Abasy Atlas (https://abasy.ccg.unam.mx (accessed on 1 January 2021)).

Acknowledgments: J.M.E.-R. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM); he received fellowship 959406 from CONACYT. We thank one anonymous reviewer for the detailed comments and suggestions that helped improve the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Escorcia-Rodríguez, J.M.; Tauch, A.; Freyre-González, J.A. Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1228–1237. [CrossRef]
2. Donovan, C.; Schauss, A.; Krämer, R.; Bramkamp, M. Chromosome Segregation Impacts on Cell Growth and Division Site Selection in Corynebacterium glutamicum. *PLoS ONE* **2013**, *8*, e55078. [CrossRef]
3. Toyoda, K.; Inui, M. Global Transcriptional Regulators Involved in Carbon, Nitrogen, Phosphorus, and Sulfur Metabolisms in Corynebacterium glutamicum. In *Corynebacterium Glutamicum: Biology and Biotechnology*; Inui, M., Toyoda, K., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 113–147. [CrossRef]
4. Brinkrolf, K.; Sandbote, J.; Pühler, A.; Tauch, A. The transcriptional regulatory repertoire of Corynebacterium glutamicum: Reconstruction of the network controlling pathways involved in lysine and glutamate production. *J. Biotechnol.* **2010**, *149*, 173–182. [CrossRef] [PubMed]
5. Pátek, M.; Dostálová, H.; Nešvera, J. Sigma Factors of RNA Polymerase in Corynebacterium glutamicum. In *Corynebacterium Glutamicum: Biology and Biotechnology*; Inui, M., Toyoda, K., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 89–111. [CrossRef]
6. Campos, A.I.; Freyre-González, J.A. Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. *Sci. Rep.* **2019**, *9*, 3618. [CrossRef] [PubMed]
7. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; DREAM5 Consortium; Kellis, M.; Collins, J.J.; et al. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804. [CrossRef] [PubMed]
8. Larsen, S.J.; Röttger, R.; Schmidt, H.H.H.W.; Baumbach, J. *E. coli* gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res.* **2019**, *47*, 85–92. [CrossRef] [PubMed]
9. Tan, K.; McCue, L.A.; Stormo, G.D. Making connections between novel transcription factors and their DNA motifs. *Genome Res.* **2005**, *15*, 312–320. [CrossRef] [PubMed]
10. Rodionov, D.A. Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria. *Chem. Rev.* **2007**, *107*, 3467–3497. [CrossRef]
11. Alkema, W.; Lenhard, B.; Wasserman, W.W. Regulog Analysis: Detection of Conserved Regulatory Networks Across Bacteria: Application to Staphylococcus aureus. *Genome Res.* **2004**, *14*, 1362–1373. [CrossRef]
12. Kılıç, S.; Erill, I. Assessment of transfer methods for comparative genomics of regulatory networks in bacteria. *BMC Bioinform.* **2016**, *17* (Suppl. S8), 277. [CrossRef]
13. Freyre-González, J.A.; Tauch, A. Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: Novel insights from a dataset with a high genomic coverage. *J. Biotechnol.* **2017**, *257*, 199–210. [CrossRef] [PubMed]
14. Nitzan, M.; Rehani, R.; Margalit, H. Integration of Bacterial Small RNAs in Regulatory Networks. *Annu. Rev. Biophys.* **2017**, *46*, 131–148. [CrossRef] [PubMed]
15. Parise, M.T.D.; Parise, D.; Kato, R.B.; Pauling, J.K.; Tauch, A.; Azevedo, V.A.D.C.; Baumbach, J. CoryneRegNet 7, the reference database and analysis platform for corynebacterial gene regulatory networks. *Sci. Data* **2020**, *7*, 142. [CrossRef] [PubMed]
16. De Witt, J.; Oetermann, S.; Parise, M.; Parise, D.; Baumbach, J.; Steinbüchel, A. Global Regulator of Rubber Degradation in Gordonia polyisoprenivorans VH2: Identification and Involvement in the Regulation Network. *Appl. Environ. Microbiol.* **2020**, *86*. [CrossRef] [PubMed]
17. Newman, M.E.J. The Structure and Function of Complex Networks. *SIAM Rev.* **2003**, *45*, 167–256. [CrossRef]
18. Dorogovtsev, S.N.; Mendes, J.F.F. The shortest path to complex networks. *arXiv* **2004**, arXiv:cond-mat/0404593.
19. Barabási, A.-L.; Oltvai, Z.N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113. [CrossRef]
20. Mason, O.; Verwoerd, M. Graph theory and networks in Biology. *IET Syst. Biol.* **2007**, *1*, 89–119. [CrossRef]
21. Costa, L.D.F.; Rodrigues, F.A.; Travieso, G.; Boas, P.R.V. Characterization of complex networks: A survey of measurements. *Adv. Phys.* **2007**, *56*, 167–242. [CrossRef]
22. Lima-Mendez, G.; Van Helden, J. The powerful law of the power law and other myths in network biology. *Mol. BioSyst.* **2009**, *5*, 1482–1493. [CrossRef]
23. Freyre-González, J.A.; Alonso-Pavón, J.A.; Treviño-Quintanilla, L.G.; Collado-Vides, J. Functional architecture of Escherichia coli: New insights provided by a natural decomposition approach. *Genome Biol.* **2008**, *9*, R154. [CrossRef] [PubMed]
24. Freyre-González, J.A.; Treviño-Quintanilla, L.G.; Valtierra-Gutiérrez, I.A.; Gutiérrez-Ríos, R.M.; Alonso-Pavón, J.A. Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and Escherichia coli unveiled by the natural decomposition approach. *J. Biotechnol.* **2012**, *161*, 278–286. [CrossRef] [PubMed]
25. Mentz, A.; Neshat, A.; Pfeifer-Sancar, K.; Pühler, A.; Rückert, C.; Kalinowski, J. Comprehensive discovery and characterization of small RNAs in Corynebacterium glutamicum ATCC 13032. *BMC Genom.* **2013**, *14*, 714. [CrossRef] [PubMed]
26. Coordinators, N.R.; Agarwala, R.; Barrett, T.; Beck, J.; Benson, D.A.; Bollin, C.; Bolton, E.; Bourexis, D.; Brister, J.R.; Bryant, S.H.; et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2018**, *46*, D8–D13. [CrossRef]

27. Nguyen, N.T.T.; Contreras-Moreira, B.; Castro-Mondragon, J.A.; Santana-Garcia, W.; Ossio, R.; Robles-Espinoza, C.D.; Bahin, M.; Collombet, S.; Vincens, P.; Thieffry, D.; et al. RSAT 2018: Regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.* **2018**, *46*, W209–W214. [CrossRef]
28. Baumbach, J.; Rahmann, S.; Tauch, A. Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Syst. Biol.* **2009**, *3*, 8. [CrossRef]
29. Krawczyk, J.; Kohl, T.A.; Goesmann, A.; Kalinowski, J.; Baumbach, J. From Corynebacterium glutamicum to Mycobacterium tuberculosis—Towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. *Nucleic Acids Res.* **2009**, *37*, e97. [CrossRef]
30. Altenhoff, A.; Levy, J.; Zarowiecki, M.; Tomiczek, B.; Vesztrocy, A.W.; Dalquen, D.A.; Müller, S.; Telford, M.J.; Glover, N.M.; Dylus, D.; et al. OMA standalone: Orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **2019**, *29*, 1152–1163. [CrossRef]
31. Bailey, T.L.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **1994**, *2*, 28–36.
32. Liu, X.; Brutlag, D.L.; Liu, J.S. BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Biocomputing* **2001**, 127–138. [CrossRef]
33. Liu, X.S.; Brutlag, D.L.; Liu, J.S. An algorithm for finding protein–DNA binding sites with applications to chromatin- immunoprecipitation microarray experiments. *Nat. Biotechnol.* **2002**, *20*, 835–839. [CrossRef] [PubMed]
34. Grant, C.E.; Bailey, T.L.; Noble, W.S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **2011**, *27*, 1017–1018. [CrossRef] [PubMed]
35. Lastiri-Pancardo, G.; Mercado-Hernández, J.S.; Kim, J.; Jiménez, J.I.; Utrilla, J. A quantitative method for proteome reallocation using minimal regulatory interventions. *Nat. Chem. Biol.* **2020**, *16*, 1026–1033. [CrossRef] [PubMed]
36. Krzywinski, M.; Schein, J.; Birol, I.; Connors, J.; Gascoyne, R.; Horsman, D.; Jones, S.J.; Marra, M.A. Circos: An information aesthetic for comparative genomics. *Genome Res.* **2009**, *19*, 1639–1645. [CrossRef]
37. Rosenfeld, N.; Elowitz, M.B.; Alon, U. Negative Autoregulation Speeds the Response Times of Transcription Networks. *J. Mol. Biol.* **2002**, *323*, 785–793. [CrossRef]
38. Ibarra-Arellano, M.A.; Campos-González, A.I.; Treviño-Quintanilla, L.G.; Tauch, A.; Freyre-González, J.A. Abasy Atlas: A comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database* **2016**, *2016*. [CrossRef]
39. Price, M.N.; Dehal, P.S.; Arkin, A.P. Orthologous Transcription Factors in Bacteria Have Different Functions and Regulate Different Genes. *PLoS Comput. Biol.* **2007**, *3*, e175. [CrossRef]
40. Novichkov, P.S.; Kazakov, A.E.; Ravcheev, D.A.; Leyn, S.A.; Kovaleva, G.Y.; Sutormin, R.A.; Kazanov, M.D.; Riehl, W.; Arkin, A.P.; Dubchak, I.; et al. RegPrecise 3.0—A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genom.* **2013**, *14*, 745. [CrossRef]
41. Baumbach, J. CoryneRegNet 4.0—A reference database for corynebacterial gene regulatory networks. *BMC Bioinform.* **2007**, *8*, 429. [CrossRef]
42. Altenhoff, A.M.; Studer, R.A.; Robinson-Rechavi, M.; Dessimoz, C. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLoS Comput. Biol.* **2012**, *8*, e1002514. [CrossRef]
43. Bott, M.; Brocker, M. Two-component signal transduction in Corynebacterium glutamicum and other corynebacteria: On the way towards stimuli and targets. *Appl. Microbiol. Biotechnol.* **2012**, *94*, 1131–1150. [CrossRef] [PubMed]
44. Glover, R.T.; Kriakov, J.; Garforth, S.J.; Baughn, A.D.; Jacobs, W.R., Jr. The Two-Component Regulatory System senX3-regX3 Regulates Phosphate-Dependent Gene Expression in Mycobacterium smegmatis. *J. Bacteriol.* **2007**, *189*, 5495–5503. [CrossRef] [PubMed]
45. Parish, T.; Smith, D.A.; Roberts, G.; Betts, J.; Stoker, N.G. The senX3–regX3 two-component regulatory system of Mycobacterium tuberculosis is required for virulence. *Microbiology* **2003**, *149*, 1423–1435. [CrossRef]
46. Rodríguez-García, A.; Sola-Landa, A.; Apel, K.; Santos-Beneit, F.; Martín, J.F. Phosphate control over nitrogen metabolism in Streptomyces coelicolor: Direct and indirect negative control of glnR, glnA, glnII and amtB expression by the response regulator PhoP. *Nucleic Acids Res.* **2009**, *37*, 3230–3242. [CrossRef] [PubMed]
47. Walter, B.; Küspert, M.; Ansorge, D.; Krämer, R.; Burkovski, A. Dissection of Ammonium Uptake Systems in Corynebacterium glutamicum: Mechanism of Action and Energetics of AmtA and AmtB. *J. Bacteriol.* **2008**, *190*, 2611–2614. [CrossRef] [PubMed]
48. He, X.; Li, H.; Pan, Y.; Wang, L.; Tan, H.; Liu, G. SCO3129, a TetR family regulator, is responsible for osmotic stress in Streptomyces coelicolor. *Synth. Syst. Biotechnol.* **2018**, *3*, 261–267. [CrossRef]
49. Cuthbertson, L.; Nodwell, J.R. The TetR Family of Regulators. *Microbiol. Mol. Biol. Rev.* **2013**, *77*, 440–475. [CrossRef]
50. Kohl, T.A.; Tauch, A. The GlxR regulon of the amino acid producer Corynebacterium glutamicum: Detection of the corynebacterial core regulon and integration into the transcriptional regulatory network model. *J. Biotechnol.* **2009**, *143*, 239–246. [CrossRef]

IV. Improving gene regulatory network inference and assessment: The importance of using network structure

(Escorcia-Rodríguez et al., 2023)

TYPE Original Research
PUBLISHED 27 February 2023
DOI 10.3389/fgene.2023.1143382

Check for updates

OPEN ACCESS

EDITED BY
Federico Zambelli,
University of Milan, Italy

REVIEWED BY
Teresa Colombo,
National Research Council (CNR), Italy
Hideo Matsuda,
Osaka University, Japan

*CORRESPONDENCE
Julio A. Freyre-González,
✉ jfreyre@ccg.unam.mx

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 16 January 2023
ACCEPTED 20 February 2023
PUBLISHED 27 February 2023

CITATION
Escorcia-Rodríguez JM, Gaytan-Nuñez E, Hernandez-Benitez EM, Zorro-Aranda A, Tello-Palencia MA and Freyre-González JA (2023), Improving gene regulatory network inference and assessment: The importance of using network structure.
Front. Genet. 14:1143382.
doi: 10.3389/fgene.2023.1143382

Improving gene regulatory network inference and assessment: The importance of using network structure

Juan M. Escorcia-Rodríguez[1], Estefani Gaytan-Nuñez[1,2†], Ericka M. Hernandez-Benitez[1,2†], Andrea Zorro-Aranda[1,3], Marco A. Tello-Palencia[1,2] and Julio A. Freyre-González[1]*

[1]Regulatory Systems Biology Research Group, Program of Systems Biology, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, [2]Undergraduate Program in Genomic Sciences, Center for Genomic Sciences, Universidad Nacional Autónoma de México, Cuernavaca, Mexico, [3]Department of Chemical Engineering, Universidad de Antioquia, Medellín, Colombia

Gene regulatory networks are graph models representing cellular transcription events. Networks are far from complete due to time and resource consumption for experimental validation and curation of the interactions. Previous assessments have shown the modest performance of the available network inference methods based on gene expression data. Here, we study several caveats on the inference of regulatory networks and methods assessment through the quality of the input data and gold standard, and the assessment approach with a focus on the global structure of the network. We used synthetic and biological data for the predictions and experimentally-validated biological networks as the gold standard (ground truth). Standard performance metrics and graph structural properties suggest that methods inferring co-expression networks should no longer be assessed equally with those inferring regulatory interactions. While methods inferring regulatory interactions perform better in global regulatory network inference than co-expression-based methods, the latter is better suited to infer function-specific regulons and co-regulation networks. When merging expression data, the size increase should outweigh the noise inclusion and graph structure should be considered when integrating the inferences. We conclude with guidelines to take advantage of inference methods and their assessment based on the applications and available expression datasets.

KEYWORDS

network inference, gene regulatory network, co-expression network, graph properties, network inference assessment, co-regulation network

Introduction

A gene regulatory network (GRN) is responsible for sensing environmental cues and responding accordingly. It represents directed regulatory interactions between genes coding transcription factors (TFs) and their target genes (TGs). Successful developments in synthetic biology require that the designed circuit properly integrates into the global and local regulatory circuits (Freyre-Gonzalez et al., 2022). This is a current challenge as there is not a single complete experimentally-validated GRN (Escorcia-Rodriguez et al., 2020), only a handful (< 4) of bacterial organisms has a known GRN having completeness > 70%, and its experimental reconstruction is a time- and resource-consuming task. Consequently, computational network inference is frequently used. Whereas previous works have evaluated network inference tools using synthetic and experimental data for several organisms (Marbach et al., 2010; Marbach et al., 2012; Chen and March 2018), they did not assess several essential criteria for the inference of GRNs such as data noise variation, and the global structure of the predictions and the gold standard (GS). Riet De Smet and Kathleen Marchal reviewed the advantages and limitations of several inference methods through the biological interpretation of the network structure but did not use the structure itself to assess the predictions (De Smet and Marchal, 2010).

Employing artificial data with varying amounts of noise, Deniz Seçilmiş et al. recently evaluated various tools and discovered that using the perturbation design matrix outperformed methods without it. (Secilmis et al., 2022). Synthetic data are the first alternative for benchmarking inference methods (Van den Bulcke et al., 2006). However, the generation of synthetic data relies on simulation parameters (e.g., dimension and noise of the dataset), which may not reflect the variability in biological data. Regarding the transcriptomic technique, most of the tools developed for GRN inference from microarray data have been indiscriminately coupled with RNA-seq (Iancu et al., 2012; Salleh et al., 2018; Zhang et al., 2019) despite tools for bulk RNA-seq data have been already developed (Proost et al., 2017; Imbert et al., 2018).

The authors of the DREAM5 network inference challenge evaluated a plethora of genome-scale transcriptional regulatory network predictions from gene expression data. Their results provided insights into the difficulty of GRN inference using correlation and mutual information between gene pairs and found that contrary to synthetic data, the dependencies between genes interacting in the cell barely exceeded the dependencies between non-interacting gene pairs in biological data. Interestingly, with synthetic and *Escherichia coli* data, the correlations between genes regulated by identical sets of TFs exceeded those between genes in the actual regulatory network (Supplementary Note S5 in Marbach et al. (2012)), but most of those interactions between co-regulated genes would be false positives (e.g., structural genes shaping a transcription unit). Recently, Simon Larsen et al. performed an in-deep analysis on this matter, their results show that the correlation of pairs of random genes is indistinguishable from those involved in known regulatory interactions in *E. coli* (Larsen et al., 2019). Doglas Parise et al. confirmed the results on *Corynebacterium glutamicum* (Parise et al., 2021).

According to the DREAM5 team, integrating predictions from different inference techniques through the Borda count method ("community network") is the best strategy because method performance is not consistent across species. (Marbach et al., 2012). Since then, the community approach has been broadly applied (Akesson et al., 2021; Zorro-Aranda et al., 2022). ComHub is a pipeline for integrating predictions from various methods to rank regulators according to their average out-degree using gene expression. (Akesson et al., 2021). Recently we inferred a GRN for *Streptomyces coelicolor* and identified the global regulators applying the NDA (natural decomposition approach) (Freyre-Gonzalez et al., 2008; Freyre-Gonzalez et al., 2012) on the across-methods community network preserving only TF-TG interactions (Zorro-Aranda et al., 2022). However, some methods are better suited to particular global topological structures (Stolovitzky et al., 2009). Thus, the hubs may differ across methods and have different biological interpretations in each global network due to the inherently different structure.

The inferences are commonly assessed using standard performance metrics such as the area under the recall vs. precision (AUPR) and true negative rate vs. recall curves. These metrics rely heavily on the ranking of the interactions (Marbach et al., 2010). Based on the ranking scheme and the cutoff value, the global network will also have a different structure. For example, using the Pearson correlation coefficient with no post-processing step as the ranking score, co-regulated genes from the same transcription unit (TU) will be at the top of the prediction and the global network will be shaped by interactions between co-expressed genes. This would be a good co-regulation network, but it will be highly penalized if it is assessed against a GRN. The edges represent different biological associations (De Smet and Marchal, 2010); therefore, the networks have a different global structure and are better suited for different purposes (Michoel et al., 2009). However, the assessment and integration of inference methods designed for co-expression are still being directly used and compared with those inferring regulation (Marbach et al., 2012; Bellot et al., 2015; Pratapa et al., 2020; Secilmis et al., 2022).

We previously explored structural properties and systems-level components to analyze curated and inferred GRNs for *Streptomyces coelicolor* (Zorro-Aranda et al., 2022). Here, we focused on the factors influencing the inference of GRNs and their assessment. Mainly, the structural characteristics of the GS and the inferred networks, the quality of the input data and the GS, and the assessment strategy. Besides synthetic data with varying noise and completeness levels, we use biological data for *Escherichia coli*, *Bacillus subtilis*, and *Pseudomonas aeruginosa* along with their experimentally-validated GRNs (Escorcia-Rodriguez et al., 2020) as the GS. Because the networks used as GS are not complete, unknown actual interactions identified in the prediction will be misclassified as a false positive. To check whether our results will hold when the GS networks are complete, we used historical snapshots with different completeness levels and evidence (Escorcia-Rodriguez et al., 2020). Figure 1 summarizes the complete workflow.

Results and discussion

We reviewed the literature to construct a collection of network inference tools. After the application of filter criteria (see Materials and methods), 15 tools were selected to be assessed along with "Community" reconstructions integrating interactions from several tools. Then, we arranged the inference tools according to the output network type into three groups (Table 1; Figure 1): 1) The COEX tools infer interactions between genes with correlated expression profiles. 2) The CAUS tools use a TFs list to infer regulatory interactions between the TFs and their TGs (i.e., GRNs) (Hecker et al., 2009). 3) The HYBR (hybrid) group contemplates ANOVA (Kuffner et al., 2012) and Friedman (Zorro-Aranda et al., 2022) which are based on analysis of variance and therefore do not infer causality. However, we used a list of TFs to keep only TF-TG

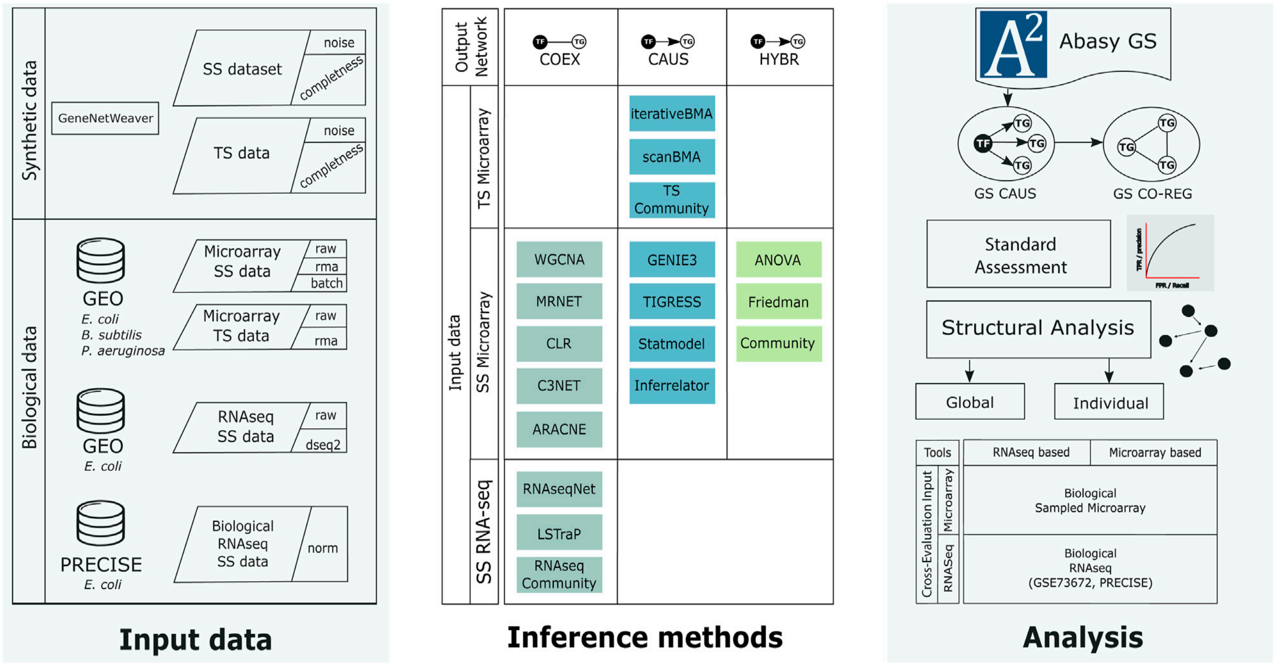


FIGURE 1
Workflow of this work. We generated synthetic data using GeneNetWeaver for *E. coli* and collected several biological microarray datasets from GEO for *E. coli, B. subtilis,* and *P. aeruginosa*, as well as RNA-seq data from GEO and PRECISE for *E. coli* (left column). The synthetic and biological datasets were used as input for the inference methods (middle row). The inference methods were classified according to their final network type. COEX tools generate undirected networks. CAUS tools generate directed networks using a list of regulators to compute the predictions as part of their algorithm. HYBR includes Friedman and ANOVA implementations (Zorro-Aranda et al., 2022) that generate co-expression networks that are trimmed to only include regulations mediated by a known transcription factor. The Community networks are classified according to the type of tools they include. We used biological networks as the gold standard to perform the assessment and analyses (right column). From the directed gold standard ("CAUS" GS) we generated a co-regulation gold standard (GS CO-REG). We performed the standard statistical and a structure-based assessment. SS: steady-state data, TS: time-series data, GS: gold standard, TF: transcription factor, TG: target gene. See Supplementary Figure S1 for further details.

interactions. The classification of Community relies on the type of interactions that it includes. It is considered HYBR when it integrates interactions from different network types, but it will be considered CAUS if it only integrates interactions from CAUS tools. Similarly, Community will be considered COEX if it only contains interactions from COEX tools. See Table 1 and the ∑ Introduction section in the Supplementary Information for a detailed description of the tools.

Tools for inferring co-expression networks should be assessed apart from those for inferring causality

We used synthetic and biological datasets to assess the tools inferring networks from microarray data (Figure 2A). We assessed the inferred networks using 30 synthetic gene expression datasets with varying noise levels and sample sizes against the biological regulatory network used to generate the synthetic data. There was an overall improvement with larger datasets with less noise (Figure 2B and Supplementary Figure S2). GENIE3 and Inferelator performed the best, even better than Community, contrasting with the results of the DREAM5 challenge where Community outperformed all the single-tool predictions on the assessment with synthetic data (Marbach et al., 2012). On the other hand, ANOVA and WGCNA showed poor performance despite the data variations. There was no clear difference among the tools at the group level.

We collected gene expression data for *E. coli*, *B. subtilis*, and *P. aeruginosa* from GEO and generated three datasets for each organism, each with different preprocessing levels: raw data, Robust Multiarray Averaging (RMA) normalization, and RMA normalization plus batch correction (R-B). For the GS, we retrieved experimentally-supported GRNs from Abasy Atlas for the three organisms. As a group, CAUS performed the best followed by HYBR. On the other hand, COEXP showed poor results. Among the CAUS tools, GENIE3, Inferelator, and TIGRESS performed the best across the three organisms. GENIE3 was the best method in *E. coli* and *P. aeruginosa*, but TIGRESS and Inferelator outperformed it in *B. subtilis*, the organism with the smallest dataset (Supplementary Figure S3). This could be due to the lower prediction stability of GENIE3 to data size variations in contrast with TIGRESS and Inferelator. Among the HYBR tools, Friedman and Community improved their performance with R-B data, while ANOVA showed inconsistent results. Most of the tools performed better with fully preprocessed R-B (Figure 2C).

For each inference tool, we averaged its prediction score with the highest-quality data: R-B for each organism and the complete synthetic dataset with the lower noise level (5%). GENIE3 obtained the highest overall score, followed by

TABLE 1 GRNs inference tools used in this work. For a detailed description of each tool, please see the Supplementary Information. COEX tools infer undirected networks, CAUS tools infer directed networks, HYBR tools infer undirected networks and the direction TF-TG is assigned with the list of known regulators to keep only the TF-mediated interactions (Zorro-Aranda et al., 2022). Community is not listed here because rather than a stand-alone tool, this method integrates the interactions from several single-tool predictions.

Method	Network type	Directed network	Main References
ARACNE	COEX	FALSE	Margolin et al. (2006)
C3NET	COEX	FALSE	Altay and Emmert-Streib (2010)
CLR	COEX	FALSE	Faith et al. (2007)
MRNET	COEX	FALSE	Meyer et al. (2007)
LSTrAP	COEX	FALSE	Proost et al. (2017)
RNA-seqNet	COEX	FALSE	Proost et al. (2017)
WGCNA	COEX	FALSE	Zhang and Horvath (2005)
GENIE3	CAUS	TRUE	Huynh-Thu et al. (2010)
INFERELATOR	CAUS	TRUE	Bonneau et al. (2006)
TIGRESS	CAUS	TRUE	Haury et al. (2012)
StatModel	CAUS	TRUE	Zorro-Aranda et al. (2022)
iBMA	CAUS	TRUE	Annest et al. (2009)
ScanBMA	CAUS	TRUE	Young et al. (2014)
ANOVA	HYBR	TRUE	Zorro-Aranda et al. (2022)
FRIEDMAN	HYBR	TRUE	Zorro-Aranda et al. (2022)

Inferelator and TIGRESS (Figure 2D). Community ranked fourth in the overall score despite it includes interactions from the COEX predictions. In concordance with the DREAM5 challenge (Marbach et al., 2012), this suggests that despite low-scored predictions integration, Community still has reliable performance. A community integration seems to be a safer choice because the rank of individual tools differs among organisms, but CAUS tools outperformed COEX tools with biological data every time (Figure 2D).

Unlike the COEX tools, the CAUS and the HYBR tools require a list of the genes coding for TFs (Table 1) to keep only TF-TG interactions and avoid TG–TG edges that are not expected in a GRN, such as the networks used as GS. On the other hand, only a few of the interactions inferred by the COEX tools include a TF, i.e., most edges are TG-TG interactions (Supplementary Figure S4). As an effort to perform a fair assessment of COEX tools, we modified the *E. coli* GS to resemble a co-regulation network where each regulon, set of co-regulated genes, is a clique (every node is interconnected). This way, COEXP outperformed the rest of the tools (Figure 2E).

The performance of every tool declined with the biological datasets in contrast to the synthetic ones. It is expected because the synthetic datasets were generated with the network used as GS. Besides, training and evaluating the tools with biological data is rare due to data accessibility (Marbach et al., 2010). There is a clear difference between the performance of CAUS and COEX tools with the biological datasets and a GRN as the GS (Figures 2C,F). On the other hand, the COEX tools succeeded with a simulated co-regulation network as the GS (Figure 2E). C3NET obtained the highest overall score, followed by CLR, ARACNE, and WGCNA. These results suggest that even though we should use CAUS tools for the inference or GRNs, tools inferring co-expression networks should be assessed apart from those inferring causality. Ignoring the direction of the GS interactions to make a fairer comparison (Chen and March 2018) is not enough. Because of the nature of the network, the interactions inferred by COEX tools will be closer to representing co-expression and co-regulation rather than regulation. Moving to regulation is not trivial, but some approaches are already trying to infer causality from co-regulation and co-expression networks (Aibar et al., 2017; Chen and Liu, 2022).

Inference methods based on Bayesian approaches take advantage of time-series data to infer causal relationships (Lo et al., 2012). We assessed two tools based on a Bayesian approach: scanBMA (Young et al., 2014) and iterativeBMA (Annest et al., 2009), along with a Community reconstruction integrating both predictions. The performance with synthetic data improved with larger datasets and less-noise levels. iterativeBMA obtained the best scores, slightly better than Community (Supplementary Figure S5). Then, we assessed the tools with biological data, one time-series experiment for *E. coli* and one for *P. aeruginosa*. We used only raw (non-normalized) and RMA pre-processing steps as batch correction is not necessary for the one-source samples. Overall, scanBMA performed better than iterativeBMA (Figure 2G). Both tools with Bayesian approaches performed poorly despite their advantage over other methods to infer causal relationships, perhaps because of the few samples available. Future data availability along with experimental annotation might improve the performance of Bayesian approaches.

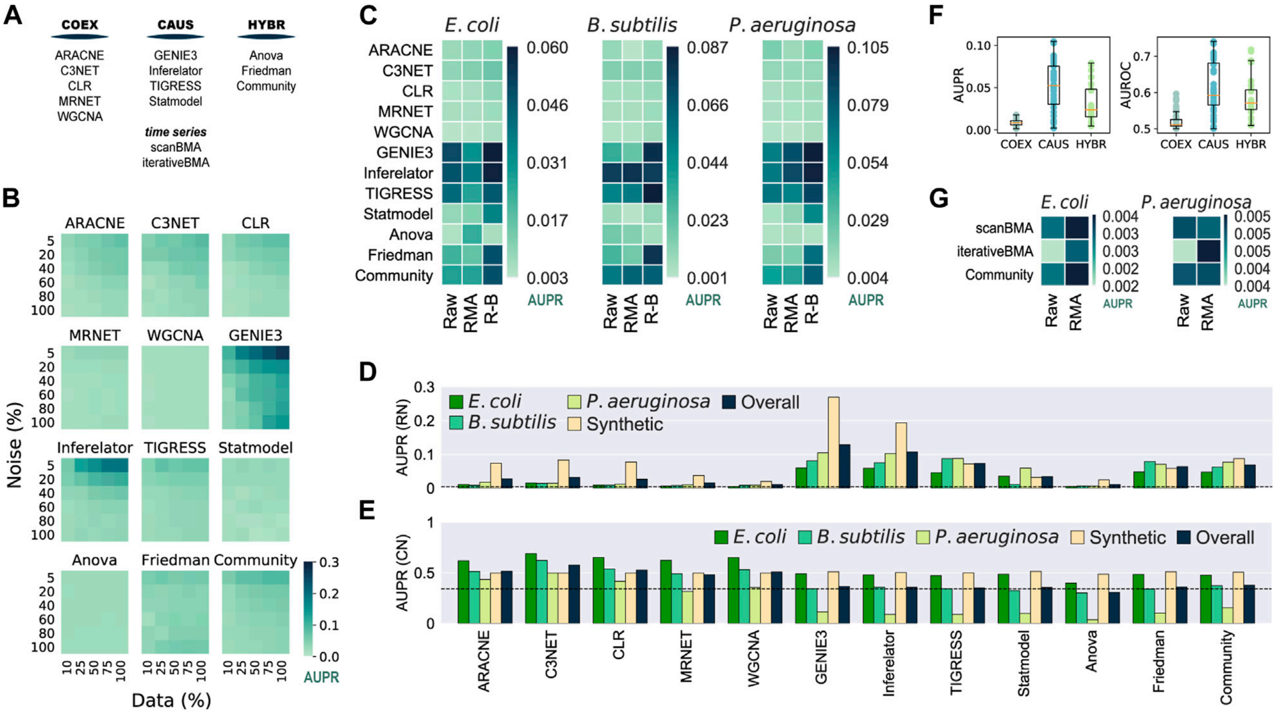


FIGURE 2
Assessment of network inference tools for microarray data. 100% of the synthetic dataset contains a total of 788 conditions. The Community Network is the integration of the single-tool predictions using the Borda count method (Marbach et al., 2012). **(A)** Network classification. Network inference tools for microarray data were classified according to the type of network they infer. **(B)** GENIE3 is the best tool for synthetic data. Synthetic gene expression datasets with different levels of noise and completeness were generated from the biological network of *E. coli* (511145_v2017_sRDB16_eStrong). The same network was used as the GS for the assessment. **(C)** Batch correction and knowledge of the transcription factors improve the inference of transcriptional GRNs. Causal and Hybrid tools outperformed Co-expression tools in the assessment of GRNs using biological data for *E. coli*, *B. subtilis,* and *P. aeruginosa* with different levels of data normalization: raw data, Robust Multiarray Averaging (RMA), and RMA plus batch correction. Inferences were assessed with experimentally-validated GRNs. **(D)** GENIE3 is the best tool for the inference of GRNs. **(E)** Assessment for the inference of co-regulation network. The COEX tools outperformed CAUS and HYBR tools. C3NET performed the best. **(F)** Boxplot representation of data in panel C to highlight the differences across tool groups. **(G)** scanBMA outperformed iterativeBMA with biological data. The Community network for this panel only integrates interactions from scanBMA and iterativeBMA.

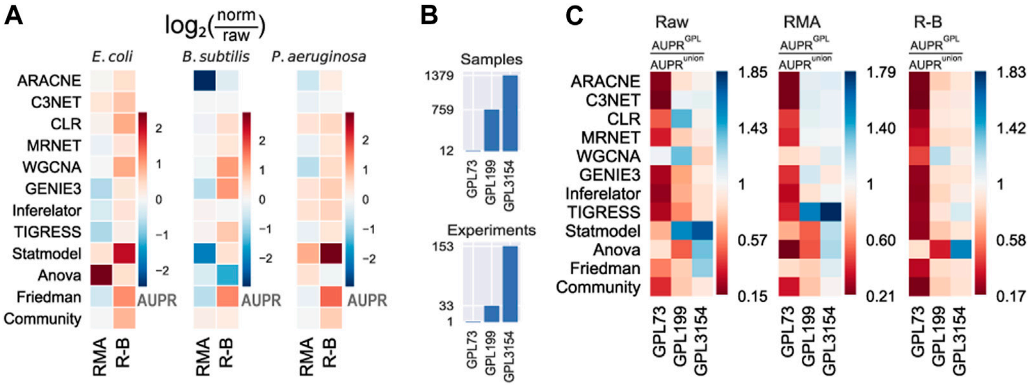


FIGURE 3
Effect of normalization and batch correction on the GRN inference with biological data. **(A)** RMA normalization with batch correction (R−B) presents a slight improvement over only RMA normalization. The values represent the log2 ratio of the AUPR with normalized data concerning the AUPR with raw data. Higher (warmer) values mean more significant improvement with normalization. **(B)** Platforms vary in the number of samples and experiments. **(C)** Methods were assessed using different Affymetrix platforms of *E. coli* and. AUPR increases with larger datasets as input data.

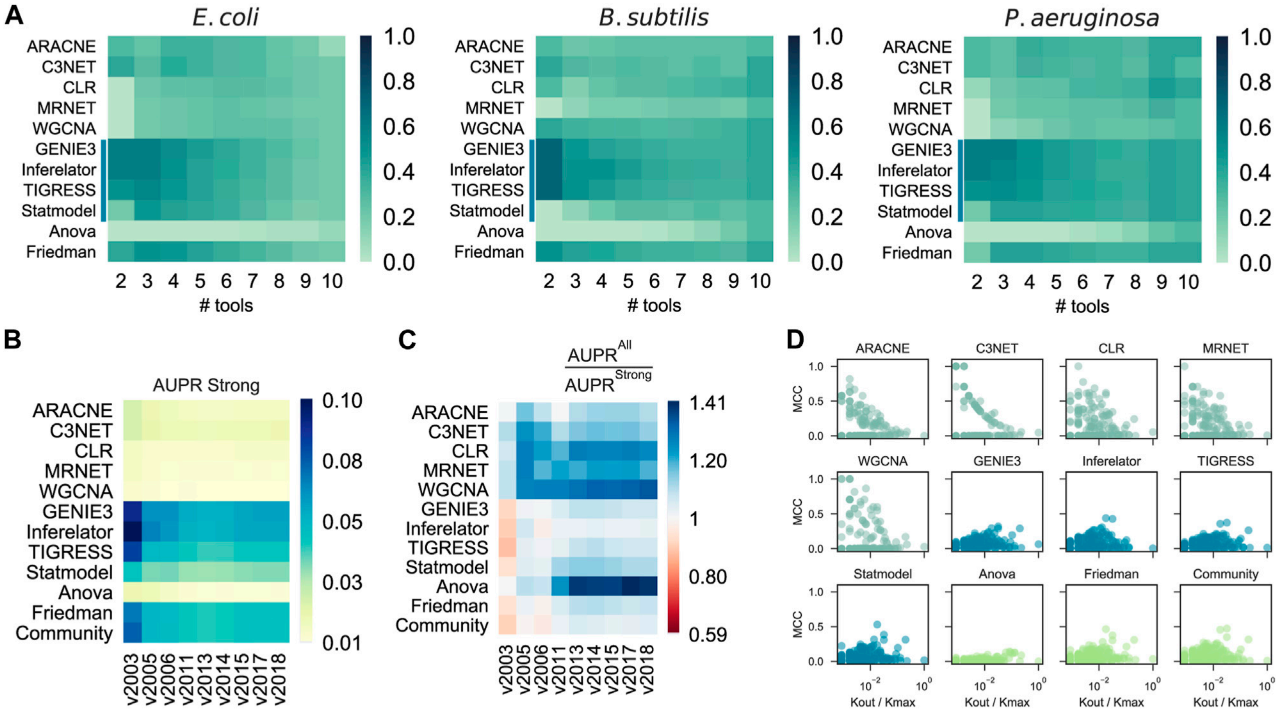


FIGURE 4
Effects of results integration, GS incompleteness, and Regulon-level assessment. **(A)** Probability of a tool to outperform Community by its integration with others (# tools) into a selective community. CAUS tools are affected rather than improved by others. **(B)** Assessment of GRN inference methods with the historical reconstruction of the *E. coli* GRN. The incompleteness of the GRN used as GS does not affect the AUPR score. **(C)** AUPR ratio between a "strong" GS and a "weak" one. In most cases, the tools performed better when a "weak" GS was used. The "weak" GS is a superset of the "strong" GS including interactions supported by non-directed experiments. **(D)** Regulon prediction assessment with Matthew's correlation coefficient (MCC). Each dot represents a regulon inference for an *E. coli* TF, higher is better. Out-degree connectivity (Kout) for the TF controlling the regulon is normalized by the maximum connectivity (Kmax) of the *E. coli* network.

