UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
DOCTORADO EN CIENCIAS BIOMÉDICAS
Centro de Ciencias Genómicas

Mapeo genómico de regulación transcripcional y metabolismo describe unidades de procesamiento de información en *Escherichia coli K-12*

TESIS

que para optar por el grado de:

Doctora en Ciencias

presenta:

Daniela Elizabeth Ledezma Tejeida

Tutor principal:
Dr. Julio Collado Vides
Centro de Ciencias Genómicas

Cuernavaca, Morelos, noviembre de 2017

A Rafa, que me decía doctora.
A Mary, que me enseñó a pensar en grande.
A mis papás, que lo hicieron posible.
Y a mi hermano, que ha sido mi héroe desde el primer día.

Agradecimientos

Académicos

A Julio Collado Vides por su guía como tutor, las invaluables discusiones de diseño del proyecto y resultados, revisiones de manuscritos y la confianza para dejarme ser cada vez más independiente.

A Agustino Martínez Antonio y Enrique Morett Sánchez por ser parte de mi comité tutor, por todas las discusiones, sugerencias, comentarios y su paciencia en épocas de trámites.

Al Programa de Doctorado en Ciencias Biomédicas por todas las facilidades prestadas a lo largo de mis estudios. En particular a Denny Peralta, Gladys Avilés y Susana Brom.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por apoyarme económicamente mediante la beca no. 275805.

Al Instituto Nacional de Salud de Estados Unidos (NIH) y FOINS CONACYT por cubrir los costos de este proyecto mediante los donativos 5R01GM110597-03 y Fronteras de la Ciencia proyecto #15.

A Socorro Gama-Castro, pionera de las GENSOR Units, por las discusiones, sugerencias y haber sentado las bases para la automatización del proceso de construcción de GENSOR Units.

A Cecilia Ishida, por curar manualmente las GENSOR Units, escribir los resúmenes y colaborar en la escritura del artículo resultante.

A Manuel Camacho, Sara Martínez y David Velázquez, por su participación en la curación manual y edición de las figuras de las 189 GENSOR Units.

A Heladia Salgado, Hilda Solano y Diego Palomares por diseñar y llevar a cabo la implementación y actualizaciones de GENSOR Units en RegulonDB.

A Luis José Muñiz por proveer los datasets de RegulonDB y por su apoyo técnico para instalar y actualizar Perlcyc y Pathway Tools.

A José Alquicira por su invaluable apoyo estadístico.

A Enrique Burguete, Anastasia Hernández y Hugo Cisneros por sus contribuciones al proyecto durante sus estancias en el laboratorio.

A Cei Abreu por sus contribuciones a este proyecto durante el examen de candidatura. Por su apoyo en el proceso de titulación.

A Fernando Pérez Villatoro y Mishael Sánchez por su ayuda para convertir este texto a Word. A Abraham Avelar por su ayuda logística.

A mi jurado de examen de grado: Ismael Sánchez, Santiago Castillo, José Utrilla y Osbaldo Reséndis. Por sus contribuciones a esta tesis y su apoyo para realizar los trámites correspondientes.

Personales

A Julio Collado, esta vez de manera personal, por su apoyo y confianza; por enseñarme que la ciencia también puede ser humana y que nunca hay que dejar de soñar pero al mismo tiempo hay que hacer que las cosas sucedan.

A todos los miembros del PGC, pasados y presentes, por formar un ambiente laboral que siempre me hizo sentir en casa. En especial a Ale López, Shirley Alquicira, Irma Martínez, Soco Gama, Heli Salgado y

Sara Martínez. A Jaime Castro por las pláticas esotéricas/filosóficas y los martes de Tepoz. A José Alquicira por tanta fritez, risas, achaques compartidos y ser pieza clave en mi carrera ukulelesca. A Conchi Hernández por todo su apoyo durante estos 5 años, por ser mi compañera de sismo y simplemente por ser el alma del laboratorio y ser tan linda con todos. A Ceci Ishida, por las pláticas, los libros y el té.

A las generaciones 10 a 15 de la Licenciatura en Ciencias Genómicas, por sus preguntas, por siempre sacarme de la zona de confort y hacerme pensar en cosas que jamás se me habrían ocurrido. Por enseñarme que nada es obvio.

A toda la comunidad de Atmaram, en especial a Tláloc Salinas, Jessy Jerman, Aura Delfín y Bindu de la Parra. Por el apoyo que me ha permitido disfrutar la montaña rusa emocional que han sido los últimos 5 años. Por hacer que todo haya cambiado aunque siga igual.

A toda la quinta generación de la Licenciatura en Ciencias Genómicas, porque cada uno sin excepción me ha dejado lecciones para toda la vida.

A todos los amigos que me escucharon, me dieron ánimos e ideas para el proyecto. Mitzy Ríos, Fernando Pérez, Abraham Avelar, Hugo Sámano, Betty Noriega, Víctor Moreno y Jazmín Ramos, que no han dejado que la distancia sea un obstáculo para que seamos amigos hasta que estemos viejitos, vivamos en Huatulco y discutamos todo el tiempo. Stefania Laddaga, que ha compartido historias, risas y café de señoras durante 17 años. Gonzalo Sarrelangue, que nunca deja de enseñarme cosas que no se pueden aprender en un libro. Óscar Migueles y Dafne Ibarra, los mejores roomies que habría podido pedir. Mario Sandoval, que siempre estuvo dispuesto a platicar y discutir sobre lo que fuera, por horas, mientras hubiera algo gordo y delicioso que comer. Laura Gómez, que siempre me recuerda no juzgar y me enseñó que personas muy distintas también pueden hacer un gran equipo. Luis Pedro Íñiguez, que lleva 10 años siendo mi ejemplo viviente de que ser feliz es muy fácil.

A mis papás y mi hermano. Por todo. No hay espacio suficiente para

mencionar todo lo que les agradezco, pero seguiré tratando de demostrarlo todos los días. A mis primos y tíos, por los consejos, las porras y el apoyo incondicional.

A Orlando Santillán, mi mejor amigo y compañero de equipo favorito. Por acompañarme, escucharme, motivarme, devolverme al mundo real en épocas de estrés y hacerme reír (e ir por pizza) cuando me tiraba al drama. Por ser mi maestro involuntario de tantas cosas, algunas imprescindibles para haber llegado a este momento.

Resúmen

Frente a cambios en el ambiente, las bacterias ajustan los niveles de expresión de sus genes para producir una respuesta apropiada. Los niveles individuales de este proceso han sido estudiados ampliamente: la red de regulación transcripcional describe las interacciones reguladoras que promueven cambios en la red metabólica, ambas coordinadas por la red de señalización. Sin embargo, la interacciones entre estos niveles nunca han sido descritas de forma sistemática. En el presente trabajo se formalizó el proceso de detección y procesamiento de señales ambientales en un concepto llamado *Genetic Sensory Response Unit* o GENSOR Unit, el cual está compuesto por cuatro componentes: (1) la señal detectada, (2) la transducción de la señal, (3) *switch* genetico y (4) la respuesta generada. Se utilizaron datasets validados experimentalmente de dos bases de datos para construir una GENSOR Unit para cada uno de los 189 factores de transcripción (TFs) locales depositados en RegulonDB. Análisis subsecuentes sugieren que la presencia de circuitos de retroalimentación es una propiedad general de la dinámica controlada por TFs individuales y que existe un gradiente de complejidad en su respuesta, contrastando con la noción paradigmática de 1 TF/1 vía metabólica. Se cuantificaron otras propiedades de las GENSOR Units y se utilizó su topología para realizar predicciones de efectores y genes blanco. Finalmente, se identificaron similitudes entre GENSOR Units para organizarlas en grupos que maximizan la cantidad de elementos compartidos, y describir interacciones de regulación indirectas entre TFs.

Índice general

Índice de figuras

Índice de cuadros

Abreviaturas

E. coli: *Escherichia coli K-12*
DNA: Ácido desoxirribonucléico
A: Adenina
C: Citosina
G: Guanina
T: Timina
U: Uracilo
RNA: Ácido ribonucléico
mRNA: ARN mensajero
TF: Factor de Transcripción
TSS: Sitio de inicio de la transcripción
TRN: Red de Regulación Transcripcional
TU: Unidad Transcripcional
GENSOR Unit: Unidad Genética de Sensor-Respuesta (Genetic Sensory Response Unit)
GO: Gene Ontology

Capítulo 1

Introducción

1.1. Modelo de estudio: *Escherichia coli K-12*

Escherichia coli K-12 es una bacteria ubicua que se encuentra principalmente en el intestino de organismos de sangre caliente o endotermos. Es una gammaproteobacteria perteneciente a la familia Enterobacteriaceae, al igual que *Salmonella enterica*. Fue descubierta en 1885 por Theodor Von Escherich, dada su facilidad para integrar DNA se utilizó como organismo modelo para técnicas de biología molecular, lo que la ha convertido en el organismo mejor estudiado.

1.2. Dogma Central de la Biología Molecular

Todas las funciones de los organismos vivos se encuentran codificadas en su DNA, el **dogma central de la biología molecular** explica el proceso mediante el cual la información del DNA es transmitida a RNA y proteína para dar lugar a funciones celulares. El **DNA** o ácido desoxirribonucéico es una molécula conformada por azúcar (desoxiribosa), fosfato y una de 4 bases nitrogenadas: **adenina** (A), **guanina** (G), **citosina** (C) o **timina** (T). En una célula, millones de moléculas individuales de DNA se unen para dar lugar a una secuencia de A's,

C's, G's y Ts. Los **genes**, la unidad de la herencia, son secuencias finitas de DNA que pueden verse como una sucesión de A's, C's, G's y T's .

El DNA puede ser pensado como un libro donde las palabras están conformadas por un alfabeto de 4 letras; aunque toda la información se encuentra ahí, el DNA no es suficiente para crear funciones, es necesario que sea leído por una proteína llamada **RNA polimerasa**. La polimerasa tiene la capacidad de copiar la secuencia de DNA de un gen a **RNA** o ácido ribonucléico, una molécula muy parecida al DNA que en lugar de desoxiribosa contiene ribosa como azúcar y, en lugar de timina, utiliza uracilo (U). La copia a RNA se da con reglas muy específicas: por cada C que se encuentre en el DNA, la polimerasa agregará una G en el RNA y viceversa; por cada T agregará una A, y por cada A agregará un uracilo (U). A este proceso de síntesis de RNA a partir de DNA se le llama **transcripción**. Al RNA resultante se le llama **RNA mensajero** (mRNA).

En bacterias, la RNA polimerasa es capaz de transcribir varios genes continuos a la vez. Al grupo de genes que es transcrito en la misma cadena de RNA mensajero se le conoce como **Unidad Transcripcional** (TU). Por ejemplo, si los genes *araB, araC y araD* son transcritos al mismo tiempo, el RNA mensajero incluirá a los tres genes y la unidad transcripcional se llamará *araBAD*.

Una vez que la transcripción ha concluido y una nueva molécula de mRNA ha sido sintetizada, estructuras protéicas llamadas **ribosomas** se unen al mRNA para interpretarlo y sintetizar una cadena de aminoácidos que, tras plegarse, se convertirá en una proteína funcional. Las células utilizan 20 aminoácidos distintos para formar proteínas. Los ribosomas se encargan de leer las bases (A's, C's, G's y T's) en el mRNA en grupos de tres y, por cada triplete, agregan un aminoácido específico. Por ejemplo, la secuencia CCG en un mRNA provocará la adición del aminoácido prolina a la cadena naciente de aminoácidos. También existen tres señales de paro (UAA, UAG y UGA) para comunicar al ribosoma que el gen ha terminado y la cadena de aminoácidos está lista para doblarse de una forma específica y ejecutar su función. Al proceso de síntesis de una cadena de aminoácidos a partir del mR-

NA se le llama **traducción**.

Las proteínas son las moléculas que hacen funcional a la célula, cada una tiene una función específica, por ejemplo, catalizar reacciones químicas para convertir alimento en energía, crear estructuras físicas que le dan formas características a las células, transportar a otras proteínas, señalizar cambios en el ambiente, etc. Existen muchos tipos y funciones protéicas, todos determinados por la secuencia de DNA de donde surgieron.

1.3. Regulación Transcripcional

El ambiente externo e interno de las células está en constante cambio. Muchas de sus funciones sólo son necesarias bajo ciertas condiciones, por ejemplo, producir las proteínas necesarias para sintetizar aminoácidos sólo será necesario cuando dichos aminoácidos no se encuentren en el medio. Producir constantemente todas las proteínas codificadas en el DNA, incluso en momentos en que no son necesarias, representaría un gasto energético muy grande que podría comprometer la capacidad de las células para crecer y reproducirse. Es por esto que las células son capaces de regular la expresión de grupos de genes según la condición en que se encuentren. Una de las forma de regulación más estudiada es aquella que se encarga de promover o inhibir la transcripción de los genes, llamada **regulación transcripcional**.

1.3.1. Factores de Transcripción

La forma más común de regulación transcripcional en bacterias está a cargo de proteínas llamadas **Factores de Transcripción** (TFs). Los TFs son capaces de unirse a moleculas señalizadoras llamadas **efectores** y, en consecuencia, alterar los niveles de transcripción de un grupo de genes. La concentración de los efectores de un TF cambia de acuerdo al ambiente en que la célula se encuentra, cuando la concentración alcanza cierto nivel, el efector y el TF se unen para formar un complejo TF-efector. Según la presencia del efector, las conformaciones de los TFs se clasifican en:

- **Conformación *holo***. Conformación en la cual el TF se encuentra en complejo con su efector. Por ejemplo, LacI-alolactosa.
- **Conformación *apo***. Conformación que adopta el TF en ausencia de su efector. Por ejemplo, LacI.

Además de unirse a metabolitos, los TFs pueden unirse a secuencias específicas de DNA llamadas **sitios de pegado**. Los sitios de pegado se encuentran en regiones cercanas al sitio de inicio de la transcripción (TSS), llamadas **promotores**. Al unirse al DNA, los TFs son capaces de promover o inhibir la transcripción de los genes subsecuentes (Figura 1.1).

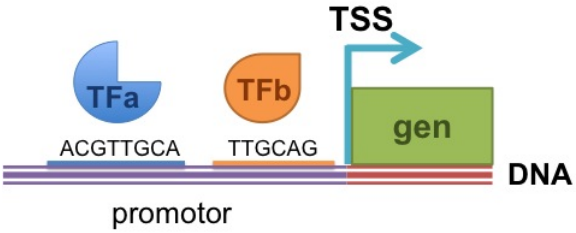


Figura 1.1: Elementos de la regulación transcripcional. Incluyen al DNA (en rojo), la secuencia del gen (en verde), el sitio de inicio de la transcripción (TSS en flecha), la región promotora (región del DNA en morado) y los sitios de pegado para TFs específicos (regiones en azul y naranja).

De acuerdo a su actividad reguladora, los TFs pueden ser:

- **Activadores**. TFs que promueven la transcripción de sus genes regulados.
- **Represores**. TFs que impiden la transcripción de sus genes regulados.
- **Duales**. TFs que son capaces de promover o impedir la transcripción de sus genes regulados, según la presencia de señales ambientales .

En algunos casos, la conformación *holo* promueve la actividad reguladora del TF. Por ejemplo, el TF TrpR es un represor, pero sólo en presencia de su efector, triptófano, y una vez que se haya formado

el complejo TrpR-triptófano, se podrá observar una reducción en la expresión de los genes que regula. En otros casos, es la conformación *apo* la responsable del efecto regulador, por ejemplo LacI: la presencia de su efector, alolactosa, y la formación del complejo LacI-alolactosa inhibe el efecto represor del TF, sólo se se puede observar una reducción en la expresión de los genes regulados por LacI en ausencia de alolactosa. Dada esta posible combinatoria, las conformaciones de los TFs también pueden clasificarse en:

- **Conformación Activa**. Aquella en la cual el TF es capaz de llevar a cabo su actividad reguladora. En el ejemplo anterior, TrpR-triptófano o LacI.
- **Conformación Inactiva**. Aquella que impide que el TF lleve a cabo su actividad reguladora. Por ejemplo, TrpR o LacI-alolactosa.

Las clasificaciones que se han enumerado hasta ahora pueden combinarse para dar lugar a 6 mecanismos de regulación distintos. Como puede observarse en el cuadro 1.1, mecanismos distintos pueden provocar el mismo efecto sobre los genes, por ejemplo, tanto un TF activador cuya conformación activa sea en *apo* como un represor cuya conformación activa sea en *holo* impedirán que sus genes regulados sean transcritos en presencia del efector. Es por ello que, desde una perspectiva más global, los sistemas que están siendo regulados por un TF se clasifican en:

- **Sistemas Inducibles**. La presencia de una molécula específica provoca un aumento en los niveles de transcripción de un grupo de genes. Por ejemplo, TFs represores con conformación activa *apo* o TFs activadores con conformación activa *holo*, en presencia de su efector. En estos casos se dice que los genes son inducidos.
- **Sistemas Reprimibles**. La presencia de una molécula específica provoca una disminución en los niveles de transcripción de un grupo de genes. Por ejemplo, TFs represores con conformación activa *holo* o TFs activadores con conformación activa *apo*, en presencia de su efector. En estos casos se dice que los genes son reprimidos.

Efecto Regulador	Conformación Activa	Resultado en presencia del efector
Activador	*holo*	Genes expresados
Activador	*apo*	Genes no expresados
Represor	*holo*	Genes no expresados
Represor	*apo*	Genes expresados
Dual	*holo*	Genes expresados y no expresados
Dual	*apo*	Genes no expresados y expresados

Cuadro 1.1: Mecanismos de Regulación de un TF

1.3.2. Regulones

Cada TF regula directamente a los genes que contienen un sitio de pegado afín a él en su promotor. Al grupo de genes regulado por el mismo TF se le llama **regulón**. Genes individuales pueden ser regulados por uno o más TFs de forma que, de acuerdo al número de TFs y sitios de pegado considerados existen 4 tipos de regulones:

- **Regulón**. Se refiere a todos los genes regulados por el mismo TF. Todos los genes que pertenecen a un regulón comparten un sitio de pegado del mismo TF. Por ejemplo, el regulón de AraC está compuesto por los genes contenidos en las TUs *araBAD, araFGH, araJ, araE-ygeA, araC, araJ, xylAB* y *ydeNM*. El de XylR incluye a *araC, xylAB* y *xylFGHR*.
- **Regulón simple**. Incluye a todos los genes que sólo son regulados por un TF y no cuentan con sitios de pegado de otros TFs. Por ejemplo, el regulón simple de XylR y AraC no incluye ningún gen pues todos son co-regulados por otros TFs como CRP y Fis.
- **Regulón complejo**. Incluye a todos los genes que son exclusivamente regulados por una combinatoria de TFs. Por ejemplo, el regulón complejo de AraC/XylR/CRP incluye a las TUs *araC* y *xylAB* porque son las únicas TUs en el genoma que cuentan con sitios de pegado sólo para AraC, XylR y CRP.

Las clasificaciones antes mencionadas no consideran el efecto del TF sobre el gen (activador o represor), para esto existen los **regulones estrictos**. Un regulón estricto puede ser un regulón simple o complejo

donde sólo se incluyen aquellos genes que son sujetos al mismo efecto regulador. Por ejemplo, el regulón complejo AraC/XylR/CRP permite 8 posibles regulones complejos estrictos basados en la combinatoria de sus efectos reguladores:

- Regulón complejo estricto AraC(+)/XylR(+)/CRP(+).
- Regulón complejo estricto AraC(+)/XylR(+)/CRP(-).
- Regulón complejo estricto AraC(+)/XylR(-)/CRP(+).
- Regulón complejo estricto AraC(+)/XylR(-)/CRP(-).
- Regulón complejo estricto AraC(-)/XylR(+)/CRP(+).
- Regulón complejo estricto AraC(-)/XylR(+)/CRP(-).
- Regulón complejo estricto AraC(-)/XylR(-)/CRP(+).
- Regulón complejo estricto AraC(-)/XylR(-)/CRP(-).

Los regulones simples permiten dos combinaciones, por ejemplo, AraC(+) y AraC(-). La combinación de todos los tipos de regulones se encuentra ejemplificada en la figura 1.2.

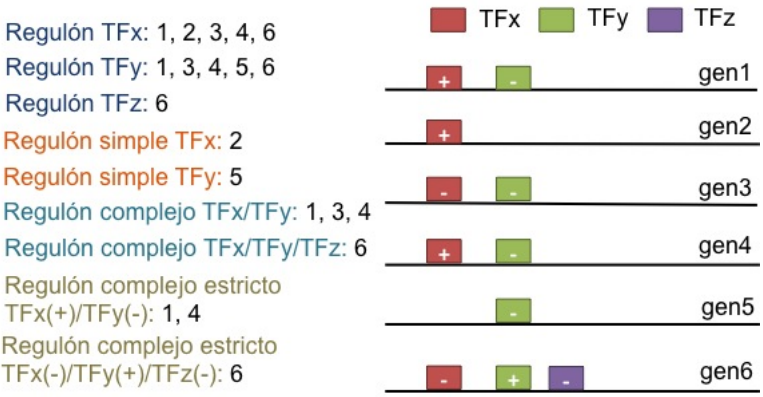


Figura 1.2: Tipos de regulones. En la imagen se muestran las regiones promotoras de 6 genes hipotéticos con sitios de pegado de 3 TFs (cubos rojos, verdes y morados). El efecto de cada TF sobre el gen se muestra en el sitio de pegado. El texto indica algunos tipos de regulones que pueden encontrarse en la imagen.

1.3.3. RegulonDB

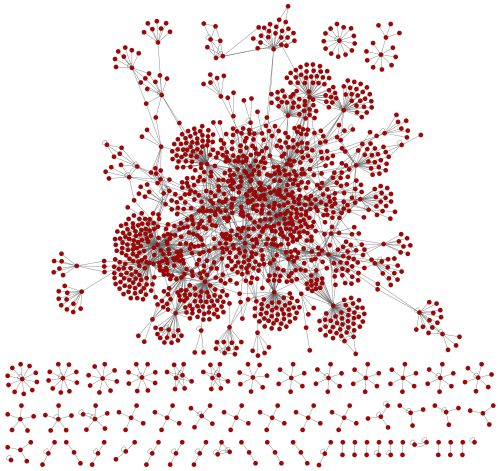
RegulonDB es la base de datos de regulación transcripcional de un solo organismo más completa que existe actualmente [Gama-Castro et al., 2016]. En su versión 9.2 cuenta con 203 TFs y 4219 interacciones de regulación entre TFs y genes de *E. coli*. Todas las interacciones reguladoras han sido extraídas manualmente de la literatura por curadores expertos y están divididas según el peso de la evidencia experimental que se encuentra en la literatura, ya sea evidencias débiles o fuertes. Por ejemplo, un experimento de *footprinting* representa una evidencia fuerte, mientras que cambios observados en la expresión de un gen en una cepa carente del TF es una evidencia débil [Weiss et al., 2013].

1.3.4. Red de Regulación Transcripcional

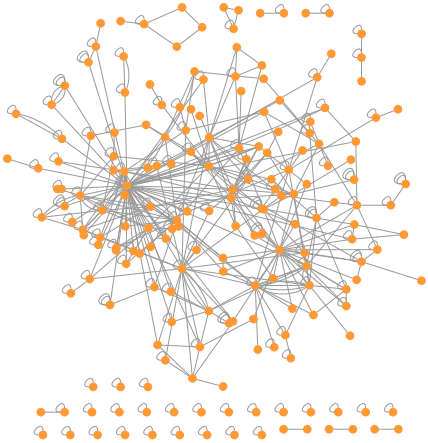
Dado que cada TF regula a un grupo de genes y cada gen puede ser regulado por más de un TF, es posible representar al conjunto de interacciones de regulación en forma de red, donde los nodos representan genes y las aristas representan interacciones de regulación entre el producto de un gen (un TF) y otros genes (Figura 1.3a). A dicha red se le denomina **Red de Regulación Transcripcional** o TRN por sus siglas en inglés. Una simplificación de la TRN es la red de regulación TF-TF donde sólo se incluyen los TFs y las relaciones entre ellos (Figura 1.3b). Ambas redes han sido sujetas a múltiples estudios que han permitido elucidar propiedades biológicas y matemáticas que rigen su estructura y evolución [Barabási and Oltvai, 2004] [Kim et al., 2009] [**?**] [Resendis-Antonio et al., 2005] [Balazsi et al., 2005], por ejemplo, la identificación de estructuras topológicas sobre-representadas que involucran tres nodos y permiten dinámicas resistentes a fluctuaciones en las señales [Shen-Orr et al., 2002] [Mangan and Alon, 2003].

1.4. Metabolismo celular

El metabolismo celular es el conjunto de reacciones que suceden en una célula y que le permiten transformar químicamente unas moléculas en otras. El metabolismo permite a una célula convertir fuentes de carbono en energía o ensamblar estructuras de soporte, como la



(a) TF-genes



(b) TF-TF

Figura 1.3: Red de Regulación Transcripcional. (a) Red de Regulación Transcripcional TF-genes. Los nodos representan genes, las aristas unen genes cuyo producto es un TF, con sus genes regulados. (b) Red de Regulación Transcripcional TF-TF. Los nodos representan TFs, las aristas unen TFs que se regulan directamente de manera transcripcional.

membrana, a partir de los metabolitos que toma del medio. Las formas más comunes de representación del metabolismo celular son las redes metabólicas y las vías metabólicas.

1.4.1. Red Metabólica

La red metabólica se refiere al conjunto de todas las reacciones enzimáticas que suceden en la célula descritas en forma de red, donde los nodos representan metabolitos y las aristas representan reacciones que transforman a un metabolito en otro. Aunque la red describe todos los posibles caminos que existen bioquímicamente entre un par de metabolitos, no hay evidencia de que todos los caminos posibles sucedan fisiológiamente en la célula en algún momento dado. La red metabólica de *E. coli* también ha sido ampliamente estudiada, principalmente para estudios de ingeniería metabólica, análisis topológicos, predicción de fenotipos e inferencias sobre relaciones evolutivas entre especies [McCloskey et al., 2014] [Kreimer et al., 2008] [Ravasz et al., 2002] [Grüning et al., 2010] [Parter et al., 2007] [Orth et al., 2010].

1.4.2. Vías Metabólicas

Las vías metabólicas son grupos de reacciones sucesivas que convierten a un metabolito inicial en otro final, de forma tal que el substrato de una reacción en la vía también funge como producto de la reacción subsecuente. Las vías metabólicas son el subconjunto de reacciones más estudiado de la red metabólica, tanto la direccionalidad como la función de cada vía son conocidas y, en su mayoría, existen experimentos que han validado la ocurrencia de las vías metabólicas en la célula. Las delimitaciones sobre el inicio y el final de una vía metabólica dependen de la fuente que se consulte. Existen diversas bases de datos que se encargan de clasificar a las vías metabólicas como KEGG [Kanehisa et al., 2017] y Ecocyc, cada una cuenta con sus propias reglas para delimitar los límites de las vías y agruparlas en elementos más generales.

1.4.2.1. Ecocyc

Ecocyc es actualmente la base de datos de metabolismo de *E. coli* más completa que existe [Keseler et al., 2017]. Su mantenimiento es res-

ponsabilidad de curadores expertos que recuperan datos directamente de la literatura, por lo que toda la información está respaldado por una referencia al artículo original. Ecocyc hospeda una clasificación del metabolismo en 362 vías metabólicas y 58 *súpervías* metabólicas, las últimas son grupos de vías metabólicas que están relacionadas funcionalmente. Las reglas para definir el principio y fin de las vías metabólicas son [Caspi et al., 2013]:

- Límites históricos definidos por los primeros grupos de investigación que estudiaron el metabolismo bacteriano.
- La aparición de cualquiera de los 13 metabolitos más comunes en la célula.
- Coincidencia con TUs.
- Coincidencia con unidades metabólicas evolutivamente conservadas.

1.5. Programas Genéticos

Francois Jacob y Jacques Monod describieron la importancia de las interacciones entre la regulación y el metabolismo en su descubrimiento del primer TF, LacI [Pardee et al., 1959]. Jacob y Monod descubrieron que LacI se une al promotor del operón *lac* e induce su expresión en presencia de lactosa. Su modelo original de regulación transcripcional también incluyó la función de las enzimas codificadas en el operón *lac*, encargadas de transportar y utilizar lactosa, la señal del sistema. Este modelo permitió explicar cómo hace la célula para administrar eficientemente sus recursos: sólo produce las enzimas necesarias para la condición ambiental en que se encuentra. El modelo de Jacob y Monod [Jacob and Monod, 1961] aún es considerado el paradigma de la regulación transcripcional, a partir de él se asume que los TFs funcionan como circuitos genéticos con una entrada (las señales ambientales que detecta un TF) y una salida (efecto metabólico que produce) (Figura 1.4), así como la suposición de que los genes regulados por el mismo TF responden ante la misma señal y, por lo tanto, participan en el mismo proceso biológico.

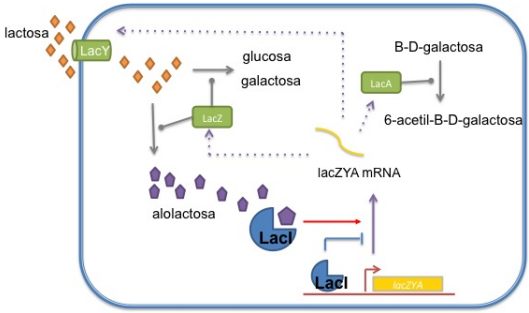


Figura 1.4: Programa genético de utilización de lactosa. La lactosa presente en el exterior de la célula es transportada al interior donde se convierte en alolactosa. La alolactosa se une al TF LacI para impedir que éste siga reprimiendo al operón *lac*. Los genes del operón *lac*, *lacA, lacZ y lacY* son transcritos y las proteínas resultantes son las encargadas de transportar y procesar la lactosa.

1.5.1. Gene Ontologies

La forma más usada de clasificar a los procesos biológicos que lleva a cabo la célula es a través de una ontología llamada **Gene Ontology** (GO). Esta ontología fue desarrollada en el año 2000 por el *Gene Ontology Consortium* como una herramienta para anotar genomas recién secuenciados de forma uniforme, rápida y eficiente [Consortium, 2000]. La GO es un vocabulario controlado y organizado jerárquicamente que busca describir, sin ambigüedad, tres propiedades de cualquier gen: proceso biológico, función molecular y compartimiento celular. La organización encargada de mantener la GO cura de manera periódica los términos asociados a cada gen, recibe aportaciones de la comunidad sobre nuevas evidencias para agregar términos a un gen y mantiene toda la información disponible en la web.

En *E. coli*, la GO cuenta con anotaciones de procesos biológicos para 3463 genes, 77 % de los genes conocidos en este organismo. Originalmente, la GO buscaba proveer información sobre genes conocidos para poder extrapolar esta información a genes ortólogos recién secuenciados. Con el tiempo, la GO se utilizó cada vez más para identificar funciones en grupos de genes. Hoy en día, con el auge de las

tecnologías masivas de análisis de expresión génica, las GOs son una herramienta muy común para encontrar sentido biológico a los genes que resultan diferencialmente expresados. Dado un grupo de genes, se recuperan todos los términos de GO asociados a cada uno de ellos y, mediante una prueba estadística, se identifican los términos sobre-representados, lo que hace posible realizar inferencias sobre la función biológica en que participan juntos los genes [Yon Rhee et al., 2008].

