



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

DISECCIÓN DE LA ARQUITECTURA FUNCIONAL DE LA RED
DE REGULACIÓN TRANSCRIPCIONAL DE *Escherichia coli*:
UN ENFOQUE DE DESCOMPOSICIÓN NATURAL

T E S I S

QUE PARA OBTENER EL GRADO DE:

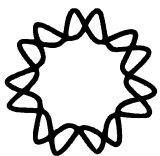
DOCTOR EN CIENCIAS

PRESENTA:

JULIO AUGUSTO FREYRE-GONZÁLEZ

DIRECTOR DE TESIS:

DR. JULIO COLLADO-VIDES



CUERNAVACA, MORELOS

NOVIEMBRE, 2008



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

**Disección de la arquitectura funcional de la red de regulación
transcripcional de *Escherichia coli*: Un enfoque de
descomposición natural**

por

Julio Augusto Freyre-González

Ing.S.C., Instituto Tecnológico de Veracruz (2000)
M.C.C., Instituto Tecnológico y de Estudios Superiores de Monterrey (2000)

Tesis presentada para obtener el grado de

Doctor en Ciencias

en el

INSTITUTO DE BIOTECNOLOGÍA

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Cuernavaca, Morelos. Noviembre, 2008

Esta obra está dedicada a todas aquellas personas, anónimas o conocidas, que a lo largo de la historia de la humanidad, y hasta nuestros días, han luchado, contra viento y marea, por hacer del librepensamiento la doctrina por excelencia para evitarnos caer víctima de los sofistas y sus sofismas.

JULIO A. FREYRE-GONZÁLEZ

Agradecimientos

En lo moral...

A mi madre, por nutrir en cada oportunidad, desde que era muy pequeño, mi inquietud e interés por la ciencia, los cuales se trocaron, eventualmente, en una profunda pasión; por ser la base de mi vida y creer en mí en todo momento. ¡Te amo madre!

A Patty, mi compañera, por haberme introducido al terreno de la biología, despertando en mí una pasión por comprenderla; por el amor, la paciencia y tolerancia con la que me has acompañado durante estos nueve años; por ser el pilar de mi vida; por tu apoyo, en todo momento desinteresado; por siempre estar a mi lado, sin importar las vicisitudes. ¡Te amo mi vida!

A mi grupo de colaboradores, más que eso, mis mejores amigos: Pepetoño Alonso y Luis Treviño; sin cuyo trabajo, apoyo, motivación y amistad esta aventura no hubiera arribado victoriosa a puerto. ¡Gracias amigos!

Al Dr. Julio Collado Vides ante todo por creer en aquel lejano computólogo egresado de Tec de Monterrey que ignoraba todo sobre biología —y quien sigue aprendiendo día a día de esta tan apasionante área—; por el tiempo y entusiasmo en aquellas añejas discusiones de viernes por la tarde, salpimentadas con exquisita filosofía de la ciencia, las cuales espero, algún día, el tiempo nos brinde —y nosotros nos demos— la oportunidad de retomar. ¡Gracias tocayo!

A mis cotutores, los Dres. Enrique Merino y Sergio Encarnación, por motivarme a siempre perseguir lo mejor de mí; así como por su apoyo y consejos.

A la Dra. Rosa María Gutiérrez y al Dr. Guillermo Gosset por invitarme a acompañarlos en sus correrías por la ciencia, las cuales contribuyeron a enriquecer mi visión de la toma de decisiones en bacterias; amén de su amistad, y sus consejos y apoyo siempre entusiastas.

Al Dr. David Romero Camarena por su amistad y don de gente; gracias David por tu apoyo desde mi intento de ingresar al Doctorado en Ciencias Biomédicas, así como tu confianza, consejos y oído siempre dispuesto.

A mis amigos Miryam Ivette, Clarita Olvera, y Héctor Avilés; así como a mis exdiscípulos y amigos (ordenados por generación): Chío, Miss Clau, Karlita, Yamile, Bere, Ilse, Beto y Hüicho; gracias a todos por sus detalles, cariño, apoyo y preocupación.