RMA with batch correction on large datasets improves the predictions

To provide deeper insights into the effects of data normalization on network inference, we contrasted the results using none (raw), RMA, and R-B preprocessing levels. The removal of batch-effect over RMA (R-B) normalization seems to slightly improve the predictions (Figure 3A and Supplementary Figure S6). RMA normalization without batch correction worsens the performance of the tools. This is because some tools might be leveraging data heterogeneity or information lost in the normalization process (Sirbu et al., 2010). Besides, the assumptions considered by normalization pipelines could be violated, resulting in spurious predictions (Evans et al., 2018). Therefore, either raw data or normalized and batch-effect-corrected data should be used for network inference with highly heterogeneous datasets.

In addition to data preprocessing, the dataset size should be considered a relevant factor in the prediction outcome. The dataset for *E. coli* was collected from three GEO platforms with a different number of samples (see Materials and methods): GPL73 (12 GSM), GPL199 (759 GSM), and GPL3154 (1379 GSM) (Figure 3B and Supplementary Figure S7). We assessed the predictions using individual GEO platforms with the three preprocessing levels as input (Figure 3C and Supplementary Figure S8). In general, there is an improvement in the prediction scores for larger datasets. The scores with GPL199 and GPL3154 are considerably higher than the score for the smallest platform (GPL73). However, there is not a remarkable difference between GPL199 and GPL3154 with RMA and R-B normalization. In the case of raw data, it seems to be an improvement as the data size increases. From these results, we can conclude that the larger the dataset the better the predictions. However, previous studies have shown that not only the dataset size but also the variability of conditions are relevant factors for network inference (Sastry et al., 2019). This is evident with the smallest platform which seems to have less heterogeneity among the platforms. In contrast, the other two platforms have better results alone than together which suggests that both have redundant information. Otherwise, normalized datasets with a size of two orders of magnitude would be good enough for network inference. These results are consistent across the three tool groups.

A network-type-driven selective community is the best choice when a GS is not available

A previous DREAM challenge suggested that integrating multiple single-tool predictions into a community network is a safe choice, especially when there is no partial network to use as GS (Marbach et al., 2012). Even though the AUPR and AUROC tend to be constrained to higher values as more single-tool predictions are

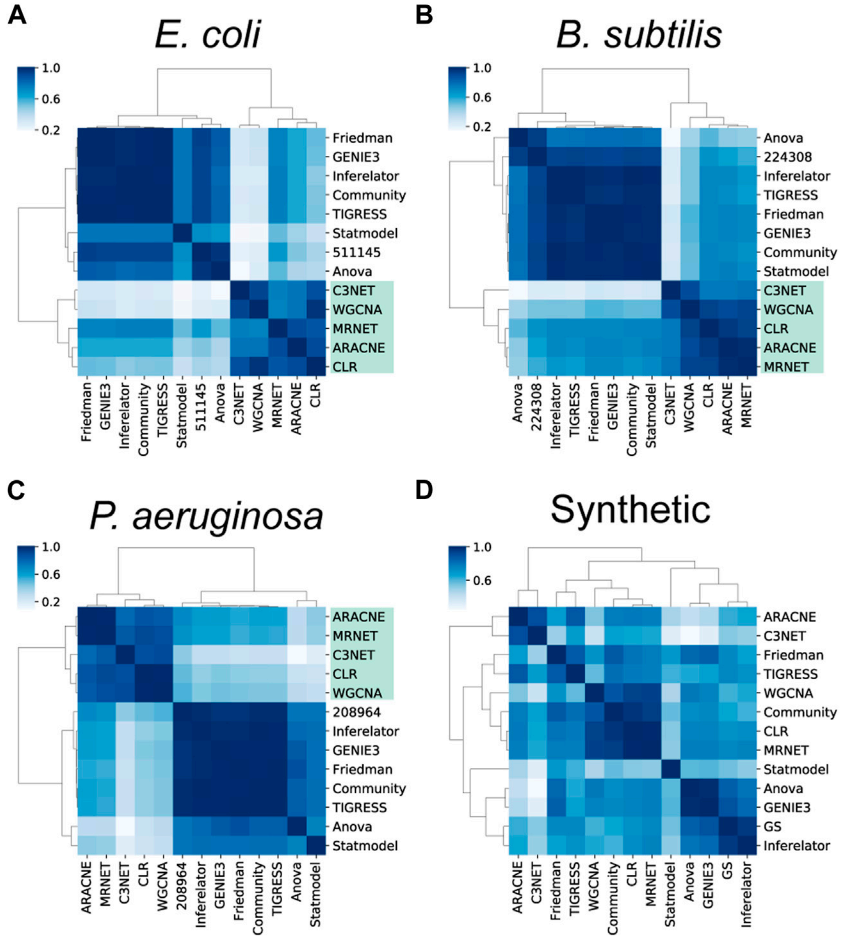


FIGURE 5
Assessment of the structural properties. Clustering of the global structural properties suggests there is a clear structural difference between causal (CAUS) and non-causal (COEX) networks. In contrast to the inferences with biological data **(A–C)**, most networks inferred from synthetic data **(D)** are more similar to each other.

integrated (Supplementary Figure S9), the probability of CAUS tools outperforming Community decreases when their predictions are merged with other single-tool predictions (Figure 4A and Supplementary Figure S10). This is due to the poor predictive power of some tools, which perform better only when integrated with several other predictions (e.g., ANOVA). The beginning of the prediction list is critical for the performance of the tools (Marbach et al., 2010). While COEX tools tend to have their true positive interactions scattered throughout the entire prediction, CAUS tools include most of their true positive interactions from the beginning (Supplementary Figure S11).

COEX tools capture function-specific regulons and non-direct interactions

We assessed the predictions with snapshots of the historical reconstruction of the *E. coli* GS, each of these networks with two versions; one with all the interactions discovered at a specific timepoint ("all") and the other one with only validated protein-DNA interactions ("strong"). The assessment methodology showed robustness to the incompleteness of the GS (Figure 4B), suggesting that CAUS tools outperform COEX tools with every snapshot of the GS, disregarding its completeness level. Moreover, even though all the tools improved the performance with the "all" GS, the difference is bigger for COEXP tools (Figure 4C). While the "strong" GS only contains direct TF-DNA interactions, the "all" GS may contain non-direct interactions (i.e., an interaction mediated by a third biological entity) (Escorcia-Rodriguez et al., 2020). Gene expression data capture both direct and non-direct regulatory events. Therefore, inference tools based solely on gene expression data tend to also infer non-direct interactions, especially COEX tools (Figure 4C). Perhaps, this consideration may shed light on the search for consistency between GRNs and gene expression data (Larsen et al., 2019; Parise et al., 2021). On the other hand, every tool performs better with the "strong" GS on AUROC (Supplementary Figure S12), but this is because of the highly unbalanced positives/negatives ratio (Saito and Rehmsmeier, 2015).

We assessed the predictions at the regulon level using the F1 score. The CAUS tools performed better on large regulons (i.e., those of global regulators) (Supplementary Figure S13). On the other hand, the COEX tools are the best alternative for local regulators, which are associated with function-specific regulons (Freyre-Gonzalez et al., 2022). To discard potential bias induced by the F1 metric (Chicco and Jurman, 2020), we also used Matthew's correlation coefficient (MCC), obtaining consistent but less meaningful patterns (Figure 4D). The explanation for this is that COEX tools distribute the interactions among all the genes sub-estimating the number of TGs for global regulators, while CAUS and HYBR tools distribute the interactions only among the TFs list provided over-estimating the number of TGs for each TFs, especially for local TFs (Supplementary Figure S14).

Unsupervised learning with global structural properties segregates COEX inferences from the rest of the networks

Beyond assessing the tools solely based on the standard statistical metrics, we analyzed global structural differences among the networks. We computed the following structural properties for the regulatory networks: density, number of regulators, maximum out-connectivity, feedforward and complex feedforward circuits (Alon, 2007; Freyre-Gonzalez and Tauch, 2017), 3-feedback loops, size of the giant component, average clustering coefficient, diameter, average shortest path length, and the coefficient of determination for the degree P(k) and clustering coefficient distribution C(k) (Albert, 2005). See Supplementary Table S2 for the definition of the structural properties. Then, we clustered the networks based on their normalized global structural properties (Materials and methods).

For the *E. coli* networks, COEX tools were clustered into one group (Figure 5A). On the other hand, CAUS and HYBR tools were clustered into a second group, excluding ANOVA. Even though the GS was not clustered with any of the two major groups, it was closer to the latter one (Figure 5A). We obtained similar results with the networks for *B. subtilis* (Figure 5B) and *P. aeruginosa* (Figure 5C), although the GS for *B. subtilis* got much closer to the CAUS and HYBR group (Figure 5B).

The clusters were not conserved with synthetic data inferences, suggesting that inferences with synthetic data were structurally similar disregarding their type of interactions (Figure 5D). Contrary to biological data, GeneNetWeaver generates the synthetic data following the topology of the network provided (Schaffter et al., 2011), making it easier for the tools to recover such topology. Several structural properties are constrained by the graph complexity and characterize the GRNs with causal interactions, despite different network completeness levels (Campos and Freyre-Gonzalez, 2019; Escorcia-Rodriguez et al., 2021). Therefore, we expect such properties to remain similar in the final GS, and the overall topological assessment of the predicted networks will be like the one performed with the current GS.

We then used an in-house Python implementation of the previously reported *D*-value (Schieber et al., 2017), which assesses network similarity based on topological evidence taking centrality into account. For the biological datasets, CAUS tools were always clustered with Community and Friedman but never with the GS (Supplementary Figure S15). Noteworthy, the GS was not clustered with the COEX tools either. Instead, it was isolated, as well as the ANOVA network. Overall, the results remain consistent across organisms, clustering CAUS networks apart from the COEX ones. Further topological analysis with all historical GS for *E. coli* showed that, despite GS completeness, the same conclusions are expected (Supplementary Figure S16).

Noteworthy, two networks might have identical global structures with no intersection between their regulations (shuffled node labels). This explains why ANOVA was repeatedly clustered with the GS, despite its poor performance with standard assessment metrics. However, between the two strategies to assess the structure of the networks, the one based on the normalized structural properties in GRNs (Figure 5) is more consistent with the standard metrics. We suggest using this approach as a complementary assessment always a GS is available. When no GS exists for the organism of interest, the structural assessment can be used along with other biological networks and random models to prove the prediction is structurally more similar to a biological network than random. We recently used this approach to assess network inferences for *Rhizobium etli* (Taboada-Castro et al., 2022).

Analyzing the structural properties individually (Supplementary Table S2 and Supplementary Figure S17), COEX tools have higher density and fraction of regulators. Given that the predictions have the same number of interactions, having a higher fraction of regulators results in lower max out-connectivity. On the other hand, synthetic predictions tend to have higher max out-connectivity values than their biological counterparts. Noteworthy, the max out-connectivity for the *P. aeruginosa* GS might be underestimated due to low genomic coverage (Escorcia-Rodriguez et al., 2020). Regarding normalized path-related properties, the COEX tools have the largest normalized diameter and average path length due to their small fraction of nodes in their giant component (Supplementary Table S2). Contrary to COEX tools that reach more than 200 components, CAUS and HYBR tools predict networks with a few components (Supplementary Figure S18) because their maximum is constrained to the number of TF-coding genes; and every interaction connecting regulons decreases the number of components. A high P(k) coefficient of determination (R^2) value was found across all biological predictions and all GSs. The C(k) R^2 was good only for COEX and HYBR biological predictions suggesting their modularity, like the one found in the GS. Regarding network motifs, the COEX inferences were the most similar to the GS. This agrees with the motifs search in DREAM5 where feed-forward loops were recovered most reliably by mutual-information and correlation-based methods (Marbach et al., 2012) (i.e., COEX tools).

GENIE3 outperformed tools developed for bulk RNA-seq

We interrogated the performance dependence of GRNs inference related to transcriptomic technique, comparing two COEX inference tools developed exclusively for bulk RNA-seq data (RNAseqNet (Imbert et al., 2018) and LSTrAP (Proost et al., 2017)) and the best CAUS microarray-based approach (GENIE3). We retrieved RNA-seq datasets for *E. coli* and performed a cross-evaluation between the tools, exchanging the input data. First, we used a subset (see Materials and methods) of our raw and RMA microarray datasets of *E. coli* to reduce the impact of data size variation and observed that GENIE3 outperformed RNASeqNet and LSTrAP significantly (Supplementary Figure S19). Next, we used the RNA-seq datasets (raw counts, normalized with DESeq2, and PRECISE (Sastry et al., 2019)) as input. The COEX RNA-seq-based tools performed better with the homogenous largest RNA-seq dataset, PRECISE (Supplementary Figure S19). Despite the improvement of RNASeqNet and LSTrAP with the RNA-seq data, GENIE3 still performed better (Supplementary Figure S19). These results agree with a previous synthetic gold standard-based benchmarking of network inference methods for scRNA-seq data where GENIE3 is still placed within the top-performing tools methods (Pratapa et al., 2020), making GENIE3 a top-performing tool regardless of the transcriptomic technique.

Furthermore, we assess the predictions based on their global structure (Supplementary Figure S20). We only considered the inferences datasets with the best MCC scores (Supplementary Figure S19), PRECISE for RNA-seq, and raw for microarray data. Both datasets and metrics showed consistent results clustering the GS with GENIE3, RNAseqNet, and Community, leaving LSTrAP out (Supplementary Figure S20). This suggests that RNAseqNet infers networks with structural properties more similar to the GS than LSTrAP does. However, non-ranked interactions might be a shortcoming for RNAseqNet.

Overall, compared to how well the tools performed with microarray data, RNA-seq data did not significantly improve their performance. It agrees with a previous assessment in *A. Thaliana,* where networks derived from simple correlations and microarray data obtained higher scores than inferences with RNA-seq data (Giorgi et al., 2013). Although RNA-seq has progressively replaced microarrays (Lowe et al., 2017), the gene coverage referred to as an advantage of RNA-seq, is less of a problem for microarrays in model prokaryotes where new microarrays have overcome the coverage issue (Swarbreck et al., 2008). Despite the amount of available RNA-seq data, most organisms do not have an appropriate annotation (Salzberg, 2019), while large microarray-based transcriptomic data have continuously grown into public databases (Barrett et al., 2013; Athar et al., 2019).

Conclusions and guidelines

All the CAUS tools (GENIE3, TIGRESS, Inferelator, and Statmodel) outperformed the COEX tools when assessed with a GRN as the GS (TF–TG interactions) with biological and synthetic data and, taking advantage of a TFs list. Even though we filtered TF-TG interactions from the co-expression inferences approaches (HYBR), the performance of CAUS tools was still better. GENIE3 and Inferelator performed the best for synthetic and biological data. GENIE3 also outperformed inference tools developed for bulk RNA-seq data. COEX tools performed better when assessed with a GS resembling a co-regulation network (interactions among co-regulated genes). Regarding time-series tools, scanBMA performed the best, although it is highly affected by dataset size.

Larger datasets result in better predictions but require a selective inferences-integration process and batch correction to mitigate technical variability; applying only RMA worsened the predictions. The probability of CAUS tools outperforming Community decreases as more tools are integrated into a community network, suggesting the use of a selective community based on the desired output network type (co-regulation or GRN). Although CAUS tools are the best alternative to infer global GRNs, COEX tools are better at inferring regulons for function-specific (i.e., local) TFs. An assessment against a GS including potential indirect interactions suggests that COEX tools might be the best alternative to identify indirect regulations. This is useful when the goal is to identify all the regulators affecting the expression of a gene, and not only DNA-binding TFs (Zorro-Aranda et al., 2022).

Based on global structural properties, COEX tools segregate from CAUS tools when using biological predictions, highlighting the differences among their output network type. Individual structural properties support the similarity between CAUS inferences and the GRNs used as GS. However, no clear clusters were found with synthetic data, suggesting that biological data is required for the structural assessment because synthetic data generation is based on the topology of the input network (Schaffter et al., 2011). Historical snapshots of the GS suggest the statistical and structural assessment to be robust to GS incompleteness.

The overall modest performance of the tools is evident and the potential inherent pitfalls to the conjecture that statistical relationships between expression profiles correspond to regulatory interactions have been previously noted (Pratapa et al., 2020; Freyre-Gonzalez et al., 2022). Recent works leveraging prior networks, structural constraints, and sequence motifs to improve transcriptomic-based GRN inference have shown promising results (Castro et al., 2019; Lim et al., 2022; Zorro-Aranda et al., 2022). Following we provide guidelines for the inference and assessment:

Inference

- Identify the best kind of tool for your purposes.
  - CAUS or Community for whole GRNs or regulons of global TFs.
  - COEX for regulons of local TFs (few targets), co-expression, or co-regulation networks.
  - Using a list of TFs to filter inferences based on co-expression (e.g., ANOVA and Friedman) to get a causal network is not enough to infer a global GRN. Integrate the TFs into the inference pipeline.
- A selective community based on the type of network required is better than an all-inclusive community.
- If you want to use one COEX tool, use C3NET but keep in mind you will obtain a co-expression network, not a GRN.
- If you want to use one CAUS tool, use GENIE3 disregarding the type of gene expression data used.
- Merge datasets only when the final size of the data outweighs the noise of merging different sources.
- In prokaryotes, dataset size and preprocessing are more important than the transcriptomic technique used to generate the data.
- Normalize your data using Batch correction if it is necessary. Using only RMA is not recommended.
- If it is feasible, take advantage of biological information such as a list of TFs.

Assessment

- Using synthetic data to assess the predictions might provide insights about the performance of the tools but expect it to worsen when assessed with biological data and the inferred networks to have a different global structure.
- Take advantage of several experimentally-validated bacterial GRNs to be used as GS (e.g., https://abasy.ccg.unam.mx/for bacteria).
- Whenever possible, use historical snapshots or network sampling to prove the results hold despite GRN incompleteness.
- Use MCC to perform a regulon-level assessment of the predictions.
- Compare network structural properties to assess the global topology of the networks inferred from biological data.
- A structural assessment of the predictions applies to biological data only. Because of the mechanisms to generate the data following the topology of an input network, predictions with synthetic data have a similar structure despite inherent differences.

Materials and methods

Selection of GRN prediction methods to be assessed

We thoroughly reviewed the literature and selected methods that were able to infer a GRN from an expression data matrix. We also considered usability, which takes into account 1) open-source availability, 2) enough documentation, and 3) the ability to be run by a command line.

Synthetic data

The synthetic datasets, all with 788 conditions (rows) and 197 genes (columns), were generated using GeneNetWeaver software (Schaffter et al., 2011) applying the standard procedure reported by the DREAM5 consortium, with the *E. coli* network (511145_v2017_sRDB16_eStrong) from Abasy Atlas (Escorcia-Rodriguez et al., 2020) being used as the seed. To explore the effects of noise levels in GRN inference, we generated datasets with 20%-step values for the noise parameter, as well as the 5% noise level selected for the DREAM5 challenge. To study the effect of sample size in GRN inference, we sampled each of the previous datasets at 10, 25, 50, 75, and 100% of experimental conditions, preserving an equal representation of each experimental condition. The same procedure was followed for time-series 4,207 conditions and 197 genes data generation.

Microarray data extraction and processing

The microarray data for *Escherichia coli* K-12 MG1655, *Bacillus subtilis* 168, and the pathogen *Pseudomonas aeruginosa* PAO1 were retrieved from the (GEO) database using four main inclusion criteria: A) records were associated with public Affymetrix platforms and had an available CEL file useful to perform Robust Multi-chip Averaging (RMA) normalization by Oligo R package (array annotation package); B) an available GEO Series Matrix, an expression matrix annotated as non-normalized data, referred here as raw data, and C) more than one available sample. In addition, we excluded GEO samples related to more than one organism. For *E. coli*, a total of 2,154 GEO samples (GSM) from 182 GEO Series (GSE) were retrieved from the GEO platforms GPL73 (1 GSE, 12 GSM), GPL199 (33 GSE, 759 GSM), and GPL3154 (153 GSE, 1379 GSM). After applying RMA, we kept with the shared genes among *E. coli* GPLs belonging to the K-12 MG1655 strain, obtaining a total of 4,003 genes, which comprise 87.7% of the genome. For *B. subtilis* we used the platform GPL343 and retrieved 7 GSE with a total of 64 GSM, obtaining a total of 4,010 genes, which comprises 88.5% of the genome. Finally, for *P. aeruginosa* we used the GPL84 platform with 125 GSE and a total of 1133 GSM, obtaining a total of 5,548 genes, which comprise de 97.4% of the genome. Microarray raw data (CEL files) were normalized through the RMA implementation in the R package *oligo* (Carvalho and Irizarry, 2010), using default parameters. Next, we removed all the conditions in which NANs or zeros were present due to normalization effects. Lastly, we performed a batch-effect correction using ComBat (Johnson et al., 2007) implementation in the sva R package with a non-parametric adjustments approach (function from the sva R package using the following parameters: mod = NULL, par.prior = FALSE, mean.only = FALSE).

Time-series microarray data and condition sampling

Since GEO does not provide a feasible way to filter TS experiments, we used all public metadata of samples to identify GSE records with a timeline progression and filtered them with our inclusion criteria. From the identified TS GSE list we selected the largest record for each organism. For *E. coli we* used GSE12411 and retrieved 28 GSM with three time-points with 4, 12, and 12 replicates respectively, regarding *P. aeruginosa* we used GSE52445 with 28 GSM representing 14 time points each one with two replicates. For the assessment, we used only raw and RMA preprocessed data, the batch correction step was not necessary for the one-source samples.

We sampled the Abasy Atlas networks (Escorcia-Rodriguez et al., 2020) to allow dimensionality reduction by the Bayesian tools (Annest et al., 2009; Young et al., 2014). We sampled the networks 511145_v2018_sRDB18_eStrong for *E. coli* and 208964_v2015_s11-RTB13 for *P. aeruginosa*. We applied snowball sampling (Heckathorn and Cameron, 2017), also known as link-tracing, using the network nodes with the highest degree of centrality as seed and 198 as the cutoff value for the sampling to get the same size of data as in the *in silico* time-series assessment. The final sample sizes were 139 samples x 198 genes for *E. coli* and 45 samples x 198 genes for *P. aeruginosa*. We used 198 genes for consistency with the time series synthetic data.

Data collection, and assessment for cross-evaluation

To compare the performance dependency of the RNA-seq-based and microarray-based inference methods, we swap their transcriptomic input data and compare it with the original correspondence input results. Due to the diversity of RNA-seq-based methods, we preselected LSTrAP, RNAseqNet, and VCNet exclusively developed for GRN inference from bulk RNA-seq. However, we excluded VCNet from the analysis since it cannot be applied to a large number of samples unless you optimize the computational complexity inherent in its loop-based code. On the other hand, RNAseqNet and LsTrAP are low-time-consuming algorithms that aim to increase the reliability of inference from biologically related datasets (Imbert et al., 2018).

Bulk RNA-seq data extraction and processing

We collected two bulk RNA-seq datasets for *E. coli* K-12 MG1655. The small one was retrieved from GEO NCBI (GSE73673) (Kim et al., 2016), we downloaded the 87 sample files with the processed reads (Kim et al., 2016) for 3,923 genes. Next, we applied the DESeq2 normalization (Love et al., 2014) a commonly used method that has been evaluated against different normalization methods (Dillies et al., 2013; Maza et al., 2013; Soneson and Delorenzi, 2013; Smid et al., 2018). For the largest one, we download the available processed (log_tpm.tsv) dataset from PRECISE 1.0 (Sastry et al., 2019), a Precision RNA-seq Expression Compendium for Independent Signal Exploration, build it with 15 studies derived with a standardized protocol from the same research group and PRECISE developer. We only kept with the genes shared between the PRECISE dataset and our microarray dataset resulting in 278 conditions and 3,557 genes.

Microarray data transformation

We sampled a subset of 87 samples from our collected *E. coli* microarray dataset. We used only the raw and RMA version since batch correction was not applicable. Unfortunately, the RNAseqNet algorithm takes as input read counts or TMM normalized counts data; thus, we avoided negative values from sampling. To the best of our knowledge, RNAseqNet is not able to work with microarray or RNA-set datasets without filter genes with at least 70% of sample coverage.

Gold standards

We used strongly-supported, meta-curated GRNs from Abasy Atlas v2.2 (Escorcia-Rodriguez et al., 2020) as GSs for *E. coli* (511145_v2018_sRDB18_eStrong), *B. subtilis* (224308_v2008_sDBTBS08_eStrong) and *P. aeruginosa* (208964_v2015_s11-RTB13). The nodes of Abasy GRNs depict either genes, regulatory sRNAs, or regulatory protein complexes. For this work, we converted networks with genes and regulatory protein complexes into gene-gene networks to use as GS since only those interactions can be inferred. We removed the genes for which no expression data was retrieved since the prediction of its interactions would not be inferred by the methods assessed in this work. We obtained a total of 4,075 interactions among 1780 genes for *E. coli*, 2294 interactions among 1,298 genes for *B. subtilis*, and 1,297 interactions among 868 genes for *P. aeruginosa*. For GS incompleteness analysis, we also retrieved from Abasy various public versions of the *E. coli* GRN (hereafter referred to as historical snapshots), with different completeness levels.

For the construction of the GS with interactions between co-regulated genes, we connected each regulon of 511145_v2018_sRDB18_eStrong so each of them forms a clique and obtained a total of 737,913 interactions between the same number of genes, overestimating the density of the network. Note that, in such network representation, the TGs from a regulon formed a clique, including the TF only if it regulates its own transcription. For the synthetic GS, we used 511145_v2017_sRDB16_eStrong as input for GeneNetWeaver (Schaffter et al., 2011) to generate datasets with 5, 20, 40, 60, 80, and 100% noise variations. From such datasets, we generated subsamples with 20, 25, 50, 75, and 100% completeness.

Integration of individual predictions into a community network

A confidence score provided by each tool (when available) was used to rank predictions and missing interactions were ranked right after the last predicted one. Therefore, longer predictions penalize more the missing interactions. Inferred interactions sharing a common score by a method were ranked equally. The average rank is used as a score for the Community. For biological data, predictions were previously trimmed to the number of interactions that the complete organism-specific GRN would have according to previous work (Campos and Freyre-Gonzalez, 2019; Escorcia-Rodriguez et al., 2020). Those values correspond to 12,000 for *E. coli* and *B. subtilis* and 16,000 for *P. aeruginosa*.

Assessment

Unless otherwise described in the analysis, network predictions larger than the expected number of interactions in the complete GRN were trimmed (Campos and Freyre-Gonzalez, 2019; Escorcia-Rodriguez et al., 2020). The first 12,000 inferred interactions were considered for the assessment with *E. coli* and *B. subtilis* and the first 16,000 inferred interactions for *P. aeruginosa*. Interactions shaping the GS were used as the positive set (P), while interactions absent in the GS were labeled as the negative set (N). Inferred interactions were considered True Positive (TP) if they were present in the GS and False positive (FP) if otherwise. Interactions in the GS that were not recovered by the algorithm were considered False Negative (FN). The Area Under Receiver Operating Characteristics (AUROC) and Area Under Precision-Recall (AUPR) curves were used to assess the predictions. While AUROC represents the specificity (FP/N) and the sensitivity

(TP/P) of the prediction compared with the whole set of potential interactions, AUPR focuses on the list of predictions and its precision (TP/(TP + FP)) as well as the sensitivity of the algorithm. We select PR as the main assessment measure, due to the imbalance between positive and negative sets (Saito and Rehmsmeier, 2015). For the overall score, we used the average score for the complete dataset with 5% of noise for the synthetic data and scores obtained with RMA plus batch effect correction for biological data. For the study of the effect of GS incompleteness, we used each historical snapshot of the *E. coli* GRNs as the GS. Inferred interactions sharing the same score were considered as equally ranked by the method and genes present neither in the GS nor in the expression data were not considered for this assessment. For the assessment of predictions not providing a score for each interaction, we used the MCC which is the best-suited coefficient for imbalanced datasets (Boughorbel et al., 2017). Note that MCC was used only for the comparative assessment between GENIE3 and RNAseqNet, and the regulon-level assessment. RNAseqNet does not score the predictions. Therefore, we considered the first 12,000 interactions to assess its prediction with MCC so the ranking of the interactions does not impact the score. Note that this is not ideal as the true positives–as well as novel interactions–may be at the bottom of the prediction making it disadvantageous for the experimental validation of such inferred interactions. For the regulon-level assessment, we trimmed the predictions to the expected number of interactions once the corresponding network is completed and compared each of the regulons against the cognate regulon in the GS using MCC and F1 score. The scores were plotting against the normalized out-connectivity.

For the prediction of the COEX methods, we duplicated every interaction in the prediction list, changing the direction. This is because the outputs provide interactions between two genes with no direction (e.g., symmetric adjacency matrix). Given the nature of the assessment with a directed network as the GS, we considered every interaction to be in both directions. While this would increase the number of predictions, consideration of the direction is required. On the other hand, for the assessment of the predictions with a co-regulation GS, we did not consider the direction of the interactions. This way, the direction of the interactions predicted by a CAUS method, was not considered.

Combinatorial

We constructed selective communities with the possible combinations of the 12 methods used in the assessment with biological data. We use the dataset normalized with RMA and batch correction for the three organisms. To measure the effect of each method on the community network, we computed the dominance score defined as the probability of a selective community network with a given tool outperforming the all-inclusive community network:

$$dominance = \frac{freq\left(AUC^{Tool} > AUC^{comm}\right)}{maxT}$$

$$maxT = \frac{(n-1)!}{(r-1)!\,(n-r)!}$$

Where $maxT$ is the theoretical maximum of selective communities with each tool, n is the number of methods (12) used for the combinatory, and r is the number of elements in the combinatory [2–11].

Structural properties

We computed several structural properties for GRNs at a global scale and normalized them as follows: Regulators, self-regulations, maximum out-connectivity, and giant component size were normalized by the network size (number of nodes). Density was used as its product with the fraction of nodes acting as regulators. Network diameter was normalized by the number of nodes—2 (as if no shortcuts would exist). Network motifs were normalized by the number of potential motifs in the network, defined as:

$$\frac{n!}{(n-r)!}\cdot\left(\frac{TF_n}{n}\right)^{TF_m}$$

Where n is the size of the network, r is the number of nodes in the motif ($r = 3$), TF_n is the number of TFs in the network, and TF_m is the number of TFs required for each type of motif ($TF_m = 2$ for feedforward and complex feedforward loops,; $TF_m = 3$ for 3-feedback loops). We scaled the property values across networks between 0 and 1. We clustered networks and properties using Ward's method. Further, we used pairwise Pearson correlation for the network property vectors and clustered them according to the Euclidean distance using Ward's method.

We used an in-house Python implementation of the dissimilarity measure proposed by Schieber et al. (2017) to quantify the differences in the structural topology between two networks considering global structural properties, node-level structural properties, and centrality. We used the parameters the authors recommend (0.45, 0.45, 0.1) and applied used to compare the networks pairwise. The dissimilarity matrix was clustered using Pearson and ward's method.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

JE-R. Investigation, Methodology, Supervision, Visualization, Writing-original draft, Writing-review and editing. EG-N Investigation, Visualization, Writing-original draft, Writing-review and editing. EH-B Investigation, Visualization, Writing-original draft, Writing-review and editing. AZ-A. Investigation, Writing-original draft, Writing-review and editing. MT-P. Investigation and. Writing-original draft. JF-G. Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing-review, and editing.

Funding

This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) IN202421 to JF-G.

Acknowledgments

We thank Roberto Olayo-Alarcón and Luis F. Altamirano-Pacheco for technical support in running some of the inference tools in a previous version of this work. JE-R is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM). He received fellowship 959406 from CONACYT. AZ-A. received a doctoral scholarship [call 647 (2014)] from the Colombian Administrative Department of Science, Technology, and Innovation (COLCIENCIAS).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1143382/full#supplementary-material

References

Aibar, S., Gonzalez-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., et al. (2017). Scenic: Single-cell regulatory network inference and clustering. *Nat. Methods* 14 (11), 1083–1086. doi:10.1038/nmeth.4463

Akesson, J., Lubovac-Pilav, Z., Magnusson, R., and Gustafsson, M. (2021). ComHub: Community predictions of hubs in gene regulatory networks. *BMC Bioinforma.* 22 (1), 58. doi:10.1186/s12859-021-03987-y

Albert, R. (2005). Scale-free networks in cell biology. *J. Cell. Sci.* 118 (21), 4947–4957. doi:10.1242/jcs.02714

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nat. Rev. Genet.* 8 (6), 450–461. doi:10.1038/nrg2102

Altay, G., and Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.* 4, 132. doi:10.1186/1752-0509-4-132

Annest, A., Bumgarner, R. E., Raftery, A. E., and Yeung, K. Y. (2009). Iterative bayesian model averaging: A method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinforma.* 10, 72. doi:10.1186/1471-2105-10-72

Athar, A., Fullgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* 47 (D1), D711–D715. doi:10.1093/nar/gky964

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. Database issue. doi:10.1093/nar/gks1193

Bellot, P., Olsen, C., Salembier, P., Oliveras-Verges, A., and Meyer, P. E. (2015). NetBenchmark: A bioconductor package for reproducible benchmarks of gene regulatory network inference. *BMC Bioinforma.* 16, 312. doi:10.1186/s12859-015-0728-4

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., et al. (2006). The inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* 7 (5), R36. doi:10.1186/gb-2006-7-5-r36

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* 12 (6), e0177678. doi:10.1371/journal.pone.0177678

Campos, A. I., and Freyre-Gonzalez, J. A. (2019). Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. *Sci. Rep.* 9 (1), 3618. doi:10.1038/s41598-019-39866-z

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics* 26 (19), 2363–2367. doi:10.1093/bioinformatics/btq431

Castro, D. M., de Veaux, N. R., Miraldi, E. R., and Bonneau, R. (2019). Multi-study inference of regulatory networks for more accurate models of gene regulation. *PLoS Comput. Biol.* 15 (1), e1006591. doi:10.1371/journal.pcbi.1006591

Chen, G., and Liu, Z. P. (2022). Inferring causal gene regulatory network via GreyNet: From dynamic grey association to causation. *Front. Bioeng. Biotechnol.* 10, 954610. doi:10.3389/fbioe.2022.954610

Chen, S., and Mar, J. C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinforma.* 19 (1), 232. doi:10.1186/s12859-018-2217-z

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21 (1), 6. doi:10.1186/s12864-019-6413-7

De Smet, R., and Marchal, K. (2010). Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8 (10), 717–729. doi:10.1038/nrmicro2419

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform* 14 (6), 671–683. doi:10.1093/bib/bbs046

Escorcia-Rodriguez, J. M., Tauch, A., and Freyre-Gonzalez, J. A. (2020). Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization. *Comput. Struct. Biotechnol. J.* 18, 1228–1237. doi:10.1016/j.csbj.2020.05.015

Escorcia-Rodriguez, J. M., Tauch, A., and Freyre-Gonzalez, J. A. (2021). Corynebacterium glutamicum regulation beyond transcription: Organizing principles and reconstruction of an extended regulatory network incorporating regulations mediated by small RNA and protein-protein interactions. *Microorganisms* 9 (7), 1395. doi:10.3390/microorganisms9071395

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform* 19 (5), 776–792. doi:10.1093/bib/bbx008

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., et al. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* 5 (1), e8. doi:10.1371/journal.pbio.0050008

Freyre-Gonzalez, J. A., Alonso-Pavon, J. A., Trevino-Quintanilla, L. G., and Collado-Vides, J. (2008). Functional architecture of *Escherichia coli*: New insights provided by a natural decomposition approach. *Genome Biol.* 9 (10), R154. doi:10.1186/gb-2008-9-10-r154

Freyre-Gonzalez, J. A., Escorcia-Rodriguez, J. M., Gutierrez-Mondragon, L. F., Marti-Vertiz, J., Torres-Franco, C. N., and Zorro-Aranda, A. (2022). System principles governing the organization, architecture, dynamics, and evolution of gene regulatory networks. *Front. Bioeng. Biotechnol.* 10, 888732. doi:10.3389/fbioe.2022.888732

Freyre-Gonzalez, J. A., and Tauch, A. (2017). Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: Novel insights from a dataset with a high genomic coverage. *J. Biotechnol.* 257, 199–210. doi:10.1016/j.jbiotec.2016.10.025

Freyre-Gonzalez, J. A., Trevino-Quintanilla, L. G., Valtierra-Gutierrez, I. A., Gutierrez-Rios, R. M., and Alonso-Pavon, J. A. (2012). Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and *Escherichia coli* unveiled by the natural decomposition approach. *J. Biotechnol.* 161 (3), 278–286. doi:10.1016/j.jbiotec.2012.03.028

Giorgi, F. M., Del Fabbro, C., and Licausi, F. (2013). Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis thaliana*. *Bioinformatics* 29 (6), 717–724. doi:10.1093/bioinformatics/btt053

Haury, A. C., Mordelet, F., Vera-Licona, P., and Vert, J. P. (2012). Tigress: Trustful inference of gene REgulation using stability selection. *BMC Syst. Biol.* 6, 145. doi:10.1186/1752-0509-6-145

Heckathorn, D. D., and Cameron, C. J. (2017). Network sampling: From snowball and multiplicity to respondent-driven sampling. *Annu. Rev. Sociol.* 43 (1), 101–119. doi:10.1146/annurev-soc-060116-053556

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: Data integration in dynamic models-a review. *Biosystems* 96 (1), 86–103. doi:10.1016/j.biosystems.2008.12.004

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5 (9), e12776. doi:10.1371/journal.pone.0012776

Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012). Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics* 28 (12), 1592–1597. doi:10.1093/bioinformatics/bts245

Imbert, A., Valsesia, A., Le Gall, C., Armenise, C., Lefebvre, G., Gourraud, P. A., et al. (2018). Multiple hot-deck imputation for network inference from RNA sequencing data. *Bioinformatics* 34 (10), 1726–1732. doi:10.1093/bioinformatics/btx819

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8 (1), 118–127. doi:10.1093/biostatistics/kxj037

Kim, M., Rai, N., Zorraquino, V., and Tagkopoulos, I. (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for *Escherichia coli*. *Nat. Commun.* 7, 13090. doi:10.1038/ncomms13090

Kuffner, R., Petri, T., Tavakkolkhah, P., Windhager, L., and Zimmer, R. (2012). Inferring gene regulatory networks by ANOVA. *Bioinformatics* 28 (10), 1376–1382. doi:10.1093/bioinformatics/bts143

Larsen, S. J., Rottger, R., Schmidt, H., and Baumbach, J. (2019). *E. coli* gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res.* 47 (1), 85–92. doi:10.1093/nar/gky1176

Lim, H. G., Rychel, K., Sastry, A. V., Bentley, G. J., Mueller, J., Schindel, H. S., et al. (2022). Machine-learning from Pseudomonas putida KT2440 transcriptomes reveals its transcriptional regulatory network. *Metab. Eng.* 72, 297–310. doi:10.1016/j.ymben.2022.04.004

Lo, K., Raftery, A. E., Dombek, K. M., Zhu, J., Schadt, E. E., Bumgarner, R. E., et al. (2012). Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Syst. Biol.* 6, 101. doi:10.1186/1752-0509-6-101

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. *PLoS Comput. Biol.* 13 (5), e1005457. doi:10.1371/journal.pcbi.1005457

Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9 (8), 796–804. doi:10.1038/nmeth.2016

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U. S. A.* 107 (14), 6286–6291. doi:10.1073/pnas.0913357107

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., et al. (2006). Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinforma.* 7 (1), S7. doi:10.1186/1471-2105-7-S1-S7

Maza, E., Frasse, P., Senin, P., Bouzayen, M., and Zouine, M. (2013). Comparison of normalization methods for differential gene expression analysis in RNA-seq experiments: A matter of relative size of studied transcriptomes. *Commun. Integr. Biol.* 6 (6), e25849. doi:10.4161/cib.25849

Meyer, P. E., Kontos, K., Lafitte, F., and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform Syst. Biol.* 2007, 79879. doi:10.1155/2007/79879

Michoel, T., De Smet, R., Joshi, A., Van de Peer, Y., and Marchal, K. (2009). Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks. *BMC Syst. Biol.* 3, 49. doi:10.1186/1752-0509-3-49