Capítulo 2

Objetivos

2.1. Planteamiento del Problema

La biología de sistemas ha logrado grandes avances en el estudio global de la TRN [Barabási and Oltvai, 2004] [**?**] [Kim et al., 2009] [Resendis-Antonio et al., 2005] [Balazsi et al., 2005], la red metabólica [McCloskey et al., 2014] [Kreimer et al., 2008] [Ravasz et al., 2002] [Grüning et al., 2010] [Parter et al., 2007] [Orth et al., 2010] y la red de transducción de señales [Papin and Palsson, 2004] [Papin et al., 2005]. Sin embargo, estos estudios se limitan a un solo nivel de organización celular y son difíciles de integrar. Existen redes pequeñas y detalladas [Alon et al., 1999] [Berthoumieux et al., 2014] mientras que otras abarcan toda la célula con poco detalle [Karr et al., 2012] [Brooks et al., 2014]. Para aumentar nuestro entendimiento de cómo la célula detecta y procesa la información de su ambiente, es necesario analizar distintos niveles de organización celular simultáneamente. La regulación es central a estos procesos de señal-respuesta, en particular los TFs, pues detectan cambios en la concentración de su efector y, en respuesta, promueven la expresión de los genes necesarios para contender con dichos cambios.

Hoy en día existe en la literatura una gran cantidad de datos que, de ser integrados de forma sistemática, podrían permitir rastrear el flujo de información a través de distintos niveles celulares y elucidar principios biológicos hasta ahora desconocidos. Una de las limitantes para

lograr este objetivo es la ausencia de formalismos que permitan dicha integración. Uno de los conceptos más utilizados para ligar distintos niveles de organización celular, en específico genes con procesos biológicos, son las Gene Ontologies [Consortium, 2000]. Sin embargo, este formalismo se centra en las propiedades individuales de los genes y no considera las interacciones entre ellos para llevar a cabo un proceso.

Es necesario generar nuevos formalismos que permitan integrar datos de forma estandarizada, en pequeña y larga escala, considerando distintos niveles de organización celular. Una nueva perspectiva con estas propiedades permitiría evaluar la validez de los paradigmas existentes y elucidar nuevos principios biológicos.

2.2. Antecedentes Directos

En la versión 7.0 de RegulonDB [Gama-Castro et al., 2011] se propuso un concepto denominado *Genetic Sensory Response Unit* o GENSOR Unit. Este concepto buscaba representar el contexto biológico en que sucede la regulación mediada por un TF. Las GENSOR Units fueron brevemente definidas como una unidad fisiológica que incluye 4 componentes:

i. Señal.
ii. Serie de reacciones concatenadas que producen un efector
iii. Cambio en la expresión de un grupo de genes
iv. Respuesta

Bajo este esquema se construyeron 25 GENSOR Units de 25 TFs relacionados con utilización de fuentes de carbono y síntesis de aminoácidos. Cada una de las GENSOR Units fue construida de forma manual por la curadora Socorro Gama consultando directamente la literatura y diversas bases de datos como RegulonDB, Ecocyc y KEGG.

2.3. Objetivo General

Elucidar de forma sistemática y estandarizada las unidades de detección y procesamiento de información mediadas por los factores de

transcripción de *Escherichia coli* K-12 y describir sus propiedades.

2.4. Objetivos Particulares

i Estandarizar el concepto de GENSOR Unit.

ii Diseñar y desarrollar un método semi-automático de ensamblado de GENSOR Units.

iii Analizar las propiedades del set de GENSOR Units.

iv Utilizar la topología de las GENSOR Units para predecir elementos faltantes.

v Identificar relaciones entre GENSOR Units.

Capítulo 3

Resultados

3.1. Objetivo i: Estandarizar el concepto de GENSOR Unit.

3.1.1. Componentes de una GENSOR Unit

Las 25 GENSOR Units depositadas en la versión 7.0 de RegulonDB cumplen con el objetivo de describir el flujo de información a través de la maquinaria celular, sin embargo, sus elementos no fueron definidos de manera formal. El primer paso fue definir los 4 componentes de una GENSOR Unit de forma que el concepto pueda ser aplicado a culquier regulador transcripcional y no sólo TFs, por ejemplo, RNAs pequeños o factores sigma (Figura 3.1). Los cuatro componentes se definieron de la siguiente forma:

i **Señal**. Molécula que señaliza un cambio en el ambiente interno o externo de la célula. Comienza la cascada de eventos que culmina en una respuesta que permite a la célula adaptarse ente dicho cambio. Por ejemplo, la presencia de homoserin-lactonas que promueve la formación de biofilms.

ii **Transducción de la señal**. Conversión de la señal en un metabolito que promueve directamente un cambio de estado en el regulador. En el caso de los TFs se refiere a la conversión de la señal al efector. Por ejemplo, la señal lactosa se convierte en alolactosa, el efector de LacI (Figura 1.4).

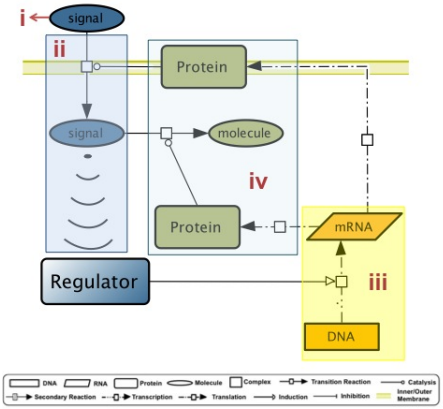


Figura 3.1: Diagrama general de una GENSOR Unit. En azul se muestra la señal (i) y su transducción (ii), que provocan un en el regulador. En amarillo (iii) el *switch* genético que altera la expresión de los genes regulados. En verde (iv) la respuesta producida por los productos génicos.

iii ***Switch* Genético**. Cambio en la expresión de los genes directamente regulados por el regulador principal de la GENSOR Unit. Por ejemplo, la activación de los genes transcritos por la polimerasa en complejo con el factor sigma 70.

iv **Respuesta**. Efecto metabólico y funcional que producen los productos de los genes regulados directamente por el regulador al sufrir un cambio en su expresión. Por ejemplo, la utilización de lactosa como fuente de carbono tras la activación del operón *lac*.

3.1.2. Objetos que conforman una GENSOR Unit

Los componentes mencionados en la sección anterior pueden entenderse como "roles" de los objetos presentes en la GENSOR Unit y sus interacciones. Para poder identificarlos es necesario reunir de antemano los objetos que incluye una GENSOR Unit. En la versión 7.0 de RegulonDB las GENSOR Units se construyeron alrededor del TF y el proceso regulado más conocido, a partir de estos datos se buscaron en bases de datos y la literatura todos los objetos que permitieran relacionar al TF con su proceso biológico regulado. En la nueva versión se buscó recopilar los objetos de cada GENSOR Unit de forma no-heurística, es decir, sin hacer suposiciones *a priori* sobre el efecto del TF en el metabolismo celular. Por el contrario, se buscó que las mismas interacciones entre los objetos reflejaran naturalmente el pro-

ceso biológico que el TF regula.

El nuevo método de ensamblado de GENSOR Units comienza con un regulador y sus unidades transcripcionales (TUs) directamente reguladas. Posteriormente se agregan los siguientes objetos y sus interacciones:

- Efectores conocidos del TF.
- Conformaciones activas e inactivas del TF.
- Efecto del TF sobre los genes regulados.
- Productos de los genes regulados.
- Complejos protéicos heteromultiméricos donde participan los productos de los genes.
- Reacciones catalizadas por los productos que tengan actividad enzimática.
- Substratos y productos de las reacciones catalizadas.

Este grupo de objetos (TUs, RNAs, proteínas, complejos y metabolitos) más sus interacciones (reacciones, catálisis de reacciones, formación de complejos, inhibición/represión de las reacciones) forman una red multinivel que puede ser representada de forma gráfica permitiendo una perspectiva global del efecto del TF en la célula. Sobre esta red se identifican los 4 componentes de cada GENSOR Unit.

3.1.3. Reacciones Secundarias

El concepto de GENSOR Unit se basa en el paradigma de Jacob y Monod que establece que cada TF regula al grupo de genes necesario para dar lugar a una capacidad celular, tales como utilización de lactosa como fuente de carbono, producción de osmoprotectores o ensamblado de flagelo. A partir de esta suposición es posible asumir que los objetos incluidos en la GENSOR Unit, que parten de un TF y sus genes regulados, son suficientes para describir los 4 componentes (señal. transducción de la señal, *switch* genético y respuesta). Sin embargo, existen cuatro escenarios que podrían generar excepciones a esta suposición:

- **Genes constitutivos**. Algunos genes son necesarios para mantener las funciones esenciales de la célula, por lo que se requiere

que estén siendo expresados de forma constante y es común que no sean regulados por TFs. Estos genes no aparecerán entre los objetos de una GENSOR Unit pero podrían tener un rol en el proceso biológico regulado por el TF.

- **Genes con función desconocida**. Sólo el 76 % de los genes en E. coli tienen una función bioquímica conocida [Karp et al., 2007]. El 24 % de los genes restantes podrían estar involucrados en el flujo de información de una GENSOR Unit, sin embargo, la falta de anotación funcional impediría que su participación se viera reflejada.
- **Interacciones reguladoras desconocidas**. Muchas interacciones TF-gen de la TRN son aún desconocidas. Estos huecos en nuestro conocimiento pueden provocar GENSOR Units incompletas donde la falta de un gen regulado podría sesgar la identificación de los 4 componentes. Por ejemplo, la falta de un gen involucrado en el transporte del efector a través de la membrana haría parecer que la señal de la GENSOR Unit es interna, cuando en realidad es externa.
- **Cooperación entre TFs**. Varios TFs pueden cooperar para regular un mismo proceso biológico [Jain and Saini, 2016]. Podría suceder que un TF regulara un subset de genes involucrados en el proceso compartido, mientras que otro TF regulara al resto de los genes, de forma que en las GENSOR Units individuales el proceso biológico aparecería incompleto.

Para compensar estas eventualidades se consideraron las vías metabólicas canónicas descritas en la base de datos Ecocyc. En cada GENSOR Units se identificaron pares de reacciones que pertenecieran a la misma vía metabólica. Si la direccionalidad de la vía permitía un flujo metabólico entre el par de reacciones, dichas reacciones se conectaron a través de una tercera reacción denominada **reacción secundaria**. Una sola reacción secundaria puede representar una o más reacciones intermedias en la vía metabólica que conecta al par de reacciones originales que pertenecen a la GENSOR Unit. Por ejemplo, el TF TrpR regula a 7 enzimas de las trece involucradas en la producción de triptófano a partir de D-eritrosa-4P (Figura 3.2a). Si se consideran sólo los genes regulados por TrpR, pareciera que las reacciones reguladas por AroH y AroL no están relacionadas, sin embargo, existen dos vías metabólicas (Figura 3.2a flechas moradas y rojas) que permiten

un flujo metabólico que convierte 3-desoxi-D-arabino-heptulosonato-7P en shikimato y, posteriormente, shikimato-3P en corismato. En la GENSOR Unit de TrpR, las reacciones catalizadas por AroB, AroD y AroE pueden ser fusionadas en una sola reacción secundaria para indicar la existencia de un flujo metabólico conocido (Figura 3.2b, líneas punteadas). Lo mismo aplica para las reacciones catalizadas por AroA y AroC. Es importante hacer énfasis en que 2 o más reacciones de la vía metabólica pueden ser fusionadas en una sola reacción secundaria en la GENSOR Unit. Las reacciones secundarias no agregan nuevos objetos a la GENSOR Unit, pues no se incluyen las enzimas que las catalizan, ni los metabolitos intermedios. Las reacciones secundarias pueden entenderse como reacciones que no son un efecto directo de la acción del TF, pero que se sabe que existen y representan interacciones posibles entre los metabolitos presentes en la GENSOR Unit. Estas reacciones permiten un mejor entendimiento del flujo de información presente en las GENSOR Units.

3.2. Objetivo ii: Diseñar y desarrollar un método semi-automático de ensamblado de GENSOR Units.

3.2.1. Herramienta de construcción de GENSOR Units

Entre las metas principales de la actualización del concepto de GENSOR Unit se encontraba eliminar sesgos de curación hacia lo ya conocido sobre el TF. Para realmente ensamblar GENSOR Units que fueran producto de una búsqueda exhaustiva de elementos asociados al TF se generó un algoritmo que automáticamente recupera, de forma secuencial, todos los objetos necesarios. Los objetos relacionados a la regulación transcripcional (TFs, genes regulados, conformaciones activas/inactivas y efectores) se obtuvieron de la base de datos RegulonDB, mientras que los objetos relacionados al metabolismo (productos de los genes, enzimas, reacciones, substratos/productos, complejos protéicos) se obtuvieron de la base de datos Ecocyc. Ambas bases de datos son las más completas en su ramo y cuentan con referencias para cada dato, por lo que cada objeto e interacción de las GENSOR Units están asociados a una referencia a la literatura.

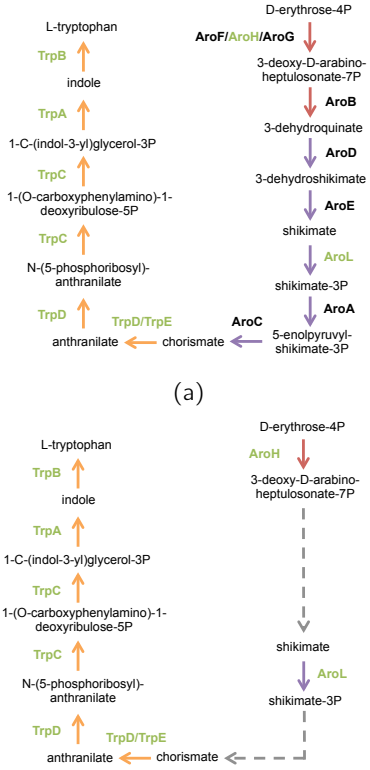


Figura 3.2: Reacciones Secundarias. (a) Vías metabólicas involucradas en la síntesis de triptófano a partir de D-eritrosa-4P. La vía de biosíntesis de 3-dehidroquinato se muestra en rojo; la vía de biosíntesis de corismato a partir de 3-dehidroquinato se muestra en morado; la vía de biosíntesis de triptófano se muestra en naranja. Enzimas que catalizan cada reacción se muestran a un lado de las reacciones. Las enzimas reguladas por TrpR se muestran en verde. (b) En la GENSOR Unit de TrpR las tres reacciones catalizadas por AroB, AroD y AroE, respectivamente, son representadas como una sola reacción secundaria, sin incluir a la enzimas que las catalizan o los metabolitos intermedios.

Para homogeneizar los nombres de objetos de las bases de datos fue necesario realizar un diccionario de sinónimos. Se obtuvieron todos los nombres de proteínas y metabolitos contenidas en ambas bases de datos, se identificaron los nombres usados como sinónimos y se seleccionó el nombre más representativo como nombre estándar en las GENSOR Units. En caso de que no existiera un nombre representativo se utilizó el más corto. Este paso fue muy importante pues nombres duplicados podrían haber sesgado cualquier análisis subsecuente. En el ejemplo más simple, una GENSOR Unit que incluya los metabolitos L-glutamato-5-semialdehído, semialdehído-glutámico, L-glutamato-gamma-semialdehído y glutamato semialdehído podría pensarse que contiene 4 metabolitos distintos cuando en realidad todos estos nombres son sinónimos de glutamato semialdehído. Cada consulta a las bases de datos incluyó un paso adicional de búsqueda en el diccionario de sinónimos para obtener el nombre estándar de los objetos encontrados.

El algoritmo de ensamblado de GENSOR Units comienza eligiendo un TF de RegulonDB y buscando sus efectores conocidos en la misma base de datos; en caso de que existan, recupera sus conformaciones activas e inactivas. Posteriormente busca los genes regulados y el efecto regulador (activación/represión/dual). A continuación se conecta a Ecocyc y recupera los productos de los genes regulados, en caso de que sean enzimas obtiene las reacciones que catalizan, así como los reactivos y productos de las reacciones y su direccionalidad (reversibles/izquierda a derecha). Finalmente, evalúa si los productos de los genes pertenecen a un complejo protéico heteromultimérico, en caso de que sí, recupera todos los monómeros de complejo (Figura 3.3).

Estos elementos se organizan en 5 tablas relacionales que posteriormente permiten generar un mapa visual de la red (Figura 3.4). Las TUs y los mRNAs automáticamente se consideran el componente de *switch* genético de la GENSOR Unit (Figura 3.4, elementos en amarillo). Los efectores conocidos se consideran la señal por default y el resto de los elementos forman parte de la respuesta (Figura 3.4, elementos en verde). Una vez que se obtiene el mapa visual, las GENSOR Units se validan de forma manual y se identifican los casos donde exis-

Figura 3.3: Algoritmo de ensamblado de GENSOR Units

ta un flujo metabólico que lleva al efector. En la figura 3.4 se puede observar que el efector, colina, es transportado a través de la membrana, por lo que se considera que los metabolitos participantes en esta reacción son parte del componente de transducción de la señal (Figura 3.4, elementos en azul).

Utilizando este algoritmo se ensambló una GENSOR Unit para cada uno de los 189 TFs locales depositados en RegulonDB. Los TFs globales [Martínez-Antonio and Collado-Vides, 2003] (ArcA, CRP, Fis, FNR, HNS, IHF y Lrp) no se consideraron pues, por definición, están involucrados en más de un proceso biológico y se sabe que parte de su función es coordinar a los TFs locales. Estas propiedades los hacen excepciones conocidas al paradigma de Jacob y Monod. Ciento cuarenta y cuatro reacciones secundarias fueron agregadas a 48 GENSOR Units. El número de reacciones intermedias en cada reacción secundaria tiene un rango de 1 a 9 (Figura 3.5). La mitad de las reacciones secundarias agregadas (50.7 %) sólo incluyen una reacción intermedia.

3.2.2. Evaluación del método de construcción de GENSOR Units

El siguiente paso fue evaluar el método de construcción de GENSOR Units. Dado que no existe un concepto similar, fue imposible comparar contra un set de datos estándar, sin embargo, se utilizaron las 25 GENSOR Units curadas manualmente en la versión 7.0 de Re-

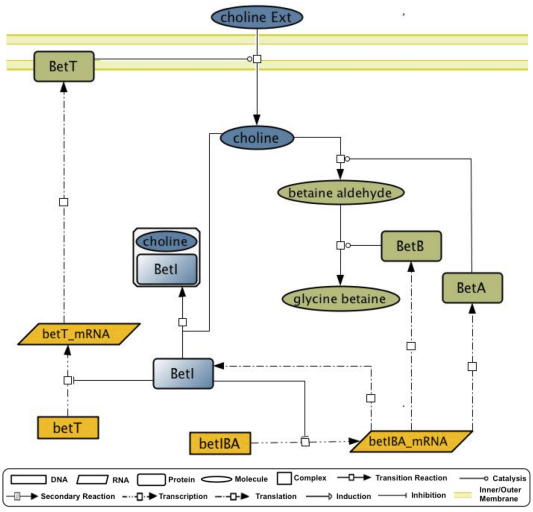


Figura 3.4: Ejemplo de GENSOR Unit. La GENSOR Unit BetI está involucrada en el transporte y utilización de colina. El flujo de información comienza afuera de la membrana: la colina es transportada y, al unirse con BetI, el TF es incapaz de continuar reprimiendo dos TUs, *betT* y *betIBA*, que codifican para tres proteínas encargadas de transportar y convertir la colina en glicina betaína.

gulonDB. Para comprobar que el método semi-automático hace una búsqueda exhaustiva de objetos relacionados con el TF, se cuantificaron las TUs, RNAs, proteínas, complejos protéicos hetermultiméricos, moléculas y reacciones de transcripción, traducción y cambio de estado presentes en las 25 GENSOR Units curadas manualmente y en sus equivalentes del set ensamblado semi-automáticamente. Las reacciones de cambio de estado incluyen reacciones enzimáticas y reacciones de formación de complejos protéicos. Posteriormente se calculó el porcentaje de elementos de cada categoría que fueron recuperados sólo en la curación manual, sólo en la curación automática o a través de ambos métodos (Figura 3.6).

Con excepción de las moléculas y las reacciones de cambio de estado, el porcentaje más alto de todas las categorías corresponde a los objetos recuperados por ambos métodos de curación. El alto porcentaje

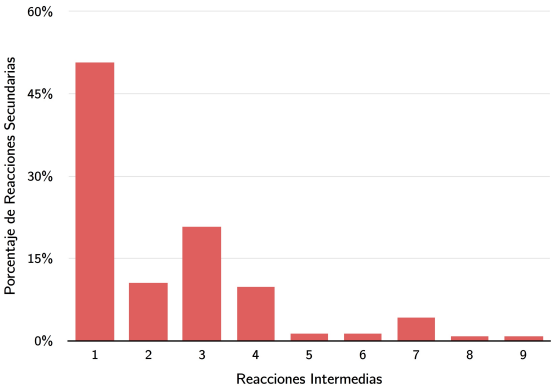


Figura 3.5: Reacciones Secundarias. Distribución de reacciones intermedias en reacciones secundarias; n=144.

de moléculas y reacciones de cambio de estado recuperadas únicamente por curación automática refleja un incremento en el nivel de detalle de las GENSOR Units. Como se mencionó anteriormente, las GENSOR Units curadas manualmente tuvieron como objetivo describir los procesos biológicos conocidos en que el TF participa, por lo que muchas reacciones fueron omitidas en pos de claridad, por ejemplo, en la GENSOR Unit de TrpR sólo se indicaba que las proteínas TrpA, TrpB, TrpC, TrpD y TrpE participaban en la vía de síntesis de triptófano, pero no se detallaban los pasos intermedios de dicha vía (Figura 3.7a). Por el contrario, en la GENSOR Unit construida automáticamente pueden observarse todos los metabolitos involucrados en cada paso intermedio de la vía (Figura 3.7b), aumentando así el conteo total de metabolitos y reacciones de cambio de estado obtenidos de forma automática. El nivel elevado de detalle amplía la utilidad de las GENSOR Units pues, por ejemplo, es imprescindible para modelar dinámicamente su comportamiento.

En todos los casos, el porcentaje de elementos únicos a la curación automática es mayor que los casos únicos a la curación manual. Este resultado apoya la hipótesis de que el método de curación automática realiza una búsqueda exhaustiva de elementos. Por ejemplo, se sabe que LacI se encarga de regular genes relacionados con el trans-

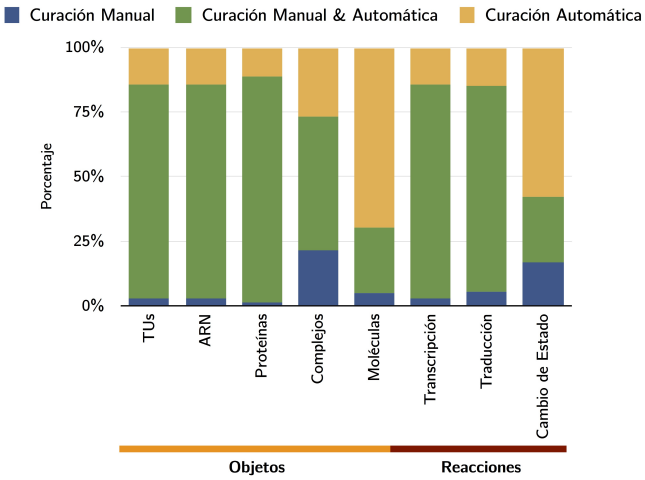
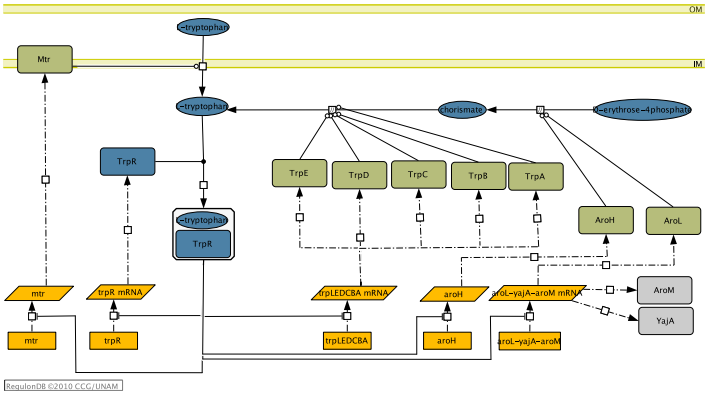
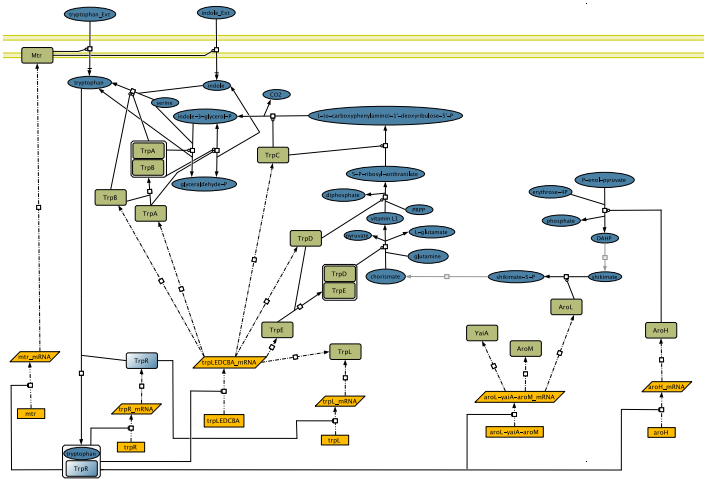


Figura 3.6: Evaluación de desempeño del método semi-automático de curación. Las barras muestran el 100 % del universo de objetos y reacciones incluidos en ambas versiones de GENSOR Units. En amarillo se muestra el porcentaje de objetos/reacciones que sólo fueron recuperados por el método de curación automática, en azul el porcentaje de objetos/reacciones que sólo fueron recuperados por el método de curación manual y en verde el porcentaje de objetos/reacciones que fueron recuperados por ambos métodos. Se evaluaron un total de 527 reacciones y 838 objetos.

porte y utilización de lactosa, en consecuencia, la GENSOR Unit de LacI refleja este efecto fisiológico (Figura 3.8a). Sin embargo, no considera que el transportador de lactosa, LacY, también se encarga de transportar melibiosa, dato que sí existe en la versión curada automáticamente (Figura 3.8b). También se agregaron las reacciones catalizadas por elementos de la GENSOR Unit cuya participación en el proceso biológico conocido no es claro, un ejemplo de esto es LacA, una acetil-transferasa cuyo rol en la utlización de lactosa no es claro. En la GENSOR Unit curada manualmente no se incluye la reacción que cataliza, mientras que en la versión automática sí aparece. Agregar esta información permite eliminar sesgos hacia lo ya conocido sobre el TF, y asociarlo a nuevas funciones que sólo son reflejadas a

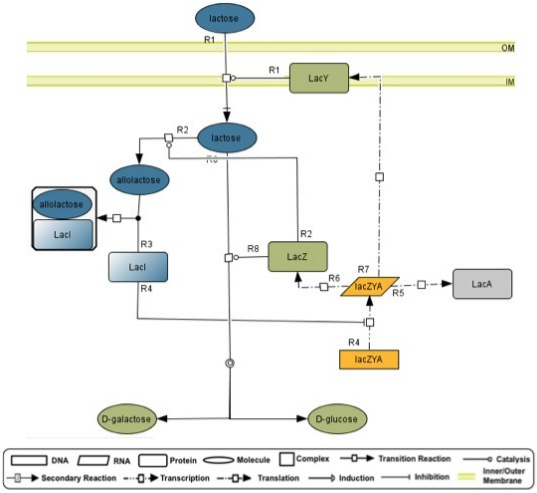


(a) Curación Manual

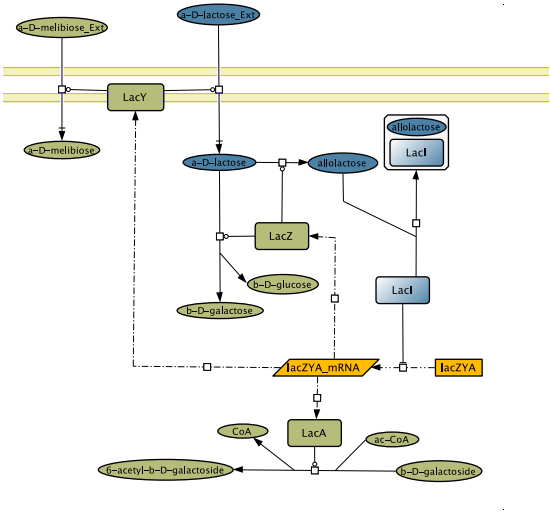


(b) Curación semi-automática

Figura 3.7: Comparación de métodos de ensamblado de GENSOR Units: TrpR. (a) GENSOR Unit de TrpR construida por curación manual. (b) GENSOR Unit de TrpR construida por curación semi-automática. La simbología es igual que en la figura 3.4.



(a) Curación manual



(b) Curación semi-automática

Figura 3.8: Comparación de métodos de ensamblado de GENSOR Units: LacI. (a) GENSOR Unit de LacI construida por curación manual. (b) GENSOR Unit de LacI construida por curación semi-automática. La simbología en (b) es la misma que en (a).

través de la integración.

Por otro lado, existe un porcentaje de elementos exclusivos a la curación manual en todas las categorías cuantificadas. Este porcentaje representa todos los elementos que el método semi-automático no fue capaz de recuperar. Existen dos razones principales por las que esto sucede. En primer lugar, como se mencionó anteriormente, el método de curación manual tenía como objetivo describir un proceso biológico conocido y no existía la restricción de sólo considerar los elementos directamente regulados por el TF. Dado que el método semi-automático se basa en esta restricción, no es capaz de recuperar estos elementos extra. La segunda razón es inherente a la automatización del método, pues obtiene todos los datos de dos bases de datos y su desempeño depende de la información disponible en dichos recursos. En contraste, la curación manual se apoyó de datos extraídos directamente de la literatura cuando las bases de datos no contaban con la información necesaria. Aunque el porcentaje de datos no recuperados existe, es importante resaltar que en 6 de 8 categorías no rebasa el 6 % de objetos totales y en el resto de categorías representa el 22 % y 17 %, un porcentaje menor al de objetos únicamente recuperados por curación automática. Conforme más datos sean agregados a las bases de datos, el porcentaje de datos no recuperados será menor. Una ventaja notable es que la automatización del proceso permite actualizar el set de GENSOR Units con cada liberación de datos.