Parise, D., Parise, M. T. D., Kataka, E., Kato, R. B., List, M., Tauch, A., et al. (2021). On the consistency between gene expression and the gene regulatory network of Corynebacterium glutamicum. *Netw. Syst. Med.* 4 (1), 51–59. doi:10.1089/nsm.2020.0014

Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A., and Murali, T. M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* 17 (2), 147–154. doi:10.1038/s41592-019-0690-6

Proost, S., Krawczyk, A., and Mutwil, M. (2017). LSTrAP: Efficiently combining RNA sequencing data into co-expression networks. *BMC Bioinforma.* 18 (1), 444. doi:10.1186/s12859-017-1861-z

Saito, T., and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 10 (3), e0118432. doi:10.1371/journal.pone.0118432

Salleh, S. M., Mazzoni, G., Lovendahl, P., and Kadarmideen, H. N. (2018). Gene co-expression networks from RNA sequencing of dairy cattle identifies genes and pathways affecting feed efficiency. *BMC Bioinforma.* 19 (1), 513. doi:10.1186/s12859-018-2553-z

Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. *Genome Biol.* 20 (1), 92. doi:10.1186/s13059-019-1715-2

Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., et al. (2019). The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat. Commun.* 10 (1), 5536. doi:10.1038/s41467-019-13483-w

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* 27 (16), 2263–2270. doi:10.1093/bioinformatics/btr373

Schieber, T. A., Carpi, L., Diaz-Guilera, A., Pardalos, P. M., Masoller, C., and Ravetti, M. G. (2017). Quantification of network structural dissimilarities. *Nat. Commun.* 8, 13928. doi:10.1038/ncomms13928

Secilmis, D., Hillerton, T., Tjarnberg, A., Nelander, S., Nordling, T. E. M., and Sonnhammer, E. L. L. (2022). Knowledge of the perturbation design is essential for accurate gene regulatory network inference. *Sci. Rep.* 12 (1), 16531. doi:10.1038/s41598-022-19005-x

Sirbu, A., Ruskin, H. J., and Crane, M. (2010). Cross-platform microarray data normalisation for regulatory network inference. *PLoS One* 5 (11), e13822. doi:10.1371/journal.pone.0013822

Smid, M., Coebergh van den Braak, R. R. J., van de Werken, H. J. G., van Riet, J., van Galen, A., de Weerd, V., et al. (2018). Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. *BMC Bioinforma.* 19 (1), 236. doi:10.1186/s12859-018-2246-7

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinforma.* 14, 91. doi:10.1186/1471-2105-14-91

Stolovitzky, G., Prill, R. J., and Califano, A. (2009). Lessons from the DREAM2 challenges. *Ann. N. Y. Acad. Sci.* 1158, 159–195. doi:10.1111/j.1749-6632.2009.04497.x

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., et al. (2008). The Arabidopsis information resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.* 36, D1009–D1014. Database issue. doi:10.1093/nar/gkm965

Taboada-Castro, H., Gil, J., Gomez-Caudillo, L., Escorcia-Rodriguez, J. M., Freyre-Gonzalez, J. A., and Encarnacion-Guevara, S. (2022). Rhizobium etli CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network. *Front. Microbiol.* 13, 947678. doi:10.3389/fmicb.2022.947678

Van den Bulcke, T., Van Leemput, K., Naudts, B., van Remortel, P., Ma, H., Verschoren, A., et al. (2006). SynTReN: A generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinforma.* 7, 43. doi:10.1186/1471-2105-7-43

Young, W. C., Raftery, A. E., and Yeung, K. Y. (2014). Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Syst. Biol.* 8, 47. doi:10.1186/1752-0509-8-47

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi:10.2202/1544-6115.1128

Zhang, J., Nie, Q., Si, C., Wang, C., Chen, Y., Sun, W., et al. (2019). Weighted gene Co-expression network analysis for RNA-sequencing data of the varicose veins transcriptome. *Front. Physiol.* 10, 278. doi:10.3389/fphys.2019.00278

Zorro-Aranda, A., Escorcia-Rodriguez, J. M., Gonzalez-Kise, J. K., and Freyre-Gonzalez, J. A. (2022). Curation, inference, and assessment of a globally reconstructed gene regulatory network for Streptomyces coelicolor. *Sci. Rep.* 12 (1), 2840. doi:10.1038/s41598-022-06658-x

V. Curation, inference, and assessment of a globally reconstructed gene regulatory network for *Streptomyces coelicolor*

(Zorro-Aranda et al., 2022)

scientific reports

OPEN

Curation, inference, and assessment of a globally reconstructed gene regulatory network for *Streptomyces coelicolor*

Andrea Zorro-Aranda[1,3], Juan Miguel Escorcia-Rodríguez[1], José Kenyi González-Kise[1,2] & Julio Augusto Freyre-González[1]✉

Streptomyces coelicolor **A3(2) is a model microorganism for the study of Streptomycetes, antibiotic production, and secondary metabolism in general. Even though *S. coelicolor* has an outstanding variety of regulators among bacteria, little effort to globally study its transcription has been made. We manually curated 29 years of literature and databases to assemble a meta-curated experimentally-validated gene regulatory network (GRN) with 5386 genes and 9707 regulatory interactions (~ 41% of the total expected interactions). This provides the most extensive and up-to-date reconstruction available for the regulatory circuitry of this organism. Only ~ 6% (534/9707) are supported by experiments confirming the binding of the transcription factor to the upstream region of the target gene, the so-called "strong" evidence. While for the remaining interactions there is no confirmation of direct binding. To tackle network incompleteness, we performed network inference using several methods (including two proposed here) for motif identification in DNA sequences and GRN inference from transcriptomics. Further, we contrasted the structural properties and functional architecture of the networks to assess the reliability of the predictions, finding the inference from DNA sequence data to be the most trustworthy approach. Finally, we show two applications of the inferred and the curated networks. The inference allowed us to propose novel transcription factors for the key *Streptomyces* antibiotic regulatory proteins (SARPs). The curated network allowed us to study the conservation of the system-level components between *S. coelicolor* and *Corynebacterium glutamicum*. There we identified the basal machinery as the common signature between the two organisms. The curated networks were deposited in Abasy Atlas (https://abasy.ccg.unam.mx/) while the inferences are available as Supplementary Material.**

Streptomycetes, the largest genus within the actinomycetes, are biotechnologically relevant organisms. They produce around half of the natural antibiotics in current use[1]. However, according to the analysis of genome mining, less than 10% of antibiotics that could be produced by actinomycetes are currently used[2]. Their production could be enhanced not only by experimental technologies such as genetic manipulation but also by a deeper knowledge of their secondary metabolism and transcriptional regulation. *Streptomyces coelicolor* A3(2) has become the model microorganism for the study of antibiotic production and secondary metabolism in general[3]. Before its sequencing, it was already known that *S. coelicolor* produces the red-pigmented antibiotic undecylprodigiosin (RED), the blue-pigmented actinorhodin (ACT), and the calcium-dependent antibiotic (CDA). However, its sequencing revealed more than 20 biosynthetic gene clusters (BGCs). Most of the metabolites produced by these clusters and their regulation are still unknown[4].

S. coelicolor secondary metabolism regulation is very complex. It is controlled by a network of regulators at many levels, from global to cluster situated regulators (CSRs). Most CSRs control their own BGC, however, some of them can bind to multiple BGCs causing a cross-cluster regulation[4]. Sequencing of *S. coelicolor* A3(2) revealed

[1]Regulatory Systems Biology Research Group, Laboratory of Systems and Synthetic Biology, Center for Genomics Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n, Col. Chamilpa, 62210 Cuernavaca, Morelos, México. [2]Undergraduate Program in Genomic Sciences, Center for Genomics Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n, Col. Chamilpa, 62210 Cuernavaca, Morelos, México. [3]Bioprocess Research Group, Department of Chemical Engineering, Universidad de Antioquia, Calle 70 No. 52-21, Medellín, Colombia. ✉email: jfreyre@ccg.unam.mx

7825 genes, 965 of them (~12%) code for proteins with a predicted regulatory function. From those, 65 genes coding for sigma factors, a remarkably high number among bacteria, of which ~70% (45/65) are ECF (extra-cytoplasmic function) sigma factors, suggesting independent regulation of diverse stress response regulons[5]. Besides, it counts with many two-component systems (TCSs), 85 sensor kinases, and 79 response regulators, also related to stress response[5]. The difference between sensor kinases and response regulators suggests a cross-talking among them. Noteworthy, *S. coelicolor* genome codes for several putative regulators that do not belong to families outside *S. coelicolor*[5]. Because of the complexity of the secondary metabolism regulation, a proper understanding of the *S. coelicolor* regulation requires it to be studied systematically at both local and global scales. On a global scale, GRNs are used to study transcription regulation. They can be represented as a directed graph where nodes represent genes, and edges represent the regulatory interactions among the transcription factors (TFs) and their target genes (TGs). Previous comprehensive reviews have been focused on specific morphological differentiation and metabolic processes[4,6–10]. However, a GRN at a global scale is still missing.

The initial approach to reconstruct a global-scale GRN will be through text mining[11]. There we would be able to collect all the information available on the literature for the microorganism. Nevertheless, it would still require manual intervention for those articles where interactions are not clearly defined. Moreover, all genes have not been studied experimentally. Therefore, alternatively, GRN inference has been applied in diverse bacteria to provide a deeper understanding of their regulatory mechanisms. Besides, it has also been applied to propose selective experimental validation of putative interactions, analyze bacterial GRN evolution, and build biological models for biotechnological processes[12–16]. A GRN inference for *S. coelicolor* was performed by Castro-Melchor, et al. in 2010 using ARACNE and applying module validation through the identification of consensus DNA sequences[17]. However, the resulting network was not assessed with any gold standard (GS) available at the time, and no thorough study of its structural properties was performed. Moreover, benchmarking studies of network inference methods have shown the poor predictive power of using a single GRN inference tool[18].

Here, we performed a collection and curation of the experimentally-validated transcriptional regulatory interactions of *S. coelicolor* A3(2) and classified them based on the confidence level of their supporting evidence. Further, we integrated this curated GRN with previous curations from DBSCR (http://dbscr.hgc.jp/) and Abasy Atlas[19]. Then, we applied the natural decomposition approach (NDA) to identify their system-level components and unveiled different biology aspects of *S. coelicolor* regulation. Next, we applied several tools to infer novel interactions, three based on DNA binding sites for the TFs, and five based on gene expression along with two modifications proposed by the authors. We integrated the predictions using a community approach, which has been reported as the best strategy to reduce the number of false positives[18]. Then we used the most reliable curated network as a GS for the validation of the inferred GRNs. We further assessed the inferred networks through their structural properties and found that the NDA [20] is a valuable tool for GRNs dissection and comparison. From the best-rated inferred network, we proposed new TF candidates for the direct regulation of some of the key S*treptomyces* antibiotic regulatory proteins (SARPs) in *S. coelicolor.* Finally, we applied the meta-curated network of *S. coelicolor* to study the conservation of the system-level components with those of its phylogenetically related *C. glutamicum* as an application of the curated network. The workflow of this work and the suggested use of the data herein reported are summarized in Fig. 1.

Results and discussion

Reconstruction of the most complete experimentally-validated regulatory network for *S. coelicolor*.

We curated a total of 124 papers retrieved from PubMed and Google Scholar queries covering a span of 29 years (from 1990 to July of 2019) (see Fig. 2 and Supplementary Table 1). We collected a total of 9714 regulatory interactions (out of the 23,908 expected interactions in the complete GRN as predicted by Abasy Atlas v2.4) among 5331 genes. We perceive a notable increment in the number of papers and interactions after the *S. coelicolor* genome was completely sequenced (2002)[5]. This eases the study of its genome and regulation, being 2012 the year with most publications (see Fig. 2). We classified the interactions according to their experimental evidence, expanding the RegulonDB scheme[21,22]. First, we label the interactions as "strong" or as "weak" according to the methodology of the experiment performed. A "strong" evidence level is assigned to experiments that prove a physical regulatory interaction between the TF and the TG. This means that the TF can bind to the upstream region of the regulated gene. Here we have experiments such as EMSA in purified proteins or in vitro transcription assays. On the other hand, a "weak" evidence level is assigned when there is no evidence of direct interaction. This means that the experiment suggests either a hypothetical DNA binding site, such as ChIP; or an effect of the TF over the gene that might be indirect, through another TF, such as microarray, RNA-Seq, RT-PCR, etc. For experiments that were not in the RegulonDB scheme, such as DNA-affinity capture assay (DACA)[23], we analyzed their methodology to classify them either as "strong" or "weak" evidence. Supplementary File 2 has the evidence classification for each interaction according to their "strongest" supporting experiment (Supplementary Table 1–2 and Supplementary Figure 1).

Afterward, we gathered these interactions along with others from the databases, RegTransBase[24] available at Abasy Atlas database (https://abasy.ccg.unam.mx) [19], and DBSCR (http://dbscr.hgc.jp/). We processed these curated interactions to construct the corresponding GRNs, removing redundancy by mapping the gene identifiers to locus tags and merging interactions while preserving the information about the effect and evidence classification of the supporting experiments, as previously reported[19]. From our curation, we reconstructed a total of seven curated networks with different evidence classification and completeness. (1) *Curated_FL* with a total of 9454 unique interactions, from which ~5% (493/9454) are "strong". (2) *Curated_FL(cS)* with 438 "strong" interactions from *Curated_FL*. (3) *Curated_DBSCR* with the 341 interactions from DBSCR and used the ~34% (115/341) "strong" interaction to reconstruct (4) *Curated_DBSCR(S)*. (5) *Curated_RTB* is the network from the RegTransBase database with 330 interactions, all of them labeled as "weak" since their experimental evidence

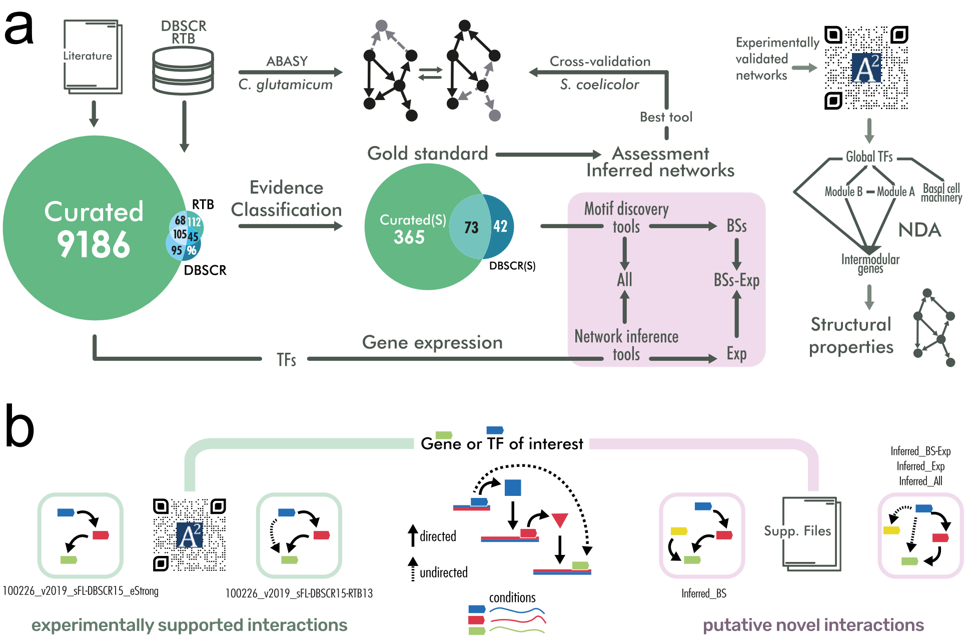


Figure 1. (**a**) Workflow of this work. The purple area covers the inference of the networks. (**b**) Type of interactions contained in the networks. The green path connects to curated regulations supported by experimental evidence. The "*strong*" network contains only the interactions that are supported by an experiment proving that the transcription factor binds a DNA site near a target gene to regulate its transcription. Curated networks without the "*strong*" label might contain indirect interactions, as they could be supported by non-directed experiments (such as gene knockout and its effect on genes transcription). The purple path connects to inferred interactions. Predictions based solely on binding sites predictions would be inferring only TF-DNA interactions. Predictions involving gene expression data might contain indirect interactions.

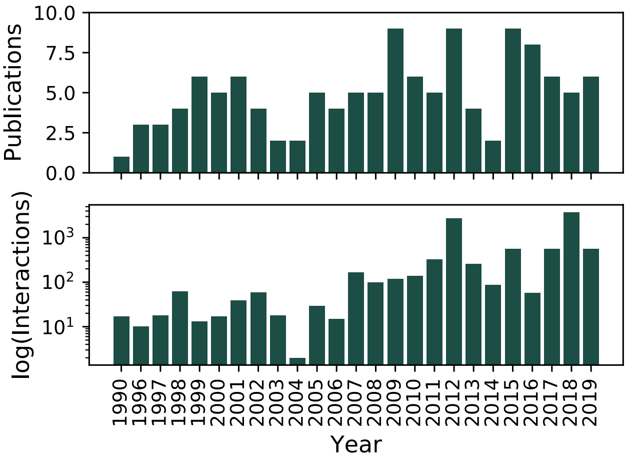


Figure 2. Interactions curated from literature for *Streptomyces coelicolor* A3(2). (**a**) Number of publications per year and (**b**) Number of interactions reported per year.

was not available. Later, we merged *Curated_FL, Curated_DBSCR,* and *Curated_RTB* into 6) *Curated_FL-DBSCR-RTB* a meta-curated network comprising a total of 5386 genes and 9707 non-redundant regulatory interactions, which is the most extensive experimental GRN of *S. coelicolor* up to date. From this meta-curation, we filtered the 480 "strong" interactions to reconstruct 7) *Curated_FL(cS)- DBSCR(S)*. All curated networks are further described in Table 1.

The functional architecture of the *S. coelicolor* GRN. To reveal the functional architecture and to elucidate the regulatory and biological function of some of the genes and interactions curated, we applied the Natural Decomposition Approach (NDA) on all the curated networks. The NDA is a biological-mathematical criterion to suggest a biological function of each gene based on the structure of the GRN[20]. It classifies the genes into one of the four structural classes: (1) global regulators (GR), coordinating genes from different metabolic

Network	Abasy ID	Genes	Interactions	Description
Curated_RTB	100226_v2015_sRTB13	311	330	Network from RegTransBase database
Curated_DBSCR	100226_v2015_sDBSCR15	273	341	Network from Database of transcriptional regulation in *Streptomyces coelicolor* and its closest relatives
Curated_DBSCR(S)	100226_v2015_sDBSCR15_eStrong	112	115	Filtration of interactions with strong evidence from the DBSCR network
Curated_FL	100226_v2019_sA22	5331	9454	Network from the collection and curation performed for this work
Curated_FL(cS)	Not reported	347	438	Filtration of interactions with strong evidence from the FL network (cS = curated strong)
Curated_FL(S)	100226_v2019_sA22_eStrong	396	493	Filtration of interactions with strong evidence from the FL network along with statistically validated interactions
Curated_FL-DBSCR-RTB	100226_v2019_sA22-DBSCR15-RTB13	5386	9707	Meta-curation of RTB, DBSCR and FL networks
Curated_FL(cS)-DBSCR(S)	Not Reported	387	480	Filtration of interactions with strong evidence from the meta-curated network
Curated_FL(S)-DBSCR(S)	100226_v2019_sA22-DBSCR15_eStrong	435	534	Filtration of interactions with strong evidence from meta-curated networks along with statistically validated interactions
Inferred_BS	Available as a Supplementary File 3	6263	23,908	Inferred GRN from binding sites prediction
Inferred_Exp	Available as a Supplementary File 3	4739	23,908	Inferred GRN from transcriptomic data
Inferred_BS-Exp	Available as a Supplementary File 3	4763	23,908	Community network from Inferred_BS and Inferred_Exp
Inferred_All	Available as a Supplementary File 3	3804	23,908	Community network from all the inference methods

Table 1. Description of networks used in this work.

pathways[25]; (2) modular genes, group of genes working together to carry out a biological function[20,26]; (3) intermodular genes, integrating at the promoter level the response from different modules[20,27]; and (4) genes constituting the basal machinery of the cell. We decided to further study the NDA analysis of *Curated_RTB-FL-DBSCR* since it is the most complete GRN.

The NDA analysis of the meta-curated network *Curated_RTB-FL-DBSCR* revealed 20 GRs (0.37% of the 5386 network genes), 502 modular genes (9.32%), 18 intermodular genes (0.33%), and 4846 basal machinery genes (89.97%). The classification of each gene can be found at https://abasy.ccg.unam.mx/genes?regnetid=100226_v2019_sA22-DBSCR15-RTB13&class=All. Through the NDA analysis, we found 35 gene modules. From them, module number 16 is a mega-module, which is divided into 12 submodules that are connected through the intermodular genes (Supplementary Figure 2a). To analyze the GRs identified in the meta-curated network, we reviewed the literature to identify the TFs that have been previously reported as global or pleiotropic regulators in *S. coelicolor*. *Martín* et al.[28] reported a detailed description of the cross-talking between the global regulators in *S. coelicolor* and other *Streptomyces*. The review provides a list of genes considered as global and wide-domain regulators, due to the hundreds of genes they regulate and the multiple effects they produce[28]. Nine out of the 20 (45%) GRs identified by the NDA were reported as such in this review. We further screened the literature to identify GRs or pleiotropic regulators reported in individual papers (Supplementary Table 4). We found 20 pleiotropic TFs or GRs reported individually, from which 13 (65%) were categorized as GRs by the NDA. See the "Global regulators" section in Supplementary File 1 for further description of the GRs identified.

This analysis also revealed 18 intermodular genes. Some of their promoters integrate the signals of different GR related to carbon, nitrogen, and phosphate metabolism. For instance, *glnA* (SCO2198), *glnII* (SCO2210), and the *amtB-glnK-glnD* (SCO5583-85) operon, which are known to be mediators between the nitrogen and phosphate metabolism through the binding of their GR (PhoP and GlnR) to these intermodular genes promoters[29]. Others integrate signals from primary and secondary metabolism, or morphological differentiation and antibiotic production. A further description of these genes can be found in Supplementary File 1 in the section "Intermodular genes". Moreover, the functional annotation of the modules identified by the NDA also provides a new functional hypothesis for genes whose function is currently unknown using a guilt-by-association strategy, as previously described[30]. From the 46 modules and submodules in the GRN, 26% (12/46) are annotated. The annotation of each module can be found at https://abasy.ccg.unam.mx/modules?regnetid=100226_v2019_sA22-DBSCR15-RTB13. Most of the annotated modules are related to cellular metabolism, organic substances metabolism, and biosynthetic processes, which are fundamental processes for every cell (Supplementary Figure 2b). We found 245 with no previous annotation in GOA[31] assigned to the annotated modules (Supplementary Table 5).

GRN inference based on binding sites identification performs better than that based on transcriptomics. Despite the exhaustive curation, the meta-curated network *Curated_FL-DBSCR-RTB* has regulatory information for only ~ 65% (5386/7825) of the *S. coelicolor* genome (network genomic coverage) and has ~ 41% (9707/23908) of its expected total interactions (network interaction coverage or completeness)[19], considering both "strong" and "weak" interactions. We leveraged the large corpus of high-throughput data available to computationally infer missing regulatory interactions to expand our GRN reconstruction. The inference was performed from two different approaches. For the first approach, we performed a regulon reconstruction through the de novo identification of TF binding sites and linked them to downstream genes. The regulon reconstruction was based on the network *Curated_FL(cS)-DBSCR(S)* using three methods for motif discovery: MEME, Bioprospector, and MDScan (see "Material and methods" section). For the second approach, we performed a GRN inference from transcriptomic data. We used seven methods for GRN inference based on the

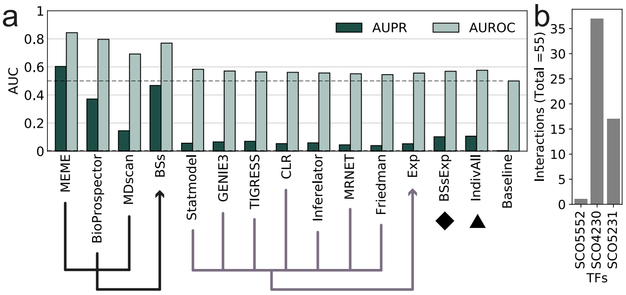


Figure 3. (**a**) AUROC and AUPR for each of the methods and the community networks. (**b**) Number of interactions statistically validated by TF.

gene expression data: CLR, Friedman, GENIE3, Inferelator, MRNET, Statmodel, and TIGRESS (see "Material and methods" section and Supplementary Table 6a). These methods were selected based on their performance in a previous benchmarking of GRN inference methods[18], their availability, complete documentation, and maintenance. For the inference, we selected an Affymetrix dataset (Platform GPL9417) from NCBI GEO[33]. See the "Transcriptomic Data" section in Supplementary File 1 for further description of the data selection. Since most of the experiments in the curation were performed on the *S. coelicolor* A3(2) strain M145, which is plasmid-free, we restricted the inference to interactions among genes of the chromosome. Next, we evaluated the inferred GRNs computing the AUROC and the AUPR of the predictions (see Fig. 3a and Supplementary Figure 4). From the AUPR, it is evident that in general GRN inference from binding sites performed better than the inference from expression data. For the inference from binding sites, MEME performed better than the other methods. For the inference from gene expression data, TIGRESS performed better, followed closely by GENIE3 and Inferelator. We assessed the inferred GRNs based mostly on the AUPR since it is more informative for imbalanced datasets[36] as it is the case of GRNs inference[37]. Binding sites for each one of the interactions identified using MEME are reported in Supplementary File 3.

For the assessment, all the inferred networks were pruned to the 23,908 best scoring interactions among genes part of the GS, since it is the number of interactions expected in the final network for *S. coelicolor*[38]. Nevertheless, the GS has only 387 interactions, which is ~1.6% of the 23,908 regulatory interactions expected in the complete regulatory network of *S. coelicolor*[19]. For this reason, the assessment only reflects the capacity of the methods to infer the interactions in the GS, while novel interactions (actual interactions not part of the incomplete GS) are labeled as false positives. Moreover, as the GS was used as prior for the regulon extension, it might provide an advantage for the network predicted by motif discovery. Because of its approach, inference by motif discovery predicts direct regulatory interactions, while inference from transcriptomic data predicts both direct and indirect ones without distinction. Thus, as the GS is only built by direct interactions, it is expected that inferred networks with the same type of interactions get a higher score. However, using the "non-strong" GRN as GS could drive to a bigger problem because indirect regulatory interactions might be spurious and are not adequate to assess causal interactions.

Given that the current GS is still quite incomplete, and we cannot do proper discrimination among the different inferred networks, instead only using the single best method, we decided to build a community network for each one of the approaches. (1) *Inferred_BS* for the prediction from binding sites; (2) *Inferred_Exp* for the prediction from expression data; (3) *Inferred_BS-Exp*, a community network from both previous community networks; and (4) *Inferred_All*, a community network built mixing individual networks from both approaches (see Table 1). For the latter, we used the three methods for binding site inference, along with Statmodel, GENIE3, and TIGRESS from expression-based GRNs (due to their superior performance) to balance both approaches. *Inferred_BS* outperformed the rest of the community networks at both AUPR and AUROC (see Fig. 3a). However, it was outperformed by MEME at both metrics. Given MEME's outstanding performance (see Fig. 3a), we used it to perform a statistical validation of "weak" interactions supported by ChIP-data, similarly as proposed in[22] (see "Material and methods" section). A total of 55 "weak" interactions were reclassified as "strong" (see Fig. 3b and Supplementary Table 6b). We found one of these interactions (SCO4230-SCO4878) already reported as "strong" in the DBSCR database (*Curated_DBSCR(S)*). These statistically validated interactions were merged with the "strong" interactions from *Curated_FL* and from the meta-curated network *Curated_FL-DBSCR-RTB* into two networks: *Curated_FL(S)* and *Curated_FL(S)-DBSCR(S).* We reassessed the network predictions with *Curated_FL(S)-DBSCR(S)* as GS and the results remained virtually the same (Supplementary Figure 5). For completeness, we also performed the assessment using *Curated_FL-DBSCR15-RTB13* as the GS (Supplementary Figure 6). However, the 72 regulators in *Curated_FL(S)-DBSCR(S)* are the only TFs that were used for the predictions based on binding sites. On the other hand, *Curated_FL-DBSCR15-RTB13* has 137 TF. This resulted in a poor recall by the predictions based on binding sites (Supplementary Figure 6), as interactions for 65 TFs are not predictable because their regulations might be carried out indirectly, with no need for a DNA binding site.

Inferred networks have a similar structure to the largest curated networks. Even though the AUPR and AUROC metrics allow the assessment of the predictions, both metrics heavily rely on the ranking of the predicted interactions. Moreover, the GS is not complete and missing interactions would be still classified

Property	Inferred_BS	Inferred_Exp	Inferred_BS-Exp	Inferred_All	Curated_FL-DBSCR-RTB
Number of nodes (N)	6263	4739	4763	3804	5386
$\ln(\ln(N))$	2.17	2.14	2.14	2.11	2.15
Average shortest path length	2.86	3.38	3.38	3.11	2.84
Average clustering coefficient	0.213	0.385	0.385	0.470	0.182
$\alpha(P(k))$	1.861	1.952	1.955	1.968	1.742
$\mathrm{R}^2_{\mathrm{adj}}$	0.87	0.92	*0.92*	0.91	0.84
$\alpha(C(k))$	0.924	0.767	0.742	0.729	1.142
$\mathrm{R}^2_{\mathrm{adj}}$	0.79	0.58	0.54	0.68	0.89

Table 2. Network properties for inferred networks.

as false positives, decreasing the score more the higher their ranking is. Therefore, we assessed the inferences in terms of their structural properties and compared them against the curated networks to compensate for such drawbacks. Note that this approach has its caveats. The global structural properties of the network might be different once the GS is complete, this can be approached by comparing the predictions to all the curated networks, each of them with different completeness. Also, two networks could have the same topology with different node entities. For this reason, we use the topological assessment in complement to the AUPR and AUROC metrics to identify the best prediction.

One of the main structural properties of biological networks is that they are scale-free and hierarchically modular. Same properties that our curated networks have been proved to possess (Supplementary File 1). Therefore, as an initial approach, we asked whether the inferred networks are scale-free too. The degree (nodes' connectivity) distribution $P(k)$ of scale-free networks follows a power law, $P(k) \sim k^{-\alpha}$, with $2 < \alpha < 3$ [20,38,39]. If $\alpha = 2$, there is a unique global regulator (hub-and-spoke network) and if $\alpha > 3$, scale-free networks lose most of their characteristic properties[39]. First, to compute this α for the inferred networks, we performed a robust linear regression over a log–log plot of the complementary cumulative degree distribution and corrected the exponent accordingly (see Table 2 and Supplementary Figures 7–11). All inferred networks' degree distribution seems to follow a power law according to the adjusted coefficient of determination. Nevertheless, the data points in *Inferred_All* appear to be divided into three regions with different tendencies, instead of the two that are present in the other networks (Supplementary Figures 7–11). Usually, this type of network is divided into two regions, the region of the nodes (genes) with a low degree, and the one with nodes with a high degree[40]. The appearance of a third region might be a consequence of merging networks from methods with different approaches. This could affect the structural properties of the merged networks, while communities from the same approach appear to have more similar structural properties. In the case of *Inferred_BS-Exp*, the construction of communities ahead by each approach create more compatible networks in terms of structure that can be conveniently mixed.

To confirm that their degree distributions follow a power law, we contrasted each distribution of the inferred networks against alternative fat-tailed probability distributions (Power-law, exponential, stretched exponential, lognormal, and truncated power-law) using Kolmogorov–Smirnov tests[38,41] (Supplementary Table 6a). We found that the degree distribution of the inferred networks adjusted better to a power-law than to an alternative distribution. Then, we computed a maximum likelihood estimation for the exponent (α) of the power laws and found that most of them are between two and three, except for *Inferred_All*. This shows it as an anomalous scale-free network (Supplementary Table 7). Perhaps caused by the mixing of networks with diverse structural properties. Nevertheless, we could consider all inferred networks to be scale-free.

Furthermore, we checked other properties of scale-free networks (see Table 2)[39,42]. The four community networks have small average shortest path lengths and a high clustering coefficient. *Inferred_BS* has the smallest average short path length, while *Inferred_All* has the highest average clustering coefficient (see Table 2). Scale-free networks also present an ultra-small world effect, which implies that the average path length is proportional to $\ln(\ln(N))$ (N is the number of nodes in the network). This is the case for all the inferred networks. Another characteristic of GRNs is their hierarchical modularity, which implies a diamond-shaped hierarchical organization as has been revealed by the NDA[20]. In a scale-free network, this implies that the clustering coefficient depending on the degree (C) follows a power law as $C(k) \sim k^{-1}$[39]. *Inferred_BS* has the exponent closest to − 1 (− 0.92), with the best R^2. Even though *Inferred_BS* seems to be the network that behaves closest to a GRN, all networks have similar values, which makes it difficult to discern the most reliable inferred network by this approach.

We included several other structural properties of the networks to perform a more thorough comparison (Supplementary Table 8). We clustered the vectors of structural properties for the curated and community-inferred networks (see Fig. 4a). The clustering partitions the networks into two major groups. The first one contains the curated networks and *Inferred_BS*, while the second group contains the other inferred networks. The first group is in turn also divided into two groups: one with the two largest curated networks and *Inferred_BS*, and the other one with the remaining curated networks. The reason for this may be due to the size of the networks (see Table 1). When standardizing the property values by max–min feature scaling, the two largest curated networks were clustered with the predictions (Supplementary Figure 12), which could be also due to the size of the network. To reduce the network size influence we used the network dissimilarity measure proposed by *Schieber* et al.[43]. We considered the third term which makes the distance measure robust to graph size in terms of the number of nodes (genes)[43] (see Fig. 4b). Even with this metric, the two largest curated networks were clustered with the

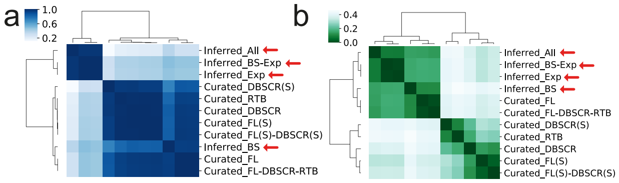


Figure 4. Network comparative by structural properties. (**a**) Pearson correlation of the profile of structural properties listed in Supplementary Figure 12. (**b**) D-value from *Schieber* et al.[43] to measure network similarity.

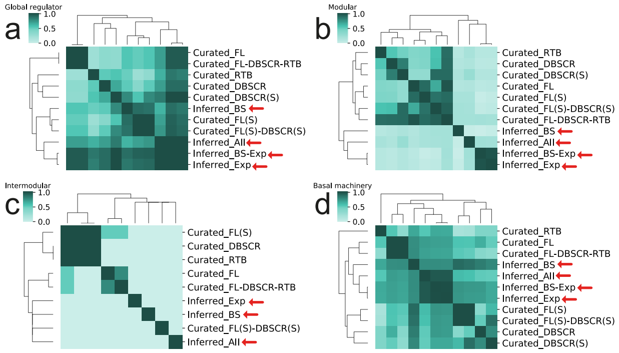


Figure 5. Simpson similarity index for NDA analysis for all curated and inferred networks. (**a**) Global regulators. (**b**) Modular genes. (**c**) Intermodular genes. (**d**) Basal machinery genes.

inferred networks. This might be a consequence of the high value of maximum out-connectivity and structural genes in the largest curated networks, like those found in the inferred networks. This shows that the inferred networks are similar that the most complete curated networks in terms of structure, suggesting their reliability.

The natural decomposition approach identifies regulatory networks similarity despite their different completeness level. We compared all the curated and community inferred networks based on the Simpson similarity index of the four components proposed by the NDA: global regulators, modular genes, intermodular genes, and the basal machinery (see Fig. 5). The Simpson similarity index measures the size of the core of two sets with reference to the smallest one[44]. Note that two identical sets, being one a subset of the other will have a score of 1. On the other hand, two completely different sets will have a score of 0.

When comparing the global regulators (see Fig. 5a), there is not a distinct division among the networks. *Infered_BS-Exp* and *Infered_Exp* have a similar correlation with all the curated networks, slightly higher with *Curated_DBSCR(S)*, *Curated_FL*, and *Curated_FL-DBSCR-RTB*. These two inferred networks have the highest amount of GR, 116, and 114 respectively; thus, the other GRs predicted could be easily a subset of them. For the case of *Infered_BS*, it has the highest correlation with the "strong" networks. This is expected since these networks have only direct regulatory interactions, as the interactions predicted in *Inferred_BS*, while there is no evidence of direct regulation for transcriptomic-based inferred interactions. This can result in an underestimation of the effect of GRs over the rest of the genes since GRs which regulate several targets from different processes might not be predicted as such in the "strong" networks. However, when their indirect influence is represented in the network, their ranking as GRs is noticeable.

When analyzing the modular genes, there are two major groups (see Fig. 5b): the major group with the curated networks is divided into two subgroups, one contains *Curated_RTB*, *Curated_DBSCR,* and its "strong" version *Curated_DBSCR(S),* and the second subgroup contains the integrations and curations proposed in this work. Interestingly, the meta-curated network *Curated_FL-DBSCR-RTB* correlates very well with all the smaller networks it contains, from which we could deduce that modular genes are conserved despite the addition of new regulatory interactions. In the second group, composed of the inferred networks, we can see there is not a high correlation among them. *Inferred_BS* is the closest to the curated networks, while Inferred_Exp and Inferred_BS-Exp have a high correlation. This tells us that the interactions in *Inferred_Exp* have a larger influence on the module configuration of *Inferred_BS-Exp* than *Inferred_BS*. The difference between the curated and inferred networks might come from the fact that inferred networks have a greater number of GRs and a much lower number of modular genes when compared with the curated networks (Supplementary Table 9).

Intermodular genes are the less conserved NDA class (see Fig. 5c). There is overlap only among the smallest curated networks, all share the intermodular gene SCO5877, which appears as a TF in the other curated networks. Moreover, there is an overlap between *Curated_FL* and *Curated_FL-DBSCR-RTB*, which share most of the interactions. Thus, is expected that they also share most of the intermodular genes. Note that the networks *Curated_DBSCR(S)* and *Inferred_BS-Exp* are not included in the clustering since they did not present any intermodular genes. Finally, when analyzing the basal machinery (see Fig. 5d), the larger curated networks are

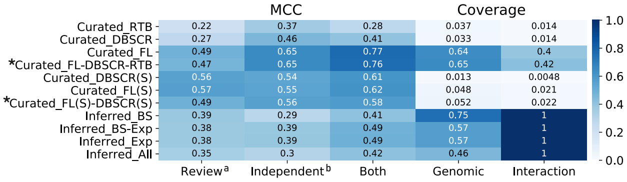


Figure 6. MCC for global regulators predicted by NDA for each of the curated and inferred networks. Scores ≥ 0.5 are represented in white numbers and meta-curations are marked with an asterisk. [a]Gold standard curated from reference[28]. [b]Gold standard curated from independent publications (Supplementary Table 4).

grouped on one side, next to the inferred networks, and finally the smallest curated networks with *Curated_RTB* as an outgroup. Even though *Inferred_BS* is grouped with the other inferred networks, it has a higher correlation with *Curated_FL(S)* and *Curated_FL(S)-DBSCR(S)*, which again evidence the similarity among these three networks.

Assessment of the global regulators' inference. In the NDA, the identification of global regulators is a key step in the classification of every node in the GRN. Previously, it has been reported a high overlap between the predictions of global regulators by the NDA and those reported in the literature for *E. coli*[20], *B. subtilis*[45], and *C. glutamicum*[27]. We used the set of GRs reported by *Martín et al.*[28], besides those reported in independent articles, and the union of both sets (Supplementary Table 4). Then we assessed the predictions of the GR using the Matthews correlation coefficient (MCC) (see Fig. 6) We used the MCC score as it is more informative and reliable than F1 for binary classification assessment[46], however when applying the F1 score we obtained consistent results (Supplementary Figure 13).

Curated_FL and *Curated_RTB-FL-DBSCR* have the best performance in GR prediction. However, the "strong" networks have a slightly smaller score even having much less genomic coverage. This shows that the GRs are very robust to perturbations in the network as previously shown[20]. On the other hand, despite the high coverage of the inferred networks, the performance of the predictions with such networks was poor. This could be, as it was mentioned before, due to the great amount of GR predicted by these networks, which would cause a high proportion of false positives affecting the score. *Inferred_BS* produced the most conservative prediction (lowest false-positives rate) among the inferred network (Supplementary Figure 14 and Supplementary Table 9).