3.3. Objetivo iii: Analizar las propiedades del set de GENSOR Units.

El set completo de GENSOR Units ensambladas comprende 189, una por cada TF local depositado en RegulonDB. Ochenta GENSOR Units incluyen al menos un efector conocido y, por lo tanto, fue posible identificar sus cuatro componentes. En estos casos, se utilizó la perspectiva global que proveen las GENSOR Unit para redactar un pequeño resumen describiendo el flujo de información observado. En las 109 GENSOR Units restantes sólo fue posible identificar los componentes de *switch* genético y respuesta, lo que refleja el efecto fisiológico

del TF. Diecisiete GENSOR Units no incluyen reacciones enzimáticas, ya sea porque la función bioquímica de los genes regulados es desconocida (BluR) o simplemente no es enzimática (ZraR). El rango de reacciones enzimáticas presentes en el set de GENSOR Units se exitiende de 0 a 129, siendo Fur la más extensa. En la figura 3.9 puede apreciarse la fracción del metabolismo que está contenida en las GENSOR Units. El metabolismo de carbono, lípidos y aminoácidos son los mejor representados.

3.3.1. Retroalimentación

El modelo de regulación transcripcional propuesto por Jacob y Monod incluye la capacidad de un TF de modular su propia actividad en respuesta a las necesidades celulares. Esta propiedad es central en sistemas sensibles, pues les permite responder rápidamente ante cambios de la señal. Alrededor del 60 % de los TFs en la TRN regulan su propio promotor, sin embargo, otra forma de retroalimentación radica en regular los genes responsables de utilizar o producir el efector del TF, de forma que el impacto de los genes regulados en el metabolismo tenga un efecto sobre la acción del TF.

Para identificar este tipo de retroalimentación en las GENSOR Units se consideraron las 80 que incluyen efectores conocidos. Se buscaron reacciones que fueran parte de la respuesta y, al mismo tiempo, tuvieran un efecto en la conversión de la señal al efector. Dos GENSOR Units fueron eliminadas dado que no incluyen enzimas y, por lo tanto, no incluyen reacciones enzimáticas. Sesenta y cinco de las 78 GENSOR Units restantes (83 %) cuentan con este tipo de retroalimentación. Los casos más sencillos consisten en transporte del efector a través de la membrana. Es interesante que TFs que cuentan con más de un efector conocido suelen tener un circuito de retroalimentación para cada uno.

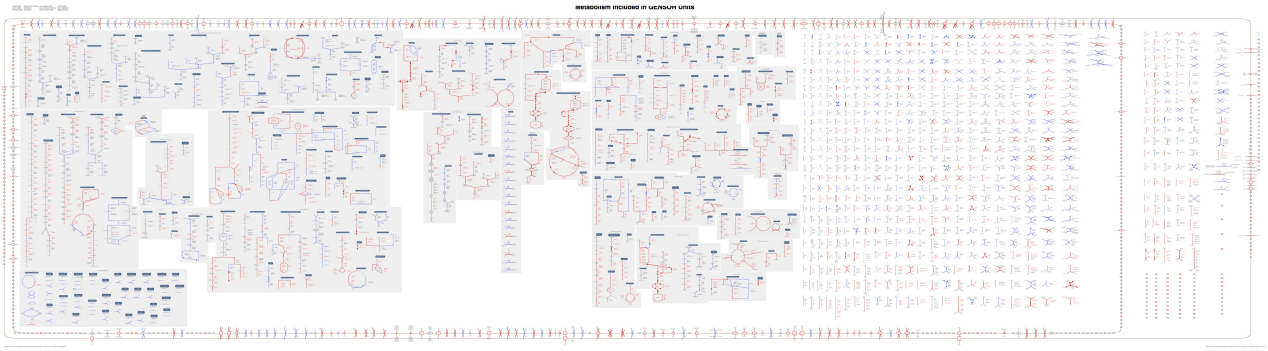


Figura 3.9: Fracción del metabolismo que contienen las GENSOR Units. En rojo se muestran las reacciones enzimáticas que están presentes en al menos una GENSOR Unit. El mapa metabólico fue tomado de Ecocyc, a la izquierda se muestran las vías metabólicas canónicas, a la derecha las reacciones que no pertencen a una vía metabólica. En las orillas del mapa se encuentran las reacciones de transporte y otras que suceden en la membrana.

La GENSOR Unit de AIIS es la única cuyo circuito de retroalimentación hace uso de reacciones secundarias, lo que sugiere que la retroalimentación subyace la capa más básica de toma de decisiones de *E. coli*, aquellas decisiones que dependen de un sólo TF que cambia de estado y promueve una respuesta ante la presencia o ausencia de su señal. Es muy probable que la presencia de circuitos de retroalimentación sea una propiedad general de las GENSOR Units y que al menos un circuito exista en cada GENSOR Unit. Los circuitos del 17 % restante de GENSOR Units podrían no estar basados en flujos metabólicos o usar metabolitos muy comunes en la célula, los cuales no considera el método de identificación usado, por ejemplo, ATP, el efector de DnaA.

3.3.2. Complejidad de la Respuesta

Los TFs son comúnmente llamados “el regulador del metabolismo de X"(por ejemplo, TrpR para triptofano; Laci para lactosa; GalR para galactosa, AraC para arabinosa), lo cual supone que los genes directamente regulados por un TF son necesarios y suficientes para llevar a cabo una capacidad celular. Para explorar esta hipótesis se utilizó el set completo de GENSOR Units y se desarrolló una métrica llamada **conectividad**, que explota las interacciones entre elementos de una GENSOR Unit, en lugar de considerar sólo las funciones de los genes individuales. La conectividad toma en cuenta el número de flujos metbaólicos individuales que están presentes en una GENSOR Unit (y que, por lo tanto, son regulados por el mismo TF). Un flujo metabólico es definido como una sucesión consecutiva de reacciones donde el producto de una reacción también actúa como substrato de otra, por ejemplo, como sucede en una vía metabólica. Las enzimas que catalizan reacciones de un flujo metabólico son consideradas “conectadas"porque se asume que enzimas involucradas en el mismo flujo metabólico participan en el mismo proceso biológico. La conectividad se calcula como el cociente de enzimas conectadas y enzimas totales de una GENSOR Unit. Si sucede que todas las enzimas de una GENSOR Unit son necesarias y suficientes para llevar a cabo un proceso, se esperaría que todas las enzimas reguladas estuvieran conectadas entre ellas formando un solo flujo metabólico, por esta razón el cálcu-

lo de conectividad penaliza flujos metabólicos extra. En resúmen, la conectividad se calcula a través de la siguiente fórmula:

$$C = \frac{Ec}{Et + (Mft - 1)}$$

Donde:
Ec = Enzimas conectadas
Et = Enzimas totales
Mft = Total de flujos metabólicos

La conectividad se mide en valores de 0 a 1. Un valor de 1 indica una GENSOR Unit paradigmática donde todas las enzimas están conectadas e involucradas en el mismo flujo metabólico, por ejemplo, una GENSOR Unit cuyo TF regula una vía metabólica completa. Por otro lado, un valor de 0 refleja una topología desconectada donde cada enzima cataliza una reacción desconectada del resto de la GENSOR Unit.

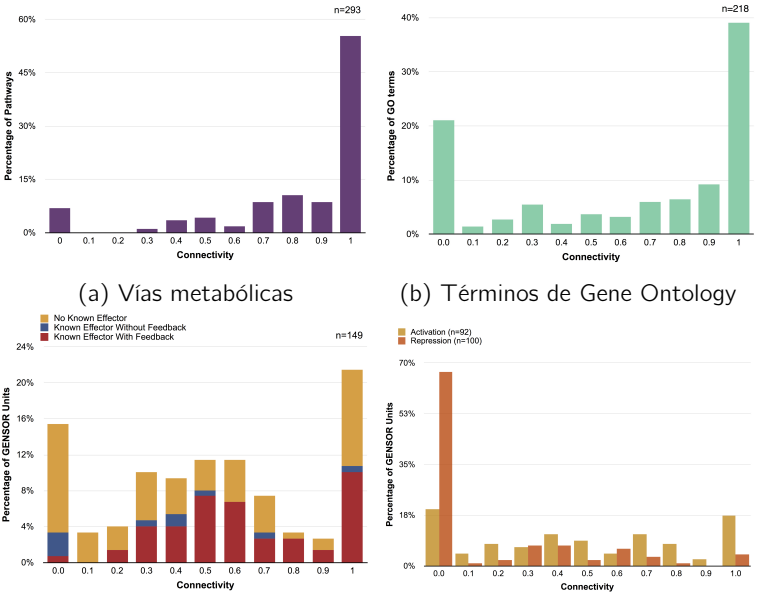
Para validar la relevancia biológica de la métrica, se calculó la conectividad de 293 vías metabólicas canónicas disponibles en Ecocyc (Figura 3.10a) y 218 términos de GO de procesos biológicos (Figura 3.10b). Ambas distribuciones de conectividad muestran una tendencia hacia valores de 1, en particular las vías metabólicas: 55 % de ellas cuenta con conectividad de 1 y 84 % cuenta con valores de 0.7 o más. Es importante notar que los términos de GO cuentan con una estructura jerárquica, lo que sugiere que los términos más generales incluyen genes menos relacionados fisiológicamente, por ejemplo, el término "transporte" incluye a muchos genes que no crean flujos metbólicos y, por lo tanto, obtendrán un valor muy bajo de conectividad. Esta propiedad de la GO explica el incremento en valores de conectividad bajos respecto a las vías metabólicas, sin embargo, el 60 % de los términos mantienen un valor igual o mayor a 0.7. En el caso de las vías metabólicas, los valores de 0 (7 %) se deben a vías metabólicas que no dependen de flujos lineales de reacciones, por ejemplo, en la vía de cargado de tRNAs participan enzimas específicas para cada tRNA cuyas reacciones no crean un flujo metabólico lineal. Estos resultados demuestran que la conectividad es capaz de reflejar una propiedad biológica, no obstante se puede esperar un margen de error del 7 % debido a vías metabólicas cuyas relaciones funcionales no se reflejan

en flujos metabólicos lineales.

Una vez validada la métrica de conectividad, se calculó la conectividad de las 149 GENSOR Units que incluyen dos o más reacciones enzimáticas, las 40 restantes se eliminaron para evitar valores de 0 no informativos. La distribución de conectividad resultante (Figura 3.10c) es significativamente diferente a la observada para las vías metabólicas (prueba Wilcoxon-Mann-Whitney; p-valor = 2.2e-16) o los términos de GO (prueba Wilcoxon-Mann-Whitney; p-valor = 2.5e-3), lo cual es interesante dado que las GENSOR Units fueron enriquecidas con vías metabólicas al agregar reacciones secundarias. Este resultado muestra que la respuesta metabólica mediada por TFs no correlaciona con las vías metabólicas canónicas ni con los procesos descritos por la GO. La proporción más grande de GENSOR Units (21 %) tiene una respuesta que involucra un sólo flujo metabólico, en contraste, la segunda proporción más grande (15 %) tiene una conectividad de 0, seguida de los valores 0.5 y 0.6 con 11 % cada uno.

Es poco probable que el gradiente de conectividad observado sea un artefacto de sitios de pegado desconocidos, pues está presente tanto en el set de GENSOR Units de TFs ampliamente estudiados (aquellos para los que se conoce el efector; figura 3.10c, barras rojas y azules) como en aquellos menos caracterizados (Figura 3.10c, barras amarillas). También es importante notar que la conectividad de una GENSOR Unit no depende del número de enzimas que contiene (Figura 3.11), de hecho, el gradiente puede seguir siendo observado si sólo se consideran GENSOR Units con 5 o más enzimas.

La conectividad de una GENSOR Unit que incluye retroalimentación (Figura 3.10c, barras rojas) puede interpretarse como una medida de la autonomía de la respuesta mediada por el TF. Un valor de conectividad de 1 significa que, en presencia de la señal, la acción del TF impactará un único flujo metabólico que tiene un efecto directo sobre la disponibilidad del efector. En consecuencia, la respuesta del TF permanecerá constante hasta que la concentración de la señal cambie. Otros elementos celulares pueden estar actuando sobre el flujo metabólico, sin embargo, desde la perspectiva del TF su efecto es claro y definido. Diecinueve GENSOR Units se ajustan a esta descripción,



(a) Vías metabólicas

(b) Términos de Gene Ontology

(c) GENSOR Units

(d) Genes bajo el mismo efecto

Figura 3.10: Conectividad. Distribución de conectividad de (a) vías metabólicas, (b) términos de GO de procesos biológicos, (c) GENSOR Units y (d) GENSOR Units construídas con genes bajo el mismo efecto regulador. En (c) se muestran las GENSOR Units con retroalimentación en rojo, en azúl aquellas con efectores conocidos sin circuitos de retroalimentación identificados y en amarillo las GENSOR Units sin efector conocido. En (d) se muestran las GENSOR Units con genes activados en amarillo y aquellas con genes reprimidos en naranja.

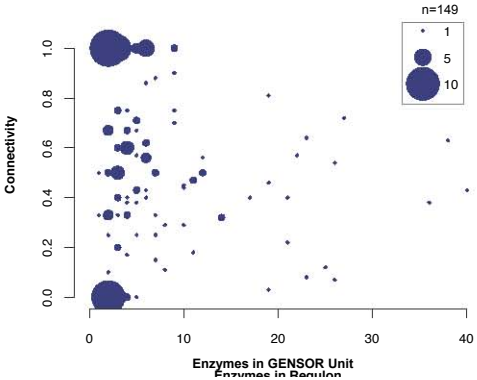


Figura 3.11: Total de enzimas en cada GENSOR Unit comparado con su conectividad.

Figure S3. Total enzymes in a GENSOR Unit compared to its connectivity value. Only enzymes where the catalyzed reaction is known were considered. GENSOR Units with less than two catalytic reactions were omitted to avoid values of 0 without biological significance.

su respuesta está relacionada a metabolismo alantoína (AllS,AllR), arsenito (ArsR), hidroxibutirato (AtoC), colina (BetI; figura 3.4), quitobiosa (ChbR), cianato (CynR), níquel (NikR), zinc (Zur, ZntR), hierro (YqjI), acetilneuraminato (NanR), propanoato (MhpR), idonato (IdnR), glicina (GcvA), citrato (PrpR), gluconato (GntR), triptófano (TrpR) y xantosina (XapR).

Los valores de conectividad bajos pueden deberse al efecto regulador del TF (activación/represión), reflejando una mayor complejidad en su respuesta. La activación de un flujo metabólico requiere la presencia de todas las enzimas necesarias, mientras que su inhibición y el subsecuente cambio de dirección del flujo metabólico pueden ser logrados a través de la represión de un sólo gen. Siguiendo esta lógica se esperaría que los genes reprimidos tuvieran valores de conectividad bajos y lo opuesto para los genes activados. Para comprobar esta hipótesis cada GENSOR Unit se dividió en dos, una sección con la respuesta de los genes reprimidos y otra derivada de los genes activados. Se calculó la conectividad de ambas partes y se graficó la distribución resultante (Figura 3.10d). Compatible con la hipótesis planteada, la distribución de conectividad de los genes reprimidos tiene un pico en 0 que incluye el 67 % de las GENSOR Units evaluadas y es significativamente diferente de la distribución de genes activados (prueba Wilcoxon-Mann-Whitney; p-valor = 7.7e-11). Las GENSOR Units con

los valores de conectividad más bajos podrían tratarse de puntos de control de la regulación que afectan de forma negativa varios flujos metabólicos independientes en respuesta a un estímulo, produciendo una respuesta más global. Fundamentalmente, la conectividad refleja la complejidad de la respuesta mediada por un TF, la cual, como se ha demostrado, presenta un gradiente con picos en ambos extremos de la escala.

3.3.3. Congruencia de Complejos Heteromultiméricos

Los complejos protéicos heteromultiméricos pueden servir como indicadores de la autonomía de una GENSOR Unit. Para que la respuesta observada en una GENSOR Unit suceda de forma inmediata, es necesario que todos los monómeros que son parte de complejos heteromultiméricos sean regulados por el mismo TF, en caso contrario puede asumirse que varios TFs requieren cooperar para producir todos los monómeros necesarios para llevar a cabo el proceso que involucra al complejo. Una GENSOR Unit donde todos los monómeros pertenecientes a complejos protéicos son directamente regulados por el TF principal tendrá una respuesta más autónoma que una GENSOR Unit donde sólo algunos monómeros sean regulados directamente y requiera de la acción de otras GENSOR Units para que el complejo sea funcional. Para cuantificar esta propiedad en las 112 GENSOR Units que incluyen complejos protéicos heteromultiméricos se utilizó la siguiente fórmula:

$$A = \frac{MC_{regulados}}{MC_{totales}}$$

Donde:
$MC_{regulados}$ = Total de monómeros que pertenecen a complejos protéicos heteromultiméricos y son directamente regulados por el TF principal de la GENSOR Unit
$MC_{totales}$ = Total de monómeros que pertenecen a complejos protéicos heteromultiméricos en la GENSOR Unit.

A la fracción calculada se le denominó valor de **autonomía por complejos**. Es importante recordar que entre los objetos que incluye una

GENSOR Unit se encuentran los complejos protéicos heteromultiméricos a los que pertenecen las proteínas cuyos genes son directamente regulados por el TF, así como el resto de los monómeros que forman parte de dichos complejos, por lo que toda la información necesaria para calcular la autonomía por complejos se encuentra en las GENSOR Units individuales.

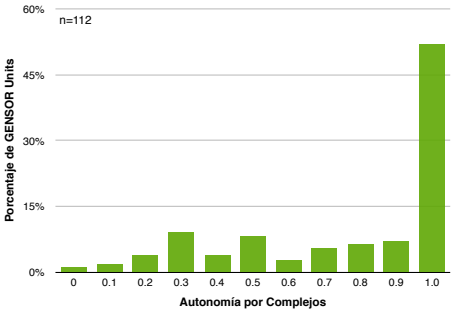
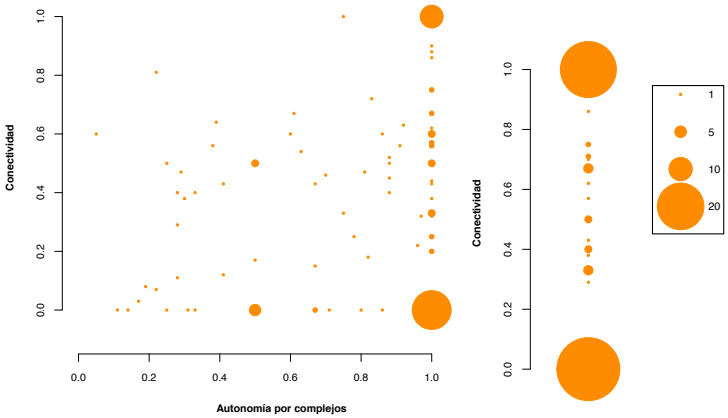


Figura 3.12: Autonomía por complejos. Distribución de valores de autonomía por complejos en las 112 GENSOR Units que incluyen complejos heteromultiméricos.

Como puede observarse en la figura 3.12, el 52 % de las GENSOR Units consideradas tiene un valor de autonomía por complejos de 1, lo que significa que un solo TF puede afectar a la vez todos los monómeros que participan en complejos. Es posible que otros TFs tengan algún efecto sobre ellos, pero en principio, el TF de la GENSOR Unit es capaz de afectarlos al mismo tiempo. Setenta por ciento de las GENSOR Units probadas cuentan con valores de autonomía por complejos de 0.7 o más, mientras que el 16 % obtuvo un valor de 0.3 o menos. Estos últimos se tratan de GENSOR Units donde el producto de un gen regulado forma parte de complejos protéicos muy grandes, por ejemplo RcsB o MlrA que regulan una porción de la subunidad 30S y 50S del ribosoma [Wada, 1986] [Thiede et al., 1998], respectivamente; en otros casos los productos de genes regulados son subunidades de varios complejos, por ejemplo Rob, que regula al gen *tolC*, cuyo producto es una porina que participa en 9 bombas de eflujo distintas [Zgurskaya et al., 2011].



(a) (b)

Figura 3.13: Autonomía por Complejos y Conectividad. (a) Distribución de valores de autonomía por complejos relacionada con distribución de conectividad de 112 GENSOR Units que contienen complejos protéicos heteromultiméricos. (b) Conectividad de 37 GENSOR Units que no incluyen complejos protéicos heteromultiméricos. El diámetro del círculo indica la cantidad de GENSOR Units que se ubican en esos valores, la escala es la misma para ambos páneles.

Para observar si el valor de autonomía por complejos de una GENSOR Unit correlaciona con el gradiente de complejidad observado en la conectividad se compararon ambos valores de las 112 GENSOR Units que incluyen complejos (Figura 3.13a) y se graficó la conectividad de las 77 GENSOR Units que no incluyen complejos heteromultiméricos (Figura 3.13b). Es claro que los valores de conectividad y autonomía por complejos no guardan una relación directa, lo que refleja la gran complejidad que subyace a los circuitos genéticos celulares. La conectividad evalúa la homogeneidad del efecto metabólico de una GENSOR Unit, mientras que la autonomía por complejos evalúa, en cierta medida, la capacidad del TF para llevar a cabo ese efecto metabólico por sí mismo. Puede pensarse que la autonomía por complejos existe en el nivel mecanístico de la GENSOR Unit, mientras que la conectividad refleja el nivel de organización fisiológico de la respuesta

celular. Aunque no existe una relación directa entre ambas métricas, es posible identificar cuatro grandes grupos de GENSOR Units:

- **Grupo 1**. Valor de conectividad de 1 y valor de autonomía por complejos de 1. Incluye 10 GENSOR Units que se enumeran a continuación incluyendo sus señales conocidas entre paréntesis: AtoC, ChbR (quitobiosa), Dan (tartrato), GatR (galactitol), HcaR (fenilpropanoato), NikR (níquel), TrpR (triptófano), UlaR (ascorbato), YiaJ (lixosa) y Zur (zínc). Estas GENSOR Units se encuentran entre las más modulares de la célula, pues al mismo tiempo tienen un efecto fisiológico muy localizado (regulan un flujo metabólico lineal) y mecanísticamente cuentan con todos los elementos necesarios para llevarlo a cabo (todos los monómeros que pertenecen a complejos).
- **Grupo 2**. Valor de conectividad de 1, sin valor de autonomía por complejos dado que no incluyen complejos heteromultiméricos. Incluye 24 GENSOR Units que se enumeran a continuación incluyendo sus señales más conocidas entre paréntesis: AllR (alantoína/glioxilato), AllS (alantoína), ArsR (arsenito), AscG (arbutina/salicina/celibiosa), BetI (colina), BirA (biotina), CadC (cadaverina), CaiF (carnitina), CsiR (aminobutirato), CynR (cianto), FabR (ácidos grasos), FeaR (succinato), GntR (gluconato), IdnR (idonato), KdgR (gluconato), MhpR (propanoato), MngR (glicerato), MtlR (manitol), NanR (ácido siálico), PrpR (metilcitrato), RtcR, XapR (xantosina), YqjI (níquel) y ZntR (zinc). Este grupo, además de compartir las características del grupo 1 y ser muy modulares, representa a las GENSOR Units más simples en la gráfica, pues su acción no requiere de la formación de complejos heteromultiméricos.
- **Grupo 3**. Valor de conectividad de 0, sin valor de autonomía por complejos dado que no incluyen complejos heteromultiméricos. Incluye 27 GENSOR Units que se enumeran a continuación, incluyendo sus señales más conocidas entre paréntesis: Ada (agentes alquilantes), AidB (agentes alquilantes), AlaS (alanina), BluR (formación de biofilm), BolA (morfología de la célula), ComR (cobre/plata/oro), DicA (ciclo celular), GlrR, LysR (lisina), MalI (maltosa), MarR (antibióticos), McbR (inductores de *quorum sensing*), NemR (etilmaleimida), NorR (especies reactivas de nitrógeno), PgrR (estrés por calor), PspF (infección por

fagos), RelB (limitación de nutrientes), RelB-RelE (limitación de nutrientes), RhaR (ramnosa), SlyA, SoxR (óxido nítrico), UhpA, UidR (glucósidos), YehT, YeiL (aerobiosis), YpdB (compuestos orgánicos volátiles) y ZraR (zinc). Este grupo muy probablemente representa GENSOR Units para las que se cuenta con muy poca información, lo cual se ve ejemplificado por el hecho de que veinte de ellas tienen valores de conectividad de 0 que no son informativos, pues incluyen menos de 2 reacciones enzimáticas. Es probable que en futuras actualizaciones estas GENSOR Units aparezcan en otro grupo.

- **Grupo 4**. Valor de conectividad de 0 y valor de autonomía por complejos de 1. Incluye 17 GENSOR Units que se enumeran a continuación, incluyendo sus señales más conocidas entre paréntesis: AgaR (galactosamina), AppY (anaerobiosis), CusR (cobre/plata), DhaR (dihidroxiacetona), DinJ-YafQ, EbgR (galactósidos), HdfR, HyfR (formato), KdpE (potasio), LrhA, MazE (limitación de nutrientes/antibioticos), MazE-MazF (limitación de nutrientes/antibioticos), QseB (inductores de *quorum sensing*), RcdA, SgrR (glucosa-6-fosfato), YefM (formación de biofilm) y YefM-YoeB (formación de biofilm). Este grupo es una mezcla de GENSOR Units de las cuales se tiene poco conocimiento (al igual que el grupo 3), y GENSOR Units que incluyen complejos transportadores. Estas últimas se encargan de transportar metabolitos a través de la membrana que funcionan como señales externas para inducir o reprimir flujos metabólicos que no incluyen al metabolito transportado pues la presencia de la señal activa mecanismos que señalizan a otros reguladores.

3.4. Objetivo iv: Utilizar la topología de las GENSOR Units para predecir elementos faltantes.

Parte del valor conceptual de las GENSOR Unit radica en la descripción de interacciones entre sus elementos. Convierten listas de genes regulados, enzimas y reacciones en redes detalladas que reflejan el efecto funcional de un regulador. Todas las interacciones entre elementos son regidas por principios biológicos, por ejemplo, una reacción

de traducción puede suceder tras una reacción de transcripción, pero nunca en la dirección opuesta. Se puede decir que las GENSOR Units proveen una perspectiva global **biológicamente coherente** y cualquier desviación de esta coherencia apunta a elementos faltantes o mecanismos desconocidos. Una vez identificados los elementos faltantes, la misma coherencia puede permitir predecirlos. Este razonamiento fue utilizado para predecir efectores y genes blanco.

3.4.1. Predicción de Efectores

Para analizar la relación entre efectores y TFs en el contexto de las GENSOR Units se identificó la posición en el flujo metabólico regulado de 77 efectores que cuentan con un circuito de retroalimentación en su GENSOR Unit. Los efectores se clasificaron en dos categorías: "substrato/producto" si el efector es el primer o último metabolito del flujo metabólico, o "intermediario" si se encuentra en cualquier otra posición (Figura 3.14). Los primeros y últimos metabolitos se agruparon en la misma categoría (substratos/productos) para eliminar ambigüedad derivada de las reacciones reversibles. Noventa y siete por ciento de los efectores conocidos (75/77) son metabolitos intermediarios en el flujo metabólico (Cuadro 7.1). La alta proporción de efectores en posición de intermediarios es relevante, pues sólo el 40 % de los metabolitos totales en las mismas GENSOR Units clasifican como intermediarios utilizando el mismo criterio. Una gran proporción de efectores intermediarios ha sido observada anteriormente en sistemas catabólicos inducibles [Savageau, 1976]. El análisis de las GENSOR Units sugiere que los efectores intermediarios son una propiedad general, independiente del modo de acción del TF.

Se ha demostrado que usar metabolitos intermedios como efectores es una estrategia efectiva para aumentar la estabilidad de un sistema [Savageau, 1974] [Savageau, 2001]. Dentro de las GENSOR Units, la estabilidad es crucial para evitar la producción innecesaria de enzimas ante señales fluctuantes, lo cual puede llegar a afectar la tasa de crecimiento celular. Aunado a esto, un efector intermediario es capaz de producir dos circuitos de retroalimentación. Las enzimas que actúan antes de la síntesis del efector participan en un circuito de retroalimentación positivo donde mayor producción de enzimas resultará en

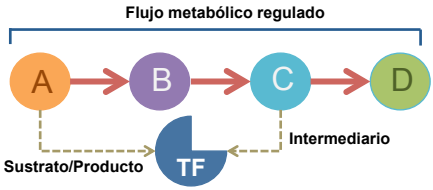


Figura 3.14: Posición del efector en el flujo metabólico regulado. A, B, C y D representan distintos metabolitos que conforman un flujo metabólico. Los genes cuyos productos catalizan las reacciones que conectan a los metabolitos son regulados directamente por el TF. Dados los límtes del flujo metabólico, si A o D actuaran como efector entrarían en la clasificación de "substratos/productos". Por el contrario, si B o C actuaran como efector se clasificarían como "intermediarios".

una mayor concentración de efector. Por otro lado, las enzimas que catalizan reacciones posteriores a la síntesis del efector, es decir, que utilizan el efector, forman un circuito de retroalimentación negativo donde sucede lo opuesto: a mayor concentración de enzimas, menor concentración del efector. Esta dinámica dual producida por efectores intermediarios ha sido demostrada comparando la expresión de enzimas que actúan sobre los metabolitos que se encuentran antes y después del efector en el flujo metabólico [Chubukov et al., 2012]. En apoyo a este resultado se observó en las GENSOR Units que las enzimas antes y después del efector en el flujo metabólico tienden a transcribirse desde distintos operones. El caso más frecuente son GENSOR Units que cuyos efectores son productos de reacciones de transporte (clasificados como "intermediarios"), por ejemplo, la GENSOR Unit de BetI (Figura 3.4) y la mayoría de GENSOR Units involucradas en utilización de fuentes de carbono y metabolismo de aminoácidos. Esta dinámica podría explicar por qué las proteínas de transporte suelen encontrarse en operones independientes. También es posible que las diferentes dinámicas presentes en la misma vía metabólica permitan modular finamente el flujo metabólico en vías metabólicas ramificadas donde el transporte de un metabolito es maximizado (a través de un circuito de retroalimentación positivo), pero su utilización es ligeramente menos inducida para permitir que otras vías metabólicas lo utilicen. Desde una perspectiva evolutiva, un efector intermediario

que produce dos dinamicas distintas permite a la célula producir respuestas metabólicas complejas sin necesidad de nuevos TFs.

La alta proporción de efectores intermediarios sugirió que es posible extrapolar esta propiedad a las GENSOR Units sin efectores conocidos y predecir sus efectores identificando las moléculas intermediarias. Dado el gradiente de complejidad observado en las GENSOR Units, es importante notar que el número de efectores candidatos (y, como consecuencia, el número de falsos positivos) incrementa de acuerdo a la complejidad de la GENSOR Unit en que el método es aplicado. La media de metabolitos intermedios en cada GENSOR Unit es de 11, la cual sigue siendo una cantidad de candidatos considerablemente más bajo que la probada en algunos estudios heurísticos [Ibañez et al., 2000]. Intuitivamente, las mejores predicciones serán aquellas en las GENSOR Units más simples, por lo que el método de predicción se aplicó en las 15 GENSOR Units con conectividad de 1 y sin efectores conocidos (Cuadro 3.1). Dado que los candidatos se toman del componente de respuesta de cada GENSOR Unit, todas las predicciones producen un circuito de retroalimentación. Para evaluar las predicciones se realizó una búsqueda en la literatura de evidencia a favor o en contra y se buscaron dominios de pegado a ligandos en la secuencia de los TFs para apoyar el mecanismo de acción.

Cuadro 3.1: Predicciones de efectores de 15 GENSOR Units con conectividad de 1.