GRN inference from transcription factor binding sites proves to be the most reliable approach and allows the prediction of new TFs for the most studied SARPs. Inferred_BS performed the best on AUPR and AUROC among the community inferences and is the most similar to the curated networks according to its structural properties and system-level components. Moreover, it has the largest genomic coverage among all the networks, which would be advantageous for a deeper study of transcriptional regulation in *S. coelicolor*. Therefore, we considered *Inferred_BS* as the most reliable inferred network, despite that similar studies suggest the integration of inference approaches as the most suitable methodology for GRN reconstruction[18,47,48].

Thus, we decided to use *Inferred_BS* to further study the regulation of the SARPs of the most studied antibiotics in *S. coelicolor*: ActII–orf4, RedD/RedZ, CpkO (also known as KasO), and CdaR, which regulate the production of ACT, RED, yCPK, and CAD, respectively[3] (Supplementary Table 10). A total of 13 new putative regulators for the SARPs were predicted, providing a great opportunity to find new targets to manipulate the *S. coelicolor* antibiotic production. As it is not possible to computationally confirm a direct binding of this regulator, it is necessary to corroborate them with wet-lab experiments.

Following, we describe some of the TFs predicted for the SARPs using the workflow suggested in Fig. 1b: For *actII-orf4* (SCO5085) only one novel regulator was inferred, MacR (SCO2120) which is the response regulator of the TCS MacRS, while the rest was already part of *Curated_FS(S)-DBSCR(S)* (Supplementary Figure 15). This TCS has been proved to activate ACT production. Nevertheless, a ChIP-qPCR analysis was not able to prove an in vivo interaction between MacR and *actII-orf4*, although a direct interaction was not tested[49]. For *redD* (SCO5877) two new regulators were predicted, LipR (SCO0712) and ActII-orf4 (SCO5085). LipR is related to AfsR (SCO4426)[50], homolog to the SARPs, and activator of the ACT and RED production[51]. Moreover, its mutant affects ACT production[50], which makes it plausible to affect RED production as well. It has been suggested that ActII-orf4 might regulate the production of other antibiotics[4], which could be by binding directly to their CSR. For *redZ* (SCO5881) five new regulators were predicted, among them is GluR (SCO5778) which has been shown to affect RED production. Nevertheless, it has been shown that GluR does not bind directly to *redZ*, thus it could an indirect regulation[52]. Another one, StgR (SCO2964) has been shown by an RT-qPCR experiment to be a repressor of redD[53]. This repression could be through the direct binding to *redZ*. HpdA (SCO2928) and HpdR (SCO2935) are related to tyrosine catabolism, which produces important precursors for antibiotic biosynthesis[54]. Moreover, HpdA has been shown to activate *actII-ORF4,* therefore might have a more direct role in RED production. In the case of *cdaR* (SCO3217), we have four predicted regulators, among them, OsdR (SCO0204) and RamR (SCO6685). Both are related to the response to stress and the development of *S. coelicolor*[55,56]. SsgR (SCO3925) regulates sporulation and morphological differentiation[57]. These all processes are highly related to antibiotic production. Finally, for *cpkO/kasO* (SCO6280) six new regulators were inferred, among them OsdR (SCO0204), LipR (SCO0712), and StgR (SCO2964) were described before. Another one is NnaR (SCO2958), which regulates

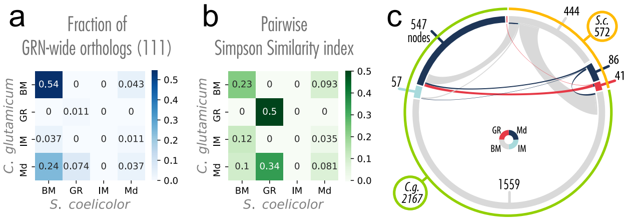


Figure 7. Conservation of the systems-level components between *S. coelicolor* and *C. glutamicum*. (**a**) NDA classification for the GRN-wide orthologs in their corresponding organism. Total matrix sum to 1. Most of the GRN-wide orthologs are classified as basal machinery in both organisms (**b**) Overlapping of the NDA classes between the two networks with reference to the smallest set. Each cell can range between 0 and 1, where 1 means one class is a subset of another, and 0 means there is no overlap at all. (**c**) Similarities between the two organisms highlight the size difference between the datasets. The color of the inner circle sections represents the NDA classes and ribbons colors represent the *S. coelicolor* NDA classes. Numbers represent the genes for each class and organism. Gray ribbons are the widest ones, representing that all the basal machinery of *S. coelicolor* with 1:1 orthology relationship with *C. glutamicum* is classified either as basal machinery or modular genes in *C. glutamicum*. This suggests that multiple basal machinery genes could be reclassified as modular components in more complete reconstructions of the *S. coelicolor* network.

spore formation and antibiotic production[58]. We refer the reader to the Supplementary material for the complete list of the predicted regulators in every predicted network (Supplementary Table 10).

We further compared the six regulators of *actII-orf4* identified by *Inferred_BS* with the 11 TFs found in Inferred_All. From the latter, only two are part of *Curated_FS(S)-DBSCR(S)*, and three more are included in *Curated_FS-DBSCR-RTB* (Supplementary Figures 15, 16). We found that the interactions included in *Inferred_All* tend to be carried out indirectly (i.e., not through a TF-DNA interaction), suggested by a poor overlap between *Inferred_BS* and *Inferred_All*, and biological insights such as the putative regulation of *actII-orf4* by *OsdR*.

Comparative analysis with *Corynebacterium glutamicum* shows coherent system-level components conservation.

The diamond-shaped structure identified by the NDA is conserved between *E. coli* and *B. subtilis*[45]. As an application of the meta-curated network, we studied the conservation of its system-level components, comparing it against the *C. glutamicum* network. *C. glutamicum* is phylogenetically related to *S. coelicolor,* and a model organism for the study of GRNs[27]. We applied the regulogs analysis[59] with one-to-one orthology relationships to alleviate network incompleteness and make them comparable. As prior networks, we used *Curated_FL(S)-DBSCR(S)* (534 interactions) for *S. coelicolor* and 196627_v2020_s21_eStrong from Abasy Atlas[19] (2941 interactions) for *C. glutamicum*[16]. After the regulogs analysis, we ended up with 2966 interactions in *C. glutamicum* and 692 interactions in *S. coelicolor*.

We used the complemented networks to identify GRN-wide orthologous relationships defined as the orthologous present in the GRN of the respective organism. We obtained a total of 188 GRN-wide orthologous relationships from a total of 995 1:1 orthologs identified by OrthoFinder[60]. We applied the NDA analysis to both GRNs to identify the system-level components and computed the fraction of the GRN-wide orthologous in each combinatory relationship between the NDA classes (see Fig. 7a). We found that most of the GRN-wide orthologous (54%) are classified as basal machinery in both organisms. This is expected since 73% and 74% of the genes correspond to the basal machinery in the complemented networks of *C. glutamicum* and *S. coelicolor*, respectively. Besides, the distribution of the genes in the chromosome of *S. coelicolor* shows a central core, where are genes likely related to primary functions such as DNA replication, transcription, translation, and amino-acid biosynthesis; and likely non-essential genes such as secondary metabolism are in the chromosome arms[5]. More than 59% (111/188) of the GRN-wide orthologs conserved the same class in both organisms (Fig. 7a) showing high conservation of the NDA classification.

We studied the pairwise Simpson similarity index between the four classes between the two organisms to remove the problem of the imbalanced classes (see Fig. 7b). GR is the class with the highest conservation rate, the orthologs of seven of the eight GRs in *C. glutamicum* are also GRs in *S. coelicolor* (see Fig. 7b and c). The conservation between the same class in the two organisms is also high for the basal machinery, while poor for the modular genes. For the case of intermodular genes, even though the networks were complemented with information from the other network, they are not conserved at all (see Fig. 7b). Previous work reported intermodular genes as the least conserved of the system-level components[27]. Intermodular genes are the most likely responsible for giving the GRN flexibility and increasing evolvability by scouting different combinations of regulatory interactions between physiological functions so the organism could adapt better to environmental changes[45]. These results agree with a previous analysis of the robustness of the NDA to a random node and edge remotion showing GR and intermodular genes as the most and least conserved classes, respectively[27].

On the other hand, 24% of the GRN-wide orthologs that are modular genes in *C. glutamicum* were classified as basal machinery in *S. coelicolor*. This could be due to three possible reasons[45]: (1) the basal machinery genes in *S. coelicolor* are misclassified and further research is needed to find the missing regulatory interactions (see Fig. 7c) that will integrate some of these genes into a module. (2) The GRs controlling *C. glutamicum* genes are not yet identified as GRs. (3) Genes in *S. coelicolor* need a more direct regulation because of their physiological function (high plasticity of transcriptional regulation). A previous genomic comparison between *S. coelicolor,* *Mycobacterium tuberculosis,* and *Corynebacterium diphtheriae* showed a synteny among the whole chromosome of these last two microorganisms and the core of the one of *S. coelicolor*[5]. *C. glutamicum* is phylogenetically closely related to *M. tuberculosis* and *C. diphtheriae*, with roughly similar genome size. Therefore, a similar result would be expected. Furthermore, as more classical experiment data become available, new regulations for the currently basal machinery would turn those genes into the modular class. However, a deeper analysis of diverse factors such as genome size, the niche of the organisms, and a wider range of organisms are required to further study the robustness of the NDA analysis.

Conclusions

A meta-curated regulatory network for *S. coelicolor* (*Curated_RTB-FL-DBSCR)* was reconstructed from a collection and curation of regulatory interactions experiments in literature and databases. From the NDA analysis of the meta-curated network, we could identify 20 global regulators, of which 95% (19/20) have already been reported as global or pleiotropic regulators. 46 functional modules were identified along with 18 intermodular genes, some of them found to be involved in more than one biological process. Functional modules annotated by GO enrichment allowed via a 'guilt by association' strategy to propose a function for 245 genes without any previous functional annotation or annotated as 'hypothetical protein'. This network is, however ~ 42% of the estimated complete regulatory network, which evidences a lack of information related to *S. coelicolor* transcriptional regulation. Especially for interactions experimentally supported by "strong" evidence, which accounts for only ~ 2% of the estimated complete network. We indeed found a low level of direct experimental validation for the regulatory interactions reported in the literature and curated in this work as only ~ 6% (533/9687) are supported by experiments confirming the binding of the TF to the upstream region of the target gene, the so-called "strong" evidence. The low level of "strong" evidence is due to the high fraction of high-throughput experiments aimed to unveil the regulatory network. This highlights the importance of carrying on classical experiments aimed to confirm the weakly supported interactions (e.g., EMSA, in vitro transcription assay, and DNA footprinting) to increase our knowledge of the transcriptional regulation in *S. coelicolor*. Notwithstanding, the meta-curated network *Curated_RTB-FL-DBSCR* provides the most extensive and up-to-date reconstruction available for the regulatory circuitry in this organism and already portrays accurately the functional organization of *S. coelicolor* regulation. GRN inference from transcriptomic and DNA sequence data was performed and the inference from TF binding sites identification showed to be the best approach according to interactions inference assessment, topological assessment, and systems-level comparison. This final inferred network is a valuable guide for wet-lab experiments, since narrows down the search space of the possible TF for each gene. Besides, it can be used in computational models of *S. coelicolor.* From this network 13 new TFs were predicted to bind in the upstream region of five of the principal SARPs, most of which previously proved to affect indirectly antibiotic production or to be related to stress response or morphological differentiation. Finally, we compared *S. coelicolor* network to *C. glutamicum* GRN, showing one of the many potential applications of the curated network. There we found high conservation only for the basal machinery, which might be a result of the high plasticity of the transcriptional regulation. To visually explore the interactions validated by experiments and identify the role of the genes in the global regulatory network (e.g., global/local regulators) and functional annotation, we strongly suggest using Abasy Atlas.

Material and methods

Transcriptional regulatory interactions curation.

We performed a comprehensive review of the literature to identify experimentally-supported transcriptional regulatory interactions in *Streptomyces coelicolor* A3(2). We searched peer-reviewed articles in Google Scholar and PubMed using the keywords "*Streptomyces coelicolor"* AND "transcriptional" and its variations AND "regulation" and its variations. In the case where reviews were found, their references were followed to the original research papers. Then, we performed the curation and organized the interactions (Supplementary File 1). Experiments were classified according to their methodology and their names were standardized for the sake of clarity and easier evidence classification. We merged these interactions with two previously curated networks, one was reconstructed from an XML provided by the DBSCR team and the other one from RegTransBase[24] available at the Abasy Atlas website. These datasets are available from http://dbscr.hgc.jp/ and https://abasy.ccg.unam.mx. Abasy Atlas is a database of meta-curated bacterial GRNs for nine species including *S. coelicolor*[19]. It also provides historical snapshots for other model organisms such as *Escherichia coli, Bacillus subtilis, Corynebacterium glutamicum,* and *Mycobacterium tuberculosis*[19].

GRN inference from transcription factor binding sites.

To extend the regulons for the TFs identified in the literature, we used the set of "strong" interactions as prior (*Curated_FL(cS)-DBSCR(S))*. We reconstructed a position weight matrix (PWM) for every TF in the "strong" network using the non-overlapping up to − 300 to + 50 bp (with reference to the translation start codon) upstream regions of their TGs as input for three motif discovery algorithms. Namely, (1) MEME, an extension of the expectation–maximization algorithm for fitting finite mixture models[61]; (2) BioProspector, based on multiple Gibbs sampling[62]; and (3) MDscan, that employs a heuristic word-enumeration approach combined with statistical modeling[63]. Then, we used FIMO[64] (*p*-value threshold $= 1 \times 10^{-4}$) to identify TF-TG interactions. As most of the interactions curated are from *S. coelicolor* A3(2) strain M145 (plasmid-free), we excluded genes that are not part of the chromosome.

GRN inference from transcriptomic data.

We downloaded first the transcriptomic dataset for *S. coelicolor* available at the COLOMBOS database [32]. Then, we also download data from the NCBI Gene Expression Omnibus (GEO)[33]. From there we download an Affymetrix dataset (Platform GPL9417) and an RNA-Seq dataset (GPL26763. Afterward, we normalize the Affymetrix data using Robust Multi-chip Averaging (RMA) with the affy package[65] and used the gPCA package[66] to identify a batch effect in the data, which was corrected with Combat from the sva package[67], all of them are packages for R. The data counts on 137 transcriptomes for 7738 genes. As in the case of GRN inference from transcription factor binding sites, we only considered genes from the chromosome. We selected the best inference methods according to their outstanding performance in the DREAM challenge[18]. Moreover, we selected methods that have an implementation in R or Matlab and were well documented. The inference methods selected were: (1) CLR[68], a method that applies mutual information; (2) GENIE3[69], which applies tree-based regression and feature selection; (3) Inferelator[70], which applies regression and variable selection; (4) MRNET[71], which applies the maximum relevance/minimum redundancy algorithm; and (5) TIGRESS[72] which applies LARS combined with stability selection. Along with these methods, we used two modifications we propose in this work, Friedman, and Statmodel (Supplementary File 1). We provided all methods with a list of 137 TFs from the meta-curated network *Curated_FL-DBSCR-RTB* to infer causality[18,72].

Integration of individual inferences into a community GRN. To increase the precision of the predictions we used a community approach[18] integrating individual predictions from different algorithms. First, the individual predictions are sorted by their confidence score, keeping the most reliable ones at the beginning of the prediction list. Then, the average of the rank positions in the predictions is given as the community score for each interaction. For missing interactions in a prediction list, the position is equal to the size of the prediction list + 1. All community networks were pruned to the 23,908 first interactions with the highest score, which is the predicted size of the complete GRN of *S. coelicolor* reported by Abasy Atlas v2.4[19] according to the model developed in[38]. The model is constantly being updated on the Abasy Atlas website by the addition of new networks and interactions, which will cause a slight variation in the number of interactions[19].

Assessment of the inferred GRN. We computed the area under the precision-recall curve (AUPR) and the area under the receiver operating characteristics curve (AUROC) to assess our predictions using in-house scripts. The AUPR depicts the precision (1) as a function of the recall (2) obtained by the predictor. The AUROC depicts the relation between the recall, also called the true positive rate (TPR) (2), and the false positive rate (FPR) (3). Note that unknown actual interactions between genes in the GS will still be considered as FP[37]. For this reason, interactions involving genes that are not part of the GS were not considered.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

Statistical validation for interactions supported by ChIP experiments. We performed a statistical validation approach for ChIP-data[22] for those interactions supported by ChIP experiments and no "strong" experiment. Because of the lack of data in some articles, instead of performing the statistical validation finding the motifs from the ChIP data directly as in[22], we used the "strong" interactions as seed to construct the matrix models for the TFs. We used MEME[61] to build a position weight matrix (PWM) for each TF with at least 3 "strong" interactions. Then, we used FIMO with the matrix for each TF to scan the non-overlapping up to − 300 to + 50 bp upstream regions with reference to the translation start sites of *S. coelicolor*. We kept those TF binding sites with p-value < 1×10^{-4} and used these interactions to compare them with those supported by ChIP technologies, keeping the intersection of both sets (interactions supported by both ChIP and motif finding approaches).

Network similarity. We computed the characteristic structural properties for GRNs reported as global properties on the Abasy Atlas database[19]. Namely, regulators ($k_{out} > 0$) (%), direct regulatory interactions, self-regulation (%), maximum out-connectivity (%), network density, weakly connected components, genes in the giant component (%), feedforward circuits, complex feedforward circuits, 3-Feedback loops, average shortest path length, network diameter, average clustering coefficient, adjusted coefficient of determination (R^2_{adj}) of $P(k)$, and R^2_{adj} of $C(k)$. Then we used pairwise Pearson correlation among the profiles of the structural properties of the networks and cluster them according to the Euclidean distance among the correlations using Ward's method. To compute the pairwise network dissimilarity we used a Python implementation adapted from that proposed by *Schieber* et al.[43]. We use it with the parameters proposed by the authors (0.45, 0.45, 0.1), being the last one required to discriminate network size in terms of nodes (genes)[43]. Then, we clustered the networks by using Euclidean distances among its dissimilarity values using Ward's variance minimization algorithm as the linkage method.

System-level components. We applied the Natural Decomposition Approach (NDA), a biological-mathematical criterion to identify the components of the diamond-shaped structure of the GRNs[20], on every curated and inferred network. See the Supplementary methods (Supplementary File 1) for a brief description of the NDA, and[20] for further details. The assessment of the prediction of GRs was performed using in-house scripts for MCC, precision, and F1-score. Scores were computed for each GRs prediction obtained by the NDA with the different networks analyzed in this work. As GS we used the GRs previously reported in the literature from a review[28], from individual publications, and both (Supplementary Table 4).

Comparative analysis against *C. glutamicum*. For the regulogs analysis, we used as prior the strong regulatory networks Curated_FL(S)-DBSCR(S) for *S. coelicolor*, and 196627_v2020_s21_eStrong from the Abasy Atlas database for *C. glutamicum*[19], considering only the interactions between two genes both mapping to a locus tag. We used MEME to construct a PWM for every TF with at least three TGs using their upstream sequences. These sequences were defined as the non-overlapping regions of up to − 300 to + 50 bp with reference to the translation start codon and were obtained with retrieve-seq from RSAT[73]. Then, we used FIMO with the PWM of the TFs from *S. coelicolor* to find individual occurrences with a p-value $< 1 \times 10^{-4}$ in the upstream sequences of *C. glutamicum*. The same was done in the opposite direction. With this, we seek to alleviate network incompleteness by extrapolating known interactions from an organism to the other[59]. Predicted interactions were sorted by p-value and only the best scoring result was conserved for redundant interactions. Afterward, we used Orthofinder to find one-to-one ortholog relationships between both organisms. We used it due to its high accuracy[60]. The orthologs were used to further filter FIMO predictions to conserve interactions in which both TF and TG have a one-to-one orthologous relationship in the other organism. We considered the original "strong" network interactions at the beginning of the interactions list. The NDA was applied to both expanded GRNs to identify ortholog systems and only the genes with one-to-one orthologs in the other organism's network (GRN-wide orthologs) were considered in the analysis.

Data availability

The data set(s) supporting the results of this article are included within the article, its additional Files, or in Abasy Atlas at https://abasy.ccg.unam.mx/.

Received: 2 September 2021; Accepted: 31 January 2022
Published online: 18 February 2022

References

1. Hoskisson, P. A. & van Wezel, G. P. Streptomyces coelicolor. *Trends Microbiol.* **27**, 468–469. https://doi.org/10.1016/j.tim.2018.12.008 (2019).
2. Mast, Y. & Stegmann, E. Actinomycetes: The antibiotics producers. *Antibiotics (Basel)* https://doi.org/10.3390/antibiotics8030105 (2019).
3. Chen, S. *et al.* Roles of two-component system AfsQ1/Q2 in regulating biosynthesis of the yellow-pigmented coelimycin P2 in Streptomyces coelicolor. *FEMS Microbiol. Lett.* https://doi.org/10.1093/femsle/fnw160 (2016).
4. McLean, T. C., Wilkinson, B., Hutchings, M. I. & Devine, R. Dissolution of the disparate: Co-ordinate regulation in antibiotic biosynthesis. *Antibiotics (Basel).* https://doi.org/10.3390/antibiotics8020083 (2019).
5. Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141–147. https://doi.org/10.1038/417141a (2002).
6. Bednarz, B., Kotowska, M. & Pawlik, K. J. Multi-level regulation of coelimycin synthesis in Streptomyces coelicolor A3(2). *Appl. Microbiol. Biotechnol.* **103**, 6423–6434. https://doi.org/10.1007/s00253-019-09975-w (2019).
7. Bibb, M. 1995 Colworth Prize Lecture. The regulation of antibiotic production in Streptomyces coelicolor A3(2). *Microbiology* **142(Pt 6)**, 1335–1344. https://doi.org/10.1099/13500872-142-6-1335 (1996).
8. Chater, K. F. Regulation of sporulation in Streptomyces coelicolor A3(2): A checkpoint multiplex?. *Curr. Opin. Microbiol.* **4**, 667–673. https://doi.org/10.1016/s1369-5274(01)00267-3 (2001).
9. Bibb, M. J. Regulation of secondary metabolism in streptomycetes. *Curr. Opin. Microbiol.* **8**, 208–215. https://doi.org/10.1016/j.mib.2005.02.016 (2005).
10. Martin, J. F. & Liras, P. Cascades and networks of regulatory genes that control antibiotic biosynthesis. *Subcell Biochem.* **64**, 115–138. https://doi.org/10.1007/978-94-007-5055-5_6 (2012).
11. Zitnik, S., Zitnik, M., Zupan, B. & Bajec, M. Sieve-based relation extraction of gene regulatory networks from biological literature. *BMC Bioinform.* **16 Suppl 16**, S1. https://doi.org/10.1186/1471-2105-16-S16-S1 (2015).
12. Novichkov, P. S. *et al.* RegPrecise 3.0—A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genom.* **14**, 745. https://doi.org/10.1186/1471-2164-14-745 (2013).
13. Leyn, S. A. *et al.* Comparative genomics and evolution of transcriptional regulons in Proteobacteria. *Microb. Genom.* **2**, e000061. https://doi.org/10.1099/mgen.0.000061 (2016).
14. Metris, A. *et al.* SalmoNet, an integrated network of ten Salmonella enterica strains reveals common and distinct pathways to host adaptation. *NPJ Syst. Biol. Appl.* **3**, 31. https://doi.org/10.1038/s41540-017-0034-z (2017).
15. Staunton, P. M., Miranda-CasoLuengo, A. A., Loftus, B. J. & Gormley, I. C. BINDER: Computationally inferring a gene regulatory network for Mycobacterium abscessus. *BMC Bioinform.* **20**, 466. https://doi.org/10.1186/s12859-019-3042-8 (2019).
16. Escorcia-Rodríguez, J. M., Tauch, A. & Freyre-González, J. A. <em>Corynebacterium glutamicum</em> regulation beyond transcription: Organizing principles and reconstruction of an extended regulatory network incorporating regulations mediated by small RNA and protein-protein interactions. *bioRxiv*, 2021.2001.2007.423633. https://doi.org/10.1101/2021.01.07.423633 (2021).
17. Castro-Melchor, M., Charaniya, S., Karypis, G., Takano, E. & Hu, W. S. Genome-wide inference of regulatory networks in Streptomyces coelicolor. *BMC Genom.* **11**, 578. https://doi.org/10.1186/1471-2164-11-578 (2010).
18. Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat. Methods* **9**, 796–804. https://doi.org/10.1038/nmeth.2016 (2012).
19. Escorcia-Rodriguez, J. M., Tauch, A. & Freyre-Gonzalez, J. A. Abasy Atlas v2.2: The most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization. *Comput. Struct. Biotechnol. J.* **18**, 1228–1237. https://doi.org/10.1016/j.csbj.2020.05.015 (2020).
20. Freyre-Gonzalez, J. A., Alonso-Pavon, J. A., Trevino-Quintanilla, L. G. & Collado-Vides, J. Functional architecture of *Escherichia coli*: New insights provided by a natural decomposition approach. *Genome Biol.* **9**, R154. https://doi.org/10.1186/gb-2008-9-10-r154 (2008).
21. Santos-Zavaleta, A. *et al.* RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucl. Acids Res.* **47**, D212–D220. https://doi.org/10.1093/nar/gky1077 (2019).
22. Weiss, V. *et al.* Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database (Oxford)* **2013**, bas059. https://doi.org/10.1093/database/bas059 (2013).

23. Park, S. S. *et al.* Mass spectrometric screening of transcriptional regulators involved in antibiotic biosynthesis in Streptomyces coelicolor A3(2). *J. Ind. Microbiol. Biotechnol.* **36**, 1073–1083. https://doi.org/10.1007/s10295-009-0591-2 (2009).
24. Cipriano, M. J. *et al.* RegTransBase–a database of regulatory sequences and interactions based on literature: A resource for investigating transcriptional regulation in prokaryotes. *BMC Genom.* **14**, 213. https://doi.org/10.1186/1471-2164-14-213 (2013).
25. Gottesman, S. Bacterial regulation: Global regulatory networks. *Annu. Rev. Genet.* **18**, 415–441. https://doi.org/10.1146/annurev.ge.18.120184.002215 (1984).
26. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52. https://doi.org/10.1038/35011540 (1999).
27. Freyre-Gonzalez, J. A. & Tauch, A. Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: Novel insights from a dataset with a high genomic coverage. *J. Biotechnol.* **257**, 199–210. https://doi.org/10.1016/j.jbiotec.2016.10.025 (2017).
28. Martín, J. F., Santos-Beneit, F., Sola-Landa, A. & Liras, P. In *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria* (ed Frans J. de Bruijn) 257–267 (Wiley, 2016).
29. Martin, J. F. *et al.* Cross-talk of global nutritional regulators in the control of primary and secondary metabolism in Streptomyces. *Microb. Biotechnol.* **4**, 165–174. https://doi.org/10.1111/j.1751-7915.2010.00235.x (2011).
30. Ibarra-Arellano, M. A., Campos-Gonzalez, A. I., Trevino-Quintanilla, L. G., Tauch, A. & Freyre-Gonzalez, J. A. Abasy Atlas: A comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database (Oxford)*. https://doi.org/10.1093/database/baw089 (2016).
31. Camon, E. *et al.* The Gene Ontology Annotation (GOA) Database: Sharing knowledge in Uniprot with Gene Ontology. *Nucl. Acids Res.* **32**, D262-266. https://doi.org/10.1093/nar/gkh021 (2004).
32. Moretto, M. *et al.* COLOMBOS v3.0: Leveraging gene expression compendia for cross-species analyses. *Nucl. Acids Res.* **44**, D620–623.https://doi.org/10.1093/nar/gkv1251 (2016).
33. Clough, E. & Barrett, T. The gene expression omnibus database. *Methods Mol. Biol.* **1418**, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5 (2016).
34. Irizarry, R. A. *et al.* Summaries of affymetrix GeneChip probe level data. *Nucl. Acids Res.* **31**, e15. https://doi.org/10.1093/nar/gng015 (2003).
35. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127. https://doi.org/10.1093/biostatistics/kxj037 (2007).
36. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432. https://doi.org/10.1371/journal.pone.0118432 (2015).
37. Siegenthaler, C. & Gunawan, R. Assessment of network inference methods: How to cope with an underdetermined problem. *PLoS ONE* **9**, e90481. https://doi.org/10.1371/journal.pone.0090481 (2014).
38. Campos, A. I. & Freyre-Gonzalez, J. A. Evolutionary constraints on the complexity of genetic regulatory networks allow predictions of the total number of genetic interactions. *Sci. Rep.* **9**, 3618. https://doi.org/10.1038/s41598-019-39866-z (2019).
39. Barabasi, A. L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113. https://doi.org/10.1038/nrg1272 (2004).
40. Barabási, A.-L. & Pósfai, M. *Network Science*. Edición: 1 edn (Cambridge University Press, 2016).
41. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law distributions in empirical data. *SIAM Rev.* **51**, 661–703. https://doi.org/10.1137/070710111 (2009).
42. Khanin, R. & Wit, E. How scale-free are biological networks. *J. Comput. Biol.* **13**, 810–818. https://doi.org/10.1089/cmb.2006.13.810 (2006).
43. Schieber, T. A. *et al.* Quantification of network structural dissimilarities. *Nat. Commun.* **8**, 13928. https://doi.org/10.1038/ncomms13928 (2017).
44. Simpson, G. G. Mammals and the nature of continents. *Am. J. Sci.* **241**, 1–31. https://doi.org/10.2475/ajs.241.1.1 (1943).
45. Freyre-Gonzalez, J. A., Trevino-Quintanilla, L. G., Valtierra-Gutierrez, I. A., Gutierrez-Rios, R. M. & Alonso-Pavon, J. A. Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and *Escherichia coli* unveiled by the natural decomposition approach. *J. Biotechnol.* **161**, 278–286. https://doi.org/10.1016/j.jbiotec.2012.03.028 (2012).
46. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 6. https://doi.org/10.1186/s12864-019-6413-7 (2020).
47. Marbach, D. *et al.* Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res.* **22**, 1334–1349. https://doi.org/10.1101/gr.127191.111 (2012).
48. Lihu, A. & Holban, S. A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Brief Bioinform.* **16**, 964–973. https://doi.org/10.1093/bib/bbv022 (2015).
49. Liu, M. *et al.* Novel two-component system MacRS is a pleiotropic regulator that controls multiple morphogenic membrane protein genes in *Streptomyces coelicolor*. *Appl. Environ. Microbiol.* https://doi.org/10.1128/AEM.02178-18 (2019).
50. Valdez, F., Gonzalez-Ceron, G., Kieser, H. M. & Servi, N. G. L. The Streptomyces coelicolor A3(2) lipAR operon encodes an extracellular lipase and a new type of transcriptional regulator. *Microbiology (Reading)* **145**(Pt 9), 2365–2374. https://doi.org/10.1099/00221287-145-9-2365 (1999).
51. Aigle, B., Wietzorrek, A., Takano, E. & Bibb, M. J. A single amino acid substitution in region 1.2 of the principal sigma factor of Streptomyces coelicolor A3(2) results in pleiotropic loss of antibiotic production. *Mol. Microbiol.* **37**, 995–1004.https://doi.org/10.1046/j.1365-2958.2000.02022.x (2000).
52. Li, L., Jiang, W. & Lu, Y. A novel two-component system, GluR-GluK, involved in glutamate sensing and uptake in *Streptomyces coelicolor*. *J. Bacteriol.* https://doi.org/10.1128/JB.00097-17 (2017).
53. Mao, X. M. *et al.* Positive feedback regulation of stgR expression for secondary metabolism in Streptomyces coelicolor. *J. Bacteriol.* **195**, 2072–2078. https://doi.org/10.1128/JB.00040-13 (2013).
54. Yang, H. *et al.* The tyrosine degradation gene hppD is transcriptionally activated by HpdA and repressed by HpdR in Streptomyces coelicolor, while hpdA is negatively autoregulated and repressed by HpdR. *Mol. Microbiol.* **65**, 1064–1077. https://doi.org/10.1111/j.1365-2958.2007.05848.x (2007).
55. Urem, M. *et al.* OsdR of Streptomyces coelicolor and the dormancy regulator DevR of Mycobacterium tuberculosis control overlapping regulons. *mSystems.* https://doi.org/10.1128/mSystems.00014-16 (2016).
56. Nguyen, K. T. *et al.* A central regulator of morphological differentiation in the multicellular bacterium Streptomyces coelicolor. *Mol. Microbiol.* **46**, 1223–1238. https://doi.org/10.1046/j.1365-2958.2002.03255.x (2002).
57. Traag, B. A., Kelemen, G. H. & Van Wezel, G. P. Transcription of the sporulation gene ssgA is activated by the IclR-type regulator SsgR in a whi-independent manner in Streptomyces coelicolor A3(2). *Mol. Microbiol.* **53**, 985–1000. https://doi.org/10.1111/j.1365-2958.2004.04186.x (2004).
58. Amin, R., Reuther, J., Bera, A., Wohlleben, W. & Mast, Y. A novel GlnR target gene, nnaR, is involved in nitrate/nitrite assimilation in Streptomyces coelicolor. *Microbiology* **158**, 1172–1182. https://doi.org/10.1099/mic.0.054817-0 (2012).
59. Alkema, W. B., Lenhard, B. & Wasserman, W. W. Regulog analysis: Detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res.* **14**, 1362–1373. https://doi.org/10.1101/gr.2242604 (2004).
60. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238. https://doi.org/10.1186/s13059-019-1832-y (2019).

61. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
62. Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 127–138 (2001).
63. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immuno-precipitation microarray experiments. *Nat. Biotechnol.* **20**, 835–839. https://doi.org/10.1038/nbt717 (2002).
64. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064 (2011).
65. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315. https://doi.org/10.1093/bioinformatics/btg405 (2004).
66. Reese, S. E. *et al.* A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* **29**, 2877–2883. https://doi.org/10.1093/bioinformatics/btt480 (2013).
67. Leek JT, J. W., Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, Torres LC. (R package version 3.38.0, 2020).
68. Faith, J. J. *et al.* Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**, e8. https://doi.org/10.1371/journal.pbio.0050008 (2007).
69. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE.* https://doi.org/10.1371/journal.pone.0012776 (2010).
70. Bonneau, R. *et al.* The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.* **7**, R36. https://doi.org/10.1186/gb-2006-7-5-r36 (2006).
71. Meyer, P. E., Kontos, K., Lafitte, F. & Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform. Syst. Biol.*, 79879. https://doi.org/10.1155/2007/79879 (2007).
72. Haury, A. C., Mordelet, F., Vera-Licona, P. & Vert, J. P. TIGRESS: Trustful inference of gene REgulation using stability selection. *BMC Syst. Biol.* **6**, 145. https://doi.org/10.1186/1752-0509-6-145 (2012).
73. Nguyen, N. T. T. *et al.* RSAT 2018: Regulatory sequence analysis tools 20th anniversary. *Nucl. Acids Res.* **46**, W209–W214. https://doi.org/10.1093/nar/gky317 (2018).

Acknowledgements

The authors thank Yuko Makita and Prof. Kenta Nakai from the RIKEN Center for Sustainable Resource Science for kindly providing data from DBSCR. This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT-UNAM) [IN205918 and IN202421 to J.A.F.-G.]. A.Z.-A. received a doctoral scholarship [call 647 (2014)] from the Colombian Administrative Department of Science, Technology, and Innovation (COLCIENCIAS). J.M.E.-R. is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM); he received fellowship 959406 from CONACYT. We thank one reviewer for helpful and encouraging suggestions.

Author contributions

Conceptualization, J.A.F.G.; Methodology, A.Z.A., J.M.E.R., J.K.G.K., and J.A.F.G.; Software, A.Z.A., J.M.E.R., J.K.G.K. and J.A.F.G; Validation, A.Z.A., J.M.E.R., and J.A.F.G; Formal Analysis, A.Z.A., J.M.E.R., J.K.G.K., and J.A.F.G; Investigation, A.Z.A., J.M.E.R., J.K.G.K., and J.A.F.G; Resources, J.A.F.G.; Data Curation, A.Z.A.; Writing – Original Draft Preparation, A.Z.A., J.M.E.R, and J.K.G.K.; Writing – Review & Editing, A.Z.A., J.M.E.R, and J.A.F.G; Visualization, J.M.E.R.; Supervision, J.A.F.G.; Project Administration, J.A.F.G.; Funding Acquisition, J.A.F.G.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-06658-x.

Correspondence and requests for materials should be addressed to J.A.F.-G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

VI. *Rhizobium etli* CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network

(Taboada-Castro et al., 2022)

frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fmicb.2022.947678

Rhizobium etli CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network

Hermenegildo Taboada-Castro[1], Jeovanis Gil[2], Leopoldo Gómez-Caudillo[1], Juan Miguel Escorcia-Rodríguez[3], Julio Augusto Freyre-González[3] and Sergio Encarnación-Guevara[1]*

[1]Proteomics Laboratory, Program of Functional Genomics of Prokaryotes, Center for Genomic Sciences, National Autonomous University of Mexico, Cuernavaca, Morelos, Mexico, [2]Division of Oncology, Section for Clinical Chemistry, Department of Translational Medicine, Lund University, Lund, Sweden, [3]Regulatory Systems Biology Research Group, Program of Systems Biology, Center for Genomic Sciences, National Autonomous University of Mexico, Mexico City, Mexico

OPEN ACCESS

EDITED BY
Reiner Rincón Rosales,
Tuxtla Gutierrez Institute of Technology,
Mexico

REVIEWED BY
Luyao Wang,
Chinese Academy of Agricultural Sciences,
China
Betsy Peña-Ocaña,
Instituto Nacional de Cardiologia Ignacio
Chavez, Mexico
Clara Ivette Rincón Molina,
Instituto Tecnológico de Tuxtla Gutiérrez/
TecNM, Mexico

*CORRESPONDENCE
Sergio Encarnacion Guevara
encarnacion@ccg.unam.mx

SPECIALTY SECTION
This article was submitted to
Microbe and Virus Interactions With Plants,
a section of the journal
Frontiers in Microbiology

RECEIVED 19 May 2022
ACCEPTED 05 September 2022
PUBLISHED 13 October 2022

CITATION
Taboada-Castro H, Gil J, Gómez-Caudillo L, Escorcia-Rodríguez JM, Freyre-González JA and Encarnación-Guevara S (2022) *Rhizobium etli* CFN42 proteomes showed isoenzymes in free-living and symbiosis with a different transcriptional regulation inferred from a transcriptional regulatory network. *Front. Microbiol.* 13:947678. doi: 10.3389/fmicb.2022.947678

A comparative proteomic study at 6h of growth in minimal medium (MM) and bacteroids at 18days of symbiosis of *Rhizobium etli* CFN42 with the *Phaseolus vulgaris* leguminous plant was performed. A gene ontology classification of proteins in MM and bacteroid, showed 31 and 10 pathways with higher or equal than 30 and 20% of proteins with respect to genome content per pathway, respectively. These pathways were for energy and environmental compound metabolism, contributing to understand how *Rhizobium* is adapted to the different conditions. Metabolic maps based on orthology of the protein profiles, showed 101 and 74 functional homologous proteins in the MM and bacteroid profiles, respectively, which were grouped in 34 different isoenzymes showing a great impact in metabolism by covering 60 metabolic pathways in MM and symbiosis. Taking advantage of co-expression of transcriptional regulators (TF's) in the profiles, by selection of genes whose matrices were clustered with matrices of TF's, Transcriptional Regulatory networks (TRN´s) were deduced by the first time for these metabolic stages. In these clustered TF-MM and clustered TF-bacteroid networks, containing 654 and 246 proteins, including 93 and 46 TFs, respectively, showing valuable information of the TF's and their regulated genes with high stringency. Isoenzymes were specific for adaptation to the different conditions and a different transcriptional regulation for MM and bacteroid was deduced. The parameters of the TRNs of these expected biological networks and biological networks of *E. coli* and *B. subtilis* segregate from the random theoretical networks. These are useful data to design experiments on TF gene–target relationships for bases to construct a TRN.

KEYWORDS

transcriptional regulatory network, *Rhizobium etli*, nitrogen fixation, *Phaseolus vulgaris*, free life, proteomics, Isoenzymes

Introduction

Rhizobium etli CFN42 is a soil bacterium classified as an alpha-proteobacterium able to establish a symbiotic relationship with leguminous plants, and this faculty is shared with some members of the beta-proteobacterium group (Andrews and Andrews 2017; Lardi and Pessi 2018; Dicenzo et al. 2019). When the seeds of the bean plant *Phaseolus vulgaris* germinate with *R. etli* CFN42 in a tropical soil, chemical communication starts in the roots to establish a symbiotic relationship. In this process, the root of the plant develops an infection thread through which the bacteria are internalized and travel with some duplications, while the root cortex gives rise to the nodule primordium. When the infection thread reaches the nodule cells, the bacteria are released into organelle-like membranes derived from the host cell plasmalemma called the symbiosome in the nodule. In this stage, *Rhizobium* has some additional duplications, but very soon they stop growing and become pleomorphic, and symbiotic biological nitrogen fixation (SNF) starts (Rascio and La Rocca 2013; Dicenzo et al. 2019). These pleomorphic bacteria, called bacteroids, carry out the expensive reduction of atmospheric N_2 to ammonium, which is exported to the plant cell, in an exchange of carbon compounds supplied from the photosynthesis of the plant cells. This photosynthate is metabolized by the bacteroid to sustain the SNF (Rascio and La Rocca 2013). Rhizobial inoculants are inexpensive alternatives to environmentally polluting industrial nitrogen fertilizers, with significant impacts on the livelihood of the community. Replacing the use of chemical fertilizers with SNF is a relevant strategy against global warming, favoring sustainable agriculture for the production of grains for human consumption, feed and pasture species (Oldroyd et al. 2011; Ferguson et al. 2019). Proteomic studies on symbiosis have been reported (Larrainzar and Wienkoop 2017; Lardi and Pessi 2018; Liu et al. 2018; Khatabi et al. 2019) and a search for binding sites (motifs) of transcriptional regulators (TFs; Fischer 1994; Tsoy et al. 2016; Rutten and Poole 2019). Now, the first study on O_2-dependent regulation of the SNF by extending known motifs by bioinformatic methods was performed to establish a regulatory network of proteins and global TFs considering 50 genomes of the Alphaproteobacteria by extending known motifs recognized by the TFs based on a phylogenetic footprinting approach, i.e., the *nifA-rpoN* regulon of nitrogen fixation in the Alphaproteobacteria group was searched, and the deduced matrix from the motifs of the TFs inferred with a strict *p*-value was used to scan with the Run profile tool in the Regpredict site (Novichkov et al. 2010). Using the same *p*-value to search for additional regulon members, 95 operons with potential NifA-binding sites comprising 280 genes were found in Alphaproteobacteria (Tsoy et al. 2016). The NifA-RpoN regulon of *R. etli* CFN42 was determined experimentally and with bioinformatic methods; it consisted of 120 genes, which indicates that the aforementioned study of the NifA regulon in Alphaproteobacteria is highly conservative, highlighting that genes not directly related to nitrogen fixation were found (Salazar et al. 2010), as was also observed (Tsoy et al. 2016). Based on the biological functions resulting from protein interactions, the symbiosis interactome of *Sinorhizobium meliloti* with its host plants was proposed by computational methods, which is composed of 440 proteins involved in 1041 unique interactions (Rodriguez-Llorente et al. 2009).

These data show that the symbiotic nitrogen fixation regulatory circuitry is suspected to be complex. Most of the symbiotic stage protein profiles in the cited literature include TFs (see above), but efforts are needed to infer the genetic circuitry between TFs and the proteins for each profile. We need to take advantage of the co-expression of TFs with potential target proteins in a proteomic profile due to the enrichment of common motif sites involved in the transcriptional regulation of these genes (Van Helden et al. 1998; McGuire et al. 2000; Aerts et al. 2003; Ihuegbu et al. 2012), considering autoregulation of the TFs and that they are involved in the transcriptional regulation of proteins of their respective profile.

We recently constructed the RhizoBindingSites database[1] (Taboada-Castro et al. 2020), a DNA-motif site collection based on the inferred motifs from each gene recognizing a site in its own promoter region, covering nine representative genomes of the taxon Rhizobiales. This algorithm aligns all the upstream regions of the orthologous genes per gene per genome to search for pairs or conserved position-specific trinucleotides (dyads) to define the motifs (Defrance et al. 2008). These dyads represented in a position-specific scoring matrix (PSSM; Hertz et al. 1990) were used to scan all the genes of a respective genome. These output data per gene per genome were fractionated at low, medium, and high stringency of *p*-value ranges. These data are used to match protein profiles from experimental or theoretical data to predict transcriptional regulatory networks at the desired *p*-value, or using the "auto" option, in which in each round the algorithm selects the data with the lowest *p*-value (high stringency) by searching from the highly strict to low strict data in the proper genome, assuring the output data are with the highly strict *p*-value as possible (Taboada-Castro et al. 2020). This database contains from one to five conserved motifs represented in matrices per

[1] http://rhizobindingsites.ccg.unam.mx/ (accessed September 8, 2022).

gene that have different significance. At the moment, it is not clear which motifs conserved in a gene are directly involved in the recognition of the ARN polymerase and which *p*-value corresponds to the biological action of the TF. The lowest *p*-values are generally used (Tsoy et al. 2016). Inferred data on transcriptional regulation in the SNF are important to accelerate experiments on transcriptional regulation to define TF gene targets, which are basic components of a regulatory network (Resendis-Antonio et al. 2005, 2012; Tsoy et al. 2016).

For *R. etli* CFN42, a systems biology of the metabolic activity during SNF integrating proteome and transcriptome data was used, i.e., 415 proteins and 689 upregulated genes, respectively. From this, 292 unique proteins were identified. This constraint-based model was used to simulate metabolic activity during SNF, and 76.83% of enzymes were justified. The metabolic pathways sustaining SNF activity were discussed compared with aerobic growth in succinate ammonium minimal medium (MM; Resendis-Antonio et al. 2011).

In this work, the study of the SNF proteome of *R. etli* CFN42 was revised with the same experimental conditions (Resendis-Antonio et al. 2011), comparing the aerobic growth at 6 h in MM and the symbiosis at 18 days post-inoculation (dpi). A total of 1730 proteins were identified in MM and 730 in bacteroids; compared to the first report (Resendis-Antonio et al. 2011), it contains 2.5 times more proteins identified in symbiosis. Similar pathways supporting the SNF and a role of the different genome compartments were identified, and new pathways related to adaptation to environmental conditions were described. A new study of the vicinity of the genes expressed in the genome of *R. etli* CFN42 showed specific zones for growth in MM and bacteroid. The chromosome has more genes for growth in MM than in bacteroid, which were more scattered, while for the SNF, the symbiotic plasmid d (p42d) has more genes than for growth in MM. The MM and bacteroid proteome profiles included 127 and 62 TFs, respectively. A potential transcriptional regulatory network for MM and bacteroid was constructed using the RhizoBindingSites database and the prediction of regulatory network approach with the auto option, proposing on average 87% of TF gen-target relationships with *p*-values ranging from 1.0e-5 to 1.0e-20, which represents a strict criterion.

Assuming that the TFs in MM and bacteroid profiles are involved in the transcription of their corresponding protein profile, a bioinformatic study with conserved motifs of TFs was used to establish a TF gen-target relationship, and a transcriptional regulatory network for MM and bacteroid was proposed.

Materials and methods

Culture of *Rhizobium etli* CFN42 strain

The *R. etli* CFN42 strain was grown in minimal medium with ammonium chloride and succinic acid as previously reported (Taboada et al. 2018), it was cultured for 6 h, and the cells were pelleted by centrifugation at 7,500 × *g* at 5°C, for 5 min.

Plant inoculation with *Rhizobium etli* CFN42

The *Phaseolus vulgaris* bean seeds were surface sterilized and placed on 0.8% agar in Petri dishes (Wacek and Brill 1976). Each seed was inoculated with 10^5 *R. etli* cells previously washed with sterilized distilled water after growing in a peptone-yeast-rich medium as described (Encarnación et al. 1995); after 18 days post-inoculation, the bacteroid were extracted on a percoll gradient as described (Romanov et al. 1994).

Sample preparation

The cell pellets from both free-living bacteria and bacteroid were lysed in a solution containing 7 M urea, 2% CHAPS, 1 mM DTT in 50 mM Tris–HCl pH 8. Cells were resuspended in the lysis buffer and sonicated on the ice for 15 microns. Samples were incubated with an additional 20 mM DTT for 30 min at 40°C to completely reduce disulfide bridges. Cysteine residues were alkylated with 50 mM IAA for 30 min at room temperature in darkness. After centrifugation, the proteins were collected in the supernatant. Proteins were precipitated overnight with cold ethanol (9 volumes) and washed with a 90% ethanol solution.

The precipitate was dissolved in sodium deoxycholate SDC 0.5%, SDS 0.5%, in 100 mM triethylammonium bicarbonate buffer (TEAB). Proteins were submitted to a chemical acetylation reaction of all lysine residues as previously described (Gil et al. 2017; Gil and Encarnación-Guevara 2022). Fully acetylated proteins were dissolved in AmBiC 50 mM, SDC 0.5%, and digested by adding trypsin to a ratio of 1:50 (enzyme:protein), and the reaction was incubated for 16 h at 37°C. SDC was removed with ethyl acetate acidified with trifluoroacetic acid (TFA) as previously reported (Gil et al. 2017; Gil and Encarnación-Guevara 2022). The peptide mixture was dried on a Speed-Vac and stored at −80°C until MS analysis.

LC–MS/MS and data analysis

Peptides were dissolved in 0.1% TFA in water and loaded on an RSLC nano UPLC system (Ultimate 3000, Dionex) coupled to a Q-Exactive high-resolution mass spectrometer (Thermo Fischer Scientific). The chromatographic conditions, as well as the MS acquisition parameters, were as previously described (Gil et al. 2017). The analysis was performed at the Proteomics Core Facility, Ecole Polytechnique Fédérale de Lausanne in Switzerland. The data presented in the study are deposited in the ProteomeXchange Consortium via the PRIDE repository

(Perez-Riverol et al. 2022), accession numbers PXD035204 and 10.6019/PXD035204.

Raw data were processed for peptide and protein identification/quantification using the MaxQuant platform. The database search parameters were as follows: Trypsin/R was selected as the digestion enzyme, up to two missed cleavages were allowed, carbamidomethylcysteine and acetylated lysine were set as fix modifications, and oxidized methionine was considered variable. The database used for protein identification was released in 2006 (González et al. 2006) and is publicly available through the UniProt repository. Three biological replicates of each condition were included in the study. Proteins and peptides were identified with an FDR of 1% based on the target-decoy strategy integrated in the software.

Statistical analysis

Only proteins identified with at least two peptides and one of these unique peptides and at least two intensity values in each condition were used for statistical analysis. The protein abundance was normalized, and missing values were imputed with the Random Forest method (missForest, R package; Stekhoven and Bühlmann 2012). The PCA was carried out on the protein intensity correlation matrix (FactoMiner, R package; Lê et al. 2008) to generate a protein abundance pattern for the cell lines. To determine whether any component could distinguish between the cell lines, the sample scores for each component were plotted. After finding the component, we identified the more correlated proteins in that component with discriminatory capacity using the square cosine of the correlation matrix between the components and the proteins (Abdi and Williams 2010). It is observed in the graphic, MM and bacteroid conditions were clustered separately but grouped by condition, recovering a great diversity of data, 64.8 and 19.3% of data for one and two dimensions, respectively, giving a total of 84.1% (Figure 1). A total of 1,730 and 735 proteins were significantly identified in the minimal medium and bacteroid, respectively. In addition, 322 proteins were without change in their expression.

Metabolic pathways analysis

Overrepresentation of pathways was performed online employing the Gene List Analysis tool on the PANTHER Classification System site.[2] Only proteins with an absolute value of association with a *p*-value equal or greater than 0.5 with data of the first two components (Abdi and Williams 2010) were selected for comparative overrepresentation analysis based on Gene Ontology (Ashburner et al. 2000). To obtain the GO terms significantly overrepresented in this experiment we used the hypergeometric test and only processes with a *p*-value less than 0.05 were selected. The presence of genes for each metabolic pathway was compared as percent respect of the background number of genes per pathway in MM and bacteroid profiles (Figures 3, 4).

Metabolic maps construction

For analysis of metabolic pathways in MM and bacteroid, and genes without changes in their expression, the Kegg mapper[3] was used (Kanehisa 2017). This mapper uses the KO Kegg Orthology, which is based on the function of the ortholog genes. The K identifiers for *R. etli* CFN42 were obtained for the entire genome with the application blastKOALA[4] (Kanehisa et al. 2016), the input was the sequences in FASTA format of the genes from *R. etli* CFN42 genome divided into two parts, then the *R. etli* CFN42 locus tag identifier was associated to the K identifiers, and a list including MM, bacteroid and with no change expression proteins was used in the Kegg mapper (Supplementery Table 2). Obtention of the EC number was from the KO Orthology application from Kegg[5] (Kanehisa et al. 2016).

Design of a regulatory network

The protein profiles of *R. etli* CFN42 grown in minimal medium (MM) at 6 h and of bacteroid isolated from nodules at 18 days post-inoculation of the bean plant *Phaseolus vulgaris*, were used to construct a transcriptional regulatory network with the application "Prediction of transcriptional regulatory networks" of the RhizoBindingSites database (Taboada-Castro et al. 2020). Briefly, this database contains predicted matrices deduced from conserved dyads (Defrance et al. 2008), composed of position-specific di or tri-nucleotides in the orthologs genes of each gene in members of the Rhizobiales taxon. These position-specific nucleotides were converted into a matrix format, which describes the conserved motifs for each gene (Hertz et al. 1990), the dyad analysis of the footprinting discovery algorithm deduced from one to five matrices per gene. The matrices of the TF's were used to scan with a matrix-scan RSAT analysis (Nguyen et al. 2018), all the upstream regulatory sequences of the genes, establishing TF gene-target relationships data, which is in the motif information window of the RhizoBindingSites database (Taboada-Castro et al. 2020; RhizoBindingSites database user guide), this information is used in the "Prediction of a transcriptional regulatory network" application.

[2] http://www.pantherdb.org/ (accessed September 8, 2022).

[3] https://www.genome.jp/kegg/tool/map_pathway.html (accessed September 8, 2022).

[4] https://www.kegg.jp/blastkoala/ (accessed September 8, 2022).

[5] https://www.genome.jp/kegg/ko.html (accessed September 8, 2022).

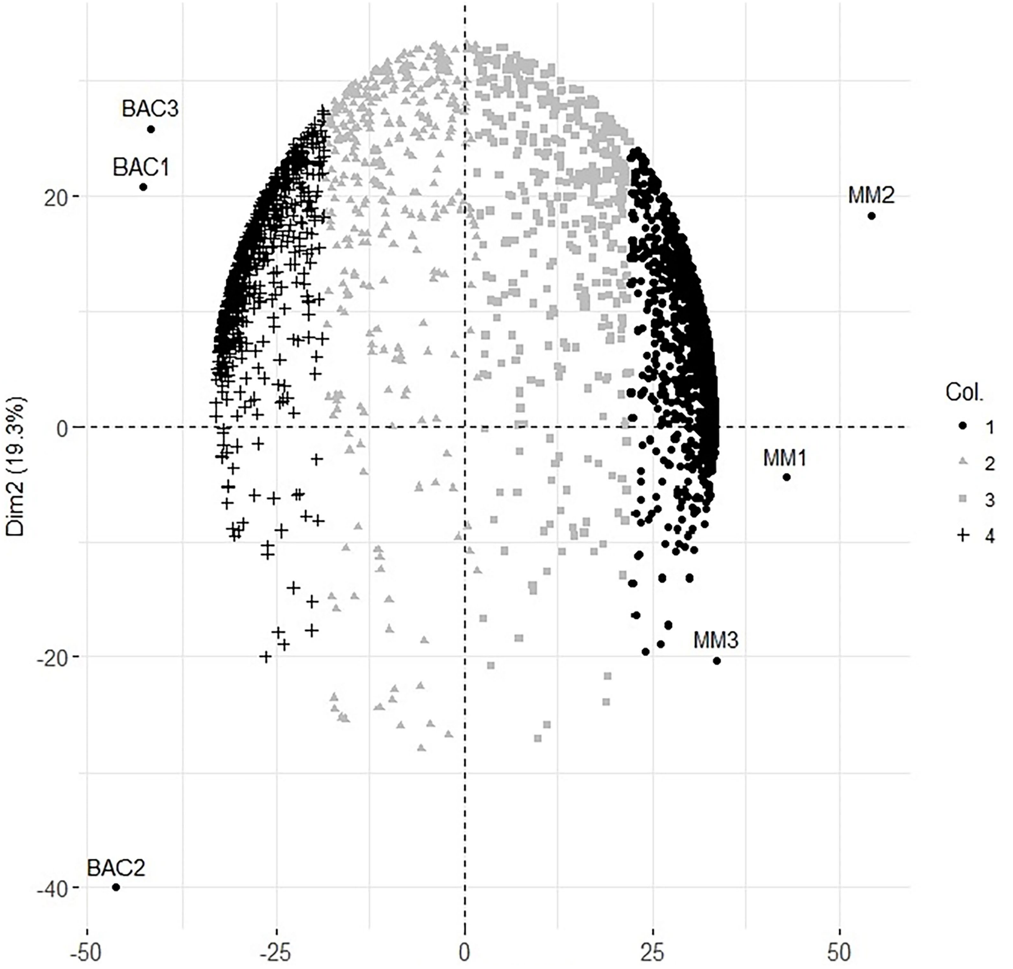


FIGURE 1
Principal component analysis of protein expression at 6 h of minimal medium growth MM1, MM2, MM3 and 18days post inoculation Bacteroids of the symbiosis from *Rhizobium etli* CFN42 with the plant Phaseolus vulgaris Bac1, Bac2, Bac3 biological replicates.

For the prediction of transcriptional regulatory networks, a three-step method was implemented. The first step consisted in to construct networks with the MM protein profile including the 127 co-expressed TF´s. As well as, the bacteroid protein profile with the 62 co-expressed TFs, with the application "Prediction of the transcriptional regulatory network" from the RhizoBindingSites database, with the "auto" option. This step, is to eliminate the genes of the proteins not recognized by any TF, and TF's whose matrices had no homology with any upstream regulatory sequences of potential target genes. With the option "auto". With this option, the application searches for TF gene-target relationships for each of the TF's co-expressed with the entered genes by looking into the motif information data from the RhizoBindingSites database. Only 1,336 genes, including 107 TF's genes from MM (Supplementary Table 1A) and 583 genes including, 50 TF's co-expressed bacteroid genes (Supplementary Table 1B), respectively, were found with a relationship, giving rise to hypothetical regulons available (Supplementary Tables 1A, B). The TF-matrices may have homology with the upstream regulatory sequences of target genes three levels of stringency, low stringency (*p*-value from 1.0e-4 to 9.9e-4), medium stringency level (*p*-value 1.0e-5 to 9.9e-5) or highly strict (from the *p*-value 1.0e-6 to lower *p*-values). In the second step, a matrix- clustering analysis for each condition, with the matrices of the 1336 genes of MM, as well as, the matrices of the 583 bacteroid genes including their respective TF´s was done (Castro-Mondragon et al. 2017). This step is to eliminate false-positive data as possibly, since the motifs are short conserved functionally compromised sequences (Ihuegbu et al. 2012), to avoid possible TF gene-target relationships by chance. In this analysis, the matrix of a TF should be grouped by homology with the nucleotide sequence of matrices of the potential target genes. Matrix-clustering algorithm creates the file clusters_motif_names.tab, which is edited to obtain all the genes whose matrices were clustered containing at least two different genes. Only the clusters, including matrices of a TF or TFs (Clustered-TF) were selected from MM (Supplementary Table 1C) and bacteroid profiles (Supplementary Table 1D), the NCBI genomic information of the genes was added to these tables as well as the Clustered-TF for each cluster (column headed "Clustered-TF" Supplementary Tables 1C, D). An alignment of MM and bacteroid matrices from matrix-clustering showed how much conserved are motifs in the clusters (Supplementary Tables 1E, F, respectively). The matrices were grouped into 207 and 92 clusters for MM and bacteroid, respectively. In this second step, additional depuration of genes after a matrix-clustering analysis was observed since only 655 genes, including 93 TF's genes from MM, and 247 genes, including 46 TF's genes, were clustered. A TF gen-target relationship with only genes of a clusters was confirmed (Results and discussion, Appendix G and H). In the third step, second networks were constructed (as in the first step) only with clustered-TF genes, called "Clustered-TF-MM" and "Clustered-TF-BACTEROID" (Figure 5). and cluster_97 and cluster_112 from bacteroid were chosen. For cluster_34, all the genes had a TF gene-target relationship. For cluster_195, 22 out of 27 genes were connected (Supplementary Table 1G). For cluster_97, 21 of 26 genes were connected and for cluster_112, 21 from 22 genes were connected (Supplementary Table 1H). It is worth noticing that, after the matrix-clustering grouping genes, all the genes for each condition had one or more relationships. Quality of MM, bacteroid, clustered-TF-MM and Clustered-TF-BACTEROID networks were analyzed (Results and discussion, Figure 5). Then, the transcriptional regulatory networks of MM and bacteroid protein profiles are constructed with motifs interspecies conserved.

These data confirmed that clustered matrices of genes are strongly related to the structure of a network, and these genes probably represent hubs.

To search for expected transcriptional regulation for isoenzymes in MM and bacteroids, the tables of the transcriptional regulatory networks described above were ordered in decreasing order by the column headed "*p*-value" (Supplementary Tables 1A, B). These tables were identified in the right column with the condition they pertain giving rise to new files ordered from MM and bacteroid and Clustered-TF-MM and Clustered-TF-BACTEROID separately. A Supplementary Table 1I containing the "K" number with a *R. etli* CFN42 locus tag identifier and the pertaining physiological condition per row was constructed. Then, the table from Supplementary Table 1I was paired with new files from the MM and bacteroid and Clustered-TF-MM and Clustered-TF-BACTEROID aforementioned. A new file with three groups of columns was produced; the first group contains information on the expected regulation with information from MM and bacteroid networks (with columns; Condition, Locus tag, K number, Upstream_region, Matrix_ID, Chain, End_motif, Start_motif, Site, Weight, *p*-value, and Significance). The second group of columns contains information of the expected transcriptional regulation with information from the Clustered-TF-MM and Clustered-TF-BACTEROID with columns headed as the MM and bacteroid data. The third group of columns contains information of the enzymatic function of the K numbers headed as; Condition, Locus tag, K number, Compartment, locus name, COG number, COG group, Function from KO orthology, and Function from NBCI. To look for the expected transcriptional regulation for the same K number with different locus tag in MM and bacteroid, it was located in the column "Matrix_ID" with a format "RHE_RS13345_m5," which means the TF is RHE_RS13345 and "_m5" means the matrix number "5" of the TF (as was mentioned, a TF has one to five matrices; Supplementary Table 4).

Properties of networks

The most recent *E. coli* and *B. subtilis* "strong" evidence networks were retrieved from Abasy Atlas v2.2 (Escorcia-Rodríguez et al. 2020). Both networks only include regulatory interactions supported by experiments showing a direct interaction between the transcription factor and the upstream region of the target gene. We contrasted the inferred networks with the *E. coli* and *B. subtilis* curated networks as a positive control, and 1000 Erdös-Rényi random networks parametrized having the same number of nodes and edges as the corresponding biological networks as a negative control.

We computed several global structural properties for regulatory networks. Namely, regulators ($k_{out} > 0$), self-regulations, maximum out-connectivity, giant component size, network density, feedforward circuits, complex feedforward circuits, 3-Feedback loops, average shortest path length, network diameter, average clustering coefficient, adjusted coefficient of determination (R^2_{adj}) of $P(k)$, and R^2_{adj} of $C(k)$. Regulators, self-regulations, maximum out-connectivity, and giant component size were normalized by the number of nodes in the network. The density was included as the product of the network density and the fraction of regulators. Network diameter was normalized by (number of nodes – 2; as if no shortcuts would exist). 3-feedback loops, feedforward loops, and complex feedforward loops were normalized by the number of potential motifs in the network, defined as:

$$\frac{n!}{(n-r)!}\cdot\left(\frac{TF_n}{n}\right)^{TF_m}$$

Where n is the number of nodes in the network, r is the number of nodes in the motif ($r = 3$), TF_n is the number of TFs in the network, and TF_m is the number of TFs required for each motif type ($TF_m = 3$ for 3-feedback loops, and $TF_m = 2$ for feedforward and complex feedforward loops). We scaled the values of each property vector across networks to the range between 0 and 1, inclusively. Then, we clustered networks and properties using Ward's method. Further, we used pairwise Pearson correlation for the network property profiles and clustered the networks according to the Euclidean distance using Ward's method.

Hierarchy reconstruction of networks

First, we removed all the structural genes (nodes having $k_{out} = 0$) and their interactions from the network. Next, we classified each network edge (a, b) as 'top-down' if $k_a^{out} > k_b^{out}$ (where k_n^{out} is the out-connectivity of node *n*), otherwise, it was classified as 'bottom-up'. Then, we removed all the 'bottom-up' edges from the network. This step removed the feedback circuits present in the network, transforming it into a directed acyclic graph. Then we applied a modified topological sorting algorithm that returned the list of layers composing the hierarchy, where each node in a layer only can regulate nodes in lower layers. As the number of 'bottom-up' edges is low (<5% in average), our strategy maintains the global structure of the network to reveal the hierarchy. Besides structural nodes, no other nodes are removed, and 'bottom-up' edges can be added back to the hierarchy to reveal the feedback among layers and reconstruct the original network.

Results and discussion

In a previous study, we identified 292 proteins of the symbiotic state at 18 days post-inoculation (Resendis-Antonio et al. 2011); now, we discuss new data covering 2.5 times more proteins from symbiosis in this work. Principal components (one and two) covered 84.1% of the total initial data (Figure 1). The update of the *R. etli* CFN42-*Phaseolus vulgaris* bean plant symbiosis is with 1,730, 738, and 323 protein profiles for MM, bacteroid, and without no change in their expression, respectively (see below). There were 39.7% of common proteins in the bacteroid between the previous report (Resendis-Antonio et al. 2011) and this study. The low coverage observed in the new data may be due to the great diversity of different experiments collected for the last study. While in the new data, the variation in the experimental condition was from only two biological replicates, because our interest was to take advantage of the TF and non-TF protein co-expression (Galán-Vásquez and Perez-Rueda 2019), under the assumption that these TFs were involved in the transcriptional regulation of these proteins, to establish a TF gene–target relationship, only new data are considered in this analysis, and our previous data are considered only for discussion.

Compartmentation of proteins in MM and bacteroid

Rhizobium etli CFN42 contains six plasmids and a chromosome (González et al. 2003). An analysis of gene location from MM and bacteroid showed that for MM proteins, most of the genes are codified in the chromosome, while for bacteroid proteins, higher participation was found for plasmids p42b, p42d, p42e, p42f than in MM (Figure 2). Of note, the symbiotic plasmid (p42d) had a 5.3% higher participation in bacteroids than in MM, in line with a wide transcription rate of the symbiotic plasmid (psym) genes of *R. etli* CFN42 under microaerobic conditions (as in symbiosis) or in aerobic conditions in the presence of genistein (Valderrama et al. 1996). Additionally, many of the genes expressed in MM (78.7%) and bacteroid (67.1%) were from the chromosome, while 21 and 32% of the expressed genes were from plasmids, respectively. The higher number of genes expressed from the chromosome agrees with the finding that exponential growth in MM and nitrogen fixation activity have in common a great demand for energy synthesis, and most of the metabolic pathways for this process are similar (see below). One of the exceptional differences is that in symbiosis, the high-affinity *cbb3* cytochrome oxidase terminal is expressed (Delgado et al. 1998; Lopez et al. 2001). *Rhizobium leguminosarum* bv. *viciae* UMP791 contain five plasmids and a chromosome, similar to *R. etli* CFN42, which contains six plasmids. A proteome analysis of *R. leguminosarum* bacteroid with its host plant *Pisum sativum* showed that most of the bacteroid proteins were from the chromosome (81.6%), showing a lower participation of the plasmids than with the *R. etli* CFN42 strain (Durán et al. 2020).

The plasmids contain essential genes for growth in MM, such as p42e (*minCDE*; Landeta et al. 2011) and plasmid p42f (*panCB*; Villaseñor et al. 2011). Moreover, a cured *R. etli* CFN42 of p42f complemented with the *panCB* genes did not restore wild-type growth, meaning that p42f has unidentified genes that are important for growth in MM (Villaseñor et al. 2011).

Metabolic pathways

A detailed view of the pathways that operate in exponential growth in MM (ammonium-succinate) and bacteroid, a non-growing state in symbiosis at 18 days post-inoculation, with a maximal peak of nitrogen fixation, showed 105 pathways according to the KEEG program with Gen Ontology (GO) gene off classification (see "Materials and Methods" section; Maere et al. 2005; Figure 3). In MM, 31 representative metabolic pathways with greater or equal to 30% of genes, and in bacteroid, 10 pathways with greater or equal to 20% of genes per pathway with respect to the genome content were found (Figure 3), which is related to the high demand for the synthesis of metabolites to sustain growth in a minimal medium compared to the non-growing bacteroid state. In contrast, in symbiosis, most of the energy for the synthesis of metabolites is dedicated to nitrogen fixation. In agreement with this, carbon metabolism, including synthesis of amino acids, sugars, purine and pyrimidine, sulfur metabolism, glycolysis-gluconeogenesis, pyruvate metabolism, TCA cycle, oxidative phosphorylation, nitrogen metabolism, fatty acid metabolism, nicotinate and nicotinamide, vitamin synthesis, DNA replication, aminoacyl-tRNA biosynthesis, ribosome synthesis, protein export, and flagellar assembly, had a higher percentage of genes in MM than in bacteroids (Resendis-Antonio et al. 2011), as was shown in a comparative proteomic study of a free-living aerobic condition and the symbiosis of the *Bradyrhizobium japonicum* USDA110 strain (Sarma and Emerich

2005, 2006). Some other pathways, such as histidine metabolism, glutathione metabolism, pentose phosphate pathway, beta-alanine, starch, and sucrose metabolism, were similar in MM and bacteroids; likely, histidine metabolism is necessary for the synthesis of inosine monophosphate, a precursor for the synthesis of purines and subsequently for the synthesis of allantoin and allantoic acids. These nitrogen compounds from nitrogen fixation are exported to the bean plant *Phaseolus vulgaris* by the *R. etli* CFN42 bacteroid (Alamillo et al. 2010; Collier and Tegeder 2012). Glutathione plays a crucial role against oxidative damage during the establishment of symbiosis (Hérouart et al. 2002); it is a precursor for cysteine synthesis, a sulfur donor for the synthesis of the Fe-S centers involved in defense against oxidative stress and in the prosthetic groups of sensory proteins. The pentose phosphate pathway is essential for the synthesis of phosphoribosyl pyrophosphate (PRPP), a precursor for purine synthesis during symbiosis (Newman et al. 1994; Miranda-Ríos et al. 1997). Beta-alanine is a precursor for the synthesis of pantothenate, which is essential for the ubiquitous compound coenzyme A (coA), subsequently used for many metabolic reactions, including phospholipid synthesis, fatty acid synthesis and degradation, and the tricarboxylic acid cycle. The *panCB* genes for the synthesis of pantothenate codified in p42f from the *R. etli* CFN42 strain were characterized (Villaseñor et al. 2011). Starch and sucrose synthesis was not detected in free-living or symbiotic conditions of *R. etli* CFN42. For the synthesis and degradation of ketone bodies, phosphonate and phosphinate metabolism were higher in bacteroids than in MM. *R. etli* CFN42 synthesizes poly-β-hydroxybutyrate granules during symbiosis with *P. vulgaris*; because this polymer is a reserve of carbon and reducing power, its accumulation is greater in symbiosis than in MM, where the energy is for supporting growth (Cevallos et al. 1996). Additionally, there is a high demand for phosphate in nitrogen-fixing nodules; it is an essential macronutrient necessary for the synthesis of proteins and nucleic acids (Liu et al. 2018), and phosphate is probably limited during symbiosis. As a response, the transcription of this pathway is raised, as was shown for bacteroids harvested from soybeans grown under field conditions (Delmotte et al. 2010). A detailed study with transcriptomic and proteomic technologies of the symbiosis compared with the aerobic growth showed 3,587 genes/proteins, expressing 43% of the predicted genome from *B. japonicum* (Delmotte et al. 2010), 807 proteins were identified in symbiosis; while in this study, 738 proteins were identified; i.e., in this study, there is a great proteomic coverage of the symbiosis *R. etli* CFN42-*Phaseolus vulgaris* bean plant considering that the *B japonicum* genome size is bigger than the *R. etli* genome. Although *R. etli* is a fast grower and *B. japonicum* is a slow grower in minimal medium, they elicit determinate nodules. In contrast to the symbiont *S. meliloti* with their host *Medicago sativa* alfalfa plant that induces indeterminate nodules, there are notable differences between the structure and composition of the symbiont in determinate and indeterminate nodules, reviewed in (Rascio and La Rocca 2013). Although *B. japonicum* and *R. etli* symbiosis occur in temperate and tropical weather, respectively, despite these differences, *R. etli* and *B. japonicum* symbiosis is more similar than *S. meliloti* symbiosis. A proteomic comparison of free-living and symbiosis from *B. japonicum* showed a greater number of proteases in free life than in symbiosis (Sarma and Emerich 2006). Similarly, in this study, 27 and two proteases were expressed. Most likely, the recycling of metabolites may be one of the factors that impacts the spending of energy in free life and symbiosis. It was suggested that bacteroids expend their energy judiciously between protein synthesis and nitrogen fixation by altering protein turnover (Sarma and Emerich 2006).

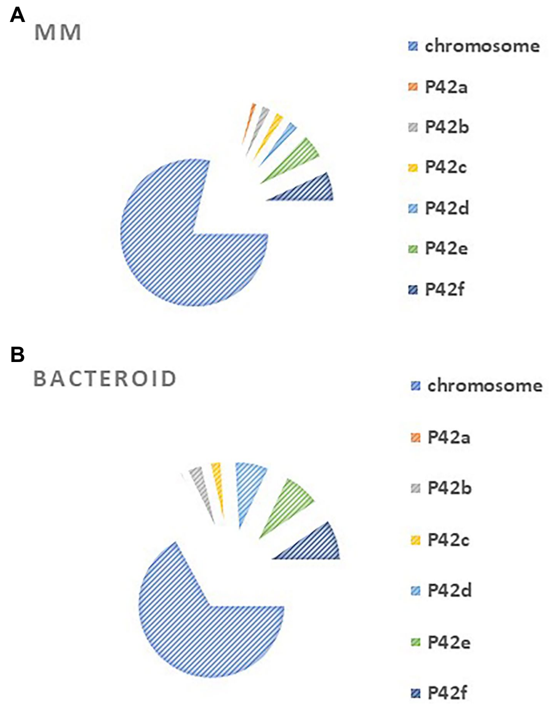


FIGURE 2
Compartmentation of **(A)** minimal medium protein profile expressed at six hours of growth in minimal medium and **(B)** 18 days post inoculation of bacteroid protein profile from *Rhizobium etli* CFN42.

Environmental metabolism

Moreover, some GO genes classified for the metabolism of environmental compounds were mapped; in MM, these genes covered approximately 39% of genes, while in bacteroids, they covered 14% with respect to the genome content (microbial metabolism in diverse environments, Figure 4). For some pathways, there is a low representation with respect to the total content of the *R. etli* CFN42 genome. The pathways for the degradation of benzoate, caprolactam, and naphthalene were more highly expressed in MM than in bacteroids. For chloroalkane and chloroalkene degradation and novobiocin biosynthesis, the

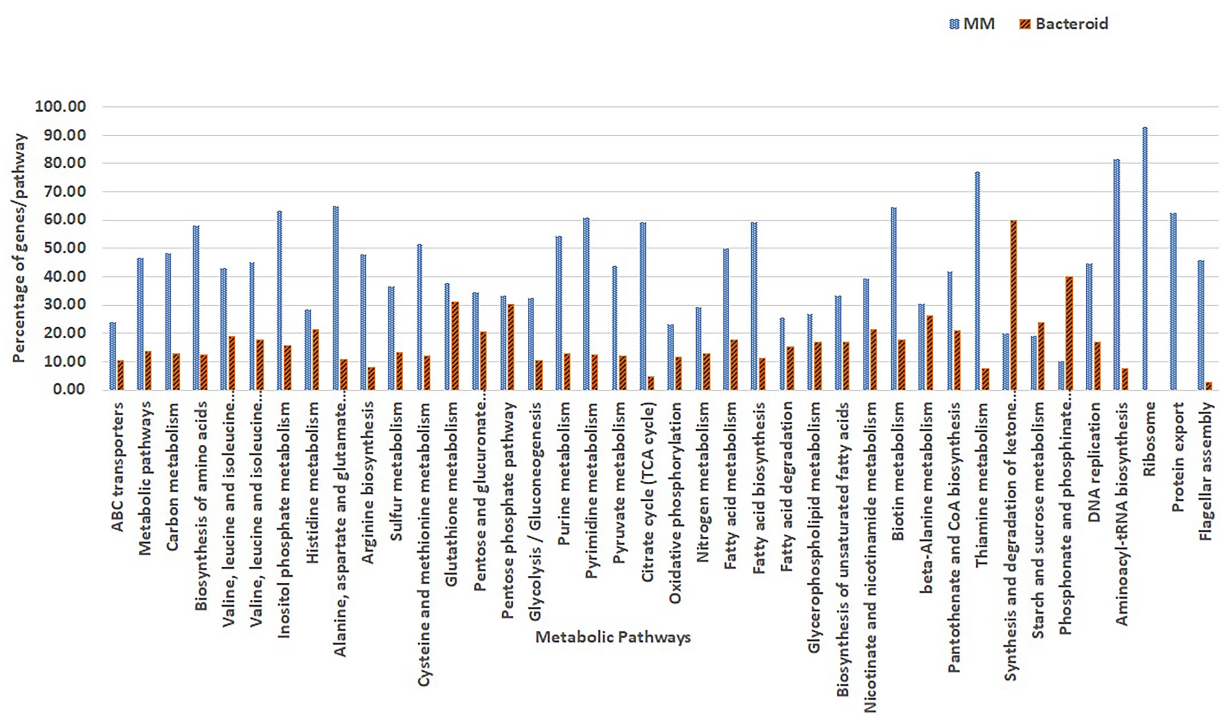


FIGURE 3
Comparison of GO classified proteins expressed per metabolic pathway in minimal medium and bacteroids from *Rhizobium etli* CFN42.

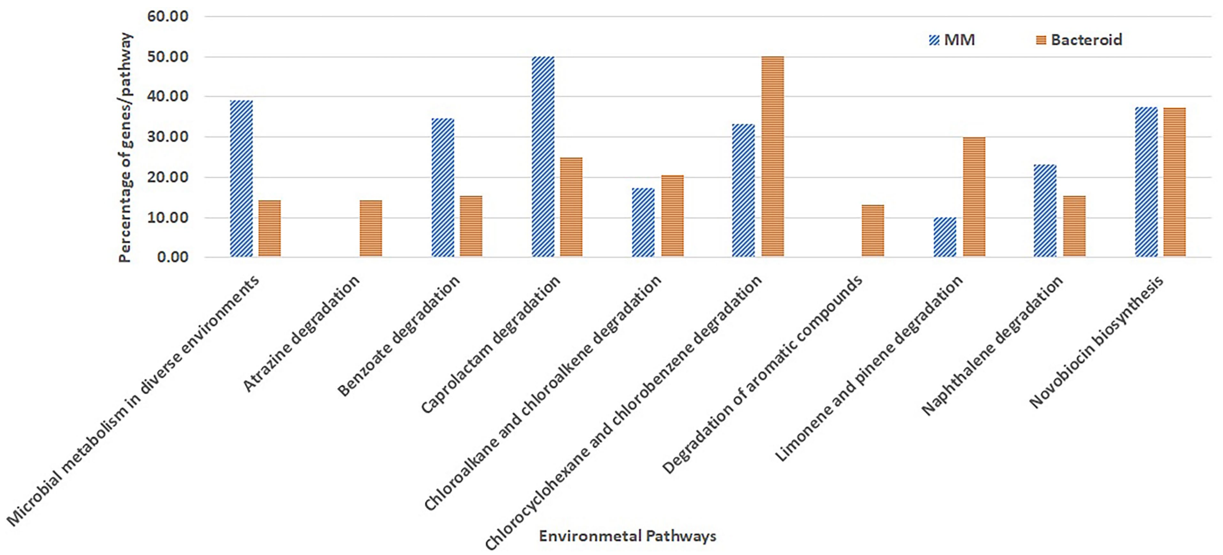


FIGURE 4
Metabolic pathways for biosynthesis and degradation of organic compounds environment related from *Rhizobium etli* CFN42.

number of proteins expressed was similar (Figure 4). Proteins for the degradation of chloroalkane and chloroalkene were also identified in a metagenomic analysis in the rhizosphere soil of a constructed wetland (Bai et al. 2014). Novobiocin is a very potent inhibitor of DNA gyrase, which works by targeting the GyrB subunit of the enzyme for energy transduction, and resistance to novobiocin of *Lotus rhizobia* was related to the effectiveness of the symbiosis with *Lotus pedunculatus* (Pankhurst 1977). For the degradation of atrazine, chlorocyclohexane and chlorobenzene, and aromatic compounds limonene and pinene, the number of genes was higher in bacteroid than in MM. Atrazine is an herbicide that may inhibit the growth of *Rhizobium* species, *P. vulgaris-Rhizobium* sp. Consortium symbiosis has been used for the bioremediation of soil contaminated with atrazine (Madariaga-Navarrete et al. 2017). Genes for the degradation of the aromatic compounds chlorocyclohexane and chlorobenzene were also reported in the genome of *Burkholderia phenoliruptrix* BR3459a, a symbiont of the *Mimosa flocculosa* leguminous plant (Zuleta et al. 2014). It was observed that for the *Rhizobium leguminosarum* E20-8 strain, limonene and pinene have antioxidant activity promoting growth under stress provoked by cadmium (Sá et al. 2020) and antibacterial activity (Ghaffari et al. 2019). In the *B. japonicum* bacteroid proteome, the NrgC protein and a gene for phenazine biosynthesis were identified for a response against microbial attack (Sarma and Emerich 2005, 2006). These genes in MM and bacteroid for degradation of metabolites of the environment are used for a fast response, competence, and better adaptation in soil conditions. Unlike *B. japonicum* (Sarma and Emerich 2006), *R. etli* bacteroid showed a wide strategy to withstand environmental stresses.