GENSOR Unit	Dominio de Unión a Ligandos	Efectores Predichos	Tipo de Predicción	Evidencia	Referencia
FabR		Ácidos grasos con acyl-ACP	Validada	Gel de Retardo	[Zhu et al., 2009]
UlaR	DeoR C terminal sensor domain (Pfam:PF00455)	Ascorbate-6P	Validada	Gel de Retardo	[Garces et al., 2008]
Dan	Bacterial regulatory helix-turn-helix protein, lysR family [Pfam: PF00126]	Tartrate	Evidencia que apoya la predicción	Cambio en expresión de genes blanco tras adición del compuesto (Ensayo de B-galactosidasa)	[Kim et al., 2009]
FeaR	AraC-binding-like domain [Pfam : PF14525]	hyacinthin (phenyl acetaldehyde)	Evidencia que apoya la predicción	Inferencia por la dinámica del operón	[Zeng and Spiro, 2013]
HcaR	LysR substrate binding domain [Pfam : PF03466]	3-(5,6+Dihydroxycyclohexa-1,3-dien-1-yl)propanoate	Evidencia que apoya la predicción	Inferencia por la dinámica del operón	[Turlin et al., 2001]
MtlR		Mannitol-1P	Evidencia que apoya la predicción	Inferencia por la dinámica del operón	[Figge et al., 1994]
KdgR	Bacterial transcriptional regulator [Pfam : PF01614]	2-Keto-3-deoxygluconate-6-P	Evidencia que apoya la predicción	2-Keto-3-deoxygluconate has been reported as effector of KdgR ortholog in *Erwinia chrysanthem*	[Nasser et al., 1992]

Continuación de cuadro 3.1

GENSOR Unit	Dominio de Unión a Ligandos	Efectores Predichos	Tipo de Predicción	Evidencia	Referencia
MngR	UTRA domain [Pfam : PF07702]	2(alpha-D-mannosyl-6-phosphate)-D-glycerate	Evidencia que apoya la predicción	Cambio en expresión de genes blanco tras adición del compuesto (microarreglo). Mutación de enzimas río abajo no afecta la inducción	[Sampaio et al., 2004]
AscG	Periplasmic binding protein-like domain [Pfam : PF13377]	arbutin-6P, beta-D-cellibiose-6P	Nueva		
CaiF		Gamma-butyrobetaine, crotonobetainyl-CoA, carnityl-CoA, gamma-butyrobetaine-CoA	Nueva. Apoyada por la dinámica de la GENSOR Unit. Evidencia en contra de otras predicciones.	Geles de retardo no reflejaron acción de efector para L-carnitina o crotonobetaína	[Buchet et al., 1999]
YiaJ	Bacterial transcriptional regulator [Pfam : PF01614]	xylulose-P5, 2-3,dioxo-L-gulonate, 3-keto-L-gulonate, 3-keto-L-gulonate 6P	Nueva. Evidencia en contra de otras predicciones	80 efectores candidatos no produjeron cambios en expresión de los genes blanco	[Ibañez et al., 2000]
CsiR	FCD domain [Pfam : PF07729]	L-glutamate, ketoglutarate, succinate semialdehyde	Nueva		

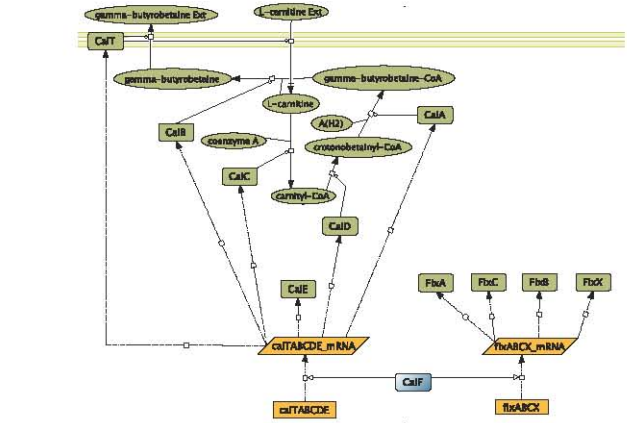
Continuación de cuadro 3.1

GENSOR Unit	Dominio de Unión a Ligandos	Efectores Predichos	Tipo de Predicción	Evidencia	Referencia
GatR	DeoR C terminal sensor domain [Pfam : PF00455]	galactitol-1P, keto-L-tagatose-6P, tagatofuranose,1,6-diphosphate	Nueva		
RtcR		RNA terminal-2',3'-cyclic-phosphate	Nueva		
CadC		cadaverine, lysine	Evidence en contra del modo de acción	Funciona como un sistema de un componente que responde a cambios en pH	[Buchner et al., 2015]

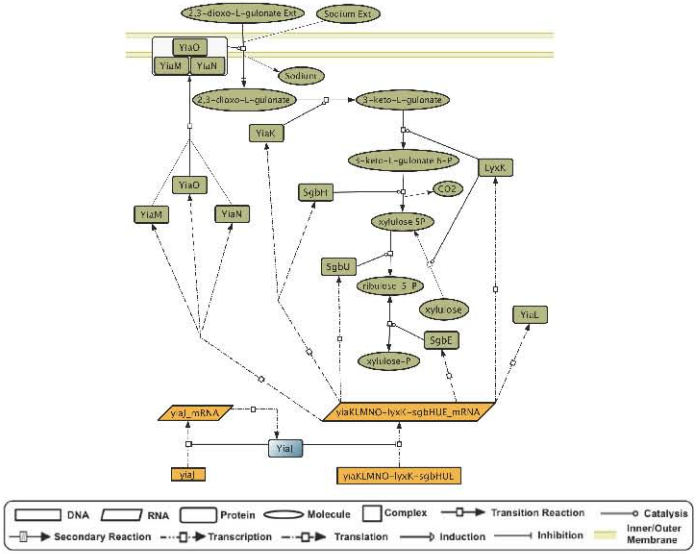
Los dominios de unión a ligandos identificados en la secuencia de los TFs se muestran junto a sus identificadores de Pfam. Los efectores fueron predichos a partir de la topología de cada GENSOR Unit. La columna "Tipo de predicción" indica si la predicción ha sido validada en la literatura, existe evidencia que la apoya/rechaza o se trata de una nueva predicción. La columna "Evidencia" indica los experimentos relevantes que han sido reportados en la literatura.

Las predicciones para FabR y UlaR fueron validadas. Evidencia que apoya las predicciones fue encontrada para Dan, FeaR, HcaR, MtlR, KdgR y MngR, lo que sugiere que son excelentes candidatos para realizar validaciones experimentales. Los efectores predichos para AscG, CsiR, GatR, RtcR, CaiF y YiaJ no han sido reportados antes en la literatura. La evidencia en la literatura para CadC apoya un mecanismo de acción que no se basa en union a ligandos [Buchner et al., 2015], lo cual es consistente con la ausencia de un dominio de pegado en su secuencia. Dos ejemplos interesantes de predicciones nuevas son CaiF y YiaJ. Se ha demostrado que CaiF funciona como un activador en la presencia de L-carnitina [Eichler et al., 1996]. Sin embargo, no se ha podido comprobar la presencia de L-carnitina en complejo con CaiF en geles de retardo [Buchet et al., 1999]. La GENSOR Unit de CaiF muestra que una de las predicciones realizadas, gamma-butirobetaína, provocaría una dinámica de inhibición por producto final que encaja con la observación de que L-carnitina actúa como inductor, específicamente porque L-carnitina es un substrato de la reacción que produce gamma-butirobetaína.

YiaJ regula negativamente la conversión de xilulosa y 2,3-dioxo-gulonato en D-xilulosa-5-fosfato. Ibañez y colegas [Ibañez et al., 2000] probaron 80 compuestos distintos, incluyendo D-xilulosa, y ninguno provocó cambios en la expresión de los genes blanco. Es posible que las predicciones propuestas aquí den mejores resultados pues están basadas en la interpretación global de las interacciones presentes en su GENSOR Unit. Por ejemplo, 2,3-dioxo-L-gulonato podría actuar como un efector cuya presencia impide el pegado de YiaJ al DNA (conformación inactiva en *holo*). En su conformación *holo*, YiaJ reprimiría las enzimas necesarias para la utilización de 2,3-dioxo-L-gulonato como fuente de carbono y éstas sólo serían producidas cuando el metabolito se encontrara en el ambiente. Dado que el 2,3-dioxo-L-gulonato es convertido en D-xylulosa-5-fosfato, probablemente sólo se tome del medio cuando D-xilulosa-5-fosfato no pueda obtenerse de otras fuentes de carbono como arabinosa, xilulosa o ascorbato. Otro efector predicho, el mismo D-xilulosa-5-fosfato, podría promover la unión de YiaJ al DNA en su presencia (conformación activa en *holo*), produciendo una dinámica de inhibición por metabolito final. La ribulosa-5-fosfato fue eliminada de las predicciones dado que es un metabolito central que



(a) CaiF



(b) YiaJ

Figura 3.15: GENSOR Units que permiten la predicción de efectores a partir de su topología. (a) GENSOR Unit de CaiF (b) GENSOR Unit de YiaJ. La simbología aplica para ambas GENSOR Units.

constantemente se produce en la célula y no existen reportes de YiaJ como un regulador con efecto en el metabolismo central.

En resúmen, 53 % de las predicciones fueron apoyadas por la literatura, 7 % fueron rechazadas y 40 % no han sido reportadas, incluyendo dos casos donde se ha demostrado que la dinámica de la GENSOR Unit apoya las predicciones. Debido a que no existen datasets similares a las GENSOR Units es imposible cuantificar la sensibilidad y especificidad del método predictivo, sin embargo, es importante notar que el enfoque utilizado no requiere de herramientas adicionales y podría ser usado para disminuir significativamente el espacio de efectores posibles antes de realizar experimentos.

3.4.2. Predicción de Genes Blanco

La perspectiva global que proveen las GENSOR Units también puede ser útil para guiar la identificación de nuevas interacciones reguladoras en la TRN. La forma más directa es a través de las reacciones secundarias, pues una de las razones que las hacen necesarias es nuestro desconocimiento de la totalidad de la TRN. Se identificaron todos los genes que catalizan las reacciones individuales que pertenece a reacciones secundarias y se consideraron genes blanco candidatos. Por este método se identificaron 267 posibles interacciones TF-gen. Para sumar evidencias independientes y obtener un set más confiable de predicciones, éstas se compararon con predicciones computacionales realizadas por dos métodos independientes cuyos resultados se encuentran disponibles en RegulonDB : TractorDB [Guía et al., 2005] y MatrixScan [Medina-Rivera et al., 2011].

El enfoque usado por TractorDB comienza recopilando todos los sitios de pegado de un TF en *E. coli* y creando una expresión regular a partir de ellos. Posteriormente identifica ortólogos de los genes regulados por el TF en 16 organismos cercanos a *E. coli* y rastrea la región río arriba de los ortólogos utilizando la expresión regular (construida en el primer paso) para identificar sitios de pegado conservados. Una vez identificados los sitios de pegado "ortólogos", todos estos se utilizan para construir una matriz de peso con la que se rastrea el genoma de

TF	Gen	Evidencia
Fur	acnA	TractorDB
Fur	gapA	matrix scan
Fur	gltA	matrix scan
Fur	icd	matrix scan
Fur	metB	matrix scan
Fur	metC	TractorDB/matrix scan
Fur	metF	matrix scan
Fur	pykA	matrix scan
GadE	fabI	matrix scan
IclR	acnA	matrix scan
IclR	mdh	matrix scan
MetR	metH	TractorDB/matrix scan
NsrR	gltA	matrix scan
PurR	purT	TractorDB/matrix scan

Cuadro 3.2: Predicciones de genes regulados obtenidas a partir de reacciones secundarias en GENSOR Units. Se muestran sólo aquellas que también se encuentran entre las predicciones de TractorDB, MatrixScan o ambos.

E. coli y se identifican nuevos sitios de pegado putativos.

MatrixScan es un enfoque que, a diferencia de TractorDB, sólo utiliza los sitios de pegado de cada TF en *E. coli*. Con ellos construye una matriz de peso que es evaluada con la herramienta *matrix-quality*, la cual ayuda a cuantificar la utilidad de una matriz de peso comparando distribuciones teóricas y empíricas de *scores* de sitios encontrados utilizándola. Una vez identificada la mejor matriz de peso, se rastrea el genoma de *E. coli* utilizando la herramienta *matrix-scan* del suite RSAT (Regulatory Sequence Analysis Tools) [Medina-Rivera et al., 2015] y se identifican nuevos posibles sitios de pegado.

Como predicciones confiables se seleccionaron aquellas que fueron generadas a través de las GENSOR Units y al menos otro método (Cuadro 3.2). Sólo una predicción fue compartida por GENSOR Units y TractorDB, 10 fueron compartidas por GENSOR Units y MatrixScan, y 3 fueron compartidas por los tres métodos predictivos. Estas últimas predicciones son las mejores candidatas para validación experimental. Dos de los genes regulados predichos, *pykA* y *purT*, no tienen reportado un sitio de pegado de ningún otro TF por lo que su validación

permitiría ampliar la TRN.

3.5. Objetivo v: Identificar relaciones entre GENSOR Units

Una de las ventajas principales de las GENSOR Units es que permiten estudios detallados a pequeña escala, por ejemplo, al analizar las propiedades de GENSOR Units con alta conectividad cuyo efecto es modular. Sin embargo, también pueden ser usadas como piezas de ensamble para descripciones de procesos que incluyan a dos o más GENSOR Units. De hecho, es posible unir las 189 GENSOR Units en un solo mapa que integra todo el conocimiento actual sobre la relación entre la regulación transcripcional y el metabolismo, codificado en las bases de datos más completas. No obstante, unir un par de GENSOR Units es suficiente para elucidar comportamientos celulares complejos, por ejemplo, la represión catabólica [Monod, 1942] [Görke and Stülke, 2008].

La represión catabólica se refiere a la capacidad de *E. coli* de consumir las fuentes de carbono disponibles en un orden determinado. Si en un momento dado se encuentra en un ambiente con dos o más fuentes de carbono disponibles, siempre las ocupará en el siguiente orden: (1) glucosa, (2) lactosa, (3) arabinosa, (4) xilosa, (5) sorbitol, (6) ramnosa y (7) ribosa [Aidelberg et al., 2014]. AraC y XylR son TFs que se unen a arabinosa y xilosa respectivamente, y se encargan de coordinar su utilización. Ambos TFs regulan a la TU *xylAB* así que sus GENSOR Units individuales pueden ser unidas a través de este elemento compartido (Figura 3.16). La GENSOR Unit compleja AraC-XylR muestra que cuando arabinosa y xilosa están presentes en el ambiente al mismo tiempo, *xylAB* será reprimida por AraC y activada por XylR. Dado que en *E. coli* la represión suele ser dominante [Collado-Vides et al., 1991], la transcripción de *xylAB* será inhibida y la arabinosa será usada preferencialmente. Una vez que toda la arabinosa haya sido metabolizada, AraC volverá a su estado inactivo, *xylAB* será inducida por XylR y la xilulosa será usada como segunda fuente de carbono. En resúmen, la regulación opuesta que es ejercida

sobre *xylAB* es el interrumptor donde *E. coli* toma la decisión de usar arabinosa antes que xilosa. La GENSOR Unit compleja AraC-XylR muestra el poder descriptivo de unir GENSOR Units individuales. Decisiones más complejas pueden involucrar más de dos GENSOR Units, pero independientemente del tamaño, las GENSOR Units complejas mantienen el nivel de detalle necesario para modelar dinámicamente o identificar mecanismos importantes en la toma de decisiones celulares.

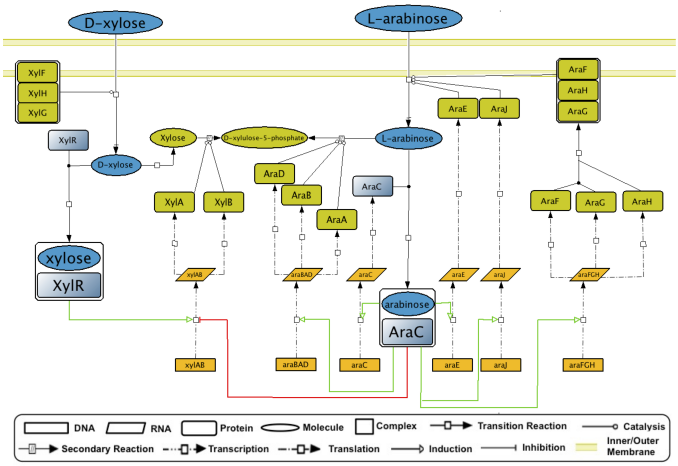


Figura 3.16: GENSOR Unit compleja: AraC-XylR. Las líneas verdes representan activación y las rojas inhibición de los genes regulados.

3.5.1. Ampliación de la TRN

Utilizar GENSOR Units como piezas de ensamble también permite elucidar mecanismos de regulación que dependen del metabolismo para modular la actividad de un TF. La presencia de circuitos de retroalimentación en una GENSOR Unit implica que la respuesta mediada por un TF tiene un efecto directo sobre la disponibilidad de su efector. Adicionalmente, la respuesta puede afectar la disponibilidad del efector de un segundo TF. Por ejemplo, AlaS se une a alanina para reprimir solamente a su propio promotor, dado que ningún otro TF regula su transcripción, AlaS aparece como un nodo aislado en la TRN. Sin embargo, las GENSOR Units de IscR, Fur y OxyR incluyen en

su respuesta la producción de alanina a través de la acción de SufS, una desulfurasa de cisteína. En presencia de clusters de hierro-azufre, óxido nítrico, hierro o estrés oxidativo (las señales de IscR, NsrR, Fur y OxyR, respectivamente) la concentración celular de alanina fluctúa y, por lo tanto, la conformación funcional de AlaS se verá afectada.

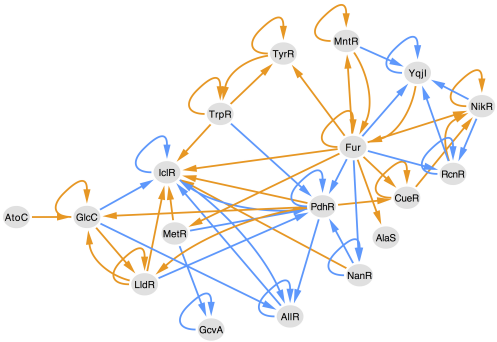


Figura 3.17: Cascada de regulación indirecta: AtoC. Los nodos representan TFs, las aristas unen TFs que producen un efector de otro TF (terminación en flecha). Las líneas naranjas indican que la presencia del efector promueve una conformación activa en el segundo TF (terminación en flecha). Líneas azules indican una conformación inactiva.

La misma lógica puede ser aplicada a gran escala para identificar cascadas de regulación TF-TF a través de sus efectores, un ejemplo se muestra en la Figura 3.17. AtoC es un TF cuya respuesta está involucrada en la producción de poliaminas, catabolismo de ácidos grasos de cadena corta, movilidad y quimiotaxis, y regula directamente sólo 4 genes. Su respuesta regulada también produce el efector de GlcC, cuya respuesta produce otros efectores que dan pie a una sucesión de cambios conformacionales de 16 TFs. El ampliar este enfoque a toda la red resulta en una Red de Regulación Indirecta TF-TF conformada por 106 nodos y 302 aristas, incluyendo retroalimentación (Figura 3.18a). Los TFs que más efectores producen son Fur, Cra y PhoB, mientras que los TFs cuyos efectores son producidos por más TFs son IclR (piruvato), GlcC (acetato) y PdhR (piruvato). Estas interacciones son relevantes porque las conformaciones de los TFs rara vez se

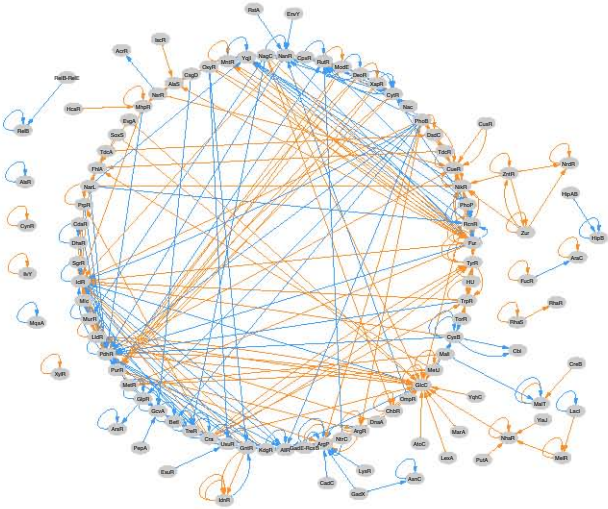
consideran en estudios de la TRN y, sin embargo, son un paso crucial para que la regulación transcripcional suceda.

La Red de Regulación Indirecta incluye 186 nuevas interacciones que no existen en la TRN, aproximadamente la mitad (48 %) de las interacciones con que la TRN cuenta actualmente, lo cual significa un gran incremento en el número de relaciones conocidas entre TFs (Figura 3.18b). Setenta y siete interacciones están presentes en ambas redes, de las cuales el 51 % son autoregulación/retroalimentación, resaltando la importancia de la automodulación en la actividad de TFs individuales. Será interesante re-evaluar las propiedades conocidas de la TRN aumentando estas interacciones reguladoras que dependen del metabolismo.

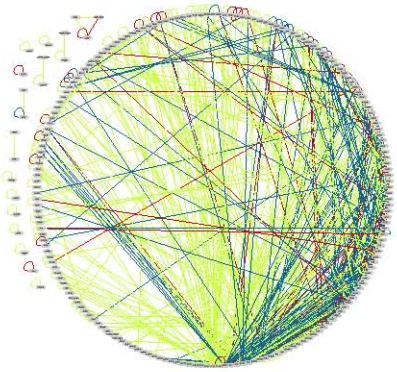
3.5.2. Unión global, no-heurística de GENSOR Units

En la GENSOR Unit compleja de AraC-XylR (Figura 3.16) el elemento común que permitió unirlas es la TU *xylAB*, sin embargo, un par de GENSOR Units pueden tener otras relaciones que permitan unirlas. Las que se consideraron más relevantes biológicamente son:

- **Unidades Transcripcionales**. Utilizadas en el ejemplo de AraC-XylR, indican que los TFs de ambas GENSOR Units tienen sitios de pegado en el mismo promotor, sugiriendo que están relacionados funcionalmente. Aunque no se cuenta con la información necesaria para determinar si se trata de una actividad sinérgica o excluyente, es un hecho que los TFs son cercanos en la TRN y sus GENSOR Units están relacionadas.
- **Metabolitos**. Metabolitos presentes en dos o más GENSOR Units indican que la respuesta mediada por el TF de cada GENSOR Unit tiene un efecto metabólico relacionado. Entre menos común sea un metabolito en la célula, más relevante será la relación entre GENSOR Units. Por ejemplo, compartir ATP no es tan relevante como compartir un aminoácido.
- **Reacciones Secundarias**. Las reacciones secundarias representan "huecos" en vías metabólicas canónicas presentes en una GENSOR Unit. En muchas ocasiones las reacciones individuales que conforman una reacción secundaria se encuentran en otra



(a)



(b)

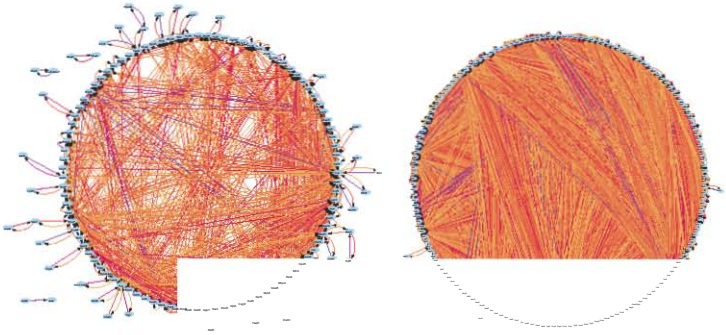
Figura 3.18: Red de Regulación Indirecta TF-TF y su similitud con la TRN. (a) Red de Regulación TF-TF a través del metabolismo. Los nodos representan TFs, las aristas unen TFs que producen un efector en su respuesta, con los TFs (terminación en flecha) que se unen a dicho efector. Las líneas naranjas indican que la presencia del efector promueve una conformación activa en el segundo TF (terminación en flecha). Líneas azules indican una conformación inactiva. (b) Comparación con la TRN TF-TF. Los nodos representan TFs. Las aristas verdes indican interacciones clásicas de la TRN, aristas azules indican aristas exclusivas de (a) y aristas rojas aparecen en ambas redes.

GENSOR Unit, de forma que al unir ambas GENSOR Units las reacciones secundarias desaparecen y puede asumirse que los TFs tienen una relación cooperativa.

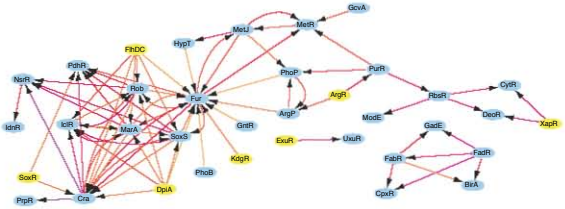
- **Principio y fin de flujos metabólicos**. Compartir metabolitos entre GENSOR Units refleja una relación funcional, sin embargo, esta se torna más relevante si el metabolito compartido es el metabolito final de un flujo metabólico en una GENSOR Unit y, al mismo tiempo, el metabolito inicial de otro flujo metabólico en la segunda GENSOR Unit. De forma que se puede asumir que cada GENSOR Unit regula una parte de un flujo metabólico más largo. Este tipo de relación es el que se espera de GENSOR Units que incluyen partes de una misma vía metabólica..
- **Monómeros de complejos protéicos heteromultiméricos**. Como se mencionó en la sección 3.3.3 (Congruencia de complejos heteromultiméricos), un TF no siempre regula todas las subunidades de un complejo heteromultimérico. Puede asumirse que GENSOR Units que incluyen subconjuntos de monómeros que son parte del mismo complejo estarán relacionadas dinámicamente, pues sus TFs deben inducir los genes responsables de producir el complejo al mismo tiempo para que éste sea funcional.

Cada enfoque permite identificar relaciones entre distintas GENSOR Units. Idealmente, las GENSOR Units cuya relación funcional sea más relevante serán aquellas que se relacionen bajo los cinco criterios, pues la respuesta de sus TFs será complementaria de varias formas. Para identificar las relaciones posibles utilizando cada enfoque por separado se utilizó el set completo de GENSOR Units y los resultados se representaron en forma de redes donde cada GENSOR Units está representada como un nodo y las aristas representan relaciones entre pares de GENSOR Units (Figuras 3.19a - 3.19f).

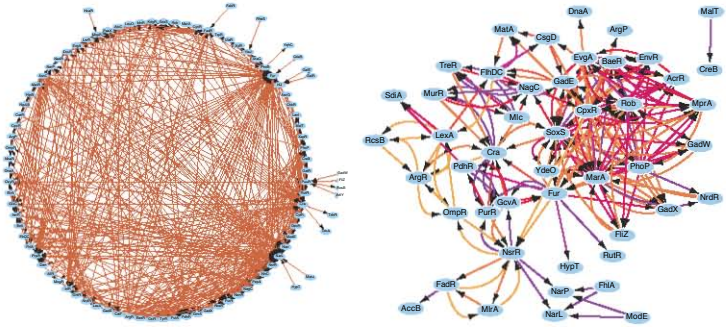
La red de TUs compartidas (Figura 3.19a) pues las aristas unen GUs que comparten una misma TU, gracias a TFs que regulan al mismo promotor. La diferencia entre redes radica en que, en la TRN (Figura 1.3a) los TFs se unen a través de los genes compartidos, es decir, se requieren dos aristas para unir un par de TFs, por ejemplo: AraC (TF) – *xylA* (gen) – XylR (TF), dos TFs que regulan a un mismo gen. Por otro lado, en la red de regulación TF-TF (Figura 1.3b) las aristas re-



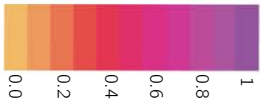
(a) TUs (b) Metabolitos



(c) Reacciones Secundarias



(d) Flujos metabólicos sucesivos (e) Complejos protéicos



(f) Pesos de aristas

Figura 3.19: Ver pie de figura en página siguiente.

Figura 3.19: Redes de elementos compartidos entre GENSOR Units. En todas las redes los nodos representan GENSOR Units, las aristas representan relaciones entre ellas bajo 5 criterios: (a) TUs compartidas, (b) metabolitos compartidos, (c) reacciones secundarias complementarias, (d) flujos metabólicos complementarios y (e) complejos protéicos complementarios. El peso de las aristas en (a), (b) y (e) es según (f). Los elementos totales incluídos en cada red se encuentran en el texto. Los nodos amarillos en (c) indican GENSOR Units que no incluyen reacciones secundarias.

presentan TFs que regulan a otros TFs. En la red de GENSOR Units por TUs compartidas, las aristas unen TFs que regulan al menos un gen compartido. Para indicar la fracción de genes compartidos entre GENSOR Units, a cada arista se le asignó un peso. Dado que el total de TUs varía en cada GENSOR Unit, los pesos tienen dirección y no son recíprocos. Por ejemplo, considérense dos GENSOR Units (a,b) que contienen 5 y 10 metabolitos respectivamente, de los cuales 3 son compartidos. El peso de la relación a ->b será 3/5, mientras que el peso de la relación b ->a será 3/10. La red cuenta con 132 nodos y 941 aristas que pueden resumirse en 495 posibles pares de GENSOR Units. Las GENSOR Units que no aparecen en la red (57) no comparten TUs reguladas con ninguna otra GENSOR Unit.

La red de GENSOR Units por metabolitos compartidos (Figura 3.19b) incluye 172 nodos y 6307 aristas, que representan 3154 posibles pares de GENSOR Units. Diecisiete GENSOR Units cuentan con metabolitos únicos y, por lo tanto, no aparecen como nodos. Las aristas de esta red también cuentan con un peso que refleja la fracción de metabolitos que comparten un par de GENSOR Units, como en la red de TUs, el peso tiene dirección y las aristas no son recíprocas. La densidad de aristas en la red da la apariencia de una bola de estambre y refleja la gran cantidad de relaciones que pueden obtenerse, apoyando que TFs lejanos en la TRN pueden ser funcionalmente cercanos al considerar el metabolismo, como se demostró en la sección anterior con la red TF-TF a través de efectores (Figura 3.18a).

Las reacciones secundarias compartidas crean una red mucho más compacta (Figura 3.19c), lo cual se explica por el bajo número de

reacciones secundarias que se agregaron (144). La red cuenta con 36 nodos y 80 aristas. Las aristas representan GENSOR Units que incluyen los genes necesarios para catalizar las reacciones secundarias presentes en la segunda GENSOR Unit (flecha en arista). El peso indica la fracción de reacciones presentes en la segunda GENSOR Unit que la primer GENSOR Unit es capaz de complementar. Por ejemplo, la GENSOR Unit de PrpR cuenta con una reacción secundaria constituida por tres reacciones individuales (Figura 3.2a), por otro lado la GENSOR unit de Cra incluye las enzimas necesarias para catalizar las tres reacciones, por lo que la arista entre Cra y PrpR apunta a PrpR con un peso de 1.0. Treinta y una de las 48 GENSOR Units que incluyen reacciones secundarias están presentes en la red. Las cinco restantes (nodos en amarillo) en la red no cuentan con reacciones secundarias pero incluyen genes que complementan reacciones secundarias en otras GENSOR Units. Esta red cuenta con un pequeño módulo compuesto por 5 GENSOR Units: GadE, FabR, FadR, BirA y CpxR, es posible asumir que los TFs de estas GENSOR Unit cooperan dinámicamente para llevar a cabo una vía metabólica canónica.

La red basada en el principio y fin de flujos metabólicos es más grande de lo esperado. Cuenta con 97 nodos y 472 ejes. En esta red las aristas no tienen dirección pues es difícil saber la dirección real del flujo metabólico y definir qué GENSOR Unit complementa a la otra. Tampoco tienen peso, esta relación se considera binaria.

La red de complejos protéicos expande el cálculo de autonomía por complejos de la sección 3.3.3, pues describe las relaciones entre GENSOR Units a través de los complejos que no son regulados por un solo TF. Cuenta con 47 nodos y 197 aristas. Las aristas unen GENSOR Units (lado con flecha) que incluyen complejos protéicos con monómeros no regulados directamente por su TF, con las GENSOR Units que incluyen los genes que codifican para dichos monómeros (lado sin flecha). El peso de la arista indica la fracción de monómeros no regulados que la segunda GENSOR Unit incluye. Por ejemplo, la GENSOR Unit de AccB cuenta con un complejo conformado por 4 monómeros, de los cuales sólo uno es regulado directamente por AccB, el resto es regulado por FadR, por lo que la arista que une a FadR con AccB tiene un peso de 1.0.

Los distintos enfoque mencionados permitieron identificar similitudes entre pares de GENSOR Units, sin embargo, también muestran la complejidad de las relaciones entre ellas a nivel global, pues es raro identificar módulos delimitados y, en general, todas las GENSOR Units presentes en cada red están conectadas entre sí. El siguiente paso fue tratar de delimitar los módulos funcionales de forma que pudieran obtenerse grupos de GENSOR Units que maximicen la cantidad de elementos compartidos entre ellas. Para esto se combinaron los 5 enfoques utilizados. Se sumaron los pesos entre cada par de GENSOR Units de las 5 redes, se realizó un clustering jerárquico a los pesos totales y se generó un mapa de calor donde es más evidente la forma en que las GENSOR Units se agrupan de acuerdo a sus similitudes (Figura 3.20). Aunque las unidades utilizadas para comparar son arbitrarias, existe un máximo de similitud que puede existir entre cualquier par de GENSOR Units, demostrado por la diagonal del mapa de calor, donde se encuentran los valores de similitud de cada GENSOR Unit con ella misma. Alrededor de la diagonal pueden identificarse grupos de GENSOR Units que comparten elementos entre ellos, el clustering jerárquico permitió posicionar más cerca las GENSOR Units que más elementos comparten y, por lo tanto, es posible identificar grupos de 3 o más GENSOR Units.

Los grupos obtenidos se muestran en el cuadro 3.3. Los grupos 1-3 reúnen a las GENSOR Units más grandes, el hecho de que tengan tantos elementos compartidos puede deberse a que cuentan con una función más global, coordinando varios puntos críticos del metabolismo en respuesta a señales importantes como hierro (Fur) o fosfato (PhoP). Es importante resaltar que AraC y XylR conforman un grupo, la relevancia de su unión fue mencionada en la sección 3.5. El grupo conformado por CysB y Cbl también ayuda a validar los agrupamientos pues se sabe que estos reguladores trabajan en forma conjunta para contender con la ausencia de sulfatos en el medio [Van der Ploeg et al., 1997]. Es posible que otros enfoques para agrupar GENSOR Units produzcan distintos grupos, por ejemplo, utilizando datos de coexpresión. Sin embargo, el enfoque aquí utilizado está basado en las similitudes entre las respuestas metabólicas mediadas por TFs. Aunque es sólo un primer paso, representa conocimiento nuevo pues es el primer estudio

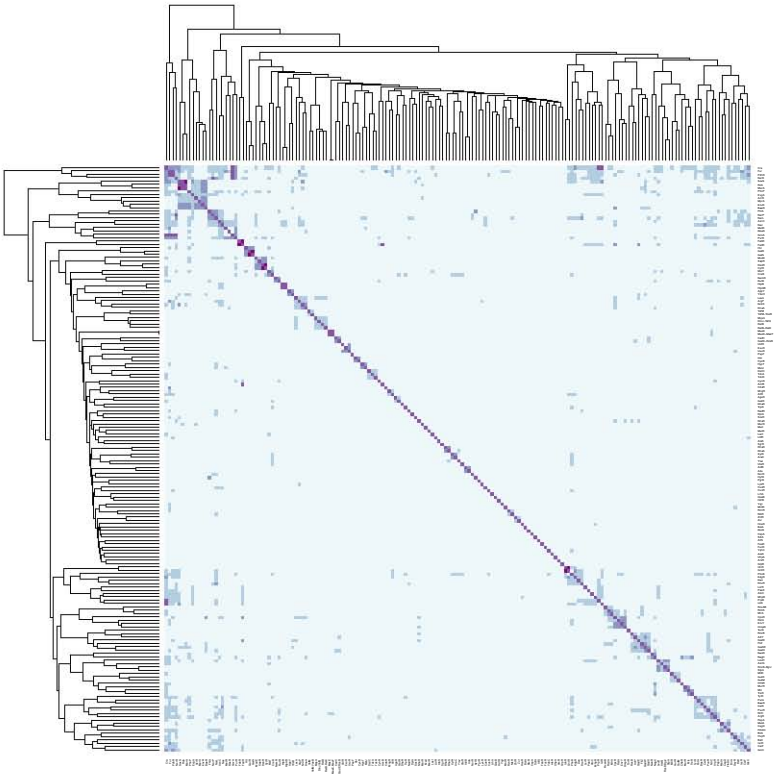


Figura 3.20: Mapa de calor de relaciones entre GENSOR Units. Los ejes muestran GENSOR Units agrupadas a través de un clustering jerárquico (dendograma) que posiciona a la GENSOR Units que más comparten elementos en posiciones más cercanas. Los colores indican el grado de similitud obtenido por la suma de los pesos de las aristas de los 5 enfoques considerados. Marrón indica alta similitud y azul cielo baja similitud.

global que incluye a todos los TFs conocidos donde se considera al mismo tiempo la regulación transcripcional y el metabolismo.

Grupo	GENSOR Units
Grupo 1	Cra, Fur, PdhR, NsrR
Grupo 2	SoxS, Rob, MarA, PhoP, EvgA, AcrR, MprA, EnvR, BaeR
Grupo 3	FhlA, NarP, NarL, AsnC, Nac, MetR, RbsR, GcvA, PurR
Grupo 4	FabR, FadR
Grupo 5	HU, GalR, GalS
Grupo 6	ModE, XapR, DeoR, CytR
Grupo 7	RcnR, NikR
Grupo 8	ZntR, Zur
Grupo 9	IdnR, GntR
Grupo 10	MalT, CreB
Grupo 11	NemR, RutR
Grupo 12	HipB, HipAB
Grupo 13	AppY, YdeO
Grupo 14	LexA, ArgP, NrdR, DnaA
Grupo 15	YefM, YefM-YoeB
Grupo 16	MqsA, DinJ-YafQ, RelB, RelB-RelE
Grupo 17	MazE, MazE-MazF
Grupo 18	CadC, GadE-RcsB
Grupo 19	UidR, ExuR, UxuR
Grupo 20	Cbl, CysB
Grupo 21	HypT, MetJ
Grupo 22	DsdC, TdcA, TdcR
Grupo 23	MngR, AllR
Grupo 24	AgaR, GatR
Grupo 25	RhaR, RhaS
Grupo 26	XylR, AraC
Grupo 27	AidB, Ada
Grupo 28	PhoB, KdgR, Dan, DcuR
Grupo 29	MhpR, PrpR
Grupo 30	RcsAB, RcdA, MlrA
Grupo 31	CpxR, RstA, EnvY, OmpR
Grupo 32	RcsB, AdiY, GadE, FliZ, GadW, GadX
Grupo 33	NanR, NagC
Grupo 34	LeuO, AscG, RcsB-BglJ, StpA
Grupo 35	MtlR, GutR, GutM
Grupo 36	MurR, Mlc, TreR
Grupo 37	TyrR, PutA, BasR, CsiR, PuuR, NtrC, ArgR, PepA
Grupo 38	MatA, CsgD, FlhDC
Grupo 39	BirA, OxyR, BetI, IscR, CaiF, GlcC

Cuadro 3.3: Grupos de GENSOR Units a partir de la similitud de elementos.

Capítulo 4

Discusión

Durante su vida las células están expuestas a innumerables cambios en su ambiente externo e interno. Para sobrevivir dependen de circuitos genéticos que detectan cambios y producen una respuesta apropiada. Entender la forma en que las células detectan y procesan información es crucial para poder identificar principios de diseño aún desconocidos. El concepto de GENSOR Unit integra relaciones entre las redes de señalización, de regulación y metabólica para describir el flujo de información detrás de procesos individuales de señal ->respuesta.

En el presente trabajo se llevó a cabo un análisis en escala genómica de la complejidad de la respuesta mediada por cada uno de los 189 TFs locales que existen actualmente en RegulonDB, poniendo a prueba el paradigma que establecieron Jacob y Monod [Jacob and Monod, 1961] con la regulación del operón *lac*, donde los reguladores funcionan como interruptores que prenden o apagan capacidades específicas de la célula. Los resultados mostraron que, a escala genómica, la relación entre los genes regulados por un TF y el metabolismo que impactan no sigue una regla general de 1 TF/1 proceso.

Las GENSOR Units utilizadas para llegar a este resultado están basadas en regulones en el sentido más amplio del palabra: en genes que son regulados por el mismo TF, independientemente de otros TFs que los regulan. Sin embargo, se sabe que los TFs cooperan entre sí de forma dinámica. Es posible que las unidades de procesamiento de

información “reales“ estén basadas en genes que son regulados exclusivamente por 1 TF (regulones simples), o por la misma combinación de 2 o más TFs (regulones complejos). De ser cierto, esto sugeriría que el gradiente de complejidad observado en las GENSOR Units (Figura 3.10c) se debe a que el uso de regulones en el sentido más amplio agrega información poco relevante que disminuye el valor de conectividad y enmascara el proceso real que está siendo regulado por una combinación de TFs. No obstante, si se construyen GENSOR Units a partir de regulones simples y complejos, su distribución de conectividad muestra valores mucho más bajos (Figura 4.1), indicando que genes regulados por la misma combinatoria de TFs no están funcionalmente más relacionados que genes que son regulados por el mismo TF. Esto apoya que la coordinación de la regulación se da al nivel de TFs individuales y no al de las combinaciones presentes en un promotor. Las combinaciones de TFs podrían estar limitadas a coordinar la dinámica (qué genes se expresan en qué momento) más que la fisiología (cuál es la función conjunta de un grupo de genes) de los genes regulados. Dado que el concepto de GENSOR Units busca reflejar propiedades fisiológicas de los TFs, este resultado justifica que sean construidas a partir de los regulones más básicos. La combinatoria de TFs podrá explorarse más a fondo cuando existan datos sobre la forma en que los TFs cooperan, por ejemplo, si son sinérgicos o excluyentes.

Como se mencionó en la sección 3.3.2, es poco probable que el gradiente de complejidad se deba a falta de informacion, pues se observa tanto en las GENSOR Units más estudiadas, como en aquellas para las que se tiene poca información. Sin embargo, es posible que suceda lo opuesto, que en la literatura estén reportadas interacciones reguladoras TF-gen espúreas, identificadas por artefactos tecnológicos, cuyo efecto metabólico no relacionado con el resto del regulón disminuye el valor de conectividad de la GENSOR Unit en la que aparecen. Para explorar esta posibilidad se construyeron GENSOR Units utilizando sólo el set de interacciones de regulación reportadas en RegulonDB con evidencia fuerte y se repitió el análisis utilizando aquellas con evidencia débil. Se calculó la distribución de conectividad de ambos sets de GENSOR Units. Si, en efecto, el gradiente de conectividad se debe a interacciones reportadas de forma errónea, se esperaría que el set de GENSOR Units derivadas de evidencias fuertes tuviera mayor

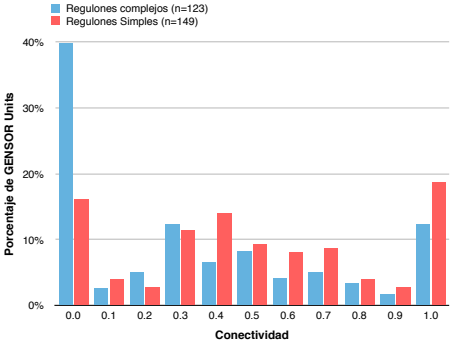


Figura 4.1: Comparación de conectividad de GENSOR Units de distintos tipos de regulones. La distribución de conectividad de regulones generales se muestra en rosa, la de regulones simples más regulones complejos se muestra en azul.

conectividad que el que incluye a todas las evidencias, sin embargo, este no es el caso (Figura 4.2). Por otro lado, la distribución de conectividad del set de GENSOR Units con evidencias débiles, aunque incluye un alto porcentaje de valores bajos, no es significativamente diferente de la distribución que incluye todas las interacciones (prueba Wilcoxon-Mann-Whitney; p-valor = 0.1819). Esto demuestra que no existe un sesgo cuantificable debido a la calidad de las interacciones de regulación utilizadas para construir GENSOR Units, y apoya que el gradiente de conectividad observado es una propiedad biológica y no un artefacto de la obtención de los datos.

Los términos de GO y las vías metabólicas son los conceptos más usados para describir unidades funcionales en bacterias. Frecuentemente son usados para obtener una visión general de la función de un grupo de genes. No obstante, algunas de sus definiciones de dónde comienza y termina un proceso biológico están basadas en razones históricas o que permiten una mejor organización de los datos, por ejemplo, definir el final de un proceso cuando aparece un metabolito muy común en la célula. Las interpretaciones derivadas de usar estos conceptos no reflejan la forma en que la célula "entiende" las funciones celulares [Bordbar et al., 2014]. Por ejemplo, 25 % de las GENSOR Units no incluyen genes presentes en vías metabólicas canónicas (Figura 4.3a).

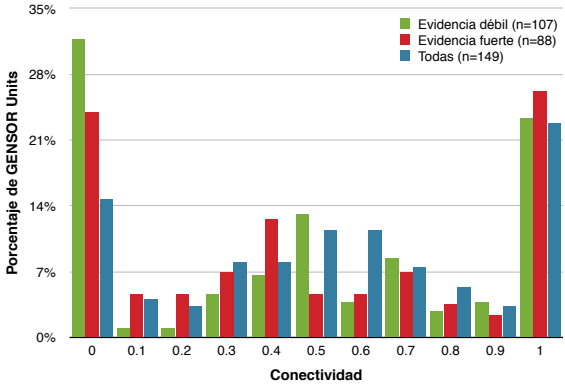


Figura 4.2: Connectividad de GENSOR Units según la evidencia de sus interacciones reguladoras.

En el caso de los términos de GO, la cantidad de términos enriquecidos en cada GENSOR Unit tiene un rango de 0 a 228. Cuarenta por ciento de las GENSOR Units tienen más de 50 términos enriquecidos. La estructura jerárquica de la ontología hace muy complicada la interpretación de este tipo de datos, pues elegir qué tanto un término padre y un término hijo en la ontología pertenecen a un mismo proceso es una decisión ambigua. Las GENSOR Units son el primer paso hacia un marco de referencia que integre transporte, señalización, regulación y metabolismo en un proceso con dirección que capture la lógica de los flujos de información que suceden en la célula. Las GENSOR Units hacen posible identificar relaciones entre genes de interés en su contexto metabólico, regulador y señalizador al mismo tiempo. Permiten responder preguntas como "¿un grupo de genes responde ante la misma señal?". Como herramienta conceptual, buscan facilitar la tarea de encontrar sentido biológico en datos provenientes de tecnologías de datos de expresión masiva, reflejando la manera en que la célula interpreta los cambios a los que está siendo sujeta.

La integración conceptual ha abierto nuevas líneas de investigación. Por ejemplo, el paradigma de homogeneidad funcional dentro de un regulón no es una propiedad general de la célula. Los valores de conectividad produjeron un gradiente que va de GENSOR units mono-

6	3		10
6	3		14
7	6		-3
7	2		2
8	1		-3
8	2		3
8	3		30
9	1		-2
9	2		16
9	2		33
10	2		-10
11	3		-2
11	2		3
12	7		5
12	10	GALACTITOLCAT-PWY(2)//	6
12	1	TRNA-CHARGING-PWY(1)//	27
13			8
13	3	PWY-4621(1)//	13
14			2
14	4	ASPARAGINESYN-PWY(1)//	6
14			15
15	2	ACETOACETATE-DEG-PWY(3)//	3
15			33
16	7	PWY0-1481(2)//	1
16			17
17	3	BETSYN-PWY(2)//	2
17			-2
17	2	PWY0-1586(1)//	13
17	7	CARNMET-PWY(4)//	25
18			12
19	5	PWY0-1309(2)//	0
19	5	NONOXIPENT-PWY(1)//	39
20			-8
20	28	PWY0-1546(1)//	-5
20			2
20	5	PWY-6823(3)//	6
20	3	CYANCAT-PWY(2)//	7
20			10
20	2	GLYCEROLMETAB-15 PWY(3)//	15
20	3	PWY0-1535(1)//	17
20			21
20	24	FERMENTATION-PWY(7)//	24
21	20	PWY0-1305(1)//	
21	13	PWY0-1305(2)//	23
22	4	GALACTITOLCAT-PWY(3)//	
22	5	PPGPPMET-PWY(1)//	18
23			9
23	6	PPGPPMET-PWY(1)//	14
23	3	GLYOXYLATE-BYPASS(2)//	15
23			40
24	3	BGALACT-PWY(1)//	37
24	17	LEUSYN-PWY(4)//	37
25			8
26	3	PWY0-1317(1	

Enriched GOs (n=113)	0.01	0.07	0.17	0.14	0.14

Percentag
7%
0%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19 20 22 27 32 42 65
Total Metabolic Pathways in Simple Regulons

28%

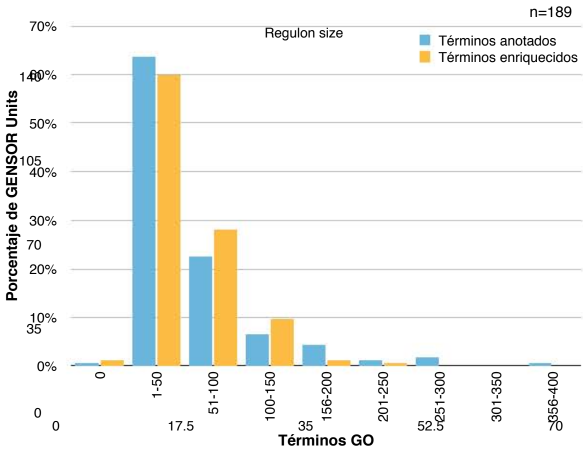
n=189

Table 2

	0	1-50	51-100	100-150	156-200	201-250
Present	1	120	42	12	8	
Enriched	2	113	53	18	2	
Términos anotados	0.01	0.63	0.22	0.06	0.04	
Términos enriquecidos	0.01	0.60	0.28	0.10	0.01	

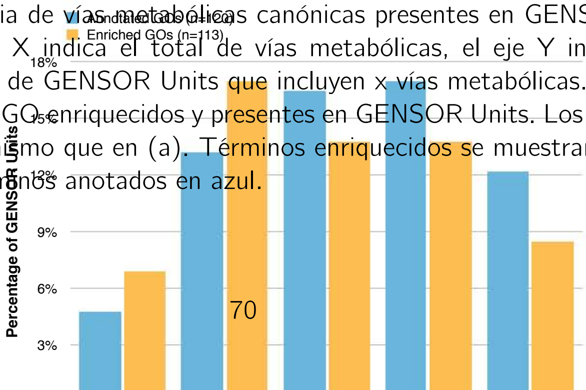
Porcentaje de GENSOR Units
21%
14%
7%
0%
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19 20 22 27 32 42 65
Vías Metabólicas Canónicas

(a) Vías Metabólicas



(b) Términos de GO

Figura 4.3: Vías metabólicas y GOs presentes en GENSOR Units. (a) Frecuencia de vías metabólicas canónicas presentes en GENSOR Units. El eje X indica el total de vías metabólicas, el eje Y indica el porcentaje de GENSOR Units que incluyen x vías metabólicas. (b) Términos de GO enriquecidos y presentes en GENSOR Units. Los ejes indican lo mismo que en (a). Términos enriquecidos se muestran en amarillo, términos anotados en azul.



temáticas a epistáticas. Es interesante notar que la mayor parte de GENSOR Units se encuentra en los extremos del gradiente de conectividad. Existen tres hipótesis no excluyentes que podrían explicar este comportamiento:

1. Diferencias en mecanismos de regulación producen distintas propiedades en la respuesta. Por ejemplo, los sistemas de dos componentes podrían ser responsables de respuestas globales con conectividades bajas.
2. Algunas funciones celulares requieren una respuesta más coordinada que otras. Utilización de fuentes de carbono y eflujo de toxinas requieren respuestas rápidas, mientras que formación de flagelo o ciclo celular son capacidades que requieren la presencia de un set más grande de señales estables.
3. TFs individuales regulan distintos sets de genes bajo distintas condiciones, creando subunidades dependientes de la condiciones ambientales dentro de una misma GENSOR Unit. GENSOR Units con valores altos de conectividad podrían estar constituidas de flujos metabólicos independientes que se activan bajo distintas condiciones.

La distintas posibilidades demuestran el reto de encontrar sentido biológico a partir de la integración de datos.

Ha sido propuesto que la coordinación de múltiples señales se encuentra embebida en una serie de circuitos de retroalimentación anidados donde las señales más generales controlan varios módulos y señales locales controlan el flujo metabólico dentro de un módulo [Chubukov et al., 2014]. En este escenario, los circuitos de retroalimentación serían una ocurrencia común. Circuitos pequeños y locales para respuestas simples y rápidas, y circuitos más grandes para repuestas centrales como división celular. Los segundos requerirían la presencia de una combinación de señales, procesadas en circuitos de retroalimentación más pequeños, en un momento específico. Esta hipótesis es soportada por las propiedades observadas en el set de GENSOR Units:

1. Los circuitos de retroalimentación son una propiedad general.
2. Existe un gradiente de modularidad en las GENSOR Units, reflejado en los valores de conectividad. De topologías dinámica-

mente autónomas a topologías con reacciones desconectadas que requieren de otras GENSOR Units para ser interpretadas.
3. Las GENSOR Units pueden ser unidas a partir de "moléculas reporteras" como son los efectores (Figura 3.17), que pueden servir como señal para otros programas de regulación.
4. Es posible unir GENSOR Units para crear un circuito de retroalimentación más grande (Figura 4.4).

Las GENSOR Units representan dos novedades metodológicas. En primer lugar, la integración automática de datos de distintas fuentes a través de una búsqueda exhaustiva que elimina sesgos hacia la función mejor conocida de cada regulador. Por ejemplo, rara vez se menciona en las descripciones del operón *lac* que LacY también es capaz de transportar melibiosa. El único sesgo inherente a la metodología usada es la disponibilidad de datos en las bases de datos utilizadas, sin embargo, las automatización permite actualizar las GENSOR Units con cada liberación de las bases de datos. La segunda novedad metodológica es la predicción de efectores únicamente a partir de la topología de las GENSOR Units. Aunque las validaciones experimentales son necesarias, la información proveída por las GENSOR Units puede reducir significativamente el número de efectores candidatos antes de diseñar un experimento.

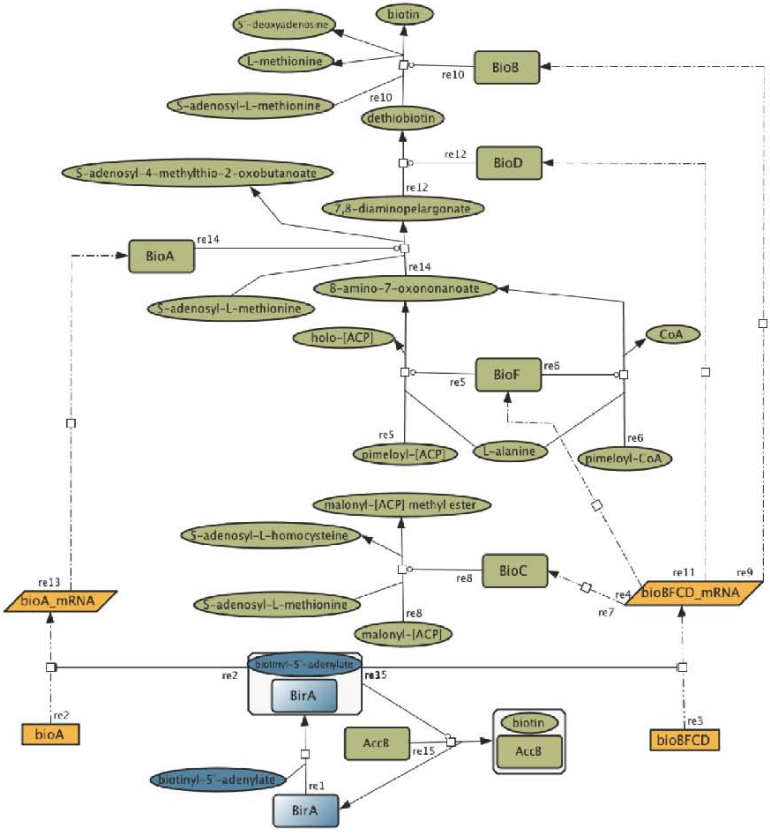


Figura 4.4: Retroalimentación uniendo GENSOR Units. BirA regula a genes encargados de la síntesis de biotina. Su conformación activa, BirA-bio-5'-AMP, promueve la formación del complejo AccB-biotina, el cual promueve el retorno de BirA a su estado inactivo.

Figure 50. BirA and AccB interactions form a feedback loop that includes two TFs. BirA regulat
biotin synthesis, and its active conformation, BirA-bio-5'-AMP, promotes AccB-biotin compl
formation, which in turn returns BirA to its inactive state.

Capítulo 5

Conclusiones

El presente trabajo puede resumirse en 7 mensajes principales:

- Las GENSOR Units son un modelo conceptual que permite describir el flujo de información que va de la detección de una señal a la generación de una respuesta apropiada, a través de distintos niveles de organización celular. Hasta este momento incluyen la regulación transcripcional y el metabolismo.
- La formalización de los elementos de las GENSOR Units y la automatización de su construcción ha permitido eliminar sesgos funcionales en su curación, realizar análisis guiados por los datos y acelerar significativamente su construcción. Por otro lado, dependen de la información presente en las bases de datos que se utilizan.
- Las GENSOR Units cuentan con un gradiente de complejidad fisiológico en la respuesta producida, contrario al paradigma de 1 TF-1 función que se asume para reguladores locales.
- Es necesario diseñar una clasificación más acertada de la actividad de los TFs, pues clasificarlos en locales y globales es una simplificación. En los TFs locales puede observarse un gradiente de complejidad tanto a nivel fisiológico como mecanístico, sin una correlación evidente entre ambos niveles.
- El nivel de integración de las GENSOR Units permite generar hipótesis y guiar el diseño de experimentos de validación de efectores y genes blanco de un TF.

- Las comparaciones entre GENSOR Units permiten identificar nuevos mecanismos de regulación indirectos y ampliar el repertorio de interacciones conocidas entre TFs.
- Existen muchos principios de diseño aún por descubrir en cuanto a la relación global de la regulación transcripcional con el metabolismo. Elucidar principios biológicos que gobiernan la cooperación entre TFs permitirá un mejor entendimiento de los procesos celulares como los interpreta la célula, en contraste con la forma en que los entendemos actualmente. Las GENSOR Units presentadas aquí son una herramienta útil para avanzar en esta dirección.

Capítulo 6

Perspectivas

6.1. Identificación de Circuitos de Retroalimentación faltantes

Para el 17 % de GENSOR Units con efectores fue imposible identificar circuitos de retroalimentación. Se propuso que estas GENSOR Units sí cuentan con ellos, sin embargo, el método automático no fue capaz de identificarlos porque no recaen en flujos metabólicos o utilizan moléculas comunes que fueron eliminadas del estudio. Para apoyar esta hipótesis y concluir de forma contundente que los circuitos de retroalimentación son una propiedad general de la dinámica de TFs individuales es necesario curar a mano estos casos e identificar los circuitos de retroalimentación.

6.2. Validación Experimental de Predicciones

Tras haber generado predicciones computacionales de efectores y genes blanco, es necesario validarlas experimentalmente. Un enfoque muy útil para hacerlo es un sistema libre de células [Pardee et al., 2014] donde se pueden introducir dos plásmidos: uno con el TF aunado a un promotor constitutivo, y el segundo con el promotor de un gen regulado por el TF más la secuencia codificante de la proteína

GFP. Los efectores predichos pueden validarse agregando el efector al medio y cuantificando la fluorescencia del sistema, que incrementará o disminuirá según el efecto del efector. Si el efector predicho no se une al TF, no habrá cambio en la fluorescencia. Para validar genes blanco es necesario utilizar el promotor del gen blanco como promotor *gfp*. Al agregar el efector del TF deben observarse cambios en la fluorescencia según el efecto del TF.

6.3. Relación entre modularidad y el ambiente

Las redes biológicas son modulares en el sentido de que son divisibles en subunidades que realizan funciones independientemente del resto de la red [Kashtan and Alon, 2005]. Esta propiedad varía entre cada red e incluso entre cada subunidad de la misma red. Los valores de conectividad también pueden ser interpretados como una medida de modularidad de las respuestas mediadas por TFs. Las GENSOR Units con conectividades altas son subsistemas cuasi-independientes que dan pie a una capacidad definida. Durante la evolución de una red, la modularidad es seleccionada en ambientes cambiantes, mientras que condiciones constantes producen topologías no modulares [Kashtan and Alon, 2005]. Ha sido demostrado que en bacterias existe una correlación entre ambientes cambiantes y la modularidad de su red metabólica: bacterias que viven en ambientes que rara vez fluctúan, como simbiontes obligados, tienen redes metabólicas altamente no-modulares y viceversa [Parter et al., 2007]. Será interesante investigar si esta propiedad puede aplicarse también a submódulos de la red, de forma que las GENSOR Units con alta conectividad sean responsables de procesar señales más fluctuantes que aquellas detectadas por GENSOR Units con menor conectividad.

6.4. Predicción de transportadores

La base de datos TDCB *Transporter Classification Database* contiene más de 10000 proteínas de transporte, las familias a las que pertenecen y los metabolitos que transportan. Para un subset de ellas en *E. coli*

no se conoce el metabolito transportado pero se conoce su regulación, por lo tanto, aparecen en GENSOR Units como enzimas con función desconocida. Utilizando la topología de las GENSOR Units puede ser posible proponer candidatos para metabolitos transportados, sobre todo para aquellas GENSOR Units con valores de conectividad altos que involucran un solo flujo metabólico.