Isoenzymes in MM and bacteroid

The KEEG mapper for visualization of the metabolic maps was used (see "Materials and Methods" section). This mapper uses the "K" number to identify the function of the gene, and it is assigned based on the orthology of the genes (Kanehisa et al. 2016). For an integral view of the metabolism in MM, bacteroid, and proteins present in both conditions with "no change" (Nch), genes were mapped (Supplementary Tables 2 and 3A). Discussion of the central metabolism involved 37 representative pathways. Analysis of mapped genes showed that for some enzymatic reactions, different genes for the same enzymatic step in MM and bacteroid were found, e.g., for the pentose phosphate pathway there were two genes for the conversion of D-ribulose phosphate to D-ribose-5P by the 6-phosphogluconate dehydrogenase enzyme; one is expressed in MM RHE_RS12615, and a different one was expressed in the bacteroid RHE_RS17825 (Table 1), suggesting the presence of a condition-dependent isoform (Supplementary Table 2 pathway 15, and Supplementary Table 3A). From here on, we will call it "multiplicity." The Fructose and mannose metabolism pathway (Supplementary Table 2 pathway 10, and Supplementary Table 3A), for the catalysis of L-fucose to L-fucolactone by the enzyme D-threo-aldose 1-dehydrogenase; the proteins in MM RHE_RS02500 and bacteroid RHE_RS28605 were expressed (Table 1), showing multiplicity. For the galactose metabolism pathway (Supplementary Table 2 pathway 11, and Supplementary Table 3A), the conversion of UDP-glucose to UDP-galactose, UDP-glucose 4-epimerase was synthesized in MM RHE_RS03845 and in bacteroid RHE_RS17845 was expressed. As well as, for the enzyme *dgoD*, galactonate dehydratase [EC:4.2.1.6] for catalysis of D-galactonate to 2-dehydro-3-deoxy-D-galactonate in MM the RHE_RS18905 and in bacteroid RHE_RS24515 proteins were expressed (Table 1), showing multiplicity for two different enzymes of the same pathway. These data showed that the same enzymatic reactions are performed in MM and bacteroids with distinct proteins, suggesting that some alternative proteins are specific for free-living aerobic conditions and others for symbiosis for the same metabolic step. The pyruvate metabolism pathway (Supplementary Table 2 pathway 18, and Supplementary Table 3A), for the conversion of acetyl-CoA to acetoacetyl-CoA in MM, RHE_RS23190 was expressed, while in bacteroid, two different genes were expressed; RHE_RS02820 and RHE_RS20545 which showed differences in metabolism from MM and in bacteroid (Table 1), and since distinct TFs were identified in MM and bacteroid, a different transcriptional regulation for isoenzymes was analyzed (see below). Moreover, for the inositol phosphate pathway (Supplementary Table 2 pathway 2, and Supplementary Table 3A), for the myo-inositol-1(or 4)-monophosphatase enzyme in MM was identified the RHE_RS10865, RHE_RS17960, and RHE_RS22570 enzymes, and in bacteroid RHE_RS22680 and RHE_RS04240 were found (Table 1), again showing that the metabolism in symbiosis compared with MM has some differences. For valine, leucine, and isoleucine biosynthesis pathway (Supplementary Table 2 pathway 6, and Supplementary Table 3A), the enzyme *ilvD*, dihydroxy-acid dehydratase [EC:4.2.1.9], and the RHE_RS08720 and RHE_RS23070 proteins were expressed in MM and bacteroid, respectively, supporting multiplicity (Table 1). Similarly, for valine, leucine, and isoleucine degradation (Supplementary Table 2 pathway 7, and Supplementary Table 3A), the enzyme *acd* acyl-CoA dehydrogenase [EC:1.3.8.7] RHE_RS20670 was expressed in MM and in bacteroid, the isoenzyme RHE_RS04555 was identified (Table 1). Additionally, for the enzyme *atoB*, acetyl-CoA C-acetyltransferase [EC:2.3.1.9] in MM the RHE_RS23190 was present, and in bacteroids, the isoenzymes RHE_RS02820 and RHE_RS20545 were found (Table 1), showing a multigenic strategy for the degradation of branched-chain amino acids. For the synthesis of the poly-β-hydroxybutyrate polymer, the enzyme β-ketothiolase (acetyl-CoA C-acetyltransferase) converts two molecules of acetyl-CoA to acetoacetyl-CoA. In MM, the enzyme RHE_RS23190 was detected, and in bacteroid, two enzymes, RHE_RS02820 and RHE_RS20545, were identified (Table 1).

The ABC components of the sugar transporters were present in MM and bacteroid; i.e., maltose/maltodextrin, galactose, raffinose/stachyose/melibiose, lactose/L-arabinose, sorbitol/mannitol, trehalose/maltose, cellobiose, chitobiose, arabinooligosaccharide. In bacteroids, for monosaccharide transporters, glucose, ribose, galactofuranose, and myo-inositol 1-phosphate were identified, while D-xylose, fructose, rhamnose, myo-inositol, and glycerol were identified in MM (Supplementary Table 2 pathway 37).

The multiplicity of ABC transporters was for seven K numbers (Supplementary Table 3A); for *afuA*, *fbpA*; iron(III) transport system substrate-binding protein; in MM,

TABLE 1 Isoenzymes in MM and bacteroid from *Rhizobium etli* CFN42.

K number	Physiological condition	Locus tag	Annotation from BlastKoala*
K02035	MM	RHE_RS10550	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS20405	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS22160	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS23500	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS23525	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS24485	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS27640	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS27665	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	MM	RHE_RS03080	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	Bacteroid	RHE_RS10750	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	Bacteroid	RHE_RS22645	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	Bacteroid	RHE_RS28255	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02035	Bacteroid	RHE_RS01120	ABC.PE.S; peptide/nickel transport system substrate-binding protein
K02052	MM	RHE_RS17470	ABC.SP.A; putative spermidine/putrescine transport system ATP-binding protein
K02052	Bacteroid	RHE_RS14790	ABC.SP.A; putative spermidine/putrescine transport system ATP-binding protein
K02052	Bacteroid	RHE_RS14870	ABC.SP.A; putative spermidine/putrescine transport system ATP-binding protein
K00249	MM	RHE_RS20670	ACADM, acd; acyl-CoA dehydrogenase [EC:1.3.8.7]
K00249	Bacteroid	RHE_RS04555	ACADM, acd; acyl-CoA dehydrogenase [EC:1.3.8.7]
K00626	MM	RHE_RS23190	ACAT, atoB; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]
K00626	Bacteroid	RHE_RS20545	ACAT, atoB; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]
K00626	Bacteroid	RHE_RS02820	ACAT, atoB; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]
K01486	MM	RHE_RS17480	ade; adenine deaminase [EC:3.5.4.2]
K01486	Bacteroid	RHE_RS15825	ade; adenine deaminase [EC:3.5.4.2]
K02012	MM	RHE_RS10880	afuA, fbpA; iron(III) transport system substrate-binding protein
K02012	Bacteroid	RHE_RS13955	afuA, fbpA; iron(III) transport system substrate-binding protein
K00759	MM	RHE_RS15525	APRT, apt; adenine phosphoribosyltransferase [EC:2.4.2.7]
K00759	Bacteroid	RHE_RS31115	APRT, apt; adenine phosphoribosyltransferase [EC:2.4.2.7]
K05349	MM	RHE_RS28885	bglX; beta-glucosidase [EC:3.2.1.21]
K05349	Bacteroid	RHE_RS29645	bglX; beta-glucosidase [EC:3.2.1.21]
K01255	MM	RHE_RS01080	CARP, pepA; leucyl aminopeptidase [EC:3.4.11.1]
K01255	Bacteroid	RHE_RS07430	CARP, pepA; leucyl aminopeptidase [EC:3.4.11.1]
K00405	MM	RHE_RS29065	ccoO; cytochrome c oxidase cbb3-type subunit II
K00405	Bacteroid	RHE_RS30885	ccoO; cytochrome c oxidase cbb3-type subunit II
K03412	MM	RHE_RS03250	cheB; two-component system, chemotaxis family, protein-glutamate methylesterase/glutaminase [EC:3.1.1.61 3.5.1.44]
K03412	Bacteroid	RHE_RS26805	cheB; two-component system, chemotaxis family, protein-glutamate methylesterase/glutaminase [EC:3.1.1.61 3.5.1.44]
K03412	Bacteroid	RHE_RS17965	cheB; two-component system, chemotaxis family, protein-glutamate methylesterase/glutaminase [EC:3.1.1.61 3.5.1.44]
K00390	MM	RHE_RS05785	cysH; phosphoadenosine phosphosulfate reductase [EC:1.8.4.8 1.8.4.10]
K00390	Bacteroid	RHE_RS05785	cysH; phosphoadenosine phosphosulfate reductase [EC:1.8.4.8 1.8.4.10]
K00285	MM	RHE_RS28700	dadA; D-amino-acid dehydrogenase [EC:1.4.5.1]
K00285	Bacteroid	RHE_RS03755	dadA; D-amino-acid dehydrogenase [EC:1.4.5.1]
K01714	MM	RHE_RS26910	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]
K01714	MM	RHE_RS27660	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]
K01714	MM	RHE_RS03065	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]
K01714	Bacteroid	RHE_RS07055	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]
K01714	Bacteroid	RHE_RS19830	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]
K01714	Bacteroid	RHE_RS22155	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]

(Continued)

TABLE 1 (Continued)

K number	Physiological condition	Locus tag	Annotation from BlastKoala*
K01714	Bacteroid	RHE_RS14280	dapA; 4-hydroxy-tetrahydrodipicolinate synthase [EC:4.3.3.7]
K02031	MM	RHE_RS24230	ddpD; peptide/nickel transport system ATP-binding protein
K02031	MM	RHE_RS24500	ddpD; peptide/nickel transport system ATP-binding protein
K02031	MM	RHE_RS25825	ddpD; peptide/nickel transport system ATP-binding protein
K02031	MM	RHE_RS27625	ddpD; peptide/nickel transport system ATP-binding protein
K02031	MM	RHE_RS20420	ddpD; peptide/nickel transport system ATP-binding protein
K02031	Bacteroid	RHE_RS28270	ddpD; peptide/nickel transport system ATP-binding protein
K01684	MM	RHE_RS18905	dgoD; galactonate dehydratase [EC:4.2.1.6]
K01684	Bacteroid	RHE_RS24515	dgoD; galactonate dehydratase [EC:4.2.1.6]
K00064	MM	RHE_RS02500	E1.1.1.122; D-threo-aldose 1-dehydrogenase [EC:1.1.1.122]
K00064	Bacteroid	RHE_RS28605	E1.1.1.122; D-threo-aldose 1-dehydrogenase [EC:1.1.1.122]
K01092	MM	RHE_RS17960	E3.1.3.25, IMPA, suhB; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]
K01092	MM	RHE_RS22570	E3.1.3.25, IMPA, suhB; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]
K01092	MM	RHE_RS10865	E3.1.3.25, IMPA, suhB; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]
K01092	Bacteroid	RHE_RS04240	E3.1.3.25, IMPA, suhB; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]
K01092	Bacteroid	RHE_RS22680	E3.1.3.25, IMPA, suhB; myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]
K01560	MM	RHE_RS05045	E3.8.1.2; 2-haloacid dehalogenase [EC:3.8.1.2]
K01560	Bacteroid	RHE_RS28210	E3.8.1.2; 2-haloacid dehalogenase [EC:3.8.1.2]
K01768	MM	RHE_RS18990	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	MM	RHE_RS24270	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	MM	RHE_RS18920	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS11150	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS12750	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS13090	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS13735	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS14395	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS18920	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K01768	Bacteroid	RHE_RS24935	E4.6.1.1; adenylate cyclase [EC:4.6.1.1]
K09458	MM	RHE_RS12650	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]
K09458	MM	RHE_RS12655	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]
K09458	MM	RHE_RS07375	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]
K09458	Bacteroid	RHE_RS10850	fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]
K00059	MM	RHE_RS06685	fabG, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]
K00059	MM	RHE_RS07365	fabG, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]
K00059	MM	RHE_RS05335	fabG, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]
K00059	Bacteroid	RHE_RS25095	fabG, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]
K00059	Bacteroid	RHE_RS19755	fabG, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]
K00135	MM	RHE_RS00470	gabD; succinate-semialdehyde dehydrogenase / glutarate-semialdehyde dehydrogenase [EC:1.2.1.16 1.2.1.79 1.2.1.20]
K00135	Bacteroid	RHE_RS28200	gabD; succinate-semialdehyde dehydrogenase / glutarate-semialdehyde dehydrogenase [EC:1.2.1.16 1.2.1.79 1.2.1.20]
K00135	Bacteroid	RHE_RS29885	gabD; succinate-semialdehyde dehydrogenase / glutarate-semialdehyde dehydrogenase [EC:1.2.1.16 1.2.1.79 1.2.1.20]
K02433	MM	RHE_RS09475	gatA, QRSL1; aspartyl-tRNA(Asn)/glutamyl-tRNA(Gln) amidotransferase subunit A [EC:6.3.5.6 6.3.5.7]
K02433	Bacteroid	RHE_RS25710	gatA, QRSL1; aspartyl-tRNA(Asn)/glutamyl-tRNA(Gln) amidotransferase subunit A [EC:6.3.5.6 6.3.5.7]
K02433	Bacteroid	RHE_RS01105	gatA, QRSL1; aspartyl-tRNA(Asn)/glutamyl-tRNA(Gln) amidotransferase subunit A [EC:6.3.5.6 6.3.5.7]

(Continued)

TABLE 1 (Continued)

K number	Physiological condition	Locus tag	Annotation from BlastKoala*
K00605	MM	RHE_RS11460	gcvT, AMT; aminomethyltransferase [EC:2.1.2.10]
K00605	Bacteroid	RHE_RS26150	gcvT, AMT; aminomethyltransferase [EC:2.1.2.10]
K00605	Bacteroid	RHE_RS26195	gcvT, AMT; aminomethyltransferase [EC:2.1.2.10]
K16147	MM	RHE_RS27870	glgE; starch synthase (maltosyl-transferring) [EC:2.4.99.16]
K16147	Bacteroid	RHE_RS27870	glgE; starch synthase (maltosyl-transferring) [EC:2.4.99.16]
K00799	MM	RHE_RS05865	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	MM	RHE_RS06130	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	MM	RHE_RS06230	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	MM	RHE_RS11855	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	MM	RHE_RS01425	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	Bacteroid	RHE_RS07560	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	Bacteroid	RHE_RS12380	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	Bacteroid	RHE_RS25110	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K00799	Bacteroid	RHE_RS05070	GST, gst; glutathione S-transferase [EC:2.5.1.18]
K02495	MM	RHE_RS30905	hemN, hemZ; oxygen-independent coproporphyrinogen III oxidase [EC:1.3.98.3]
K02495	MM	RHE_RS29140	hemN, hemZ; oxygen-independent coproporphyrinogen III oxidase [EC:1.3.98.3]
K02495	Bacteroid	RHE_RS30730	hemN, hemZ; oxygen-independent coproporphyrinogen III oxidase [EC:1.3.98.3]
K00817	MM	RHE_RS19480	hisC; histidinol-phosphate aminotransferase [EC:2.6.1.9]
K00817	Bacteroid	RHE_RS30550	hisC; histidinol-phosphate aminotransferase [EC:2.6.1.9]
K00817	Bacteroid	RHE_RS06810	hisC; histidinol-phosphate aminotransferase [EC:2.6.1.9]
K00457	MM	RHE_RS23940	HPD, hppD; 4-hydroxyphenylpyruvate dioxygenase [EC:1.13.11.27]
K00457	Bacteroid	RHE_RS08930	HPD, hppD; 4-hydroxyphenylpyruvate dioxygenase [EC:1.13.11.27]
K01745	MM	RHE_RS24440	hutH, HAL; histidine ammonia-lyase [EC:4.3.1.3]
K01745	Bacteroid	RHE_RS01780	hutH, HAL; histidine ammonia-lyase [EC:4.3.1.3]
K10191	MM	RHE_RS22750	lacK; lactose/L-arabinose transport system ATP-binding protein
K10191	Bacteroid	RHE_RS19645	lacK; lactose/L-arabinose transport system ATP-binding protein
K10111	MM	RHE_RS14795	malK, mtlK, thuK; multiple sugar transport system ATP-binding protein [EC:7.5.2.-]
K10111	MM	RHE_RS27505	malK, mtlK, thuK; multiple sugar transport system ATP-binding protein [EC:7.5.2.-]
K10111	MM	RHE_RS10605	malK, mtlK, thuK; multiple sugar transport system ATP-binding protein [EC:7.5.2.-]
K10111	Bacteroid	RHE_RS25965	malK, mtlK, thuK; multiple sugar transport system ATP-binding protein [EC:7.5.2.-]
K03406	MM	RHE_RS02080	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS02690	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS03220	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS03580	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS03585	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS04470	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS04590	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS04920	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS05950	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS06430	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS17765	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS17980	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS17990	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS27980	mcp; methyl-accepting chemotaxis protein
K03406	MM	RHE_RS02065	mcp; methyl-accepting chemotaxis protein
K03406	Bacteroid	RHE_RS27360	mcp; methyl-accepting chemotaxis protein
K10112	MM	RHE_RS18950	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	MM	RHE_RS22575	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	MM	RHE_RS23370	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein

(Continued)

TABLE 1 (Continued)

K number	Physiological condition	Locus tag	Annotation from BlastKoala*
K10112	MM	RHE_RS26890	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	MM	RHE_RS28085	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	MM	RHE_RS29410	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	MM	RHE_RS12565	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	Bacteroid	RHE_RS24950	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	Bacteroid	RHE_RS28400	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K10112	Bacteroid	RHE_RS24520	msmX, msmK, malK, sugC, ggtA, msiK; multiple sugar transport system ATP-binding protein
K01916	MM	RHE_RS06125	nadE; NAD+ synthase [EC:6.3.1.5]
K01916	Bacteroid	RHE_RS06125	nadE; NAD+ synthase [EC:6.3.1.5]
K00459	MM	RHE_RS29235	ncd2, npd; nitronate monooxygenase [EC:1.13.12.16]
K00459	Bacteroid	RHE_RS02555	ncd2, npd; nitronate monooxygenase [EC:1.13.12.16]
K23537	MM	RHE_RS10660	nupA; general nucleoside transport system ATP-binding protein
K23537	Bacteroid	RHE_RS00955	nupA; general nucleoside transport system ATP-binding protein
K10018	MM	RHE_RS24420	occT, nocT; octopine/nopaline transport system substrate-binding protein
K10018	Bacteroid	RHE_RS30295	occT, nocT; octopine/nopaline transport system substrate-binding protein
K00033	MM	RHE_RS12615	PGD, gnd, gntZ; 6-phosphogluconate dehydrogenase [EC:1.1.1.44 1.1.1.343]
K00033	Bacteroid	RHE_RS17825	PGD, gnd, gntZ; 6-phosphogluconate dehydrogenase [EC:1.1.1.44 1.1.1.343]
K22468	MM	RHE_RS02565	ppk2; polyphosphate kinase [EC:2.7.4.1]
K22468	Bacteroid	RHE_RS23870	ppk2; polyphosphate kinase [EC:2.7.4.1]
K00286	MM	RHE_RS15425	proC; pyrroline-5-carboxylate reductase [EC:1.5.1.2]
K00286	Bacteroid	RHE_RS28670	proC; pyrroline-5-carboxylate reductase [EC:1.5.1.2]
K10439	MM	RHE_RS22400	rbsB; ribose transport system substrate-binding protein
K10439	MM	RHE_RS27555	rbsB; ribose transport system substrate-binding protein
K10439	MM	RHE_RS30010	rbsB; ribose transport system substrate-binding protein
K10439	MM	RHE_RS30060	rbsB; ribose transport system substrate-binding protein
K10439	MM	RHE_RS09135	rbsB; ribose transport system substrate-binding protein
K10439	Bacteroid	RHE_RS29865	rbsB; ribose transport system substrate-binding protein
K02968	MM	RHE_RS01805	RP-S20, rpsT; small subunit ribosomal protein S20
K02968	Bacteroid	RHE_RS01805	RP-S20, rpsT; small subunit ribosomal protein S20
K01609	MM	RHE_RS11125	trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]
K01609	Bacteroid	RHE_RS11125	trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]

*Annotation of genes in the program BlastKoala (Kanehisa et al. 2016), based on the orthology assigns a *K* number.

RHE_RS10880 was identified, and RHE_RS13955 in bacteroid (Table 1). For *occT*, *nocT*, octopine/nopaline transport system substrate-binding protein; in MM, RHE_RS24420 was identified and RHE_RS30295 was expressed in bacteroid (Table 1). The *malK*, *mtlK*, *thuK*; multiple sugar transport system ATP-binding protein [EC:3.6.3.-]; in MM, the proteins RHE_RS10605, RHE_RS14795, RHE_RS27505 were identified, and RHE_RS25965 was expressed in bacteroid (Table 1). The *msmX*, *msmK*, *malK*, *sugC*, *ggtA*, *msiK*; multiple sugar transport system ATP-binding protein, in MM represented by RHE_RS12565, RHE_RS18950, RHE_RS22575, RHE_RS23370, RHE_RS26890, RHE_RS28085, RHE_RS29410 were found, while the RHE_RS24520, RHE_RS24950 and RHE_RS28400 were identified in bacteroid (Table 1). For the *lacK*; lactose/L-arabinose transport system ATP-binding protein, sn-glycerol-3-phosphate ABC transporter, the ATP-binding protein UgpC in MM RHE_RS22750 and in bacteroid RHE_RS19645 were identified (Table 1). The *rbsB*; ribose transport system substrate-binding protein is represented by the isoenzymes RHE_RS09135, RHE_RS22400, RHE_RS27555, RHE_RS30010, RHE_RS30060 in MM, and RHE_RS29865 was expressed in bacteroid (Table 1). For the *nupA*, general nucleoside transport system ATP-binding protein in MM RHE_RS10660 and in bacteroid RHE_RS00955 were identified (Table 1; Supplementary Table 2, pathway 37 and Supplementary Table 3A). Once multiplicity was detected in MM and bacteroid, a wide search for multiplicity in data was performed. Interestingly, from the 101 proteins representing 48 unique K numbers (a K number may have more than one protein), 34 isoenzymes were identified that cover 60 metabolic pathways (Table 1; Supplementary Table 3A). In synthesis, multiplicity in only one enzyme was equally found for other metabolic processes such as for peptidases, inhibitors, amino acids and related enzymes, messenger ARN biogenesis, ribosome, ribosome biogenesis, transfer ARN biogenesis, translation factors, chaperones and folding catalysis, DNA replication proteins, DNA repair, and recombination proteins. While other pathways had multiplicity in two different enzymes, e.g., lipid biosynthesis proteins, mitochondrial biogenesis, two-component system, and bacterial motility proteins. Furthermore, multiplicity for three enzymes in a pathway was also detected, e.g., glutathione metabolism (Supplementary Table 2 pathway 3, and Supplementary Table 3A), for *pepA*, leucyl aminopeptidase [EC:3.4.11.1] enzyme, the RHE_RS01080 was expressed in MM, while in bacteroid the RHE_RS07430 was identified (Table 1). For the *gst*, glutathione S-transferase [EC:2.5.1.18] in MM RHE_RS01425, RHE_RS05865, RHE_RS06130, RHE_RS06230, and RHE_RS11855 were identified and in bacteroids, RHE_RS05070, RHE_RS07560, RHE_RS25110, and RHE_RS12380 were identified (Table 1). As well as, the *gntZ*, 6-phosphogluconate dehydrogenase [EC:1.1.1.44 1.1.1.343] in MM RHE_RS12615 and in the bacteroid RHE_RS17825 were expressed (Table 1). Multiplicity was also found in 5 transcription regulators and 16 transporters (Supplementary Table 3A).

From this data, there are some relevant points; it has been shown that during symbiosis of *R. leguminosarum*, bacteroids become auxotrophic for branched-chain amino acids, and their supply depends on the leguminous pea plant (Prell et al. 2009). In contrast, in *R. etli* CFN42, for valine, leucine, and isoleucine biosynthesis, 9, 3, and 3 enzymes were detected in MM, bacteroid, and Nch, respectively (Supplementary Table 2 pathway 6, and Supplementary Table 3A), suggesting a functional pathway in *R. etli* CFN42. Multiplicity was also found for the β-hydroxybutyrate dehydrogenase enzyme in *B. japonicum* USDA110, two isoforms were exclusively expressed in free-living conditions and a new isoform was expressed in nodule proteomes (Sarma and Emerich 2006). Another difference between the symbiosis of *R. etli* CFN42 is the expression of a great number of ABC sugars transporters which does not seem to be expressed in the symbiosis of *B. japonicum* and *S. meliloti*, reviewed in (Sarma and Emerich 2006). Also, this data confirmed two different systems for defense against oxidative stress for *R. etli* CFN42 (Resendis-Antonio et al. 2011), which is also observed in *S. meliloti* 1021 (see below), one prevailing in free-living conditions and the other in symbiosis. As shown, the multiplicity of genes for an enzyme is a generality in the cellular functioning of *R. etli* CFN42 in free-living conditions and symbiosis, clearly showing a greater genetic redundancy for enzymes expressed in MM than in symbiosis that may or may not be paralogous genes (Supplementary Table 3A). Additionally, a contrasting analysis of function assigned to the genes between the KO Orthology database (Kanehisa et al. 2016; Kanehisa 2017) and the NCBI database[6] was performed from the 48 unique K numbers covering 101 and 74 proteins for MM and bacteroid, respectively (Table 1); only four K numbers from *R. etli* CFN42; K01684, K02433, K00459, and K10439 were different, showing a great coincidence between the two methods (see shaded green rows; Supplementary Table 3A). When genes with the same annotated function exist, phenotypic change of a bacterium is not present by loss of function of a gene copy; it is called "Robustness," which is the ability to maintain the function when there is a change, as it was from free life to symbiosis (González et al. 2006; Diss et al. 2014), and they are maintained by context-dependent differences (Putty et al. 2013). These data suggest that when *R. etli* CFN42 is in free life and under symbiotic conditions, there is a metabolic adaptation, implying distinct transcriptional regulation for these genes.

Isoenzymes in *Sinorhizobium meliloti* 1021

An identical analysis was performed with a peptone yeast-rich medium and bacteroid transcriptome data from *S. meliloti* 1021 to search for isoenzymes. Significant data were selected with two parameters, log ≥0.96 and with software with $p \geq 0.05$ (Barnett et al. 2004). In contrast to *R. etli* CFN42, *S. meliloti* only showed 7K genes for isoenzymes; SMc03978 *tkt2* for transketolase was expressed in TY, while in bacteroids, SMc00270 was expressed (Supplementary Table 3B). The protein SMc03994 for the 30S ribosomal protein S21 was present in TY medium, while SMc04320 for the 30S ribosomal protein was present in bacteroids. The SMa0744 protein GroEL was translated in TY and was substituted by the SMa0124 GroEL protein in bacteroids. Moreover, SMc02897 for the cytochrome C transmembrane protein was expressed in TY medium, and the equivalent activity was substituted by the SMc01981 cytochrome C protein in the bacteroid. Moreover, as shown for *R. etli* CFN42 for defense against oxidative stress in MM and bacteroids, *S. meliloti* 1,021 in TY medium expressed five glutathione-S transferases, SMc00097 (gst2), SMc00383 (gst3), SMc00407 (gst4), SMc03082 (gst8), and SMc00036 (gst1). This activity was performed by the SMc01443 (gst6) glutathione-S transferase protein in bacteroids (Supplementary Table 3B). These data suggest that an alternative system for defense against oxidative stress also exists in *S. meliloti* 1021 bacteroids. There were contrasting low K numbers in *S. meliloti* 1021 compared with the *R. etli* CFN42 genome, and these data probably have a bias from a different method for the selection of significant data between these bacteria.

Transcriptional regulatory network

Taking advantage of the RhizoBindingSites database[7] (Taboada-Castro et al. 2020), networks were constructed for MM and bacteroid protein profiles with the application "Prediction of regulatory network" (see "Materials and Methods" section). A

[6] https://www.ncbi.nlm.nih.gov/genome/browse/#!/proteins/827/383937%7CRhizobium%20etli/ (accessed September 8, 2022).

[7] http://rhizobindingsites.ccg.unam.mx/ (accessed September 8, 2022).

three-step method to build a network was implemented (see "Materials and Methods" section). In the second step (see methods), Clustered-TF genes obtained with the matrix-clustering analysis, were used as input in the application of RhizoBindingSites database "Prediction of regulatory networks" with the option "auto", to corroborated potential TF gen-target relationships. The cluster_34 and cluster_195 from MM, and cluster_97 and cluster_112 from bacteroid were chosen. For cluster_34, all the genes had a TF gene-target relationship. Indeed, for cluster_195, 22 out of 27 genes were connected (Supplementary Table 1G). For cluster_97, 21 of 26 genes were connected and for cluster_112, 21 from 22 genes were connected (Supplementary Table 1H). These data showed that the matrix of a clustered-TF, has homology to a matrix of the target gene. In consequence, the matrices from both, the TF´s and the gene-target are conserved in their respective orthologs genes, because upstream regulatory regions of the orthologs genes were used to deduce the matrices (Taboada-Castro et al. 2020). Suggesting, this conservation is by a compromised function of the motifs for the TF and the target genes and not by chance. Then, the transcriptional regulatory networks of MM and bacteroid protein profiles are constructed with motifs interspecies conserved. The quality of the MM, bacteroid, clustered-TF-MM and clustered-TF-BACTEROID networks, which are data of the three-step method, was compared by analyzing the number of interactions per p-value range. The number of interactions of p-values with low stringency decreased, and those with a higher stringency in the network from clustered-TF-MM and clustered-TF-BACTEROID increased, meaning that there was an enrichment of interactions with high stringency p-value levels (see "Materials and Methods" section, Figure 5), emphasizing that most of the TF gene–target interactions eliminated from clustered-TF-MM and clustered-TF-BACTEROID had low stringency p-values. These data confirmed that clustered matrices of genes are strongly related to the structure of a network, and these genes probably represent hubs. We expect this new method will be helpful for the depuration of regulons from any potential TF gene-target data, since it provides data with the highest level of restriction as possible, based on coexpression of the TF´s, instead of arbitrarily imposing a threshold to determine the significance of data. The number of clusters per network was 654 and 92 for Clustered-TF-MM and Clustered-TF-BACTEROID, respectively. Moreover, 654 proteins, including 93 TFs for Clustered-TF-MM, and 246 TFs for Clustered-TF-BACTEROID, including 46 TF proteins, were identified (Supplementary Tables 4A-B). These expected regulatory networks had 5,091 and 1,114 TF gene–target relationships for MM and bacteroid, respectively, the hypothetical regulons are available (Supplementary Tables 4 A-B). Additionally, to determine whether the matrices of these networks detect motifs in the upstream regulatory region of their corresponding orthologous genes in the order Rhizobiales, an analysis with a footprint-scan method was conducted (Nguyen et al. 2018). These data showed a great number of motifs detected with these matrices even for phylogenetically distant species of *R. etli* CFN42 (data not shown), suggesting that this conservation of motifs occurs by a functional compromise.

We wondered how our inferred networks assess against known curated networks. As no curated network is available for *R. etli*, inspired by recent work showing that assessing using network structural properties provides results consistent with using a gold-standard (Zorro-Aranda et al. 2022), we performed a pairwise comparison *via* correlation of the normalized structural profiles of two well-curated regulatory networks, *E. coli* and *B. subtilis*, as positive control and a background of Erdös-Rényi parametrized random networks as a negative control (Figure 6A; "Materials and Methods" section).

Comparing these properties showed that negative control networks were clearly segregated from the experimental and

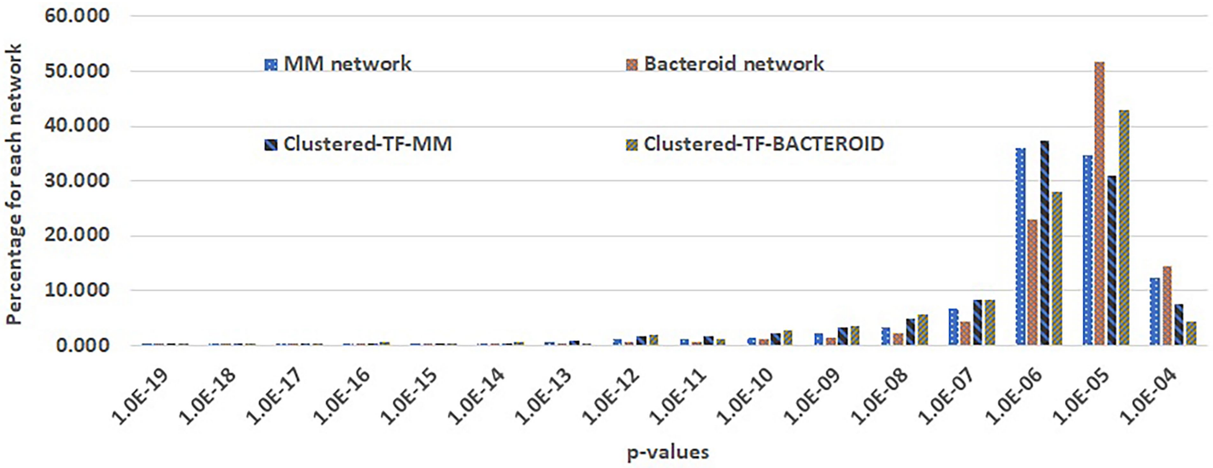


FIGURE 5
Comparison of genes per p-value range of the first and second transcriptional regulatory networks of minimal medium and bacteroid from *Rhizobium etli* CFN42.

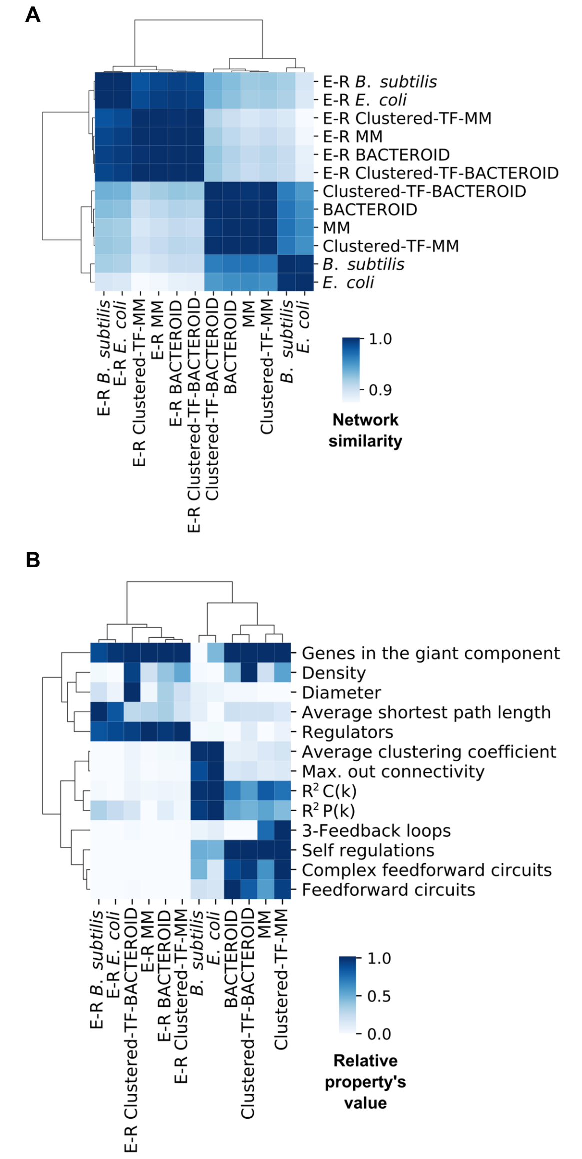


FIGURE 6
(A) Network similarity and **(B)** Symmetric properties of experimental networks from *Escherichia coli*, *Bacillus subtilis* and predicted networks from *Rhizobium etli* CFN42; MM, Clustered-TF-MM, BACTEROID and Clustered-TF-BACTEROID compared with 1000 Erdös-Rényi random networks.

inferred biological networks, showing that experimental and inferred biological networks were more similar (Figure 6A). Consequently, the inferred biological networks were not random. We then analyzed these structural property profiles of the networks using mix-max scaling across networks to maximize the differences (Figure 6B). We confirmed the segregation of the negative controls from the biological and experimental networks, which means that our networks are not random and that the experimental and inferred networks were more similar. The density was higher for the inferred networks than for the experimental networks. Between the inferred networks, the density of bacteroid and clustered-TF-BACTEROID was higher than that of the MM and clustered-TF-MM networks. The density of Clustered-TF-MM and Clustered-TF-BACTEROID could be increased due to grouping the matrices with the aforementioned matrix-clustering strategy (Figure 6B; "Materials and Methods" section).

The scale-free properties of the inferred networks were contrasted to the experimental networks of *E. coli* and *B. subtilis* by two alternative methods: robust linear regression and maximum likelihood estimation. Currently, the transcriptional regulatory network of a *Rhizobium* strain is unknown. A bioinformatic study based on functional relationships from the PROLINKS and STRING databases showed a scale-free interaction network and modularity for *Sinorhizobium meliloti* (Rodriguez-Llorente et al. 2009). However, they considered greater, more significant proteins than this study. Consistently, many genes showed a modular organization in a metabolic network of *R. etli* CFN42 with proteomic, transcriptomic, and metabolomic data (Resendis-Antonio et al. 2012). Additionally, there are more 3-feedback loops in the MM and Clustered-TF-MM networks than in the bacteroid, Clustered-TF-BACTEROID, *E. coli*, and *B. subtilis* networks. The self-regulation, complex feed-forward circuits, and feed-forward circuits from inferred networks were higher than the experimental ones (Figure 6B). Self-regulation is higher for inferred networks than experimental networks because the RhizoBindingSites database was built only with genes whose matrices could recognize a motif in their upstream promoter region.

The average clustering coefficient, maximum out connectivity, cluster coefficient R^2 C(k), and connectivity distribution R^2 P(k) were higher for the experimental than for inferred biological networks, implying that the inferred networks have an atypical very low modularity. As previously shown in several organisms (Freyre-González et al. 2008; 2012; Freyre-González and Tauch 2017; Escorcia-Rodríguez et al. 2021), the Natural Decomposition Approach (NDA) reveals that bacterial regulatory networks shape a diamond-like, three-tier, hierarchy where global TFs govern modules, and the local response of these modules is integrated at the promoter level by intermodular genes, whereas modules are shaped by local TFs and structural genes (Freyre-González et al. 2022). An analysis of our predicted networks using the NDA showed a hierarchy only composed of global TF and basal machinery, where neither modules nor intermodular genes could be identified (data not shown). These could be a consequence of the atypical high density of the inferred network, as this causes the networks to be more interconnected than usual.