6.5. Modelado matemático de GENSOR Units

El nivel de detalle de las GENSOR Units permite modelarlas matemáticamente. Enfoques utilizados por Michael Savageau basados en subsistemas no lineales [Lomnitz and Savageau, 2013] permiten identificar rangos de concentraciones de cada elemento de una GENSOR Unit en los cuales la GENSOR Unit fisiológicamente puede existir como una unidad. Validar experimentalmente estas concentraciones permitiría probar que las GENSOR Units funcionan como una unidad fisiológica.

6.6. Comparación de GENSOR Units entre especies

Aunque la información organizada en bases de datos sobre regulación en otras bacterias no es tan abundante como en *E. coli*, se utilizaron los datos disponibles de *Salmonella typhimurium serovar enterica* para ensamblar GENSOR Units de este organismo. Se sabe que la secuencias reguladoras evolucionan más rápido que las secuencias codificantes [Borneman et al., 2007], por lo que se esperaría que las topologias de GENSOR Units entre especies varíen más que las secuencias codificantes. Será interesante evaluar las similitudes y diferencias entre estos organismos filogenéticamente cercanos para identificar genes con la misma función en ambos organismos, cuyo contexto de regulación es distinto y observar propiedades que emergen de estas diferencias. Por ejemplo, genes cuyo contexto de regulación los haga participar en procesos de patogénesis. Conforme más información se haga disponible para otros organismos será interesante identificar relaciones evolutivas utilizando topologías de GENSOR Units.

Capítulo 7

Métodos

7.1. Construcción de GENSOR Units

Los elementos e interacciones de las GENSOR Units fueron recuperados automáticamente usando códigos *ad hoc* en lenguaje Perl. Las conformaciones activas e inactivas, efectores y genes regulados de 189 TFs con evidencia experimental de su actividad reguladora fueron obtenidas automáticamente de datasets de RegulonDB (`http://regulondb.ccg.unam.mx/`). El API *Perlcyc* del software *Pathway Tools* fue utilizado para obtener automáticamente productos de genes, reacciones catalizadas, substratos, productos y direccionalidad de las reacciones, complejos protéicos heteromultiméricos y sus subunidades. Fueron eliminadas del análisis 9 moléculas ubicuas en la célula: protones, AMP, ADP, ATP, agua, NAD, NADH, NADPH y fosfato. Todos los datos recuperados fueron organizados en 5 tablas relacionales usadas posteriormente para generar un archivo SBML para cada GENSOR Units usando el software CellDesigner 4.4 [Funahashi et al., 2008]. La red resultante fue inspeccionada manualmente y los componentes disponibles en cada GENSOR Unit fueron identificados. Las reacciones secundarias fueron agregadas identificando pares de reacciones en cada GENSOR Unit que pertenecen a la misma vía metabólica. Esto se hizo utilizando el software *Pathway Tools* y su API *Perlcyc*. El grupo de reacciones en la vía metabólica que conectan al par de reacciones originales en la GENSOR Unit fueron agregadas a la GENSOR Unit como una sola reacción secundaria, siempre

y cuando la direccionalidad de todas las reacciones individuales fuera en el mismo sentido. Las redes resultantes e información sobre cada elemento e interacción de las GENSOR Units fueron agregadas a RegulonDB (`http://regulondb.ccg.unam.mx/central_panel_menu/integrated_views_and_tools/gensor_unit_groups`).

7.2. Propiedades de GENSOR Units

Todas las propiedades fueron identificadas automáticamente utilizando códigos *ad hoc* escritos en lenguaje Perl y disponibles en GitHub. (`https://github.com/dledezma/gensor_units/tree/master/perl_scripts`).

7.2.1. Flujos metabólicos en GENSOR Units

Se consideró "flujo metabólico" a la transformación química de un metabolito en otro a través de una o más reacciones enzimáticas. Un flujo metabólico de dos o más reacciones enzimáticas es creado cuando el substrato de una reacción es substrato o producto de otra. La direccionalidad de las reacciones fue considerada para inferir la direccionalidad del flujo metabólico.

7.2.2. Circuitos de Retroalimentación

Se calcularon automáticamente todos los flujos metabólicos presentes en cada GENSOR Unit. Se consideró la existencia de un circuito de retroalimentación en una GENSOR Unit cuando al menos uno de los efectores reportados se encontró en un flujo metabólico. Los sistemas de dos componentes fueron omitidos del estudio pues su molécula efectora está anotada como fosfato y esta molécula es ubicua en la célula.

7.2.3. Conectividad

La conectividad fue calculada en todas sus instancias con la siguiente fórmula:

$$C = \frac{Ec}{Et + (Mft - 1)}$$

Donde Ec es el total de enzimas conectadas en la GENSOR Unit; Et es el total de enzimas y MFt es el total de flujos metabólicos independientes en una GENSOR Unit. Dos enzimas se consideraron "conectadas" si cualquiera de las reacciones que catalizan se encontraban en el mismo flujo metabólico. En otras palabras, dos reacciones se consideraron "conectadas" si compartían al menos un substrato o producto. Un flujo metabólico independiente es aquel cuyas enzimas están conectadas entre sí, pero desconectadas del resto del componente de respuesta de la GENSOR Unit. Por ejemplo, dos flujos metabólicos independientes cuentan cada uno con reacciones conectadas entre sí, pero las reacciones del primer grupo y las del segundo no están conectadas. En todas las gráficas de distribución de conectividad se omitieron los grupos que incluían menos de dos reacciones enzimáticas para evitar valores de 0 no informativos.

7.2.3.1. Conectividad de Vías Metabólicas Canónicas

Las vías metabólicas canónicas y los genes involucrados en cada una de ellas fueron obtenidos de Ecocyc a través de la herramienta *Pathway Tools* y su API *Perlcyc*. Los genes de 362 vías metabólicas canónicas se utilizaron en una versión modificada del algoritmo de construcción de GENSOR Units (disponible en GitHub). El resto de la tubería de datos se utilizó en estas GENSOR Units de vías metabólicas para calcular su conectividad. La regulación de los genes no fue considerada.

7.2.3.2. Conectividad de GO

Se obtuvieron los genes asociados a 311 términos de GO utilizando datasets disponibles en la página del *Gene Ontology Consortium* (`http://www.geneontology.org/page/download-annotations`). Se utilizó la misma tubería de datos que para las vías metabólicas canónicas y se graficó la distribución de conectividad.

7.2.3.3. Conectividad de GENSOR Units según su efecto regulador

Las interacciones reguladoras de cada GENSOR Unit se dividieron según el efecto regulador del TF (activación/represión). Las interacciones duales y con efecto desconocido fueron consideradas en ambos grupos. Para cada grupo (activación/represión) se utilizó la misma tubería de datos que para el set original de GENSOR Units.

7.2.3.4. Conectividad de GENSOR Units según su evidencia en RegulonDB

Las interacciones reguladoras TF-TU incluidas en cada GENSOR Unit se dividieron según la evidencia de la interacción en RegulonDB en dos grupos: evidencia fuerte y evidencia débil. Las evidencias fueron clasificadas según el mismo criterio desarrollado por RegulonDB [Weiss et al., 2013]. Se calculó la distribución de conectividad de ambos sets utilizando la tubería de datos modificada que se desarrolló para calcular la conectividad de las vías metabólicas. Para que la comparación con el resto de las GENSOR Units fuera válida, se repitió el análisis del set original de GENSOR Units utilizando la tubería modificada.

7.2.3.5. Conectividad de GENSOR Units de regulones complejos

Se utilizó el dataset TF-gen de RegulonDB y se extrajeron automáticamente todas las combinaciones de TFs que se encuentran en cada promotor de la TRN. Se agruparon los genes regulados por la misma combinatoria de 1 (regulones simples) o más TFs (regulones complejos) para dar lugar a 766 grupos que incluyen a los 7 reguladores globales (ArcA, CRP, Fis, FNR, HNS, IHF y Lrp). Se utilizó la tubería de datos original para calcular la distribución de conectividad.

7.2.4. Autonomía por Complejos

Se obtuvieron automáticamente todos los monómeros de complejos de cada GENSOR Unit y el total de aquellos directamente regulados por el TF. Estos valores se utilizaron para calcular el valor de autonomía por complejos utilizando la siguiente fórmula:

$$A = \frac{MC_{regulados}}{MC_{totales}}$$

Donde $MC_{regulados}$ es el total de monómeros que pertenecen a complejos protéicos heteromultiméricos y son directamente regulados por el TF principal de la GENSOR Unit. $MC_{totales}$ es el total de monómeros que pertenecen a complejos protéicos heteromultiméricos en la GENSOR Unit. El valor se calculó para cada GENSOR Unit y se graficó la distribución.

7.3. Predicciones

7.3.1. Predicción de efectores

7.3.1.1. Posición del efector

Setenta y ocho GENSOR Units con efectores conocidos fueron analizadas (Cuadro 7.1). La posición de cada efector en la vía metabólica regulada fue recuperada automáticamente usando la siguiente clasificación:

- Substrato/Producto. Efectores que tienen rol de substrato o producto en una sola reacción en la GENSOR Unit. Se agrupó la clasificación de substratos y producto para eliminar ambigüedad derivada de reacciones reversibles.
- Intermediario. Efectores que tienen rol de substrato o producto en dos o más reacciones enzimáticas en la GENSOR Unit. Efectores que son producto de reacciones de transporte fueron considerados intermediarios para ajustarse a la clasificación propuesta por Savageau [Savageau, 1976].

7.3.1.2. Selección de efectores hipotéticos

Las 15 GENSOR Units con conectividad de 1 sin efectores reportados en RegulonDB fueron utilizadas para predecir efectores a partir de su topología. Todos los metabolitos en las 15 GENSOR Units fueron clasificados como substrato/producto o intermediario utilizando el criterio mencionado en la sección anterior. Los metabolitos intermediarios resultantes se buscaron en la literatura para identificar evidencia experimental de función de ligando. Los metabolitos con evidencia

que soporta la predicción fueron considerados efectores hipotéticos. En caso de que no se encontrara información para apoyar alguna predicción, todos los metabolitos intermedios fueron reportados como candidatos. Se buscaron dominos de pegado a ligando en la secuencia de los 15 TFs utilizando herramientas de las bases de datos *NCBI Conserved Domain* y Pfam [Finn et al., 2016].

7.3.2. Predicción de genes blanco

Se recuperaron todas las reacciones individuales que forman parte de reacciones secundarias presentes en el set original de 189 GENSOR Units. Se identificaron los genes que codifican para las enzimas que catalizan las reacciones individuales. Estos genes fueron propuestos como candidatos a genes blanco del TF central de la GENSOR Unit en que se encontraron las reacciones secundarias. Los genes resultantes se compararon con los datasets de RegulonDB de predicciones computacionales de genes blanco derivados de TractorDB y MatrixScan (`http://regulondb.ccg.unam.mx/menu/download/computational_predictions/index.jsp`). Se reportaron los genes que fueron predichos a partir de las GENSOR Units y al menos un método más.

7.4. Análisis Estadísticos

Todas las pruebas Wilcoxon-Mann-Whitney se realizaron en el software R v3.3.2 usando una suma Wilcoxon de rangos con corrección de continuidad (*Wilcoxon rank sum with continuity correction*).

7.5. Unión de GENSOR Units

7.5.1. Red de efectores

Se identificaron todas las GENSOR Units que incluyen reacciones que producen su efector o el efector de otra GENSOR Unit. La cascada de AtoC fue identificada manualmente usando AtoC como punto de partida. Todas las GENSOR Units cuyos TFs se unen a metabolitos producidos en la GENSOR Unit de AtoC fueron incluídas en la cascada y posteriormente fueron utilizadas como puntos de partida. Este

algoritmo se repitió de forma recursiva hasta que no existieron efectores presentes en las últimas GENSOR Units de la cascada. Ambas redes se dibujaron utilizando el software Cytoscape v3.1.1 [Shannon et al., 2003].

7.5.2. Grupos de GENSOR Units

Todos los elementos comunes entre GENSOR Units fueron identificados automáticamente de la siguiente forma:

- **Unidades Transcripcionales**. Para cada GENSOR Unit se calculó la fracción de las TUs que incluye que comparte con el resto de las GENSOR Units, generando una matriz de adyacencia que fue usada para generar una red. El peso de cada arista en la red refleja la fracción de TUs compartidas.
- **Metabolitos**. Al igual que en las TUs compartidas, para cada GENSOR Unit se calculó la fracción de sus metabolitos que aparecen en cada una de las otras GENSOR Units, generando una matriz de adyacencia que fue usada para generar una red. El peso de cada arista en la red refleja la fracción de metabolitos compartidos.
- **Reacciones Secundarias**. Se identificaron las reacciones individuales que forman parte de reacciones secundarias en cada GENSOR Unit. Posteriormente se obtuvieron los genes que codifican para las enzimas que catalizan las reacciones individuales y se identificaron las GENSOR Units donde se regula directamente a estos genes. Con esta información se construyó una matriz de adyacencia binaria donde 1 significa que la GENSOR Unit X incluye reacciones que forman parte de reacciones secundarias en la GENSOR Unit Y. La matriz se utilizó para generar una red cuyas aristas no tienen peso pero sí dirección.
- **Principio y fin de flujos metabólicos**. Se identificaron todos los flujos metabólicos independientes presentes en cada GENSOR Unit. Se tomaron los metabolitos iniciales y finales de cada flujo metabólico y se identificaron todas las GENSOR Units que completaran relaciones "metabolito final en GENSOR Unit 1/metabolito inicial en GENSOR Unit 2" o "metabolito inicial en GENSOR Unit 1/metabolito final en GENSOR Unit 2". Se generó una matriz de adyacencia binaria donde 1 significa que sí

existe una relación de este tipo entre un par de GENSOR Units y 0 significa que no existe tal. La matriz se utilizó para construir una red cuyas aristas no cuentan con peso.

- **Monómeros de complejos protéicos heteromultiméricos**. Se identificaron todos los monómeros de complejos que pertenecen a una GENSOR Unit y no son regulados directamente por el TF. Posteriormente se identificaron todas las GENSOR Units donde dichos monómeros son regulados directamente. Se construyó una matriz de adyacencia donde, para cada par de GENSOR Units (X,Y), se indicó la fracción de monómeros no regulados directamente por la GENSOR Unit X que sí regula directamente la GENSOR Unit Y. La matriz fue utilizada para generar una red donde el peso de las aristas representa esta fracción.

Todas las redes fueron construidas utilizando el software Cytoscape v3.1.1 [Shannon et al., 2003]. Todas las redes, a excepción de la red de flujos metabólicos sucesivos, son direccionadas. Los valores entre cada par de GENSOR Units de las 5 matrices se sumaron, en total se sumaron 10 valores ([X,Y + Y,X] en las 5 matrices) y se obtuvo una matriz diagonal con pesos totales que se llamaron valores de similitud. Esta última matriz se utilizó para construir un mapa de calor utilizando el software R v3.3.2. El mapa de calor refleja el valor de similitud entre cada par de GENSOR Units y agrupa a las GENSOR Units más similares utilizando un clustering jerárquico. Los valores de corte para definir cada grupo se realizaron de manera manual buscando maximizar la similitud entre miembros del grupo.

7.6. Disponibilidad de datos y códigos generados

Todos los datos generados, los códigos utilizados para los análisis y las tuberías de datos utilizadas se encuentran disponibles en GutHub (`https://github.com/dledezma/gensor_units/`). Las 189 GENSOR Units en formato texto también están disponibles en GitHub (`https://github.com/dledezma/gensor_units/tree/master/datasets`), la versión gráfica está disponible en RegulonDB (`http://regulondb.ccg.unam.mx/central_panel_menu/integrated_views_and_tools/gensor_unit_groups`).

Apéndice

Cuadro 7.1: Posición de efectores en GENSOR Units con retroalimentación

GENSOR Unit	Efector	Posición del Efector
AlaS	alanine	intermediario
AllR	glyoxylate	intermediario
AllS	allantoin	intermediario
AlsR	allose	intermediario
AraC	arabinose	intermediario
ArgP	arginine	intermediario
ArgP	lysine	intermediario
ArgR	arginine	intermediario
ArsR	arsenite	intermediario
AsnC	asparagine	intermediario
BetI	choline	intermediario
ChbR	N-monoacetylchitobiose 6'-phosphate	intermediario
Cra	fructose 1,6-biphosphate	intermediario
Cra	fructose 1-phosphate	intermediario
CueR	Cu(I)	intermediario
CynR	cyanate	intermediario
CysB	acetylserine	intermediario
CysB	sulfide	intermediario
CysB	thiosulfate	intermediario
CytR	cytidine	intermediario
DeoR	2-deoxy-D-ribose 5-phosphate	intermediario
DhaR	DhaK	intermediario
DsdC	serine	intermediario

Continuación de cuadro 7.1

GENSOR Unit	Efector	Posición del Efector
FhlA	formate	intermediario
FucR	fuculose-1-P	intermediario
Fur	Fe+2	intermediario
Fur	Mn(II)	intermediario
GalR	galactose	intermediario
GalS	galactose	intermediario
GcvA	glycine	intermediario
GlcC	glycolate	intermediario
GlpR	glycerol-3-phosphate	intermediario
GntR	gluconate	intermediario
GutR	gulitol	substrato/producto
HipB	HipA	intermediario
IclR	glyoxylate	intermediario
IdnR	5-ketogluconate	intermediario
IdnR	idonate	intermediario
IlvY	alpha-acetolactate	intermediario
LacI	allolactose	intermediario
LldR	lactate	intermediario
LsrR	AI-2	intermediario
MalT	MalK	intermediario
MalT	maltotriose	intermediario
MelR	melibiose	intermediario
MetJ	SAM	intermediario
MetR	homo-cys	intermediario
MhpR	2,3-DHP	intermediario
MhpR	3HPP	intermediario
MntR	Mn(II)	intermediario
ModE	molybdate	intermediario
MqsA	MqsR	intermediario
MurR	MurNAc-6-P	intermediario
NanR	N-acetylneuraminate	intermediario
NhaR	Sodium	intermediario
NikR	nickel	intermediario
NrdR	dATP	intermediario
PdhR	pyruvate	intermediario
PrpR	(2S,3S)-2-methylcitrate	intermediario
RcnR	cobalt ion	intermediario
RcnR	nickel	intermediario
RelB	RelE	intermediario
RhaS	rhamnose	intermediario
RutR	thymine	intermediario
RutR	uracil	intermediario

Continuación de cuadro 7.1

GENSOR Unit	Efector	Posición del Efector
TreR	alpha,alpha-trehalose 6-phosphate	intermediario
TreR	trehalose	substrato/producto
TrpR	tryptophan	intermediario
TyrR	phenylalanine	intermediario
TyrR	tryptophan	intermediario
TyrR	tyrosine	intermediario
UxuR	fructuronate	intermediario
XapR	xanthosine	intermediario
XylR	xylose	intermediario
YqjI	Fe+2	intermediario
ZntR	Zinc	intermediario
Zur	Zinc	intermediario

frontiers in Microbiology

ORIGINAL RESEARCH
published: 03 August 2017
doi: 10.3389/fmicb.2017.01466

Genome-Wide Mapping of Transcriptional Regulation and Metabolism Describes Information-Processing Units in *Escherichia coli*

Daniela Ledezma-Tejeida, Cecilia Ishida and Julio Collado-Vides**

Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Mexico

OPEN ACCESS

Edited by:
Dionysios A. Antonopoulos, Argonne National Laboratory (DOE), United States

Reviewed by:
*Dave Siak-Wei Ow, Bioprocessing Technology Institute (A*STAR), Singapore*
Alberto Marin-Sanguino, Technische Universität München, Germany

****Correspondence:***
Daniela Ledezma-Tejeida
dledezma@lcg.unam.mx
Julio Collado-Vides
collado@ccg.unam.mx

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *03 March 2017*
Accepted: *20 July 2017*
Published: *03 August 2017*

Citation:
Ledezma-Tejeida D, Ishida C and Collado-Vides J (2017) Genome-Wide Mapping of Transcriptional Regulation and Metabolism Describes Information-Processing Units in Escherichia coli. Front. Microbiol. 8:1466. doi: 10.3389/fmicb.2017.01466

In the face of changes in their environment, bacteria adjust gene expression levels and produce appropriate responses. The individual layers of this process have been widely studied: the transcriptional regulatory network describes the regulatory interactions that produce changes in the metabolic network, both of which are coordinated by the signaling network, but the interplay between them has never been described in a systematic fashion. Here, we formalize the process of detection and processing of environmental information mediated by individual transcription factors (TFs), utilizing a concept termed genetic sensory response units (GENSOR units), which are composed of four components: (1) a signal, (2) signal transduction, (3) genetic switch, and (4) a response. We used experimentally validated data sets from two databases to assemble a GENSOR unit for each of the 189 local TFs of *Escherichia coli* K-12 contained in the RegulonDB database. Further analysis suggested that feedback is a common occurrence in signal processing, and there is a gradient of functional complexity in the response mediated by each TF, as opposed to a one regulator/one pathway rule. Finally, we provide examples of other GENSOR unit applications, such as hypothesis generation, detailed description of cellular decision making, and elucidation of indirect regulatory mechanisms.

Keywords: data integration, networks, transcriptional regulation, effector prediction, metabolism, genotype-to-phenotype mapping, information flow

INTRODUCTION

Jacob and Monod outlined the relevance of coupling between regulation and metabolism in their discovery of transcriptional regulators (Pardee et al., 1959). They discovered LacI, a protein later termed a transcription factor (TF), that binds to the *lac* operon promoter and represses its expression unless lactose is available. Their model of regulatory activity stated that TFs bind to signaling molecules called effectors, which promote changes in the expression of genes involved in

Abbreviations: CCR, carbon catabolite repression; GENSOR unit, genetic sensory response unit; TF, transcription factor; TRN, transcriptional regulatory network.

the processing of said molecules. They explained how the cell efficiently manages its resources by only producing specific enzymes when environmental conditions make them necessary, and this is still considered the paradigm for transcriptional regulation: co-regulated genes are assumed to be involved in the same biological process.

Currently, 189 local TFs are listed in RegulonDB, the largest database of transcriptional regulation in *Escherichia coli* K-12 (Gama-Castro et al., 2016), but a genome-scale description of the functional effects of their regulatory activities is still lacking. Previous formalisms have analyzed properties of gene regulation through genetic circuits. Efforts have spanned dynamical modeling (Thomas and D'Ari, 1990) to identifying general network properties (Savageau, 1976, 2001; Kauffman, 1993; Gerosa and Sauer, 2011), but none of them has studied the complete set of regulatory interactions in their functional context.

In the current genomic era, in which "we are drowning in information but starved for knowledge" (Naisbitt, 1984), there is a need for concepts that (a) integrate numerous and different types of molecules and their interactions, (b) reflect biological properties of the cooperation between elements, and (c) can be applied on a small or large scale (Hyduke and Palsson, 2010). Great strides have been made in network analysis to understand how cellular behavior arises from interacting molecules (Ravasz et al., 2002; Shen-Orr et al., 2002; Balazsi et al., 2005). However, the best-studied networks tend to focus on individual layers of interactions, such as TF–gene interactions (Gama-Castro et al., 2016), metabolic reactions (Forster et al., 2003), and signaling pathways (Papin and Palsson, 2004; Papin et al., 2005), which portray an incomplete vision of the information that promotes phenotypes. The goal is to integrate different layers and obtain a thorough picture of the way that functions emerge from combinations of individual mechanisms. This poses a methodological challenge, since some networks are compact and detailed (Alon et al., 1999; Berthoumieux et al., 2014) and others are large and less precise (Karr et al., 2012; Brooks et al., 2014), making it difficult to integrate this information into a single framework that can be used to generate new knowledge. Functional descriptions of the integration also require conceptual improvements. Gene ontologies (GOs) have been the reference for gene classification into biological processes for the past 16 years, but they were conceived to describe individual components rather than the interactions among them (The Gene Ontology Consortium, 2000).

Here, we formalized the process of signal detection to the outset of a functional response, mediated by an individual TF, into four components: (1) signal, (2) conversion of signal into the effector, (3) genetic switch, and (4) response. The integration product of the four components is termed a genetic sensory response unit (GENSOR unit). Ideally, GENSOR units describe the information that flows through different layers of cellular organization to produce an appropriate response (Supplementary Figure S1). We assembled a GENSOR unit for each of the 189 local TFs present in RegulonDB by integrating experimentally validated data from the literature using simple regulons as starting points. Further analysis of the GENSOR unit set showed that less than a quarter of the TFs regulate genes that belong to the same metabolic flux, but feedback is a common occurrence. A gradient of response complexity can be observed and is partially explained by the regulatory effect of the corresponding TF. Beyond the biological insights presented here, we provide the set of GENSOR units as a standardized framework for small- and large-scale analyses of the interplay between transcriptional regulation and metabolism. Last, we show examples of practical applications, such as hypothesis generation, detailed description of cellular decision making, and elucidation of indirect regulatory mechanisms.

RESULTS

GENSOR Units of Local TFs in *E. coli* K-12

GENSOR units are integrative networks that describe in detail the information flow that goes along the molecular circuitry from signal detection to generation of a response (**Figure 1A**). They formalize the signal-response process in four components. (1) Signal: the molecule that begins the information flow by reflecting a change in the external or internal environment. (2) Signal transduction: the conversion of the signal into a molecule that will prompt a regulator. In the case of TFs, it refers to the conversion of the signal into the effector molecule that binds to the TF. For example, the signal lactose is transformed into allolactose, the molecule that binds to the regulator LacI. (3) Genetic switch: the repression/activation of the specific set of genes needed to contend with the signaled change. (4) Response: the effect of the gene products, which together produce a new phenotype, a change in metabolism, or signal other regulators.

To assemble a GENSOR unit for each of the 189 local TFs, we used a data-driven approach. We automatically retrieved from the RegulonDB database (Gama-Castro et al., 2016) the genes directly regulated by a TF (regulon), its known effectors, its active/inactive conformations, and the TF's regulatory effect over the regulated genes. From the EcoCyc database (Keseler et al., 2013), we automatically obtained the gene products of the regulated genes, the reactions catalyzed by the gene products, the substrates and products of the catalyzed reactions, and the protein complexes in which gene products participate. It is important to note that the only heuristics included in our method is to include no more genes than those directly regulated by the TF. An exhaustive search is performed to retrieve all the available elements that have been experimentally validated, so all the interactions in GENSOR units have been proved to occur naturally.

According to the Jacob and Monod paradigm, each TF will directly regulate genes that together give rise to a cellular capacity, for example, uptake of lactose as the carbon source, production of osmoprotectants, or flagellar assembly. From this assumption, it follows that the four components of the GENSOR unit can be identified within the retrieved set of elements and their interactions (**Figure 1B**). Eighty GENSOR units included a known effector, and it was possible

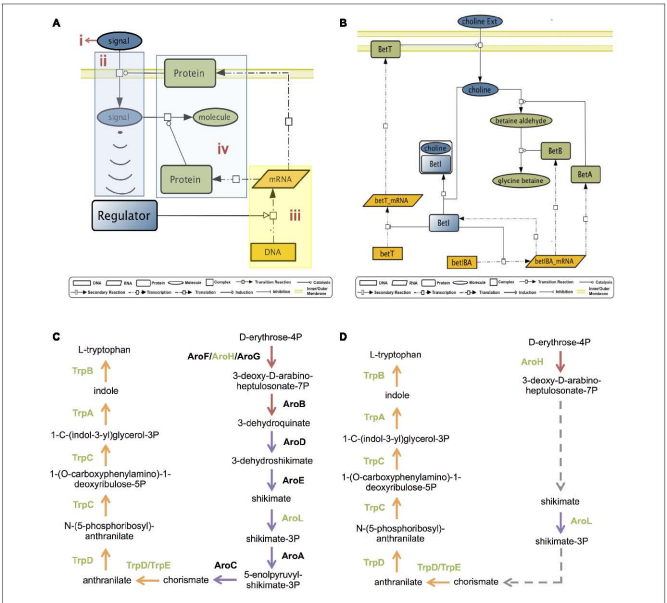


FIGURE 1 | GENSOR units. **(A)** The general GENSOR unit concept includes four components: (i) signal, (ii) signal transduction (shown in blue), (iii) a genetic switch (shown in yellow), and (iv) the response (shown in green). The concept can be applied to any regulator that produces a genetic switch. **(B)** The BetI GENSOR unit. This is an example of a GENSOR unit based on a bacterial TF; color coding for components is as depicted for panel **(A)**. Signal transduction involves transport of choline through the membrane, a genetic switch causes the repression of two transcription units, *betT* and *betIBA*, and the response involves the transformation of choline into glycine betaine. From the higher perspective provided, it is possible to infer that mechanistically in presence of choline, BetI will cease to repress the necessary enzymes for its transport and transformation into glycine betaine. Physiologically, BetI negatively regulates the expression of genes involved in the response to osmotic stress in the absence of choline. The presence of choline induces the expression of the genes required for transport and conversion of choline to glycine betaine, an osmoprotectant. **(C)** Pathways involved in production of L-tryptophan from D-erythrose-4P. The 3-dehydroquinate biosynthesis I pathway is shown in red; the chorismate biosynthesis from 3-dehydroquinate pathway is shown in purple; the L-tryptophan biosynthesis pathway is shown in orange. Enzymes that catalyze reactions are shown beside the reactions. Enzymes regulated by TrpR are shown in green. **(D)** In the TrpR GENSOR unit, the three reactions catalyzed by AroB, AroD, and AroE are summarized in one secondary reaction without indication of the involved enzymes or intermediate metabolites. Another secondary reaction includes reactions catalyzed by AroA and AroC.

to identify their four components. The resulting integrative networks described in detail the steps from signal detection to metabolic impact, which were summarized in a short sentence (**Figure 1B**). For the remaining 109 GENSOR units, only genetic switches and response information was pinpointed, reflecting the physiological effect of the TF. It is relevant to note that our previous assumption excluded three occurrences: constitutive enzymes, cooperation between

TFs, and yet-unknown regulatory interactions. Any of them could account for genes involved in the biological process depicted by a GENSOR unit, but they are not present because they are not directly regulated by the TF. In order to enrich GENSOR units with known functional interactions, we considered canonical metabolic pathways. Links were added between metabolites already present in a GENSOR unit if they belonged to the same metabolic pathway and the directionality of the pathway permitted metabolic flux between them. For example, TrpR regulates seven enzymes of the 13 enzymes involved in the production of L-tryptophan from D-erythrose-4P (**Figure 1C**). Considering only TrpR's direct targets, it would appear that the reactions catalyzed by AroH and AroL are not related. However, the metabolic pathways of chorismate biosynthesis from 3-dehydroquinate (**Figure 1C**, purple arrows) and 3-dehydroquinate biosynthesis I (**Figure 1C**, red arrows) indicate the existence of a metabolic flux that converts 3-deoxy-D-arabino-heptulosonate-7P into shikimate and then shikimate-3P into chorismate, respectively. In the TrpR GENSOR unit, the reactions catalyzed by AroB, AroD, and AroE are summarized in one link, termed a secondary reaction, without indication of the involved enzymes or intermediate metabolites (**Figure 1D**, dashed lines). The same happens for reactions catalyzed by AroA and AroC. A total of 144 secondary reactions were added to 48 GENSOR units (Supplementary Figure S2). The 189 GENSOR units are publicly available in the RegulonDB database[1].