As we found that in our networks the integrative layer composed of the intermodular genes is absent, we leverage that it has been previously shown that regulatory networks are mainly descendent (Ma et al. 2004) but there are still some feedback circuits (Freyre-González et al. 2008, 2012). We unveil a hierarchy of the inferred networks by removing the top-down edges, thus eliminating feedback, and applying a topological sorting algorithm to the predicted network (Figure 7; Supplementary Table 5 and "Materials and Methods" section). Our strategy maintains the global structure of the network to reveal the hierarchy. Besides structural nodes, no other nodes are removed, and 'bottom-up' edges can be added back to the hierarchy to reveal the feedback among layers and reconstruct the original network.

The hierarchy of the MM network showed that RHE_RS06405 (MucR family) is at the top. Under the top, there are four genes: RHE_RS25725 (LysR family), RHE_RS16230 (ROK family), RHE_RS05945 (LuxR family), and RHE_RS02355 (ROK family; Figure 7A; Supplementary Table 5). In contrast, for the Clustered-TF-MM network (Figure 7B; Supplementary Table 5), RHE_RS06405 (MucR family) and RHE_RS05945 (LuxR family) were at the top, and under the top seven TFs identified, RHE_RS25725 (LysR family) and RHE_RS20575 (*carD* family), a CarD protein, pertaining to the CarD_CdnL_TRCF family of TFs described in *Mycobacterium tuberculosis* 2018, binds to RNA polymerase and activates transcription by stabilizing the transcription initiation complex, elongation or termination steps, and deletion of N-terminal residues hampers amyloid formation (Kaur et al. 2018). It was shown that the interaction of CarD with the RNAP beta-subunit is responsible for mediating *M. tuberculosis* viability, rifampicin resistance, and pathogenesis. It is a highly expressed protein, also induced by multiple stresses. Transient depletion of CarD makes *M. tuberculosis* more sensitive to being killed by reactive oxygen species, and its mutation abolishes persistence in mice (Weiss et al. 2012). In addition, RHE_RS16230 (ROK family), RHE_RS02355 (ROK family), RHE_RS17050 (response regulator), RHE_RS01875 (helix-turn-helix transcriptional regulator), and RHE_RS00415 (TetR family) were identified. For bacteroid and clustered TF-BACTEROID, the hierarchy of transcriptional regulatory networks showed the same three TFs at the top; RHE_RS23775 (NAC, nitrogen assimilation transcriptional regulator). In *Escherichia coli,* the *nac* and *glnK* promoters were strongly activated when cells stopped growing, and ammonium became scarce (Atkinson et al., 2002), as well as RHE_RS03515 (substrate-binding domain) and RHE_RS10580 (LacI family DNA-binding transcriptional regulator; Figures 7C,D; Supplementary Table 5). For MM and clustered-TF-MM transcriptional regulatory networks, seven and six different levels of regulation are shown, respectively (Figures 7A,B; Supplementary Table 5). In contrast, for bacteroid and clustered-TF_BACTEROID, only three and four levels of regulation were shown, respectively (Figures 7C,D; Supplementary Table 5).

Inferred transcriptional regulation of isoenzymes in MM and bacteroids

TF gene–target relationships for genes coding for isoenzymes in the MM, Clustered-TF-MM and bacteroid, Clustered-TF-BACTEROID networks were inferred (Supplementary Table 6). The transcriptional regulator per locus tag is found in the column E headed "Matrix_ID" with the RHE_RS13345_m5 format (Supplementary Table 6), see "Materials and Methods" section. It is shown by the enzyme PGD, *gnd*, *gntZ*; 6-phosphogluconate dehydrogenase [EC:1.1.1.44 1.1.1.343] (Supplementary Table 6, column AK), the isoenzymes RHE_RS12615 and RHE_RS17825 were expressed in MM and in the bacteroid (Supplementary Table 6, column B), respectively. In MM, the transcriptional regulator is RHE_RS13345_m5 (Supplementary Table 6, column E) with a *p*-value of 1.7e-5 (Supplementary Table 6, column K), and in the bacteroid, it is RHE_RS27925_m4 (Supplementary Table 6, column E) with a *p*-value of 0.18e-4 (Supplementary Table 6, column K). However, there are no data with respect to the Clustered-TF-MM and Clustered-TF-BACTEROID networks (Supplementary Table 6, columns S–Z) due to a reduction of TF's by the matrix-clustering analysis. Therefore, *fabG*, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100] (Supplementary Table 6, columns A–K) enzyme for fatty acid biosynthesis, we identified RHE_RS05335 and RHE_RS06685 in MM, and RHE_RS19755 in bacteroid (Supplementary Table 6, column B), which are potentially regulated by the TFs RHE_17755_m2, RHE_RS30790_m1 and RHE_RS23180_m2 (Supplementary Table 6, columns E and S), with *p*-value of 7.30E-07, 1.20E-06 and 2.60E-06 (Supplementary Table 6, columns K and Y), respectively, for MM, bacteroid and Clustered-TF-MM and Clustered-TF-BACTEROID networks, showing for this enzymatic step that the multiplicity also corresponds with a different TF involved in transcriptional regulation. Note that each network contains its own *p*-value (see Supplementary Table 6, columns K and Y). For a better choice of a TF gene-target, data from a clustered-TF network and a low *p*-value as possible is desirable. From here on, in this discussion, the *p*-value located in Supplementary Table 6, column K, for not clustered networks and Supplementary Table 6, column Y, for clustered networks will be omitted. Concerning the enzyme D-threo-aldose 1-dehydrogenase [EC:1.1.1.122], in MM and bacteroid, the isoenzymes RHE_RS02500 and RHE_RS28605 were expressed, and the inferred TFs were RHE_RS22090_m3 and RHE_RS03515_m5, respectively, showing a potentially distinct TF-dependent physiological condition, but incomplete data were obtained for Clustered-TF networks (Supplementary Table 6, columns S–Z). In the case of *gcvT* and AMT, aminomethyltransferase enzyme [EC:2.1.2.10] in MM expressed RHE_RS28340 and two isoenzymes in bacteroid; RHE_RS26195 and RHE_RS26150 were expressed, and their corresponding TFs RHE_RS28340_m3, RHE_RS00285_m4, and RHE_RS05730_m3 were deduced, respectively, for both non and clustered-TF networks, supporting the suggestion of distinct regulation of these genes in MM and bacteroid (Supplementary Table 6).

Most likely, the microaerobic conditions and the metabolic functions prevailing in the bacteroid (fixing nitrogen), in comparison with the bacteria cultivated in MM (free life), induce specific strategies against oxidative stress, e.g., for the case of the enzyme GST, *gst*; glutathione S-transferase

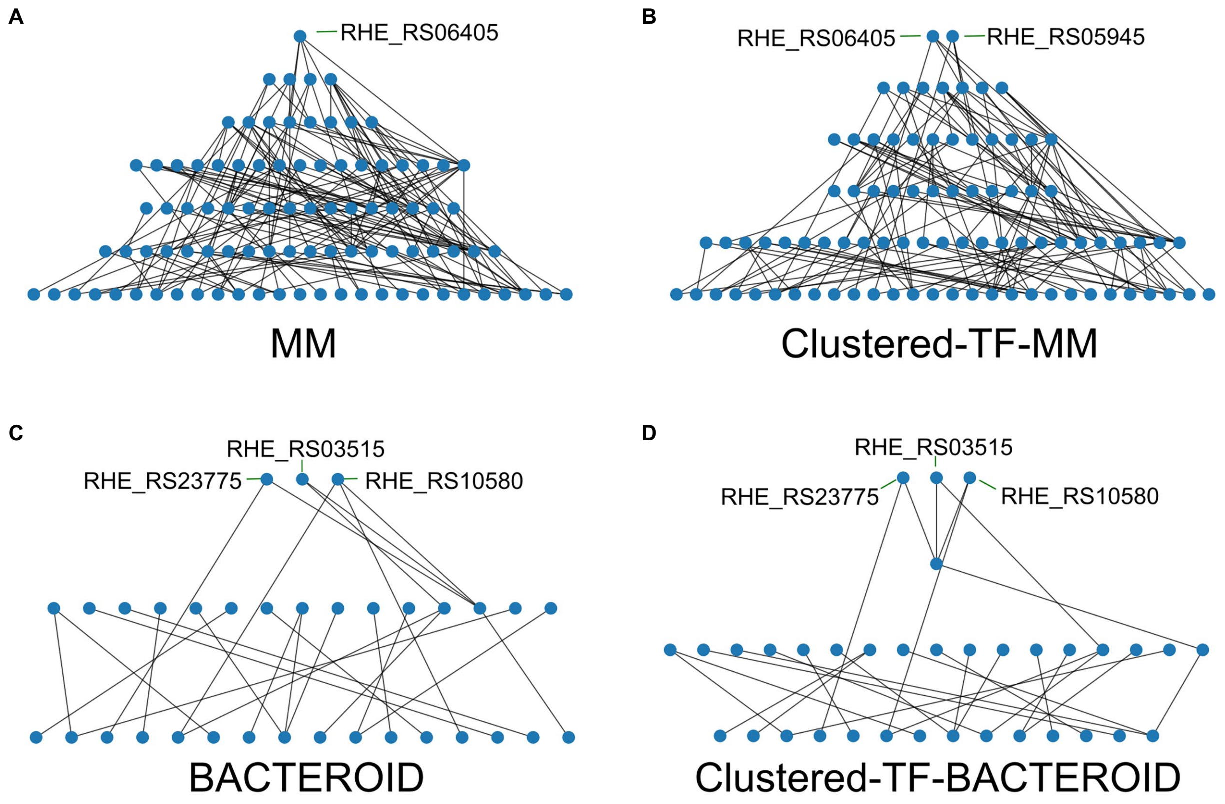


FIGURE 7
Hierarchy of the transcriptional regulators (TF's) of the networks; **(A)** Minimal medium, **(B)** Clustered-TF-MM, **(C)** BACTEROID, and **(D)** Clustered-TF-BACTEROID from *Rhizobium etli* CFN42.

[EC:2.5.1.18], in MM RHE_RS0630 and RHE_RS11855 proteins were expressed for the MM network, while in bacteroid, the RHE_RS12380 was identified with a Clustered-TF network, with the TFs RHE_RS06135_m4, RHE_RS27645_m3 in MM and RHE_RS08350_m3 in the bacteroid (Supplementary Table 6). Regarding *yghU* and *yfcG*, GSH-dependent disulfide-bond oxidoreductase [EC:1.8.4.-] in MM and bacteroid isoenzymes RHE_RS22490 and RHE_RS04155, respectively, were expressed, potentially under the transcriptional control of TFs RHE_RS12670_m4 and RHE_RS12205_m4, respectively (Supplementary Table 6). For these proteins involved in the repair of oxidized proteins, a different transcriptional regulation is suggested in MM and bacteroid and clustered-TF-MM and clustered-TF-BACTEROID networks. Iron transport is relevant for metabolism regarding *afuA* and *fbpA,* which encode the iron(III) transport system substrate-binding protein and express the isoenzymes RHE_RS10880 and RHE_RS13955 in MM and bacteroid, respectively, with the TFs RHE_RS28340_m4 and RHE_RS16205_m5, respectively, for MM and bacteroid networks (Supplementary Table 6), our data suggest two distinct metabolic strategies for transport of iron in MM (free life) and bacteroid (nitrogen fixing) conditions. It has been discussed that transport is specific for these metabolic stages (Sarma and Emerich 2006); indeed, this was supported for amino acid transport regarding ABC.PA. S; the polar amino acid transport system substrate-binding protein, in MM RHE_RS02695, RHE_RS11720, and RHE_RS27400, and in bacteroid RHE_RS07475 and RHE_RS27430 were expressed, potentially regulated by the TFs RHE_RS30745_m3, RHE_RS24110_m2, RHE_RS14135_m3 and RHE_RS18525_m2, RHE_RS26505_m5, respectively. All these data were clustering TF-associated, showing distinct TFs for each metabolic condition (Supplementary Table 6). Concerning transcriptional regulators, *lacI* and *galR* belonging to the LacI family in Clustered-TF-MM RHE_RS03090, RHE_RS12585, RHE_RS17450, RHE_RS23055, RHE_RS23350, and RHE_RS27560 were expressed in comparison with Clustered-TF-BACTEROID, where the following proteins were identified: RHE_RS03515, RHE_RS15245, and RHE_RS27525. Probably some genes are expressed because they respond to different physiological conditions with the aim of regulating different groups of genes. The inferred TFs for these genes were RHE_RS03090_m2, RHE_RS12585_m4, RHE_RS17450_m4, RHE_RS23055_m3, RHE_RS23350_m1, RHE_RS27560_m3, and RHE_RS03515_m5 for cluster-TF-MM, as well as, RHE_RS24095_m3 for bacteroid network, and RHE_RS03515_m5, RHE_RS27525_m2 for clustered-TF-BACTEROID, respectively (Supplementary Table 6). These data support the idea that isoenzymes have distinct regulations. For the ABCB-BAC ATP-binding cassette, subfamily B, bacterial beta-(1 –> 2) glucan export ATP-binding/permease NdvA protein, the proteins RHE_RS20455 and RHE_RS10390 were expressed in MM and bacteroid, respectively, with the TFs RHE_RS23325_m5 and RHE_RS26875_m3, for both not and clustered-TF were inferred, respectively, supporting a differential transcriptional regulation (Supplementary Table 6). Multiple *rbsB*; ribose transport system substrate-binding protein transporters, RHE_RS09135, RHE_RS22400, RHE_RS27555, RHE_RS30060, and RHE_RS30060 were expressed in MM, while RHE_RS29865 was identified in bacteroid; the data suggested that they were under the Clustered-TF transcriptional control of RHE_RS22090_m2, RHE_11740_m2, RHE_RS27560_m3, RHE_RS04690_m3 and RHE_RS02355_m4 and the not clustering TF associated RHE_RS10580_m1, respectively (Supplementary Table 6). Currently, it is not clear whether the plant supplies sugar to the bacteroid. A metabolome study showed that GDP-mannose and GDP-galactose were identified to be 7.4 times higher in bacteroids than in bacteria grown in MM (data not shown); in the opposite sense, proteins for these pathways were significantly higher in MM than in bacteroids (Supplementary Table 2, pathway 10). The two-component system, OmpR family response regulator proteins RHE_RS06580, RHE_RS10890, RHE_RS12325, and RHE_RS21355, were detected in MM and RHE_RS29195 in bacteroid, with TFs RHE_RS06580_m4, RHE_RS05790_m3, RHE_RS12325_m2, and RHE_RS21355_m3, for the proteins expressed in MM and RHE_RS29195_m2 for bacteroid, respectively (Supplementary Table 6), for non and clustered-TF networks, showing that multiplicity has a distinct potentially transcriptional regulation. The *nodD* LysR family transcriptional regulator recognizes a *nod-box* for transcriptional activation (Mao et al. 1994). We have demonstrated the function of the *nodD* transcriptional regulators by supplementation of MM with the flavonoid naringenin, which induced the synthesis of the nodulation factor (Meneses et al. 2017). The *nodD* genes RHE_RS30790, RHE_RS31010, and RHE_RS31005 proteins were expressed in Clustered-TF-MM and Clustered-TF-BACTEROID, respectively, probably under the transcriptional control of the inferred TFs RHE_RS30790_m2, RHE_RS12670_m4 Clustered-TF and RHE_RS20460_m2, not Clustered-TF, respectively (Supplementary Table 6). It was demonstrated that lysR *nodD* genes were autoregulated (Hu et al. 2000), as was *in silico* shown for the NodD RHE_RS30790 (Taboada-Castro et al. 2020); in addition, the *nodD* genes may be regulated by other TFs (Barnett and Long 2015), as was inferred for *nodD* RHE_RS31005 (Taboada-Castro et al. 2020). Altogether, these data suggested that in addition to specific isoenzymes expressed in a condition-dependent manner, they are potentially under specific transcriptional regulatory control. This data suggested how *R. etli* CFN42 re-program its transcriptional regulatory network to be metabolically adapted for growth in MM or in the symbiosis with the leguminous plant.

Conclusion

A free-living and symbiotic proteomic study from *R. etli* CFN42 were performed. A lower number of proteins per pathway in bacteroids than in MM was found, and approximately 30 and 20% of proteins for some metabolic pathways were detected in MM and bacteroids with respect to the genomic content, respectively. A mapping of classified proteins based on orthology allowed us to discover the presence of isoenzymes specific for growth in minimal medium and symbiosis with deduced specific transcriptional regulation. In addition to the metabolic pathways identified, genes for the degradation of environmental compounds were detected in MM and symbiotic proteomes. In contrast, a low number of isoenzymes were found in the *S. meliloti* transcriptome data. Taking advantage of the RhizoBindingSites database, which contains inferred TF gene–target relationships of *R. etli* CFN42 and eight additional symbiotic species, a method was implemented to construct transcriptional regulatory networks for these metabolic conditions. An inferred clustered TF gene network was constructed with motifs highly conserved in the upstream regulatory regions of the genes that are also conserved in the orthologous genes from each gene.

This pioneer bioinformatic framework is an important reference to obtain basic information on the genetic circuitry to increase knowledge about an experimental transcriptional regulatory network. Given the changing climate conditions, experimental validation of these genetic circuits for remodeling the metabolic pathways to optimize the SNF of *R. etli* CFN42 is the next step.

Data availability statement

The authors acknowledge that the data presented in this study must be deposited and made publicly available in an acceptable repository, prior to publication. Frontiers cannot accept a manuscript that does not adhere to our open data policies.

Author contributions

HT-C, JG, and SE-G conceived the idea. JG, HT-C, JF-G, JE-R, LG-C, and SE-G designed the analysis. HT-C, JG, JF-G, and SE-G analyzed the results and drafted the manuscript. SE-G revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Part of this work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica

(PAPIIT-UNAM), grants IN 213522 to SE-G and IN202421 to JF-G. JE-R is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM); he received fellowship 959406 from CONACYT.

Acknowledgments

The authors wish to thank María del Carmen Vargas-Lagunas and Yolanda Mora for their technical support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.947678/full#supplementary-material

SUPPLEMENTARY TABLE 1
Method for network construction A-I.

SUPPLEMENTARY TABLE 2
Metabolic pathways of the proteins expressed in Minimal medium, bacteroid and present in both conditions from *Rhizobium etli* CFN42.

SUPPLEMENTARY TABLE 3
Isoenzymes from *Rhizobium etli* CFN42 and *Sinorhizobium meliloti* 1021.

SUPPLEMENTARY TABLE 4
Predicted second transcriptional regulatory network with matrix-clustering of MM and Bacteroid profiles from *Rhizobium etli* CFN42.

SUPPLEMENTARY TABLE 5
Predicted Transcription Factor hierarchy in networks of MM, Clustered-TF-MM, BACTEROID and Clustered-TF-BACTEROID from *Rhizobium etli* CFN42.

SUPPLEMENTARY TABLE 6
Isoenzymes with inferred Transcriptional regulation of non-clustered and clustered networks from *Rhizobium etli* CFN42.

References

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* 2, 433–459. doi: 10.1002/WICS.101

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., and De Moor, B. (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res.* 31, 1753–1764. doi: 10.1093/nar/gkg268

Alamillo, J. M., Dí Az-Leal, J. L., Sánchez-Moran, M. A. V., and Pineda, M. (2010). Molecular analysis of ureide accumulation under drought stress in *Phaseolus vulgaris* L. *Plant Cell Environ.* 33, 1828–1837. doi: 10.1111/j.1365-3040.2010.02187.x

Andrews, M. E. M., and Andrews, M. E. M. (2017). Specificity in legume-rhizobia symbioses. *Int. J. Mol. Sci.* 18. doi: 10.3390/ijms18040705

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Atkinson, M. R., Blauwkamp, T. A., Bondarenko, V., Studitsky, V., and Ninfa, A. J. (2002). Activation of the glnA, glnK, and nac promoters as Escherichia coli undergoes the transition from nitrogen excess growth to nitrogen starvation. *J. Bacteriol.* 184, 5358–5363. doi: 10.1128/JB.184.19.5358-5363.2002

Bai, Y., Liang, J., Liu, R., Hu, C., and Qu, J. (2014). Metagenomic analysis reveals microbial diversity and function in the rhizosphere soil of a constructed wetland. *Environ. Technol.* 35, 2521–2527. doi: 10.1080/09593330.2014.911361

Barnett, M. J., and Long, S. R. (2015). The *sinorhizobium meliloti* SyrM Regulon: effects on global gene expression are mediated by *syrA* and *nodD3*. *J. Bacteriol.* 197, 1792–1806. doi: 10.1128/JB.02626-14

Barnett, M. J., Toman, C. J., Fisher, R. F., and Long, S. R. (2004). A dual-genome Symbiosis Chip for coordinate study of signal exchange and development in a prokaryote-host interaction. *Proc. Natl. Acad. Sci. U. S. A.* 101, 16636–16641. doi: 10.1073/pnas.0407269101

Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M., and van Helden, J. (2017). RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.* 45:e119. doi: 10.1093/nar/gkx314

Cevallos, M. A., Encarnación, S., Leija, A., Mora, Y., and Mora, J. (1996). Genetic and physiological characterization of a *rhizobium etli* mutant strain unable to synthesize poly-beta-hydroxybutyrate. *J. Bacteriol.* 178, 1646–1654. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8626293 [], doi: 10.1128/jb.178.6.1646-1654.1996

Collier, R., and Tegeder, M. (2012). Soybean ureide transporters play a critical role in nodule development, function and nitrogen export. *Plant J.* 72, 355–367. doi: 10.1111/j.1365-313X.2012.05086.x

Defrance, M., Janky, R., Sand, O., and van Helden, J. (2008). Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protoc.* 3, 1589–1603. doi: 10.1038/nprot.2008.98

Delgado, M. J., Bedmar, E. J., and Downie, J. A. (1998). Genes involved in the formation and assembly of rhizobial cytochromes and their role in symbiotic nitrogen fixation. *Adv. Microb. Physiol.* 40, 191–231. doi: 10.1016/s0065-2911(08)60132-0

Delmotte, N., Ahrens, C. H., Knief, C., Qeli, E., Koch, M., Fischer, H. M., et al. (2010). An integrated proteomics and transcriptomics reference data set provides new insights into the *Bradyrhizobium japonicum* bacteroid metabolism in soybean root nodules. *Proteomics* 10, 1391–1400. doi: 10.1002/pmic.200900710

Dicenzo, G. C., Zamani, M., Checcucci, A., Fondi, M., Griffitts, J. S., Finan, T. M., et al. (2019). Multidisciplinary approaches for studying rhizobium–legume symbioses. *Can. J. Microbiol.* 65, 1–33. doi: 10.1139/cjm-2018-0377

Diss, G., Ascencio, D., Deluna, A., and Landry, C. R. (2014). Molecular mechanisms of paralogous compensation and the robustness of cellular networks. *J. Exp. Zool. Part B Mol. Dev. Evol.* 322, 488–499. doi: 10.1002/jez.b.22555

Durán, D., Albareda, M., Marina, A., García, C., Ruiz-Argüeso, T., and Palacios, J. (2020). Proteome analysis reveals a significant host-specific response in *rhizobium leguminosarum* bv *viciae* endosymbiotic cells. *Mol. Cell. Proteomics* 20:100009. doi: 10.1074/mcp.RA120.002276

Encarnación, S., Dunn, M., Willms, K., Mora, J., Dunn, M., Willms, K., et al. (1995). Fermentative and aerobic metabolism in *rhizobium etli*. *J. Bacteriol.* 177, 3058–3066. doi: 10.1128/jb.177.11.3058-3066.1995

Escorcia-Rodríguez, J. M., Tauch, A., and Freyre-González, J. A. (2020). Abasy atlas v2.2: the most comprehensive and up-to-date inventory of meta-curated, historical, bacterial regulatory networks, their completeness and system-level characterization. *Comput. Struct. Biotechnol. J.* 18, 1228–1237. doi: 10.1016/j.csbj.2020.05.015

Escorcia-Rodríguez, J. M., Tauch, A., and Freyre-González, J. A. (2021). *Corynebacterium glutamicum* regulation beyond transcription: organizing principles and reconstruction of an extended regulatory network incorporating regulations mediated by small RNA and protein-protein interactions. *Microorganisms* 9. doi: 10.3390/microorganisms9071395

Ferguson, B. J., Mens, C., Hastwell, A. H., Zhang, M., Su, H., Jones, C. H., et al. (2019). Legume nodulation: the host controls the party. *Blackwell Publishing Ltd* 42, 41–51. doi: 10.1111/pce.13348

Fischer, H. M. (1994). Genetic regulation of nitrogen fixation in rhizobia. *Microbiol. Rev.* 58, 352–386. doi: 10.1128/mr.58.3.352-386.1994

Freyre-González, J. A., Alonso-Pavón, J. A., Treviño-Quintanilla, L. G., and Collado-Vides, J. (2008). Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach. *Genome Biol.* 9:R154. doi: 10.1186/gb-2008-9-10-r154

Freyre-González, J. A., Escorcia-Rodríguez, J. M., Gutiérrez-Mondragón, L. F., Martí-Vértiz, J., Torres-Franco, C. N., and Zorro-Aranda, A. (2022). System principles governing the organization, architecture, dynamics, and evolution of gene regulatory networks. *Front. Bioeng. Biotechnol.* 10:888732. doi: 10.3389/FBIOE.2022.888732

Freyre-González, J. A., and Tauch, A. (2017). Functional architecture and global properties of the Corynebacterium glutamicum regulatory network: novel insights from a dataset with a high genomic coverage. *J. Biotechnol.* 257, 199–210. doi: 10.1016/j.jbiotec.2016.10.025

Freyre-González, J. A., Treviño-Quintanilla, L. G., Valtierra-Gutiérrez, I. A., Gutiérrez-Ríos, R. M., and Alonso-Pavón, J. A. (2012). Prokaryotic regulatory systems biology: common principles governing the functional architectures of *Bacillus subtilis* and *Escherichia coli* unveiled by the natural decomposition approach. *J. Biotechnol.* 161, 278–286. doi: 10.1016/j.jbiotec.2012.03.028

Galán-Vásquez, E., and Perez-Rueda, E. (2019). Identification of modules with similar gene regulation and metabolic functions based on co-expression data. *Front. Mol. Biosci.* 6:139. doi: 10.3389/fmolb.2019.00139

Ghaffari, T., Kafil, H. S., Asnaashari, S., Farajnia, S., Delazar, A., Baek, S. C., et al. (2019). Chemical composition and antimicrobial activity of essential oils from the aerial parts of *Pinus eldarica* grown in northwestern Iran. *Molecules* 24:3203. doi: 10.3390/molecules24173203

Gil, J., and Encarnación-Guevara, S. (2022). "Lysine acetylation stoichiometry analysis at the proteome level", in *Clinical Proteomics. Methods in Molecular Biology. Vol. 2420*. eds. F. J. Corrales, A. Paradela and A. Marcilla New York, NY: Humana.

Gil, J., Ramírez-Torres, A., Chiappe, D., Luna-Penãloza, J., Fernandez-Reyes, F. C., Arcos-Encarnación, B., et al. (2017). Lysine acetylation stoichiometry and proteomics analyses reveal pathways regulated by sirtuin 1 in human cells. *J. Biol. Chem.* 292, 18129–18144. doi: 10.1074/jbc.M117.784546

González, V., Bustos, P., Ramírez-Romero, M., Medrano-Soto, A., Salgado, H., Hernández-González, I., et al. (2003). The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. *Genome Biol.* 4:R36. doi: 10.1186/gb-2003-4-6-r36

González, V., Santamaría, R. I., Bustos, P., Hernández-González, I., Medrano-Soto, A., Moreno-Hagelsieb, G., et al. (2006). The partitioned *Rhizobium etli* genome: genetic and metabolic redundancy in seven interacting replicons. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3834–3839. doi: 10.1073/pnas.0508502103

Hérouart, D., Baudouin, E., Frendo, P., Harrison, J., Santos, R., Jamet, A., et al. (2002). Reactive oxygen species, nitric oxide and glutathione: a key role in the establishment of the legume-rhizobium symbiosis? *Plant Physiol. Biochem.* 40, 40, 619–624. doi: 10.1016/S0981-9428(02)01415-8

Hertz, G. Z., Hartzell, G. W., and Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* 6, 81–92. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2193692 (Accessed March 14, 2017).

Hu, H., Liu, S., Yang, Y., Chang, W., and Hong, G. (2000). In *Rhizobium leguminosarum*, NodD represses its own transcription by competing with RNA polymerase for binding sites. *Nucleic Acids Res.* 28, 2784–2793. doi: 10.1093/nar/28.14.2784

Ihuegbu, N. E., Stormo, G. D., and Buhler, J. (2012). Fast, sensitive discovery of conserved genome-wide motifs. *J. Comput. Biol.* 19, 139–147. doi: 10.1089/cmb.2011.0249

Kanehisa, M. (2017). "Enzyme annotation and metabolic reconstruction using KEGG," in *Protein Function Prediction. Methods in Molecular Biology. Vol. 1611*. ed. Kihara, D. (New York, NY: Humana Press).

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Kaur, G., Kaundal, S., Kapoor, S., Grimes, J. M., Huiskonen, J. T., and Thakur, K. G. (2018). *Mycobacterium tuberculosis* CarD, an essential global transcriptional regulator forms amyloid-like fibrils. *Sci. Rep.* 8:10124. doi: 10.1038/s41598-018-28290-4

Khatabi, B., Gharechahi, J., Ghaffari, M. R., Liu, D., Haynes, P. A., McKay, M. J., et al. (2019). Plant–microbe symbiosis: what has proteomics taught us? *Proteomics* 19:1800105. doi: 10.1002/pmic.201800105

Landeta, C., Dávalos, A., Cevallos, M. Á., Geiger, O., Brom, S., and Romero, D. (2011). Plasmids with a chromosome-like role in rhizobia. *J. Bacteriol.* 193, 1317–1326. doi: 10.1128/JB.01184-10

Lardi, M., and Pessi, G. (2018). Functional genomics approaches to studying symbioses between legumes and nitrogen-fixing rhizobia. *High Throughput* 7:15. doi: 10.3390/ht7020015

Larrainzar, E., and Wienkoop, S. (2017). A proteomic view on the role of legume symbiotic interactions. *Front. Plant Sci.* 8:1267. doi: 10.3389/fpls.2017.01267

Lê, S., Josse, J., Rennes, A., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/jss.v025.i01

Liu, A., Contador, C. A., Fan, K., and Lam, H.-M. (2018). Interaction and regulation of carbon, nitrogen, and phosphorus metabolisms in root nodules of legumes. *Front. Plant Sci.* 9:1860. doi: 10.3389/fpls.2018.01860

Lopez, O., Morera, C., Miranda-Rios, J., Girard, L., Romero, D., and Soberon, M. (2001). Regulation of gene expression in response to oxygen in *Rhizobium etli*: role of FnrN in *fixNOQP* expression and in symbiotic nitrogen fixation. *J. Bacteriol.* 183, 6999–7006. doi: 10.1128/JB.183.24.6999-7006.2001

Ma, H. W., Buer, J., and Zeng, A. P. (2004). Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinform.* 5, 1–10. doi: 10.1186/1471-2105-5-199

Madariaga-Navarrete, A., Rodríguez-Pastrana, B. R., Villagómez-Ibarra, J. R., Acevedo-Sandoval, O. A., Perry, G., and Islas-Pelcastre, M. (2017). Bioremediation model for atrazine contaminated agricultural soils using phytoremediation (using *Phaseolus vulgaris* L.) and a locally adapted microbial consortium. *J. Environ. Sci. Heal. Part B Pestic. Food Contam. Agric. Wastes* 52, 367–375. doi: 10.1080/03601234.2017.1292092

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21, 3448–3449. doi: 10.1093/bioinformatics/bti551

Mao, C., Downie, J. A., and Hong, G. (1994). Two inverted repeats in the *nodD* promoter region are involved in *nodD* regulation in *Rhizobium leguminosarum*. *Gene* 145, 87–90. doi: 10.1016/0378-1119(94)90327-1

McGuire, A. M., Hughes, J. D., and Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 10, 744–757. doi: 10.1101/GR.10.6.744

Meneses, N., Taboada, H., Dunn, M. F. M. F., Vargas, M., Del, C. M. C., Buchs, N., et al. (2017). The naringenin-induced exoproteome of *Rhizobium etli* CE3. *Arch. Microbiol.* 199, 737–755. doi: 10.1007/s00203-017-1351-8

Miranda-Ríos, J., Morera, C., Taboada, H., Dávalos, A., Encarnación, S., Mora, J., et al. (1997). Expression of thiamin biosynthetic genes (*thiCOGE*) and production of symbiotic terminal oxidase *cbb3* in *Rhizobium etli*. *J. Bacteriol.* 179, 6887–6893. doi: 10.1128/jb.179.22.6887-6893.1997

Newman, J. D., Diebold, R. J., Schultz, B. W., and Noel, K. D. (1994). Infection of soybean and pea nodules by *Rhizobium* spp. purine auxotrophs in the presence of 5-aminoimidazole-4-carboxamide riboside. *J. Bacteriol.* 176, 3286–3294. doi: 10.1128/jb.176.11.3286-3294.1994

Nguyen, N. T. T., Contreras-Moreira, B., Castro-Mondragon, J. A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C. D. D. D., et al. (2018). RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.* 46, W209–W214. doi: 10.1093/nar/gky317

Novichkov, P. S., Rodionov, D. A., Stavrovskaya, E. D., Novichkova, E. S., Kazakov, A. E., Gelfand, M. S., et al. (2010). RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Res.* 38, W299–W307. doi: 10.1093/nar/gkq531

Oldroyd, G. E. D., Murray, J. D., Poole, P. S., and Downie, J. A. (2011). The rules of engagement in the legume-rhizobial symbiosis. *Annu. Rev. Genet.* 45, 119–144. doi: 10.1146/annurev-genet-110410-132549

Pankhurst, C. E. (1977). Symbiotic effectiveness of antibiotic-resistant mutants of fast-and slow-growing strains of *Rhizobium* nodulating lotus species. *Can. J. Microbiol.* 23, 1026–1033. doi: 10.1139/m77-152

Perez-Riverol, Y., Bai, J., Bandla, C., Hewapathirana, S., García-Seisdedos, D., Kamatchinathan, S., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.* 50, D543–D552. doi: 10.1093/nar/gkab1038

Prell, J., White, J. P., Bourdes, A., Bunnewell, S., Bongaerts, R. J., and Poole, P. S. (2009). Legumes regulate *rhizobium* bacteroid development and persistence by the supply of branched-chain amino acids. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12477–12482. doi: 10.1073/pnas.0903653106

Putty, K., Marcus, S. A., Mittl, P. R. E., Bogadi, L. E., Hunter, A. M., Arur, S., et al. (2013). Robustness of *Helicobacter pylori* infection conferred by context-variable redundancy among cysteine-rich paralogs. *PLoS One* 8:e59560. doi: 10.1371/journal.pone.0059560

Rascio, N., and La Rocca, N. (2013). *Biological Nitrogen Fixation. Reference Module in Earth Systems and Environmental Sciences* (New, York: Elsevier). doi:10.1016/B978-0-12-409548-9.00685-0.

Resendis-Antonio, O., Freyre-González, J. A., Menchaca-Méndez, R., Gutiérrez-Ríos, R. M., Martínez-Antonio, A., Ávila-Sánchez, C., et al. (2005). Modular analysis of the transcriptional regulatory network of *E. coli*. *Trends Genet.* 21, 16–20. doi: 10.1016/j.tig.2004.11.010

Resendis-Antonio, O., Hernández, M., Mora, Y., and Encarnación, S. (2012). Functional modules, structural topology, and optimal activity in metabolic networks. *PLoS Comput. Biol.* 8:e1002720. doi: 10.1371/journal.pcbi.1002720

Resendis-Antonio, O., Hernández, M., Salazar, E., Contreras, S., Batallar, G. M., Mora, Y., et al. (2011). Systems biology of bacterial nitrogen fixation: high-throughput technology and its integrative description with constraint-based modeling. *BMC Syst. Biol.* 5:120. doi: 10.1186/1752-0509-5-120

Rodriguez-Llorente, I., Caviedes, M. A., Dary, M., Palomares, A. J., Cánovas, F. M., and Peregrín-Alvarez, J. M. (2009). The Symbiosis Interactome: a computational approach reveals novel components, functional interactions and modules in *Sinorhizobium meliloti*. *BMC Syst. Biol.* 3, 1–18. doi: 10.1186/1752-0509-3-63

Romanov, V. I., Hernandez-Lucas, I., and Martinez-Romero, E. (1994). Carbon metabolism enzymes of *Rhizobium tropici* cultures and bacteroids. *Appl. Environ. Microbiol.* 60, 2339–2342. doi: 10.1128/aem.60.7.2339-2342.1994

Rutten, P. J., and Poole, P. S. (2019). Oxygen regulatory mechanisms of nitrogen fixation in rhizobia. *Adv. Microb. Physiol.* 75, 325–389. doi: 10.1016/bs.ampbs.2019.08.001

Sá, C., Matos, D., Pires, A., Cardoso, P., and Figueira, E. (2020). Airborne exposure of *Rhizobium leguminosarum* strain E20-8 to volatile monoterpenes: effects on cells challenged by cadmium. *J. Hazard. Mater.* 388:121783. doi: 10.1016/j.jhazmat.2019.121783

Salazar, E., Javier Díaz-Mejía, J., Moreno-Hagelsieb, G., Martínez-Batallar, G., Mora, Y., Mora, J., et al. (2010). Characterization of the Nif A-RpoN regulon in *Rhizobium etli* in free life and in symbiosis with *Phaseolus vulgaris*. *Appl. Environ. Microbiol.* 76, 4510–4520. doi: 10.1128/AEM.02007-09

Sarma, A. D., and Emerich, D. W. (2005). Global protein expression pattern of *Bradyrhizobium japonicum* bacteroids: a prelude to functional proteomics. *Proteomics* 5, 4170–4184. doi: 10.1002/pmic.200401296

Sarma, A. D., and Emerich, D. W. (2006). A comparative proteomic evaluation of culture grown vs nodule isolated *Bradyrhizobium japonicum*. *Proteomics* 6, 3008–3028. doi: 10.1002/pmic.200500783

Stekhoven, D. J., and Bühlmann, P. (2012). MissForest: non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/BIOINFORMATICS/BTR597

Taboada, H., Meneses, N., Dunn, M. F., Vargas-Lagunas, C., Buchs, N., Castro-Mondragon, J. A., et al. (2018). Proteins in the periplasmic space and outer membrane vesicles of *Rhizobium etli* CE3 grown in minimal medium are largely distinct and change with growth phase. *Microbiology* 165, 638–650. doi: 10.1099/mic.0.000720

Taboada-Castro, H., Castro-Mondragón, J. A., Aguilar-Vera, A., Hernández-Álvarez, A. J., van Helden, J., and Encarnación-Guevara, S. (2020). RhizoBindingSites, a database of DNA-binding motifs in nitrogen-fixing bacteria inferred using a footprint discovery approach. *Front. Microbiol.* 11:567471. doi: 10.3389/fmicb.2020.567471

Tsoy, O. V., Ravcheev, D. A., Cuklina, J., and Gelfand, M. S. (2016). Nitrogen fixation and molecular oxygen: comparative genomic reconstruction of transcription regulation in Alphaproteobacteria. *Front. Microbiol.* 7:1343. doi: 10.3389/fmicb.2016.01343

Valderrama, B., Dávalos, A., Girard, L., Morett, E., Mora, J., Valderrama, B., et al. (1996). Regulatory proteins and cis-acting elements involved in the transcriptional control of *Rhizobium etli* reiterated *nifH* genes. *J. Bacteriol.* 178, 3119–3126. doi: 10.1128/jb.178.11.3119-3126.1996

Van Helden, J., André, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827–842. doi: 10.1006/jmbi.1998.1947

Villaseñor, T., Brom, S., Dávalos, A., Lozano, L., Romero, D., Los Santos, A. G., et al. (2011). Housekeeping genes essential for pantothenate biosynthesis are plasmid-encoded in *Rhizobium etli* and *rhizobium leguminosarum*. *BMC Microbiol.* 11:66. doi: 10.1186/1471-2180-11-66

Wacek, T. J., and Brill, W. J. (1976). Simple, rapid assay for screening nitrogen-fixing ability in soybean1. *Crop Sci.* 16, 519–523. doi: 10.2135/cropsci1976.0011183X001600040020x

Weiss, L. A., Harrison, P. G., Nickels, B. E., Glickman, M. S., Campbell, E. A., Darst, S. A., et al. (2012). Interaction of CarD with RNA polymerase mediates *Mycobacterium tuberculosis* viability, rifampin resistance, and pathogenesis. *J. Bacteriol.* 194, 5621–5631. doi: 10.1128/JB.00879-12

Zorro-Aranda, A., Escorcia-Rodríguez, J. M., González-Kise, J. K., and Freyre-González, J. A. (2022). Curation, inference, and assessment of a globally reconstructed gene regulatory network for *Streptomyces coelicolor*. *Sci. Rep.* 12:2840. doi: 10.1038/S41598-022-06658-X

Zuleta, L. F. G., Cunha, C. D. O., de Carvalho, F. M., Ciapina, L. P., Souza, R. C., Mercante, F. M., et al. (2014). The complete genome of *Burkholderia phenoliruptrix* strain BR3459a, a symbiont of *Mimosa flocculosa*: highlighting the coexistence of symbiotic and pathogenic genes. *BMC Genomics* 15, 1–19. doi: 10.1186/1471-2164-15-535

VII. Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty

(Escorcia-Rodriguez et al., 2022)

Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty

Juan M. Escorcia-Rodríguez[1], Mario Esposito[2],
Julio A. Freyre-González[1] and Gabriel Moreno-Hagelsieb[2]

[1] Regulatory Systems Biology Research Group, Program of Systems Biology, Center for Genomic Sciences, Universidad Nacional Autonóma de México, Cuernavaca, Morelos, México
[2] Department of Biology, Wilfrid Laurier University, Waterloo, Canada

Submitted 25 August 2020
Accepted 14 July 2022
Published 31 August 2022

Corresponding author
Gabriel Moreno-Hagelsieb,
gmoreno@wlu.ca

Academic editor
Joseph Gillespie

Additional Information and Declarations can be found on page 11

DOI 10.7717/peerj.13843

ABSTRACT

Orthologs separate after lineages split from each other and paralogs after gene duplications. Thus, orthologs are expected to remain more functionally coherent across lineages, while paralogs have been proposed as a source of new functions. Because protein functional divergence follows from non-synonymous substitutions, we performed an analysis based on the ratio of non-synonymous to synonymous substitutions (dN/dS), as proxy for functional divergence. We used five working definitions of orthology, including reciprocal best hits (RBH), among other definitions based on network analyses and clustering. The results showed that orthologs, by all definitions tested, had values of dN/dS noticeably lower than those of paralogs, suggesting that orthologs generally tend to be more functionally stable than paralogs. The differences in dN/dS ratios remained suggesting the functional stability of orthologs after eliminating gene comparisons with potential problems, such as genes with high codon usage biases, low coverage of either of the aligned sequences, or sequences with very high similarities. Separation by percent identity of the encoded proteins showed that the differences between the dN/dS ratios of orthologs and paralogs were more evident at high sequence identity, less so as identity dropped. The last results suggest that the differences between dN/dS ratios were partially related to differences in protein identity. However, they also suggested that paralogs undergo functional divergence relatively early after duplication. Our analyses indicate that choosing orthologs as probably functionally coherent remains the right approach in comparative genomics.

Subjects Bioinformatics, Computational Biology, Evolutionary Studies, Genomics, Microbiology
Keywords Orthologs, Paralogs, Nonsynonymous to synonymous substitutions, dN/dS, Functional divergence, Positive selection

INTRODUCTION

Since the beginning of comparative genomics, the assumption was made that orthologs could be expected to conserve their functions more often than paralogs (*Mushegian & Koonin, 1996*; *Huynen & Bork, 1998*; *Bork et al., 1998*; *Tatusov et al., 2000*).
The expectation is based on the definitions of each homolog type: orthologs are characters

How to cite this article Escorcia-Rodríguez JM, Esposito M, Freyre-González JA, Moreno-Hagelsieb G. 2022. Non-synonymous to synonymous substitutions suggest that orthologs tend to keep their functions, while paralogs are a source of functional novelty. PeerJ 10:e13843 DOI 10.7717/peerj.13843

separating after speciation events, while paralogs are characters separating after duplication events (Fitch, 2000). Given those definitions, orthologs could be considered the "same" genes in different species, while paralogy has been proposed as a mechanism for the evolution of new functions, under the argument, in very simplified terms, that one of the copies could maintain the original function, while the other copy would have some freedom to functionally change (Ohno, 1970). This neither means that orthologs cannot evolve new functions, nor that paralogs necessarily evolve new functions. However, a scenario whereby most orthologs would diverge in functions at a higher rate than paralogs seems far from parsimonious, thus very unlikely. Therefore, it has been customary to use some working definition of orthology to infer the genes whose products most likely perform the same functions across different lineages (Mushegian & Koonin, 1996; Huynen & Bork, 1998; Bork et al., 1998; Tatusov et al., 2000; Gabaldón & Koonin, 2013).

Despite such a straightforward expectation, a report was published making the surprising claim that orthologs diverged in function more often than paralogs (Nehrt et al., 2011). The controversial article was mainly based on the comparison of Gene Ontology annotations among orthologs and paralogs from two species: humans and mice (Nehrt et al., 2011). If the report were correct, it would mean, for example, that mice myoglobin could be performing the function that human alpha-haemoglobin performs. However, data in the article showed that paralogs found within a genome, had more consistent gene ontology annotations than any homologs between both genomes. This was true even for identical proteins. Thus, rather than functional differences, it was possible that annotations of homologous genes were more consistent within a genome than between genomes. Accordingly, later work showed that gene ontologies suffered from "ascertainment bias", which made annotations more consistent within an organism than without (Thomas et al., 2012; Altenhoff et al., 2012). Later work showed gene expression data suggesting that orthologs had more coherent functions than paralogs (Kryuchkova-Mostacci & Robinson-Rechavi, 2016).

We thus wondered whether we could perform some analyses that did not suffer from annotation bias, and that could cover most of the homologs found between any pair of genomes, even if they had no functional annotations. Given that changes in protein function require changes in amino acids, analyses of non-synonymous to synonymous substitution rates, which compare the relative rates of positive and negative (purifying) selection (Ohta, 1995; Yang & Nielsen, 2000), might serve as proxies for functional divergence. The most functionally stable homologs would be expected to have lower dN/dS ratios compared to less functionally stable homologs. Thus, comparisons between the dN/dS distributions of orthologs and paralogs could show differences in their tendencies to conserve their functions. Since most of the related works have focused on eukaryotes, we centered our analyzes on prokaryotes (Bacteria and Archaea). We used five working definitions of orthology, including RBH, which is the foundation of most graph-based orthology prediction methods, besides arguably being the most usual working definition of orthology (Altenhoff & Dessimoz, 2009; Wolf & Koonin, 2012; Galperin et al., 2019).

Table 1 Genomes used in this study.

Genome ID	Class	Order	Species
Phylum proteobacteria			
GCF_000005845	Gammaproteobacteria	Enterobacterales	*Escherichia coli*
GCF_002370525	Gammaproteobacteria	Pseudomonadales	*Acinetobacter guillouiae*
GCF_002847445	Alphaproteobacteria	Rhodobacterales	*Paracoccus zhejiangensis*
GCF_004194535	Betaproteobacteria	Neisseriales	*Iodobacter fluviatilis*
GCF_013085545	Deltaproteobacteria	Desulfovibrionales	*Desulfovibrio marinus*
GCF_013283835	Epsilonproteobacteria	Campylobacterales	*Poseidonibacter lekithochrous*
GCF_000317895	Oligoflexia	Bdellovibrionales	*Bdellovibrio bacteriovorus*
GCF_009662475	Acidithiobacillia	Acidithiobacillales	*Acidithiobacillus thiooxidans*
GCF_002795805	Zetaproteobacteria	Mariprofundales	*Mariprofundus aestuarium*
GCF_003574215	Hydrogenophilalia	Hydrogenophilales	*Hydrogenophilus thermoluteolus*
Phylum firmcutes			
GCF_000009045	Bacilli	Bacillales	*Bacillus subtilis*
GCF_002197645	Bacilli	Lactobacillales	*Enterococcus wangshanyuanii*
GCF_000218855	Clostridia	Eubacteriales	*Clostridium acetobutylicum*
GCF_003991135	Clostridia	Halanaerobiales	*Anoxybacter fermentans*
GCF_000020005	Clostridia	Natranaerobiales	*Natranaerobius thermophilus*
GCF_003966895	Negativicutes	Selenomonadales	*Methylomusa anaerophila*
GCF_003367905	Negativicutes	Veillonellales	*Megasphaera stantonii*
GCF_012317185	Erysipelotrichia	Erysipelotrichales	*Erysipelatoclostridium innocuum*
GCF_000299355	Tissierellia	Tissierellales	*Gottschalkia acidurici*
GCF_001544015	Limnochordia	Limnochordales	*Limnochorda pilosa*
Phylum euryarchaeota			
GCF_000025625	Halobacteria	Natrialbales	*Natrialba magadii*
GCF_000011085	Halobacteria	Halobacteriales	*Haloarcula marismortui*
GCF_000025685	Halobacteria	Haloferacales	*Haloferax volcanii*
GCF_000195895	Methanomicrobia	Methanosarcinales	*Methanosarcina barkeri*
GCF_000013445	Methanomicrobia	Methanomicrobiales	*Methanospirillum hungatei*
GCF_001433455	Thermococci	Thermococcales	*Thermococcus barophilus*
GCF_000024185	Methanobacteria	Methanobacteriales	*Methanobrevibacter ruminantium*
GCF_000006175	Methanococci	Methanococcales	*Methanococcus voltae*
GCF_000734035	Archaeoglobi	Archaeoglobales	*Archaeoglobus fulgidus*
GCF_000007185	Methanopyri	Methanopyrales	*Methanopyrus kandleri*

Note:
The query genomes were the first in each group.

MATERIALS AND METHODS

Genome data

We downloaded the analyzed genomes from NCBI's RefSeq Genome database (*Haft et al., 2018*). We performed our analyses by selecting genomes from three taxonomic phyla, using one genome within each phylum as a query genome (Table 1): *Escherichia coli* K12 MG1655 (phylum Proteobacteria, domain Bacteria, assembly ID: GCF_000005845), *Bacillus subtilis* 168 (Firmicutes, Bacteria, GCF_000009045), and *Natrialba magadii* ATCC43099 (Euryarchaeota, Archaea, GCF_000025625).

Orthologs

We used five working definitions of orthology:

Reciprocal best hits (RBH)

We compared the proteomes of each of these genomes against those of other members of their taxonomic phylum using diamond (*Buchfink, Xie & Huson, 2015*), with the – – *very – sensitive* option, and a maximum e-value of 1×10^{-6} (-evalue 1e–6) (*Hernández-Salmerón & Moreno-Hagelsieb, 2020*). We also required a minimum alignment coverage of 60% of the shortest sequence. Orthologs were defined as reciprocal best hits (RBH) as described previously (*Moreno-Hagelsieb & Latimer, 2008*; *Ward & Moreno-Hagelsieb, 2014*; *Hernández-Salmerón & Moreno-Hagelsieb, 2020*). Except where noted, paralogs were all matches left after finding RBH.

Ortholog groups with inparalogs (InParanoid)

InParanoid is a graph-based tool to identify orthologs and in-paralogs from pairwise sequence comparisons (*Sonnhammer & Östlund, 2015*). InParanoid first runs all-*vs*-all blastp and identifies RBH. Then, it uses the RBH as seeds to identify co-orthologs for each gene (which the authors define as in-paralogs), proteins from the same organism that obtain better bits score than the RBH. Finally, through a series of rules, InParanoid cluster the co-orthologs to return non-overlapping groups. The authors define outparalogs as those blast-hits outside of the co-ortholog clusters (*Sonnhammer & Östlund, 2015*).

We ran InParanoid for each query genome against those of other members of their taxonomic order. InParanoid was run with the following parameters: double blast and 40 bits as score cutoff. The first pass run with compositional adjustment on and soft masking. This removes low complexity matches but truncates alignments (*Sonnhammer & Östlund, 2015*). The second pass run with compositional adjustment off to get full-length alignments. We used as in-paralogs the combinatorial of the genes of the same organism from the same cluster, and as out-paralogs those blast-hits outside of the co-ortholog clusters.

Orthologous MAtrix (OMA)

OMA is a pipeline and database that provides three different types of orthologs: pairwise orthologs, OMA groups (orthogroups), and hierarchical orthologous groups (*Zahn-Zabal, Dessimoz & Glover, 2020*). OMA makes an effort to remove xenologs by using a third proteome as witness of non-orthology (*Roth, Gonnet & Dessimoz, 2008*). To the best of our knowledge, OMA is the only orthology prediction method, still being maintained, able to deal with xenology. The OMA pipeline for the identification of orthologs is based on best reciprocal Smith-Waterman hits and some tolerance for evolutionary distance that allows for co-orthology. For pairwise orthology identification, a verification step to detect xenologs is applied using a third proteome that retained both pseudo-orthologous genes (*Train et al., 2017*).

We ran the OMA standalone (version 2.5.0) with all the proteomes for each taxonomic group using the default parameters (*Train et al., 2017*). We used the pairwise orthology outputs considering the query organisms. For the identification of in-paralogs for the query organism, we used the co-orthologous genes mapping to one or more orthologs in the rest of the organisms. OMA also generates pairwise paralogy outputs, including former candidates for orthologs that did not reach the thresholds or were discarded by a third organism retaining both genes (*Zahn-Zabal, Dessimoz & Glover, 2020*).

OrthoFinder

OrthoFinder defines an orthogroup as the set of genes derived from a single gene in the last common ancestor of the all species under consideration (*Emms & Kelly, 2015*). First, OrthoFinder performs all-*vs*-all blastp (*Camacho et al., 2009*) comparisons and uses an e-value of 1×10^{-3} as a threshold. Then, it normalizes the gene length and phylogenetic distance of the BLAST bit scores. It uses the lowest normalized value of the RBH for either gene in a gene pair as the threshold for their inclusion in an orthogroup. Finally, it weights the orthogroup graph with the normalized bit scores and clusters it using MCL. OrthoFinder outputs the orthogroups and orthology relationships, which can be many to many (co-orthology).

We ran OrthoFinder with all the proteomes for each of the taxonomic groups listed (Table 1). From the OrthoFinder outputs with the orthology relationships between every two species, we used those considering the query organism. From an orthogroup containing one or more orthology relationships, we identified the outparalogs as those genes belonging to the same orthogroup but not to the same orthology relationship. We identified the inparalogs for the query organisms as its genes belonging to the same orthogroup since they derived from a single ancestor gene.

ProteinOrtho

ProteinOrtho is a graph-based tool that implements an extended version of the RBH heuristic and is intended for the identification of ortholog groups between many organisms (*Lechner et al., 2011*). First, from all-*vs*-all blast results, ProteinOrtho creates subnetworks using the RBH at the seed. Then, if the second best hit for each protein is almost as good as the RBH, it is added to the graph. The algorithm claims to recover false negatives and to avoid the inclusion of false positives (*Lechner et al., 2011*).

We ran ProteinOrtho pairwise since we needed to identify orthologs and paralogs between the query organisms and the other members of their taxonomic data, not those orthologs shared between all the organisms. ProteinOrtho run blast with the following parameters: an e-value cutoff of 1×10^{6}, minimal alignment coverage of 50% of the shortest sequence, and 25% of identity. Orthologs were the genes from different genomes that belonged to the same orthogroup, and inparalogs genes from the same genome that belong to the same orthogroup. We reran ProteinOrtho with a similarity value of 75% instead of 90%, to identify outparalogs as those interactions not identified in the first run.

Non-synonymous to synonymous substitutions

To perform *dN*/*dS* estimates, we used the CODEML program from the PAML software suite (*Yang, 2007*). The DNA alignments were derived from the protein sequence alignments using an *ad hoc* program written in PERL. The same program ran pairwise comparisons using CODEML to produce Bayesian estimates of *dN*/*dS* (*Angelis, dos Reis & Yang, 2014*; *Anisimova, Bielawski & Yang, 2002*). The results were separated between ortholog and paralog pairs, and the density distributions were plotted using R (*R Core Team, 2020*). Statistical analyses were also performed with R.

Codon adaptation index

To calculate the Codon Adaptation Index (CAI) (*Sharp & Li, 1987*), we used ribosomal proteins as representatives of highly expressed genes. To find ribosomal proteins we matched the COG ribosomal protein families described by *Yutin et al. (2012)* to the proteins in the genomes under analysis using RPSBLAST (part of NCBI's BLAST+ suite) (*Camacho et al., 2009*). RPSBLAST was run with soft-masking (-seg yes -soft_masking true), a Smith-Waterman final alignment (-use_sw_tback), and a maximum e-value threshold of 1×10^{-3} (-evalue 1e–3). A minimum coverage of 60% of the COG domain model was required. To produce the codon usage tables of the ribosomal protein-coding genes, we used the program *cusp* from the EMBOSS software suite (*Rice, Longden & Bleasby, 2000*). These codon usage tables were then used to calculate the CAI for each protein-coding gene within the appropriate genome using the *cai* program also from the EMBOSS software suite (*Rice, Longden & Bleasby, 2000*).

RESULTS AND DISCUSSION

While we have been working on this report, an article following the same basic idea, comparing *dN*/*dS* distributions between orthologs and paralogs, though focusing on vertebrates, was published (*David, Oaks & Halanych, 2020*). Their results were consistent with those described below.

Reciprocal best hits showed lower *dN*/*dS* ratios than paralogs

These studies used Bayesian *dN*/*dS* estimates, because they are considered the most robust and accurate (*Anisimova, Bielawski & Yang, 2002*; *Angelis, dos Reis & Yang, 2014*). To compare the distribution of *dN*/*dS* values between orthologs and paralogs, we plotted *dN*/*dS* density distributions using violin plots (Fig. 1). These plots demonstrated evident differences, with orthologs showing lower *dN*/*dS* ratios than paralogs, thus indicating that orthologs have diverged in function less frequently than paralogs. In line with the noticeable differences, Wilcoxon rank tests showed that the differences were statistically significant, with probabilities much lower than 1×10^{-9} (Table S1). Since most comparative genomics work is done using reciprocal best hits (RBH) as a working definition for orthology (*Wolf & Koonin, 2012*; *Galperin et al., 2019*), this result suggests that most research in comparative genomics has used the proteins/genes that most likely share their functions.

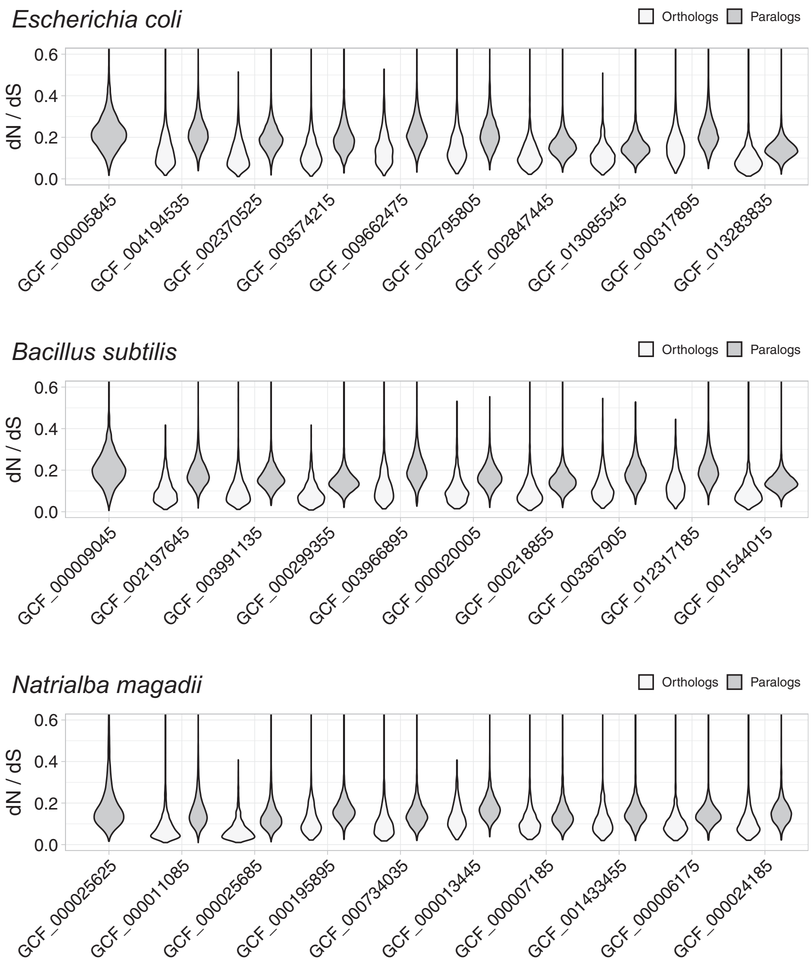


Figure 1 **Non-synonymous to synonymous substitutions (*dN/dS*).** The *dN/dS* ratios correspond to genes compared between query organisms against genomes from organisms in the same taxonomic phylum, namely: *E. coli* against other Proteobacteria, *B. subtilis* against other Firmicutes, and *N. magadii* against other Euryarchaeota. Genome identifiers are ordered from most similar to least similar to the query genome. The *dN/dS* distribution is higher for paralogs, suggesting that a higher proportion of orthologs have retained their functions. Full-size DOI: 10.7717/peerj.13843/fig-1

Differences in *dN*/*dS* resisted other working definitions of orthology

A concern with our analyses might arise from our initial focus on reciprocal best hits (RBH). However, RBH might arguably be the most usual working definition of orthology (*Altenhoff & Dessimoz, 2009*; *Wolf & Koonin, 2012*; *Galperin et al., 2019*). Thus, it is important to start these analyses with RBH, at least to test whether RBH are a good choice for the purpose of inferring genes most likely to have similar functions.

Analyses of the quality of RBH for inferring orthology, based on synteny, showed that RBH error rates were lower than 5% (*Moreno-Hagelsieb & Latimer, 2008*; *Wolf & Koonin, 2012*; *Hernández-Salmerón & Moreno-Hagelsieb, 2020*). Other analyses showed that the problem with RBH, was a slightly higher rate of false positives (paralogs mistaken for orthologs), than databases based on phylogenetic and network analyses (*Altenhoff &*

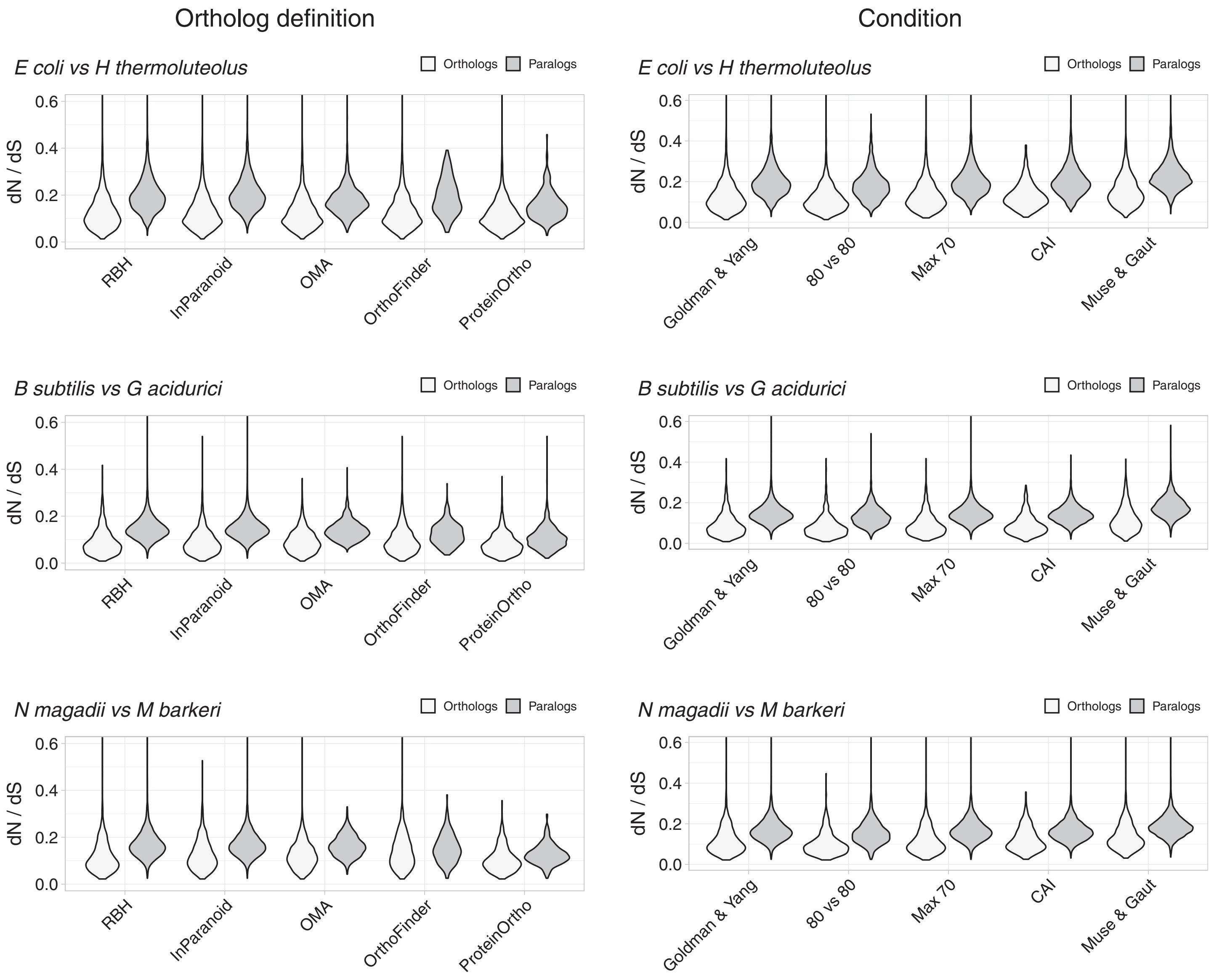


Figure 2 Control experiments. Left: values of *dN*/*dS* ratios were higher for different definitions of orthology than for their paralogs. RBH were included as reference. Right: examples of *dN*/*dS* values obtained testing for potential biases. The Goldman and Yang model for estimating codon frequencies (*Goldman & Yang, 1994*), included as reference, is the default. The 80 *vs* 80 test used data for orthologs and paralogs filtered to contain only alignments covering at least 80% of both proteins. The maximum identity test filtered out sequences more than 70% identical. The CAI test filtered out sequences having Codon Adaptation Indexes (CAI) from the top and bottom 15 percentile of the genome's CAI distribution. We also tested the effect of the Muse and Gaut model for estimating background codon frequencies (*Muse & Gaut, 1994*).

Full-size DOI: 10.7717/peerj.13843/fig-2

Dessimoz, 2009). Therefore, we can assume that orthologs dominate the RBH *dN*/*dS* distributions.

Despite the above justification for focusing on RBH, we considered four other definitions of orthology (Fig. 2, Figs. S1–S4). Orthologs obtained with different working definitions, including one method dealing with xenologs (OMA), showed *dN*/*dS* ratio distributions that suggest that a higher proportion of orthologs have similar functions compared to paralogs (Fig. 2).

Differences in *dN*/*dS* persisted after testing for potential biases

While the tests above suggest that RBH separate homologs with higher tendencies to preserve their functions than other homologs, we tested for some potential biases. A potential problem could arise from comparing proteins of very different lengths. We thus filtered the *dN*/*dS* results to keep those where the pairwise alignments covered at least 80% of the length of both proteins. The results showed shorted tails in both density distributions, but the tendency for orthologs to have lower *dN*/*dS* values remained (Fig. 2, Fig. S5).

Another parameter that could bias the *dN*/*dS* results is high sequence similarity. In this case, the programs tend to produce high *dN*/*dS* ratios. While we should expect this issue to have a larger effect on orthologs, we still filtered both datasets, orthologs and paralogs, to contain proteins less than 70% identical. This filter had very little effect (Fig. 2, Fig. S6).

Lateral gene transfer events might be a problem with orthology predictions. However, proper genome-wide identification of lateral gene transfer events is difficult, as xenologs are hard to distinguish from duplications events (*Roth, Gonnet & Dessimoz, 2008*). Additionally, there is no good agreement between the output of different xenolog prediction methods benchmarked against real data (*Ravenhall et al., 2015*). In an attempt to deal with xenologs we used two approaches: We removed genes with atypical codon usage bias (see below), besides including an orthology working definition (OMA), that attempts to deal with xenologs. OMA uses a verification step to help reduce the number of xenologs by using a third proteome as witness of non-orthology (*Roth, Gonnet & Dessimoz, 2008*).

As mentioned above, to try and avoid the effect of sequences with unusual compositions, we filtered out sequences with extreme codon usages as measured using the Codon Adaptation Index (CAI) (*Sharp & Li, 1987*). For this test, we eliminated sequences with CAI values from the top and the bottom 15 percentile of the respective genome's CAI distribution. After filtering, orthologs still exhibited *dN*/*dS* values below those of paralogs (Fig. 2, Fig. S7).

Different models for background codon frequencies can also alter the *dN*/*dS* results (*Bielawski, 2013*). Thus, we performed the same tests using the Muse and Gaut model for estimating background codon frequencies (*Muse & Gaut, 1994*), as advised in (*Bielawski, 2013*). Again, the results showed orthologs to have lower *dN*/*dS* ratios than paralogs (Fig. 2, Fig. S8).

Differences in *dN*/*dS* ratios were more evident for genes encoding for less divergent proteins

Orthologs will normally contain more similar proteins than paralogs. Thus, a similarity test alone would naturally make orthologs appear less divergent and, apparently, less likely to have evolved new functions. While synonymous substitutions attest for the strength of negative/purifying selection in *dN*/*dS* analyses, seemingly making these ratios independent of the similarity between proteins, we still wondered whether the data changed with protein sequence divergence.

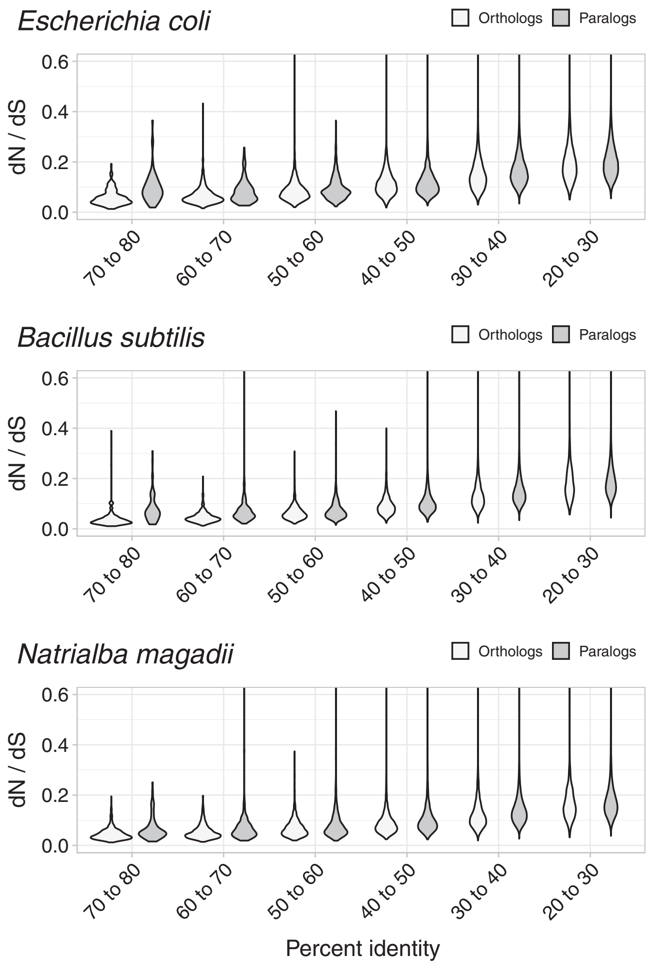


Figure 3 Non-synonymous to synonymous substitutions *dN/dS* and divergence. The difference between *dN/dS* ratios became less apparent as protein identity decreased.

Full-size DOI: 10.7717/peerj.13843/fig-3

To test whether *dN/dS* increased against sequence divergence, we separated orthologs and paralogs into ranges of divergence of the encoded protein's percent identity. The more similar the protein sequences, the more evident were the differences between the *dN/dS* of orthologs and paralogs (Fig. 3). Since protein sequence identity plays a role in most working definitions of orthology, the latter results partially explained the evident disparity in *dN/dS* ratios between orthologs and paralogs. However, that the ratio differences were more evident at low protein sequence divergence supports the hypothesis that paralogs might be an immediate source of functional novelty. Given that redundant duplications would be expected to eventually erode (*Ochman & Davalos, 2006*), early functional divergence might provide paralogs with the selective pressure to survive genetic erosion.

CONCLUSION

The results shown above used a measure of divergence that relates to the tendencies of sequences to diverge in amino-acid composition, against their tendencies to remain unchanged; namely, non-synonymous to synonymous substitution rates (*dN/dS*). Since changes in function require changes in amino-acids, this measure might suggest which sequence datasets have higher proportions that remain functionally coherent. Such proportions would show as a tendency towards lower *dN*/*dS* values. Orthologs showed evidently lower values of *dN*/*dS* than paralogs. Thus, orthologs could be though as more functionally stable than paralogs, with paralogs being a main source of novel functions.

ACKNOWLEDGEMENTS

We are grateful to Joe Bielawski for helpful advice. We thank The Shared Hierarchical Academic Research Computing Network (SHARCNET) for computing facilities.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Work supported with a Discovery Grant to Gabriel Moreno-Hagelsieb from the Natural Sciences and Engineering Research Council of Canada (NSERC). This work was also supported by the Programa de Apoyo a Proyectos de Investigacion e Innovacion Tecnologica (PAPIIT-UNAM) (IN205918 and IN202421) to Julio A. Freyre-González. Juan M. Escorcia-Rodríguez is supported by PhD fellowship 959406 from Consejo Nacional de Ciencia y Tecnología (CONACyT-Mexico). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Natural Sciences and Engineering Research Council of Canada (NSERC).
Programa de Apoyo a Proyectos de Investigacion e Innovacion Tecnologica (PAPIIT-UNAM): IN205918 and IN202421.
PhD Fellowship 959406 from Consejo Nacional de Ciencia y Tecnología (CONACyT-Mexico).

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Juan M. Escorcia-Rodríguez performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Mario Esposito performed the experiments, analyzed the data, prepared figures and/or tables, and approved the final draft.
- Julio A. Freyre-González analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Gabriel Moreno-Hagelsieb conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The programs for obtaining orthologs and dN/dS values as tested in this study are available at GitHub: https://github.com/Computational-conSequences/SequenceTools.

Genomes used in this study are available in Table 1.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13843#supplemental-information.

REFERENCES

Altenhoff AM, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLOS Computational Biology* **5(1)**:e1000262 DOI 10.1371/journal.pcbi.1000262.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLOS Computational Biology* **8(5)**:e1002514 DOI 10.1371/journal.pcbi.1002514.

Angelis K, dos Reis M, Yang Z. 2014. Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. *Molecular Biology and Evolution* **31(7)**:1902–1913 DOI 10.1093/molbev/msu142.

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Molecular Biology and Evolution* **19(6)**:950–958 DOI 10.1093/oxfordjournals.molbev.a004152.

Bielawski JP. 2013. Detecting the signatures of adaptive evolution in protein-coding genes. *Current Protocols in Molecular Biology* **101(1)**:19.1.1–19.1.21 DOI 10.1002/0471142727.mb1901s101.

Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, Yuan Y. 1998. Predicting function: from genes to genomes and back 1 1Edited by P. E. Wright. *Journal of Molecular Biology* **283(4)**:707–725 DOI 10.1006/jmbi.1998.2144.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12(1)**:59–60 DOI 10.1038/nmeth.3176.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10(1)**:421 DOI 10.1186/1471-2105-10-421.

David KT, Oaks JR, Halanych KM. 2020. Patterns of gene evolution following duplications and speciations in vertebrates. *PeerJ* **8(5)**:e8813 DOI 10.7717/peerj.8813.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16(1)**:157 DOI 10.1186/s13059-015-0721-2.

Fitch WM. 2000. Homology a personal view on some of the problems. *Trends in Genetics: TIG* **16(5)**:227–231 DOI 10.1016/S0168-9525(00)02005-9.

Gabaldón T, Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* **14(5)**:360–366 DOI 10.1038/nrg3456.

Galperin MY, Kristensen DM, Makarova KS, Wolf YI, Koonin EV. 2019. Microbial genome analysis: the COG approach. *Briefings in Bioinformatics* **20(4)**:1063–1070 DOI 10.1093/bib/bbx117.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11(5)**:725–736 DOI 10.1093/oxfordjournals.molbev.a040153.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, Gwadz M, Lu F, Marchler GH, Song JS, Thanki N, Yamashita RA, Zheng C, Thibaud-Nissen F, Geer LY, Marchler-Bauer A, Pruitt KD. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research* **46(D1)**:gkx1068 DOI 10.1093/nar/gkx1068.

Hernández-Salmerón JE, Moreno-Hagelsieb G. 2020. Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. *BMC Genomics* **21(1)**:741 DOI 10.1186/s12864-020-07132-6.

Huynen MA, Bork P. 1998. Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America* **95(11)**:5849–5856 DOI 10.1073/pnas.95.11.5849.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2016. Tissue-specificity of gene expression diverges slowly between orthologs, and rapidly between paralogs. *PLOS Computational Biology* **12(12)**:e1005274 DOI 10.1371/journal.pcbi.1005274.

Lechner M, FindeiB S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12(1)**:124 DOI 10.1186/1471-2105-12-124.

Moreno-Hagelsieb G, Latimer K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics* **24(3)**:319–324 DOI 10.1093/bioinformatics/btm585.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11(5)**:715–724 DOI 10.1093/oxfordjournals.molbev.a040152.

Mushegian AR, Koonin EV. 1996. Gene order is not conserved in bacterial evolution. *Trends in Genetics: TIG* **12(8)**:289–290 DOI 10.1016/0168-9525(96)20006-X.

Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLOS Computational Biology* **7(6)**:e1002073 DOI 10.1371/journal.pcbi.1002073.

Ochman H, Davalos LM. 2006. The nature and dynamics of bacterial genomes. *Science* **311(5768)**:1730–1733 DOI 10.1126/science.1119966.

Ohno S. 1970. *Evolution by gene duplication*. Berlin: Springer-Verlag.

Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution* **40(1)**:56–63 DOI 10.1007/BF00166595.

R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. *Available at http://www.R-project.org/*.

Ravenhall M, Škunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. *PLOS Computational Biology* **11(5)**:e1004095 DOI 10.1371/journal.pcbi.1004095.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics: TIG* **16(6)**:276–277 DOI 10.1016/S0168-9525(00)02024-2.

Roth AC, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9(1)**:518 DOI 10.1186/1471-2105-9-518.

Sharp PM, Li WH. 1987. The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* **15(3)**:1281–1295 DOI 10.1093/nar/15.3.1281.

Sonnhammer EL, Östlund G. 2015. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research* **43(D1)**:D234–D239 DOI 10.1093/nar/gku1203.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28(1)**:33–36 DOI 10.1093/nar/28.1.33.

Thomas PD, Wood V, Mungall CJ, Lewis SE, Blake JA, On Behalf of the Gene Ontology ConsortiumBourne PE. 2012. On the use of gene ontology annotations to assess functional similarity among orthologs and paralogs: a short report. *PLOS Computational Biology* **8(2)**:e1002386 DOI 10.1371/journal.pcbi.1002386.

Train C-M, Glover NM, Gonnet GH, Altenhoff AM, Dessimoz C. 2017. Orthologous matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **33(14)**:i75–i82 DOI 10.1093/bioinformatics/btx229.

Ward N, Moreno-Hagelsieb G. 2014. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? *PLOS ONE* **9**:e101850 DOI 10.1371/journal.pone.0101850.

Wolf YI, Koonin EV. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biology and Evolution* **4(12)**:1286–1294 DOI 10.1093/gbe/evs100.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24(8)**:1586–1591 DOI 10.1093/molbev/msm088.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17(1)**:32–43 DOI 10.1093/oxfordjournals.molbev.a026236.

Yutin N, Puigbò P, Koonin EV, Wolf YI. 2012. Phylogenomics of prokaryotic ribosomal proteins. *PLOS ONE* **7(5)**:e36972 DOI 10.1371/journal.pone.0036972.

Zahn-Zabal M, Dessimoz C, Glover NM. 2020. Identifying orthologs with OMA: a primer. *F1000Research* **9**:27 DOI 10.12688/f1000research.21508.1.