Feedback Is a Common Occurrence in GENSOR Units

The next step was to identify general properties of GENSOR units. The regulation model by Jacob and Monod (Pardee et al., 1959; Monod et al., 1963) includes the ability of TFs to autotune their activity according to cellular needs and is explained by a direct effect of the regulated response on effector availability. We considered the 80 GENSOR units with known effectors and looked for the presence of reactions that were part of the response while also having a role in the conversion of the signal into the effector. Sixty-five of 78 GENSOR units (83%) included this type of feedback. Two GENSOR units were excluded from the analysis because they did not include regulated enzymes and therefore had no reactions. The simplest feedback consisted of an effector transport through the membrane, and it is interesting that TFs with two or more known effectors had as many feedback loops. The AllS GENSOR unit was the only one whose feedback involved a secondary reaction, suggesting that feedback loops underlie the most basic layer of bacterial decision-making by relying on a single TF switch that senses and responds to these changes. It is possible that feedback loops are a general property of GENSOR units and at least one exists in each, but our method was unable to identify the remaining 17% because they did not rely on metabolic fluxes or use common metabolites that our methodology excludes from the analysis, like DnaA, whose effector is ATP. We expect to identify more feedback loops as new information is included in GENSOR units.

[1]http://regulondb.ccg.unam.mx/

Complexity of TF Responses Covers a Continuum

We would expect that all of the genes directly regulated by a TF are necessary and sufficient to give rise to a cellular capacity. In fact, bacterial regulators are often referred to as "regulator of X metabolism." We used the complete set of GENSOR units to obtain a genome-wide distribution of the functional homogeneity of TF responses. In order to exploit the interactions between elements rather than the individual functions of genes, we developed a metric termed connectivity. Connectivity considers the number of individual metabolic fluxes present in a GENSOR unit (and therefore regulated by an individual TF). A metabolic flux is defined as a consecutive set of reactions where the product of a reaction is the reactant of the next one, for example, as in a metabolic pathway (see Materials and Methods). Enzymes that catalyze individual reactions in a metabolic flux are considered "connected," because we assume that enzymes present in the same metabolic flux will be part of the same functional process. Connectivity is calculated as the ratio of connected enzymes (Ec) to total enzymes (Et). If all enzymes are indeed sufficient and necessary for a functional process, we would expect all the regulated enzymes to be connected, and so we penalize deviations by calculating the total independent metabolic fluxes in the GENSOR unit (MFt) and adding the extra fluxes to the denominator. Hence, connectivity is calculated as:

$$C = \frac{\mathrm{Ec}}{\mathrm{Et} + (\mathrm{MFt} - 1)}$$

Connectivity values range from 0 to 1. A value of 1 indicates a paradigmatic GENSOR unit where all the enzymes are connected and involved in a single metabolic flux. On the other hand, a value of 0 reflects a disconnected topology where each enzyme catalyzes a reaction disconnected from the rest of the GENSOR unit. To validate the biological significance of our metric, we calculated the connectivity of the 293 base pathways reported in EcoCyc (**Figure 2A**). The majority of metabolic pathways (55%) scored a value of 1, and 84% scored 0.7 or higher. A value of 0 was present for pathways such as tRNA charging, in which metabolic reactions are not successive but are functionally related. Results showed that connectivity does reflect a biological property, albeit with an expected 7% margin of error due to pathways that are not linear.

We calculated the connectivity distribution of 149 GENSOR units. Forty were excluded from the analysis because they included less than two catalytic reactions and would produce artificial values of 0. The resulting connectivity distribution (**Figure 2B**) was significantly different (Wilcoxon–Mann–Whitney; p-value $<$ 2.2e-16) from the metabolic pathway distribution (**Figure 2A**), which is noteworthy because we enriched the GENSOR unit set to include known metabolic pathways through the addition of secondary reactions. This result showed that the metabolic response mediated by TFs does not correlate with canonical metabolic pathways. The largest proportion of GENSOR units (21%) had a response involved in an individual metabolic flux, including 23% of GENSOR units for which feedback was present. In contrast,

in the second largest proportion, 15% of GENSOR units had a connectivity of 0, followed by 11% with values of 0.5 and 0.6. The resulting gradient of connectivity is not likely to be an artifact of unknown binding sites, since it is present in the set of the most extensively studied TFs (TFs with known effectors; **Figure 2B**, red and blue bars), as well as in the set of less-studied TFs (TFs without known effectors; **Figure 2B**, yellow bars). It is important to note that connectivity of a GENSOR unit did not depend on the number of enzymes present in it (Supplementary Figure S3). Moreover, the gradient is still observed if only GENSOR units with five enzymes or less are considered.

The connectivity of GENSOR units in the presence of feedback (**Figure 2B**, red bars) can be interpreted as a measure of autonomy of the TF response. Having a value of 1 means that, in the presence of the signal, the TF will impact a single metabolic flux that has an effect on said signal availability. Therefore, the response will continue until the signal concentration changes. Other forces can act on the metabolic flux, but from the TF perspective its effect is straightforward. A total of 19 TFs fall into this category, including those with responses involved in allantoin (AllS, AllR), arsenite (ArsR), hydroxybutyrate (AtoC), choline (BetI), chitobiose (ChbR), cyanate (CynR), nickel (NikR), zinc (Zur, ZntR), Fe^{2+} (YqjI), acetylneuraminate (NanR), 3-(3-hydroxyphenyl)propanoate (MhpR), idonate (IdnR), glycine (GcvA), citrate (PrpR), gluconate (GntR), tryptophan (TrpR), and xanthosine (XapR) metabolism.

The TF regulatory effect could account for low connectivity values and a higher complexity of a GENSOR unit response. Activation of a metabolic flux needs the presence of all the necessary enzymes, but inhibition of a pathway and redirection of the metabolic flux can be achieved by repressing a single gene. Following this logic, we would expect lower connectivity values for repressed enzymes. To test this hypothesis, we calculated connectivity of activated and repressed genes in each GENSOR unit separately. Consistent with our hypothesis, the connectivity distribution of repressed genes had a peak at 0 that included 67% of tested GENSOR units (**Figure 2C**) and was significantly different from the distribution of activated genes (Wilcoxon–Mann–Whitney; p-value = 7.724e-11). GENSOR units with the lowest connectivity values might be regulatory checkpoints that affect several independent metabolic fluxes in response to a stimulus, producing a more global response. Ultimately, connectivity reflects the complexity of the response mediated by a TF and, as we have shown, it is a continuum with peaks on both sides of the scale. Physiologically relevant metrics like connectivity might aid in more accurate functional classifications for regulators.

Prediction of Effectors Using GENSOR Unit Topology

The value of GENSOR units lies partially in the depiction of the interactions between their elements. They turn lists of regulated genes, enzymes, and reactions into a comprehensive network that reflects the functional effect of a TF. We proceeded to analyze topological properties of the GENSOR units regarding the relationship between effectors and TFs. We considered the set

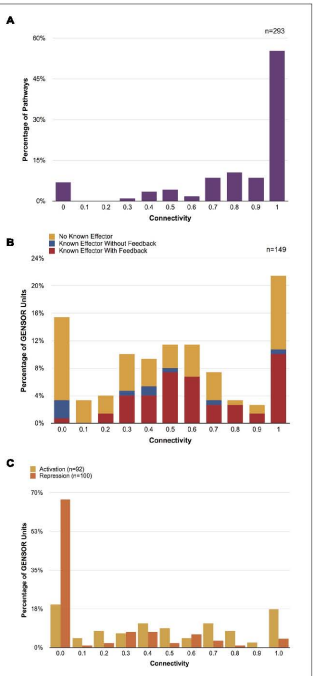


FIGURE 2 | Connectivity distributions. **(A)** Distribution of connectivity values of 293 metabolic pathways. Only sets of genes that catalyze two or more reactions were considered. **(B)** Distribution of connectivity values of 149 GENSOR units. GENSOR units with known effectors and identified feedback loops are shown in red, GENSOR units with known effectors and no identified feedback loops are shown in blue, GENSOR units with no known effectors where only a genetic switch and response have been identified are shown in yellow. The distributions of values in panels **(A)** and **(B)** were significantly different (Wilcoxon–Mann–Whitney; p-value $< 2.2e\text{-}16$). **(C)** Distribution of connectivity values of activated (yellow) and repressed (orange) genes in GENSOR units. Only sets of genes that catalyze two or more reactions were considered. Activated and repressed distributions were significantly different (Wilcoxon–Mann–Whitney; p-value = 7.724e-11).

of GENSOR units with known effectors and feedback (**Figure 2B**, red bars), and we identified the position of the effector in the regulated metabolic flux. Effectors were classified into two categories: "substrate/product" if the effector was the first or last metabolite in the metabolic flux, or "intermediate" if it was in any other position. First and last metabolites were grouped in the same category to eliminate ambiguity due to reversible reactions. Ninety-seven percent of the known effectors (75/77) are intermediate metabolites of a metabolic flux (Supplementary Table S1). The high proportion of intermediate effectors is relevant given that only 40% of all metabolites in these GENSOR units with known effectors classify as intermediates under the same criteria. A high proportion of intermediate effectors had been previously observed in inducible catabolic systems (Savageau, 1976). The global analysis presented here suggests that intermediate effectors are a general property, irrespective of the TF mode of action.

It has been shown that using intermediate metabolites as effectors is an effective strategy to increase the stability of a system (Savageau, 1974, 2001). In GENSOR units, stability is crucial to avoid unnecessary production of enzymes under fluctuating signals, which can affect cellular growth rate. Additionally, an intermediate effector will produce two feedback loops. The enzymes upstream of the effector will create a positive feedback loop, where more enzymatic activity will produce more effector. The enzymes downstream of the effector will be involved in a negative feedback loop where the opposite will happen: the more enzymatic activity, the less effector will be present. This dual dynamic produced by intermediate metabolites has been demonstrated by comparing the expression patterns of upstream and downstream enzymes (Chubukov et al., 2012). Accordingly, we observed that upstream and downstream enzymes in GENSOR units tend to be present in different operons. The most frequent case in GENSOR units are effectors positioned as products of transport reactions (here considered within the set of intermediate metabolites), for example, BetI GENSOR unit (**Figure 1B**), and most GENSOR units involved in carbon source utilization and amino acid metabolism. This dynamic might explain why transport enzymes are commonly encoded in a different operon. It is also possible that different dynamics present in the same pathway account for fine-tuning of the metabolic flux in branched pathways, where transport of a metabolite is maximized (through the positive feedback loop), but utilization is modulated so that other pathways can use the metabolite as well. From an evolutionary perspective, an intermediate effector producing two different dynamics would allow the cell to produce more complex metabolic behaviors without the need of new TFs.

The high proportion of intermediate effectors suggested that it would be possible to extrapolate this property to the GENSOR units with no known effector, and to predict effector candidates by retrieving their intermediate molecules. Given the gradient of complexity observed in the GENSOR unit set, it is important to note that the number of effector candidates (and as a result, the number of false positives) will increase according to the complexity of the GENSOR unit in which the method is applied. The median of intermediate metabolites in each GENSOR unit is 11, which is considerably lower than the number of metabolites screened in some heuristic studies (Ibañez et al., 2000). Intuitively, the best predictions would be derived from the simplest GENSOR units, so we applied our method to the 15 GENSOR units with a connectivity value of 1 and no known effector (**Table 1**). Since candidates are taken from the response component of the GENSOR unit, all putative effectors are expected to produce feedback. To support the predictions, experimental evidence was searched in the literature and TFs were searched for ligand-binding domains to further support the mechanism of action.

Predictions for FabR and UlaR were validated. Supporting evidence was found for hypothetical effectors of Dan, FeaR, HcaR, MtlR, KdgR, and MngR, suggesting that they are excellent candidates for validation experiments. Predictions for AscG, CsiR, GatR, RtcR, CaiF, and YiaJ have not been previously reported in the literature. Evidence for CadC supports a mechanism of action that does not rely on ligand binding (Buchner et al., 2015), which is consistent with its lack of a molecule-binding domain. Two interesting examples of new predictions are CaiF and YiaJ. It has been shown that CaiF acts as an activator in the presence of L-carnitine (Eichler et al., 1996). However, no binding of L-carnitine was identified in a mobility shift assay (Buchet et al., 1999). CaiF GENSOR unit (**Figure 3A**) shows that one of our predictions, gamma-butyrobetaine, would render an end-product inhibition dynamic that fits with the observation of L-carnitine acting as inducer, particularly because it is a substrate in the reaction producing gamma-butyrobetaine. YiaJ negatively regulates the conversion of xylulose and 2,3-dioxo-L-gulonate into D-xylulose 5-phosphate (**Figure 3B**). Ibañez et al. (2000) tested 80 different compounds, including D-xylulose, and none showed a significant increase in expression of target genes. It is possible that our predictions might yield different results, given that they rely on the global interpretation of the interactions in the GENSOR unit. For example, 2,3-dioxo-L-gulonate might act as an effector whose presence unbinds YiaJ from DNA. When bound to DNA, YiaJ would repress the enzymes needed for 2,3-dioxo-L-gulonate utilization as the carbon source, and enzymes would be produced only when it is present in the environment. Since 2,3-dioxo-L-gulonate is converted into D-xylulose 5-phosphate, its uptake would probably only take place when D-xylulose 5-phosphate is not obtained from other carbon sources, like arabinose, xylulose, or ascorbate. Another predicted effector, D-xylulose 5-phosphate, might also act as an effector promoting YiaJ binding to DNA in its presence, rendering an end product inhibition dynamic. We omitted ribulose-5-P from the predictions because it is a central metabolite constantly present in the cell, and the activity of YiaJ has not been reported as central to its metabolism.

In summary, 53% of predictions were supported, 7% were rejected, and 40% have never been reported, including two cases where the GENSOR unit dynamics support the prediction. Because larger data sets are not available, it was not possible to assess our method using receiver operating characteristic curves, but it is important to note that our approach did not require additional tools and it could be used to significantly

TABLE 1 | Effector predictions in 15 GENSOR units with no reported effector in RegulonDB, and a connectivity value of 1.

GENSOR unit	Ligand binding domain	Predicted effectors	Prediction status	Evidence	Reference
FabR		Fatty acids attached to acyl-ACP	Validated	Gel mobility shift assay	Zhu et al., 2009
UlaR	DeoR C terminal sensor domain (Pfam:PF00455)	Ascorbate-6P	Validated	Gel mobility shift assay	Garces et al., 2008
Dan	Bacterial regulatory helix-turn-helix protein, lysR family (Pfam:PF00126)	Tartrate	Supporting evidence	Change in gene expression due to addition of the compound (β-galactosidase assay)	Kim et al., 2009
FeaR	AraC-binding-like domain (Pfam:PF14525)	Hyacinthin (phenyl acetaldehyde)	Supporting evidence	Inference from operon dynamics	Zeng and Spiro, 2013
HcaR	LysR substrate binding domain (Pfam:PF03466)	3-(5,6-Dihydroxycyclohexa-1,3-dien-1-yl)propanoate	Supporting evidence	Inference from operon dynamics	Turlin et al., 2001
MtlR		Mannitol-1P	Supporting evidence	Inference from operon dynamics	Figge et al., 1994
KdgR	Bacterial transcriptional regulator (Pfam:PF01614)	2-Keto-3-deoxygluconate-6-P	Supporting evidence	2-Keto-3-deoxyguconate has been reported as effector of KdgR ortholog in *Erwinia chrysanthemi*	Nasser et al., 1992
MngR	UTRA domain (Pfam:PF07702)	2 (Alpha-D-mannosyl-6-phosphate)-D-glycerate	Supporting evidence	Change in gene expression due to addition of the external form of the compound (microarray). Mutation of downstream enzymes did not affect induction.	Sampaio et al., 2004
AscG	Periplasmic binding protein-like domain (Pfam:PF13377)	Arbutin-6P, beta-D-cellibiose-6P	New		
CaiF		Gamma-butyrobetaine, crotonobetainyl-CoA, carnityl-CoA, gamma-butyrobetaine-CoA	New. Supported by dynamics of the GENSOR unit (see text). Evidence against other predictions.	Mobility shift assay reflected no binding of L-carnitine or crotonobetaine	Buchet et al., 1999
YiaJ	Bacterial transcriptional regulator (Pfam:PF01614)	Xylulose-5P, 2-3,dioxo-L-gulonate, 3-keto-L-gulonate, 3-keto-L-gulonate 6-P	New. Evidence against other predictions (see text).	80 candidate effectors did not show changes in target gene expression	Ibañez et al., 2000
CsiR	FCD domain (Pfam:PF07729)	L-Glutamate, ketoglutarate, succinate semialdehyde	New		
GatR	DeoR C terminal sensor domain (Pfam:PF00455)	Galactitol 1-phosphate, keto-L-tagatose 6-phosphate, tagatofuranose 1,6-diphosphate	New		
RtcR		RNA terminal-2′,3′-cyclic-phosphate	New		
CadC		Cadaverine, lysine	Evidence against mode of action	Anchored to the membrane; works as a one-component system. Responds to PH stress.	Buchner et al., 2015

Ligand-binding domains identified in TF sequences are shown alongside their Pfam identifiers. Predicted effectors were inferred from the GENSOR unit topology; the prediction status indicates whether the prediction has been validated in the literature, evidence exists that supports or contradicts the hypotheses, or that predictions are new. 'Evidence' column indicates the relevant experiments that have been reported in the literature.

reduce the search space for possible effectors before experimental procedures.

GENSOR Units Can Be Used as a Standardized Framework for Integrative Studies

The collection of regulatory interactions in RegulonDB has been widely used as the gold standard to test new algorithms and as a reliable data set for analysis of the transcriptional regulatory network (TRN). GENSOR units reflect the metabolic impact of that gold standard. Since they were assembled through a data-driven, exhaustive approach, they describe a new layer of biologically relevant knowledge and can be used by the community as a standardized framework for the study of the interplay between transcriptional regulation and metabolism in *E. coli* K-12. One of the main advantages is that GENSOR units are useful for small-scale studies, for example, by analyzing those with high connectivity whose effect is modular. In addition, GENSOR units can be used as building blocks for higher-level descriptions, as high as a whole-cell description that integrates the complete set of GENSOR units through their overlapping elements (Supplementary Figure S4). By merging individual GENSOR units, it is possible to elucidate complex cellular behaviors that involve more than one TF, for example, carbon catabolite repression (Monod, 1942; Görke and Stülke, 2008). When presented with two or more different carbon sources, *E. coli* will begin uptake and utilization in a fixed order: first glucose, then lactose, arabinose, xylose, sorbitol, or rhamnose, and finally ribose (Aidelberg et al., 2014). AraC and XylR are TFs that bind to arabinose and xylose, respectively, and coordinate their utilization. Both regulate the *xylAB* transcription unit, so their GENSOR units can be merged into a complex GENSOR unit (**Figure 4A**). The AraC–XylR complex GENSOR unit shows that when arabinose and xylose are present at the same time, *xylAB* will be repressed by AraC and activated by XylR. Given that in *E. coli* repression tends to be dominant (Collado-Vides et al., 1991), transcription of *xylAB* will be halted and arabinose will be used preferentially. Once arabinose is depleted from the environment, AraC will return to an inactive state. *xylAB* can be induced by XylR, and xylulose will be used as the second carbon source. In summary, the opposite regulation of *xylAB* is the switch where *E. coli* decides on the uptake of arabinose over xylose. AraC–XylR complex GENSOR unit shows the descriptive power of merging individual GENSOR units. More complex decisions can involve more than two GENSOR units, but regardless of the size they retain the same level of detail. Availability of gold standard data sets is a current limitation for studies that integrate regulation and metabolism (Imam et al., 2015). The GENSOR units presented here seek to fill that gap and can also be used for dynamic modeling or analysis of general properties from the complete set of interactions.

Using GENSOR units as building blocks can also shed light onto indirect regulatory mechanisms that rely on metabolism to tune the activity of a TF in the presence of a signal sensed by another TF. The presence of feedback in a GENSOR unit conveys that the response mediated by a TF has a direct effect on its effector availability. Nevertheless, the response can also act on the availability of the effector of a second TF. For example, AlaS binds to L-alanine to solely repress its own promoter; since no other TFs regulate its transcription, it appears as an isolated node in the TRN. However, IscR, NsrR, Fur, and OxyR GENSOR units include the production of alanine in their response through the action of SufS, an L-cysteine desulfurase. In the presence of iron–sulfur clusters, nitric oxide, iron, or oxidative stress (IscR, NsrR, Fur, and OxyR signals), the cellular concentration of alanine will fluctuate, and in turn the AlaS functional conformation will be affected. This rationale can be applied on a larger scale to identify cascades of indirect TF–TF regulation (**Figure 4B**). It will be interesting to couple metabolic and transcriptional regulation cascades, since conformational changes are rarely considered in large-scale analyses of the TRN.

DISCUSSION

During their lifetimes, cells are challenged with a plethora of perturbations from the environment and their internal machinery. To survive, they rely on genetic circuits that sense changes and give an appropriate response. Understanding how the cell processes information is crucial for advancing the elucidation of design principles. The GENSOR unit concept presented here integrates relationships between the signaling, regulatory, and metabolic networks to depict the information flow behind individual signal → response processes. We have performed a genome-scale analysis of the complexity of the response mediated by each TF, testing the paradigm set by Jacob and Monod (1961) with the regulation of the *lac* operon, whereby regulators work as on/off switches for a particular capacity. Our results showed that, at the genome scale, the relationship between regulated genes and metabolism does not follow an evident one TF/one process rule (**Figures 2A,B**), but design principles can still be observed (**Figure 2C**).

GOs and metabolic pathways are the most used concepts to describe functional units in bacteria. They are mostly used for obtaining functional overviews of groups of genes. However, their definitions of where a process begins and ends are sometimes based on historical and organizational reasoning (such as the presence of common metabolites). Interpretations derived from them are not likely to reflect the way bacteria "understand" function (Bordbar et al., 2014). As an example, 25% of GENSOR units do not include any genes present in a canonical pathway. As for GO enrichment analysis, enriched GO terms in individual GENSOR units range from 0 to 228 (Supplementary Figure S5). Forty percent of GENSOR units have more than 50 enriched GO terms. The hierarchical tree structure of the ontology makes difficult the interpretation of such data, since researchers have to guess to what extent parent–child terms can be thought of as being the same process. Clusters of Orthologous Groups functional categories are also widely used to describe gene function. Given the broadness of the functional terms, they are complementary to GENSOR units. They could be used as a guide to merge individual GENSOR units and describe higher level interactions between their elements. GENSOR units are

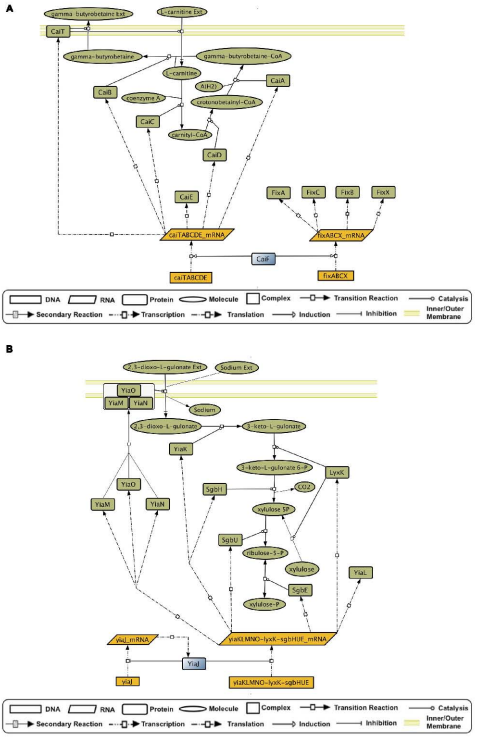


FIGURE 3 | GENSOR unit topologies that allow the inference of hypothetical effectors. **(A)** The CaiF GENSOR unit is active under anaerobic conditions. It is involved in transport of L-carnitine, efflux of gamma-butyrobetaine, and the synthesis of the latter from L-carnitine and gamma-butyrobetaine–CoA. Gamma-butyrobetaine could work as an effector that creates a complex with CaiF and promotes its unbinding from DNA, thus inhibiting the activation of the necessary enzymes for its synthesis and yielding an end product inhibition dynamic that stops the production of the enzymes if the gamma-butyrobetaine concentration inside the cell reaches a threshold. **(B)** The YiaJ GENSOR unit is involved in the transport of 2,3-dioxo-L-gulonate and its transformation into ribulose-5P. The presence of 2,3-dioxo-L-gulonate could promote unbinding of YiaJ from DNA, thus inhibiting the repression of the necessary enzymes for its utilization. Alternatively, xylulose-5P in complex with YiaJ could bind to DNA and promote the inhibition of the necessary enzymes for its synthesis in an end product inhibition dynamic.

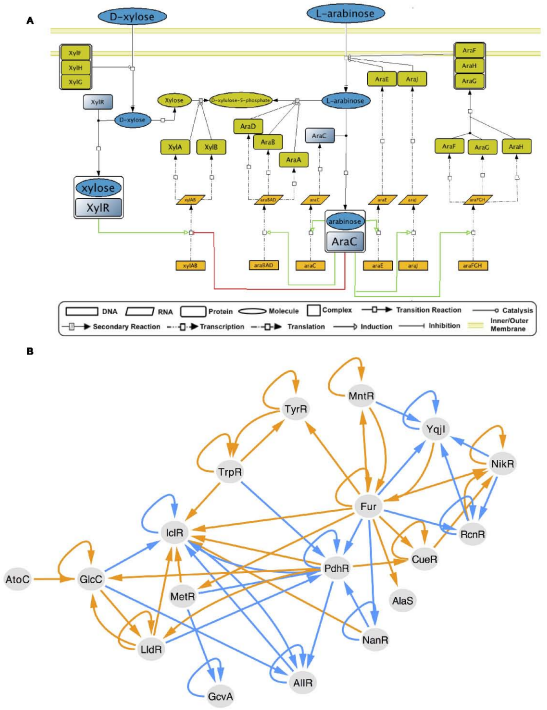


FIGURE 4 | GENSOR unit merging. **(A)** The XylR-AraC complex GENSOR unit. Activation of transcription units is shown in green, and repression is shown in red. Coregulation of *xylAB* with opposite effects creates a switch in carbon source utilization; if arabinose is present and xylose is absent, *xylAB* will be repressed and arabinose will be used as the main carbon source. If xylose is present and arabinose is absent, AraC will stop repressing *xylAB*, XylR will activate it, and xylose will be used. When both carbon sources are present, AraC repression will produce the preferential use of arabinose. **(B)** Cascade of TF–TF regulation through the production of effectors. The cascade is elicited by the activation of the two-component system AtoC in the presence of acetoacetate. AtoC is involved in biosynthesis of polyamines, catabolism of short-chain fatty acids, motility, and chemotaxis and directly regulates only four genes. The metabolic effect of its response produces the effector GlcC, whose response produces other effectors that give rise to a succession of changes in 16 other TF conformations. Edges link TFs that produce an effector in their response to TFs (arrow in edge) that bind to that effector. Orange edges indicate the presence of an effector produces an active conformation in the second TF (arrow in edge), and blue edges indicate inactive conformations.

a stepping stone toward building a framework that integrates transport, signal transduction, gene regulation and metabolism, in a directional regulated process that should capture the logics of information flow as it happens in the cell. They make it possible to trace relationships between genes of interest in their transport, signaling, metabolic, and regulatory contexts at the same time. GENSOR units can address questions such as "Is a group of genes responding to the same signal?" As a conceptual tool, they aim at facilitating the task of making biological sense of high-throughput expression data by reflecting the way that the cell is interpreting the changes that it is being subjected to.

The conceptual integration presented here has opened new questions. For instance, the paradigmatic functional homogeneity among regulons is not a general property. Connectivity values yield a gradient that goes from monothematic to epistatic GENSOR units. It is noteworthy that most GENSOR units are on opposite sides of the connectivity gradient. We propose three non-exclusive hypotheses to explain this. (1) Differences in regulatory mechanisms might produce differences in response properties. For example, two-component systems might account for global responses with low connectivity. (2) Some cellular functions require a more coordinated response. Nutrient uptake and efflux of toxins need fast responses, but flagellum assembly or cell cycle are capacities that depend on the presence of a large set of stable signals. (3) TFs regulate different subgroups of genes under different conditions, creating condition-specific subunits within GENSOR units. Higher-connectivity GENSOR units might pool independent metabolic fluxes that become active under different conditions. This shows the challenge ahead of making sense of the cell circuitry at higher levels of integration.

It has been proposed that coordination of multiple signals is embedded in a series of nested loops where general signals control several modules and local signals control the metabolic flux within modules (Chubukov et al., 2014). In this scenario, feedback loops would be a common occurrence. Small, local loops are for simple and fast responses, like carbon uptake, and larger loops are for central responses, like cell division. The latter requires a combination of signals, encompassed in smaller loops, to be present at a given time. This hypothesis agrees with our observations of GENSOR units. (1) Feedback loops are a common occurrence. (2) There is a gradient of modularity in GENSOR units reflected by connectivity values, from evident self-sufficient topologies to unrelated reactions that need the presence of other GENSOR units to be interpreted. (3) GENSOR units can be linked through "reporter molecules," such as effectors that can signal to general regulatory programs (**Figure 4B**). (4) GENSOR units can be merged to create a broader feedback loop (Supplementary Figure S6).

To the best of our knowledge, there is no other framework that comprises the complete catalog of natural genetic circuits mediated by TFs. Since GENSOR units place regulation in its natural cellular context, the concept is in line with operons and regulons and can be interpreted as a higher-level natural unit. GENSOR units also provide twofold methodological novelty. First, there is the automatic integration of data through an exhaustive search that eliminates any bias toward what the curator knows about the regulator. For example, rarely is it mentioned in descriptions of the *lac* operon that LacY can also transport melibiose. The only inherent bias is the availability of data in the two databases used. However, GENSOR units could be updated, as for any data set, with each new database release. The second methodological novelty is the prediction of effectors from the topology of the GENSOR unit alone. Although further validations are needed, we have proved that the information provided by GENSOR units can meaningfully reduce the number of effector candidates before designing an experiment.

GENSOR units can be used as templates for dynamic modeling. Either individually or merged to trace metabolic fluxes of interest, the GENSOR units can be used to predict the effects of adding molecules in the medium and to identify functional modules. They can also be used as a gold standard for new methodologies that predict properties of the interplay between transcriptional regulation and metabolism. Efforts are currently being made to assemble GENSOR units for other bacteria. It will be interesting to compare rewiring of their components, considering that TF binding sites diverge faster than coding sequences (Borneman et al., 2007). It might be possible to identify "orthologous" topologies that produce the same functional output using different network architectures. Identifying GENSOR units in pathogenic strains could also help in antibiotic design. Finally, the conceptual framework of GENSOR units can be expanded to other types of regulators. We have assembled a GENSOR unit of sigma factor 19 from *E. coli* that shows the four components and a feedback loop (**Figure 5**). Eventually, GENSOR units could be applied to eukaryotic regulators involved in disease to understand the mechanisms that cause disruptions in cellular dynamics, for example, the disappearance of feedback loops.

MATERIALS AND METHODS

GENSOR Unit Assembly

Active and inactive conformations, effectors, and regulated genes of the 189 local TFs with experimental evidence for their regulatory activities were obtained automatically from RegulonDB data sets; the Perlcyc API of Pathway Tools was used for automatic retrieval of gene products, catalyzed reactions, substrates, products, and directionality of reactions, heteromultimeric protein complexes in which gene products participate, and the rest of the monomers involved in the complex. Data were used to automatically generate an SBML file using CellDesigner 4.4 (Funahashi et al., 2008), the resulting network that was manually inspected and components in GENSOR units were identified. Secondary reactions were included by identifying pairs of reactions in the GENSOR unit that belonged to the same metabolic pathway using Perlcyc API of Pathway Tools software, the connecting reactions were added as a single secondary reaction only if directionality was maintained. Reversible reactions were considered. The resulting network and the information on each element and interaction were added to RegulonDB.

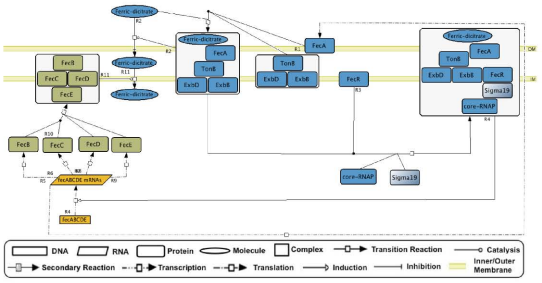


FIGURE 5 | The sigma 19 (FecI) GENSOR unit. In the presence of ferric dicitrate, FecI activates transcription of genes that code for proteins involved in transport of ferric dicitrate. Events occur as follows: external ferric dicitrate binds to FecA and causes a conformational change that allows FecA binding to TonB. FecA, TonB, and ExbD, and ExbB forms a complex that transports external ferric dicitrate into the periplasmic space. Alternatively, this complex binds to FecR, a protein that binds to sigma 19 and stimulates the binding of sigma 19 to RNA polymerase. Finally, the complex stimulates the transcription of the genes that code for proteins involved in the transport of ferric dicitrate from the periplasmic space to the cytoplasm. The signal and signal transduction is shown in blue, the genetic switch is shown in yellow, and the response is shown in green.

Properties of GENSOR Units

All properties were automatically identified and/or quantified using custom perl scripts available at GitHub[2].

Metabolic Fluxes in GENSOR Units

A metabolic flux was considered a chemical transformation of one metabolite into another; it could comprise one or more enzymatic reactions. A metabolic flux of two or more enzymatic reactions was created when substrates of one reaction were present in another either as substrates or products. Directionality of the reactions was considered to infer directionality of metabolic flux.

Feedback

All the possible metabolic fluxes in each GENSOR unit were obtained. Feedback was considered present when an effector was involved in one or more metabolic fluxes. Two-component systems were not considered, because their effector molecule was annotated as phosphate.

Connectivity

Connectivity was calculated through the following formula:

$$C = \frac{\mathrm{Ec}}{\mathrm{Et} + (\mathrm{MFt} - 1)}$$

where Ec is the connected enzymes in the GENSOR unit; Et is the total enzymes in the GENSOR unit; and MFt is the total independent metabolic fluxes in the GENSOR unit. Two enzymes were connected if any of their catalyzed reactions were also connected. Two reactions were connected if they shared substrates or the product of a reaction was also the reactant of a second reaction. Metabolic fluxes are independent groups of connected reactions (by the criteria described above), e.g., two components were present in a GENSOR unit if two groups of reactions were not connected between them but reactions within each group created a continuous flux.

Connectivity of Metabolic Pathways

Canonical pathways and the genes involved in each were obtained from EcoCyc by using the Pathway Tools software, Perlcyc API, and custom Perl scripts. A total of 362 base pathways were introduced to a modified version of the GENSOR unit pipeline; only 293 included two or more enzymatic reactions, and connectivity was calculated for these. Regulation of the genes was not considered. The pipeline used for the analysis can be found at GitHub[3].

Connectivity of GENSOR Units, Considering Their Regulatory Effects

Metabolic fluxes identified in each GENSOR unit only considered reactions catalyzed by enzymes whose genes were subject to the same type of regulation (activation/repression). Dual and unknown regulatory interactions were considered in both sets. Connectivity was obtained with the same algorithm, considering

[2]https://github.com/dledezma/gensor_units/tree/master/perl_scripts

[3]https://github.com/dledezma/gensor_units/blob/master/pipelines/pipeline_connectivity_pathways.sh

activated and repressed metabolic fluxes separately. The pipeline used for the analysis can be found at GitHub[4].

Effector Predictions

Position of Effector in Pathway

Seventy-eight GENSOR units with known effectors were analyzed. The position of each effector in the regulated pathway was automatically retrieved using a custom Perl script. The classification criteria were as follows:

- Substrate/product. Effectors that had a role as reactant in only one enzymatic reaction in the GENSOR unit. First and last positions were grouped to decrease ambiguity due to reversible reactions.

- Intermediate. Effectors that had a role as reactant in two or more enzymatic reactions in the GENSOR unit. Effectors that were products of transport reactions were considered intermediates to follow the classification defined by Savageau (1976).

Selection of Effector Candidates

GENSOR units with a connectivity value of 1 and no reported effectors in RegulonDB were used to predict effectors from their topology. All metabolites in the GENSOR unit were classified as substrate/product or intermediate according to the criteria mentioned above. Resulting intermediate metabolites were searched in the literature for experimental evidence of ligand function. The metabolites with supporting evidence on each GENSOR unit were considered hypothetical effectors; if no information was available for any of the molecules, all were reported as new candidates. Ligand-binding domains of the 15 TFs were identified using the NCBI Conserved Domains and Pfam (Finn et al., 2016) databases.

Statistical Analyses

Wilcoxon–Mann–Whitney tests were performed through a Wilcoxon rank sum test with continuity correction, using R software.

AtoC Cascade of Indirect TF Regulation

GENSOR units that included a reaction producing the effector of their own reaction of or another GENSOR unit were identified using custom Perl scripts on the relational tables of the GENSOR unit data set. The AtoC cascade was identified manually using AtoC as the starting point; all GENSOR units whose TFs bound to a metabolite produced in the AtoC GENSOR unit were included in the cascade and used as new starting points. The algorithm was run recursively until no more effectors were present in the lowest-level GENSOR unit response. The cascade network was produced using Cytoscape v3.1.1 (Shannon et al., 2003).

Metabolic Pathways and Gene Ontologies in GENSOR Units

Metabolic pathways of all genes were obtained from EcoCyc using Pathway Tools software. GOs of all genes were obtained from the Gene Ontology Consortium[5]. GO enrichments were obtained using the SmartTables tool in EcoCyc. Enrichments were calculated using all the genes in a GENSOR unit, along with analysis via the Fisher exact statistic with Bonferroni correction; *p*-values of <0.05 were considered statistically significant. A pathway/GO term was considered "present" in all the GENSOR units that included at least one gene from the pathway/GO term.

Data and Code Availability

GENSOR units in network form are publicly available in RegulonDB[6], and GENSOR units in tab-delimited files are publicly available at GitHub[7]. All scripts and files generated in the analysis are also available at GitHub[8].

AUTHOR CONTRIBUTIONS

JC-V conceived, designed, and supervised the study, and critically edited the manuscript. DL-T designed the study, performed all computational analysis, and wrote the manuscript. CI performed all manual curation, validated and summarized the GENSOR units, and wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Institutes of Health (grant number R01GM110597); and FOINS CONACYT Fronteras de la Ciencia (project 15). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. DL-T is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional Autónoma de México (UNAM) and received fellowship 275805 from CONACYT.

ACKNOWLEDGMENTS

Authors wish to thank Michael A. Savageau, Socorro Gama-Castro, and José Alquicira-Hernández for insightful discussions. Luis Muñiz-Rascado, Heladia Salgado, César Bonavides-Martínez, and Hilda Solano-Lira for their skillful technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01466/full#supplementary-material

[4]https://github.com/dledezma/gensor_units/blob/master/pipelines/pipeline_connectivity_with_effect.sh

[5]http://geneontology.org/page/download-annotations

[6]http://regulondb.ccg.unam.mx/central_panel_menu/integrated_views_and_tools/gensor_unit_groups

[7]https://github.com/dledezma/gensor_units/tree/master/datasets

[8]https://github.com/dledezma/gensor_units/

REFERENCES

Aidelberg, G., Towbin, B. D., Rothschild, D., Dekel, E., Bren, A., and Alon, U. (2014). Hierarchy of non-glucose sugars in *Escherichia Coli*. *BMC Syst. Biol.* 8:133. doi: 10.1186/s12918-014-0133-z

Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature* 397, 168–171. doi: 10.1038/16483

Balazsi, G., Barabasi, A. L., and Oltvai, Z. N. (2005). Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia Coli*. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7841–7846. doi: 10.1073/pnas.0500365102

Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D., et al. (2014). Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.* 9, 634. doi: 10.1038/msb.2012.70

Bordbar, A., Nagarajan, H., Lewis, N. E., Latif, H., Ebrahim, A., Federowicz, S., et al. (2014). Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Mol. Syst. Biol.* 10, 737. doi: 10.15252/msb.20145243

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., et al. (2007). Divergence of transcription factor binding sites across related yeast species. *Science (New York, N.Y.)* 317, 815–819. doi: 10.1126/science.1140748

Brooks, A. N., Reiss, D. J., Allard, A., Wu, W.-J., Salvanha, D. M., Plaisier, C. L., et al. (2014). A system-level model for the microbial regulatory genome. *Mol. Syst. Biol.* 10, 740. doi: 10.15252/msb.20145160

Buchet, A., Nasser, W., Eichler, K., and Mandrand-Berthelot, M. A. (1999). Positive co-regulation of the *Escherichia Coli* carnitine pathway cai and fix operons by CRP and the CaiF Activator. *Mol. Microbiol.* 34, 562–575. doi: 10.1046/j.1365-2958.1999.01622.x

Buchner, S., Schlundt, A., Lassak, J., Sattler, M., and Jung, K. (2015). Structural and functional analysis of the signal-transducing linker in the pH-responsive one-component system CadC of *Escherichia Coli*. *J. Mol. Biol.* 427, 2548–2561. doi: 10.1016/j.jmb.2015.05.001

Chubukov, V., Gerosa, L., Kochanowski, K., and Sauer, U. (2014). Coordination of microbial metabolism. *Nat. Rev. Microbiol.* 12, 327–340. doi: 10.1038/nrmicro3238

Chubukov, V., Zuleta, I. A., and Li, H. (2012). Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5127–5132. doi: 10.1073/pnas.1114235109

Collado-Vides, J., Magasanik, B., and Gralla, J. D. (1991). Control site location and transcriptional regulation in *Escherichia Coli*. *Microbiol. Rev.* 55, 371–394.

Eichler, K., Buchet, A., Lemke, R., Kleber, H. P., and Mandrand-Berthelot, M. A. (1996). Identification and characterization of the caiF gene encoding a potential transcription activator of carnitine metabolism in *Escherichia Coli*. *J. Bacteriol.* 178, 1248–1257. doi: 10.1128/jb.178.5.1248-1257.1996

Figge, R. M., Ramseier, T. M., and Saier, M. H. (1994). The mannitol repressor (MtlR) of *Escherichia Coli*. *J. Bacteriol.* 176, 840–847. doi: 10.1128/jb.176.3.840-847.1994

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Forster, J., Famili, I., Palsson, B. O., and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 13, 244–253. doi: 10.1101/gr.234503.complex

Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE* 96, 1254–1265. doi: 10.1109/JPROC.2008.925458

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., et al. (2016). RegulonDB Version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* 44, D133–D143. doi: 10.1093/nar/gkv1156

Garces, F., Fernández, F. J., Gómez, A. M., Pérez-Luque, R., Campos, E., Prohens, R., et al. (2008). Quaternary structural transitions in the DeoR-type repressor ular control transcriptional readout from the L-ascorbate utilization regulon in *Escherichia Coli*. *Biochemistry* 47, 11424–11433. doi: 10.1021/bi800748x

Gerosa, L., and Sauer, U. (2011). Regulation and control of metabolic fluxes in microbes. *Curr. Opin. Biotechnol.* 22, 566–575. doi: 10.1016/j.copbio.2011.04.016

Görke, B., and Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat. Rev. Microbiol.* 6, 613–624. doi: 10.1038/nrmicro1932

Hyduke, D. R., and Palsson, B. Ø (2010). Towards genome-scale signalling-network reconstructions. *Nat. Rev. Genet.* 11, 297–307. doi: 10.1038/nrg2750

Ibañez, E., Campos, E., Baldoma, L., Aguilar, J., and Badia, J. (2000). Regulation of expression of the yiaKLMNOPQRS Operon for Carbohydrate Utilization in *Escherichia Coli*: involvement of the Main Transcriptional Factors. *J. Bacteriol.* 182, 4617–4624. doi: 10.1128/JB.182.16.4617-4624.2000

Imam, S., Schäuble, S., Brooks, A. N., Baliga, N. S., and Price, N. D. (2015). Data-driven integration of genome-scale regulatory and metabolic network models. *Front. Microbiol.* 6:409. doi: 10.3389/fmicb.2015.00409

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

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolivar, B. Jr., et al. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell* 150, 389–401. doi: 10.1016/j.cell.2012.05.044.A

Kauffman, S. A. (1993). *The Origins of Order?: Self-Organization and Selection in Evolution*. Oxford: Oxford University Press.

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., et al. (2013). EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* 41, 605–612. doi: 10.1093/nar/gks1027

Kim, O. B., Reimann, J., Lukas, H., Schumacher, U., Grimpo, J., Dünnwald, P., et al. (2009). Regulation of tartrate metabolism by TtdR and relation to the DcuS-DcuR-regulated C4-Dicarboxylate metabolism of *Escherichia Coli*. *Microbiology* 155, 3632–3640. doi: 10.1099/mic.0.031401-0

Monod, J. (1942). *Recherches Sur La Croissance Des Cultures Bacteriennes*. Available at: http://agris.fao.org/agris-search/search.do?recordID = US201300336259

Monod, J., Changeux, J.-P., and Jacob, F. (1963). Allosteric proteins and cellular control systems. *J. Mol. Biol.* 6, 306–329. doi: 10.1016/S0022-2836(63)80091-1

Naisbitt, J. (1984). *Megatrends?: Ten New Directions Transforming Our Lives*. London: Macdonald.

Nasser, W., Reverchon, S., and Robert-Baudouy, J. (1992). Purification and functional characterization of the KdgR protein, a major repressor of pectinolysis genes of erwinia chrysanthemi. *Mol. Microbiol.* 6, 257–265. doi: 10.1111/j.1365-2958.1992.tb02007.x

Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.* 6, 99–111. doi: 10.1038/nrm1570

Papin, J. A., and Palsson, B. O. (2004). Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.* 227, 283–297. doi: 10.1016/j.jtbi.2003.11.016

Pardee, A. B., Jacob, F., and Monod, J. (1959). The genetic control and cytoplasmic expression of 'inducibility' in the synthesis of β-Galactosidase by E. *Coli*. *J. Mol. Biol.* 1, 165–178. doi: 10.1016/S0022-2836(59)80045-0

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374

Sampaio, M. M., Chevance, F., Dippel, R., Eppler, T., Schlegel, A., Boos, W., et al. (2004). Phosphotransferase-mediated transport of the osmolyte 2-O-α -Mannosyl-D-Glycerate in *Escherichia Coli* occurs by the product of the mngA (hrsA) gene and is regulated by the mngR (farR) gene product acting as repressor. *J. Biol. Chem.* 279, 5537–5548. doi: 10.1074/jbc.M310980200

Savageau, M. A. (1974). Comparison of classical and autogenous systems of regulation in inducible operons. *Nature* 252, 546–549. doi: 10.1038/252546a0

Savageau, M. A. (1976). *Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology*, 1st Edn. Boston, MA: Addison-Wesley Publishing Company.

Savageau, M. A. (2001). Design principles for elementary gene circuits: elements, methods, and examples. *Chaos* 11, 142–159. doi: 10.1063/1.1349892

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of

biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia Coli*. *Nat. Genet.* 31, 64–68. doi: 10.1038/ng881

The Gene Ontology Consortium (2000). Gene ontology: tool for the identification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Thomas, R., and D'Ari, R. (1990). *Biological Feedback*. Boca Raton, FL: CRC Press.

Turlin, E., Perrotte-piquemal, M., Danchin, A., and Biville, F. (2001). Regulation of the early steps of 3-Phenylpropionate catabolism in *Escherichia Coli*. *J. Mol. Microbiol. Biotechnol.* 3, 127–133.

Zeng, J., and Spiro, S. (2013). Finely tuned regulation of the aromatic amine degradation pathway in *Escherichia Coli*. *J. Bacteriol* 195, 5141–5150. doi: 10.1128/JB.00837-13

Zhu, K., Zhang, Y. M., and Rock, C. O. (2009). Transcriptional regulation of membrane lipid homeostasis in *Escherichia Coli*. *J. Biol. Chem.* 284, 34880–34888. doi: 10.1074/jbc.M109.068239

Conflict of Interest Statement: 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.

Bibliografía

[Aidelberg et al., 2014] Aidelberg, G., Towbin, B. D., Rothschild, D., Dekel, E., Bren, A., and Alon, U. (2014). Hierarchy of non-glucose sugars in Escherichia coli. *BMC Systems Biology*, 8(1):133.

[Alon et al., 1999] Alon, U., Surette, M. G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171.

[Balazsi et al., 2005] Balazsi, G., Barabasi, A. L., and Oltvai, Z. N. (2005). Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proc Natl Acad Sci U S A*, 102(22):7841–7846.

[Barabási and Oltvai, 2004] Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113.

[Berthoumieux et al., 2014] Berthoumieux, S., de Jong, H., Baptist, G., Pinel, C., Ranquet, C., Ropers, D., and Geiselmann, J. (2014). Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Molecular Systems Biology*, 9(1):634–634.

[Bordbar et al., 2014] Bordbar, A., Nagarajan, H., Lewis, N. E., Latif, H., Ebrahim, A., Federowicz, S., Schellenberger, J., and Palsson, B. O. (2014). Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Molecular Systems Biology*, 10(7):737–737.

[Borneman et al., 2007] Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science (New York, N.Y.)*, 317(5839):815–9.

[Brooks et al., 2014] Brooks, A. N., Reiss, D. J., Allard, A., Wu, W.-J., Salvanha, D. M., Plaisier, C. L., Chandrasekaran, S., Pan, M., Kaur, A., and Baliga, N. S. (2014). A system-level model for the microbial regulatory genome. *Molecular Systems Biology*, 10(7):740–740.

[Buchet et al., 1999] Buchet, A., Nasser, W., Eichler, K., and Mandrand-Berthelot, M. A. (1999). Positive co-regulation of the Escherichia coli carnitine pathway cai and fix operons by CRP and the CaiF activator. *Molecular Microbiology*, 34(3):562–575.

[Buchner et al., 2015] Buchner, S., Schlundt, A., Lassak, J., Sattler, M., and Jung, K. (2015). Structural and Functional Analysis of the Signal-Transducing Linker in the pH-Responsive One-Component System CadC of Escherichia coli. *Journal of Molecular Biology*, 427(15):2548–2561.

[Caspi et al., 2013] Caspi, R., Dreher, K., and Karp, P. D. (2013). The challenge of constructing, classifying, and representing metabolic pathways. *FEMS Microbiology Letters*, 345(2):85–93.

[Chubukov et al., 2014] Chubukov, V., Gerosa, L., Kochanowski, K., and Sauer, U. (2014). Coordination of microbial metabolism. *Nature Reviews Microbiology*, 12(5):327–340.

[Chubukov et al., 2012] Chubukov, V., Zuleta, I. A., and Li, H. (2012). Regulatory architecture determines optimal regulation of gene expression in metabolic pathways. *Proceedings of the National Academy of Sciences*, 109(13):5127–5132.

[Collado-Vides et al., 1991] Collado-Vides, J., Magasanik, B., and Gralla, J. D. (1991). Control site location and transcriptional regulation in Escherichia coli. *Microbiol. Rev.*, 55(3):371–394.

[Consortium, 2000] Consortium, T. G. O. (2000). Gene ontology: Tool for the identification of biology. *Natural Genetics*, 25(may):25–29.

[Eichler et al., 1996] Eichler, K., Buchet, A., Lemke, R., Kleber, H. P., and Mandrand-Berthelot, M. A. (1996). Identification and characterization of the caiF gene encoding a potential transcription activator of carnitine metabolism in Escherichia coli. *J Bacteriol*, 178(5):1248–1257.

[Figge et al., 1994] Figge, R. M., Ramseier, T. M., and Saier, M. H. (1994). The mannitol repressor (MtlR) of Escherichia coli. *Journal of Bacteriology*, 176(3):840–847.

[Finn et al., 2016] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–85.

[Funahashi et al., 2008] Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N., and Kitano, H. (2008). CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks. *Proceedings of the IEEE*, 96(8):1254–1265.

[Gama-Castro et al., 2011] Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muniz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J. S., López-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, R., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A. M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E., and Collado-Vides, J. (2011). RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research*, 39(SUPPL. 1):98–105.

[Gama-Castro et al., 2016] Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., Alquicira-Hernández, K., Martínez-Flores, I., Pannier, L., Castro-Mondragón, J. A., Medina-Rivera, A., Solano-Lira, H., Bonavides-Martínez, C., Pérez-Rueda, E., Alquicira-Hernández, S., Porrón-Sotelo, L., López-Fuentes, A., Hernández-Koutoucheva, A., Moral-Chávez, V. D., Rinaldi, F., and Collado-Vides, J. (2016).

RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic acids research*, 44(D1):D133–43.

[Garces et al., 2008] Garces, F., Fern??ndez, F. J., G??mez, A. M., P??rez-Luque, R., Campos, E., Prohens, R., Aguilar, J., Baldom??, L., Coll, M., Bad??a, J., and Vega, M. C. (2008). Quaternary structural transitions in the DeoR-type repressor ular control transcriptional readout from the L-ascorbate utilization regulon in Escherichia coli. *Biochemistry*, 47(44):11424–11433.

[Görke and Stülke, 2008] Görke, B. and Stülke, J. (2008). Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature reviews. Microbiology*, 6(8):613–24.

[Grüning et al., 2010] Grüning, N. M., Lehrach, H., and Ralser, M. (2010). Regulatory crosstalk of the metabolic network. *Trends in Biochemical Sciences*, 35(4):220–227.

[Guía et al., 2005] Guía, M. H., Pérez, A. G., Espinosa, V., Vasconcelos, A. T. R., and Collado-Vides, J. (2005). Complementing computationally predicted regulatory sites in Tractor_DB using a pattern matching approach. *In silico biology*, 5(2):209–219.

[Ibañez et al., 2000] Ibañez, E., Campos, E., Baldoma, L., Aguilar, J., and Badia, J. (2000). Regulation of expression of the yiaKLMNOPQRS operon for carbohydrate utilization in Escherichia coli: Involvement of the main transcriptional factors. *Journal of Bacteriology*, 182(16):4617–4624.

[Jacob and Monod, 1961] Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356.

[Jain and Saini, 2016] Jain, K. and Saini, S. (2016). MarRA, SoxSR, and Rob encode a signal dependent regulatory network in Escherichia coli. *Mol. BioSyst.*, 12(6):1901–1912.

[Kanehisa et al., 2017] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361.

[Karp et al., 2007] Karp, P. D., Keseler, I. M., Shearer, A., Latendresse, M., Krummenacker, M., Paley, S. M., Paulsen, I., Collado-Vides, J., Gama-Castro, S., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spínola, M. I., Bonavides-martinez, C., and Ingraham, J. (2007). Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Research*, 35(22):7577–7590.

[Karr et al., 2012] Karr, J., Sanghvi, J., Macklin, D., Gutschow, M., Jacobs, J., Bolivar Jr, B., Assad-Garcia, N., Glass, J., and Covert, M. (2012). A whole-cell computational model predicts phenotype from genotype.

[Kashtan and Alon, 2005] Kashtan, N. and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778.

[Keseler et al., 2017] Keseler, I. M., Mackie, A., Santos-Zavaleta, A., Billington, R., Bonavides-Martínez, C., Caspi, R., Fulcher, C., Gama-Castro, S., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Peralta-Gil, M., Subhraveti, P., Velázquez-Ramírez, D. A., Weaver, D., Collado-Vides, J., Paulsen, I., and Karp, P. D. (2017). The EcoCyc database: Reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research*, 45(D1):D543–D550.

[Kim et al., 2009] Kim, H. D., Shay, T., O'Shea, E. K., and Regev, A. (2009). Transcriptional regulatory circuits: predicting numbers from alphabets. *Science (New York, N.Y.)*, 325(5939):429–432.

[Kreimer et al., 2008] Kreimer, A., Borenstein, E., Gophna, U., and Ruppin, E. (2008). The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 105(19):6976–6981.

[Lomnitz and Savageau, 2013] Lomnitz, J. G. and Savageau, M. A. (2013). Phenotypic deconstruction of gene circuitry. *Chaos*, 23(2):1–10.

[Mangan and Alon, 2003] Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of*

the National Academy of Sciences of the United States of America, 100(21):11980–11985.

[Martínez-Antonio and Collado-Vides, 2003] Martínez-Antonio, A. and Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 6(5):482–489.

[McCloskey et al., 2014] McCloskey, D., Palsson, B. O., and Feist, A. M. (2014). Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Molecular Systems Biology*, 9(1):661–661.

[Medina-Rivera et al., 2011] Medina-Rivera, A., Abreu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J., and Van Helden, J. (2011). Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Research*, 39(3):808–824.

[Medina-Rivera et al., 2015] Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J. A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C., Staines, D. M., Contreras-Moreira, B., Artufel, M., Charbonnier-Khamvongsa, L., Hernandez, C., Thieffry, D., Thomas-Chollier, M., and Van Helden, J. (2015). RSAT 2015: Regulatory sequence analysis tools. *Nucleic Acids Research*, 43(W1):W50–W56.

[Monod, 1942] Monod, J. (1942). Recherches sur la croissance des cultures bacteriennes.

[Nasser et al., 1992] Nasser, W., Reverchon, S., and Robert-Baudouy, J. (1992). Purification and functional characterization of the KdgR protein, a major repressor of pectinolysis genes of Erwinia chrysanthemi. *Molecular Microbiology*, 6(2):257–265.

[Orth et al., 2010] Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat Biotechnol*, 28(3):245–248.

[Papin et al., 2005] Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature Reviews Molecular Cell Biology*, 6(2):99–111.

[Papin and Palsson, 2004] Papin, J. A. and Palsson, B. O. (2004). Topological analysis of mass-balanced signaling networks: A framework to obtain network properties including crosstalk. *Journal of Theoretical Biology*, 227(2):283–297.

[Pardee et al., 1959] Pardee, A. B., Jacob, F., and Monod, J. (1959). The genetic control and cytoplasmic expression of "Inducibility" in the synthesis of β-galactosidase by E. coli. *Journal of Molecular Biology*, 1(2):165–178.

[Pardee et al., 2014] Pardee, K., Green, A. A., Ferrante, T., Cameron, D. E., Daleykeyser, A., Yin, P., and Collins, J. J. (2014). Paper-based synthetic gene networks. *Cell*, 159(4):940–954.

[Parter et al., 2007] Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC evolutionary biology*, 7:169.

[Ravasz et al., 2002] Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., and Barabási, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science*, 297(5586):1551–1555.

[Resendis-Antonio et al., 2005] Resendis-Antonio, O., Freyre-González, J. A., Menchaca-Méndez, R., Gutiérrez-Ríos, R. M., Martínez-Antonio, A., Avila-Sánchez, C., and Collado-Vides, J. (2005). Modular analysis of the transcriptional regulatory network of E. coli. *Trends in genetics : TIG*, 21(1):16–20.

[Sampaio et al., 2004] Sampaio, M. M., Chevance, F., Dippel, R., Eppler, T., Schlegel, A., Boos, W., Lu, Y. J., and Rock, C. O. (2004). Phosphotransferase-mediated Transport of the Osmolyte 2-O-α -Mannosyl-D-glycerate in Escherichia coli Occurs by the Product of the mngA (hrsA) Gene and Is Regulated by the mngR (farR) Gene Product Acting as Repressor. *Journal of Biological Chemistry*, 279(7):5537–5548.

[Savageau, 1974] Savageau, M. A. (1974). Comparison of classical and autogenous systems of regulation in inducible operons. *Nature*, 252(5484):546–549.

[Savageau, 1976] Savageau, M. A. (1976). *Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology*. Addison-Wesley Publishing Company, 1 edition.

[Savageau, 2001] Savageau, M. A. (2001). Design principles for elementary gene circuits: Elements, methods, and examples. *Chaos*, 11(1):142–159.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–504.

[Shen-Orr et al., 2002] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–8.

[Thiede et al., 1998] Thiede, B., Urlaub, H., and Neubauer, H. (1998). Precise determination of RNA–protein contact sites in the 50 S ribosomal subunit of Escherichia coli. *Biochem. J*, 42:39–42.

[Turlin et al., 2001] Turlin, E., Perrotte-piquemal, M., Danchin, a., and Biville, F. (2001). Regulation of the early steps of 3-phenylpropionate catabolism in Escherichia coli. *Journal of molecular microbiology and biotechnology*, 3(1):127–133.

[Van der Ploeg et al., 1997] Van der Ploeg, J., Iwanicka-Nowicka, R., Kertesz, M., Leisinger, T., and Hryniewicz, M. (1997). Involvement of CysB and Cbl Regulatory Proteins in Expression of the tauABCD Operon and Other Sulfate Starvation-Inducible Genes in Escherichia coli. *Journal of Bacteriology*, 179(24):7671–7678.

[Wada, 1986] Wada, A. (1986). Analysis of Escherichia coli Ribosomal Proteins by an Improved Two Dimensional Gel Electrophoresis . Characterization of Four New Proteins. *J. Biochem*, 100(6):1595–1605.

[Weiss et al., 2013] Weiss, V., Medina-Rivera, A., Huerta, A. M., Santos-Zavaleta, A., Salgado, H., Morett, E., and Collado-Vides, J. (2013). Evidence classification of high-throughput protocols and confidence integration in RegulonDB. *Database*, 2013(September):1–15.

[Yon Rhee et al., 2008] Yon Rhee, S., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7):509–515.

[Zeng and Spiro, 2013] Zeng, J. and Spiro, S. (2013). Finely tuned regulation of the aromatic amine degradation pathway in Escherichia coli. *Journal of Bacteriology*, 195(22):5141–5150.

[Zgurskaya et al., 2011] Zgurskaya, H. I., Krishnamoorthy, G., Ntreh, A., and Lu, S. (2011). Mechanism and function of the outer membrane channel TolC in multidrug resistance and physiology of enterobacteria. *Frontiers in Microbiology*, 2(SEP):1–13.

[Zhu et al., 2009] Zhu, K., Zhang, Y.-M., and Rock, C. O. (2009). Transcriptional regulation of membrane lipid homeostasis in Escherichia coli. *The Journal of biological chemistry*, 284(50):34880–34888.