En lo técnico...

A Concepción Hernandez por su trato siempre atento y servicial; gracias Conchita.

A Romualdo y Vic por su amable y atento servicio al frente del clúster de computo y el correo electrónico, dos herramientas muy importantes para lo que hacemos; gracias a ambos.

Al Profesor Donald E. Knuth de la Universidad de Stanford por crear el excelente sistema de composición tipográfica T_EX, el cual me permitió impregnar esta obra con una rica apariencia estética y profesional.

Al Dr. David S. Goodsell del Instituto de Investigación Scripps por su excelente trabajo ilustrando de manera fiel la complejidad molecular en las células, así como por las facilidades prestadas para incorporar a ésta obra una de sus extraordinarias ilustraciones.

Al Dr. Dennis Kunkel de Dennis Kunkel Microscopy, Inc. (www.denniskunkel.com) por otorgarme permiso para reproducir tres de sus excelentes microfotografías; además de su interés y disposición para contribuir a esta obra al amablemente brindarse a facilitarme imágenes de alta calidad.

En lo económico...

A mi padre por su “apoyo”, aunque sólo económico y el cual al último momento me negó; y por al fin mostrarse como el criminal que siempre indicí al intentar asesinar a mi madre por la espalda ante familiares y ajenos.

Al Conacyt por la beca doctoral 176341, así como a la DGEP-UNAM por otorgarme una beca complementaria. Este trabajo fue financiado de forma parcial por los donativos 47609-A del Conacyt, IN214905 del PAPIIT-UNAM, y NIH R01 GM071962-04 otorgados al Dr. Julio Collado Vides.

Índice general

Prólogo	XII
Resumen	XVII
Abstract	XIX
1. Prolegómenos	1
1.1. ¿Qué es <i>Escherichia coli</i> ?	2
1.2. ¿Qué es la regulación transcripcional?	4
1.3. ¿Por qué un enfoque de descomposición natural?	7
1.4. Descripción de la investigación	10
2. Control Transcripcional de la Expresión Genética	12
2.1. Los componentes químicos de la célula	12
2.1.1. Ácidos nucleicos	13
2.1.2. Proteínas	15
2.1.3. Metabolitos	16
2.2. Control del inicio de la transcripción	17
2.2.1. Arquitectura del promotor bacteriano	18
2.2.2. ¿Cómo se inicia la transcripción?	18
2.2.3. Elementos de control de la transcripción	20
2.3. Niveles de organización de las redes de regulación	20
2.3.1. Operón	22
2.3.2. Regulón	24

2.3.3. Modulón	25
2.4. Factores transcripcionales globales	26
2.4.1. Proteínas regulatorias globales	26
2.4.2. Factores σ	27
3. Teoría de Redes	28
3.1. El Uroboros	28
3.1.1. El Ubuntu	29
3.2. El problema de los siete puentes de Königsberg	30
3.3. Conceptos básicos de teoría de grafos	33
3.4. ¿Unos cuantos principios gobiernan el todo?	37
3.5. ¡El mundo es un pañuelo!	37
3.6. La importancia de la popularidad	42
3.7. La tragedia de los cosmopolitas	45
3.8. El nivel mesoscópico	46
3.8.1. Circuitos de retroalimentación	46
3.8.2. Motivos topológicos	48
4. Modularidad y Jerarquía en Redes Biológicas	52
4.1. ¿Cómo ensamblar un reloj sin morir en el intento?	53
4.1.1. Agrupamiento jerárquico aglomerativo	54
4.1.2. Maximizando la modularidad	60
4.1.3. Técnicas actuales son inadecuadas para redes jerárquico-modulares	63
4.2. ¿Cómo coordinar las partes para darle sentido al todo?	63
4.2.1. Agregación de motivos topológicos	63
4.2.2. Estructuras jerárquicas piramidales	64
5. Disectando la Arquitectura Funcional de <i>E. coli</i>	68
5.1. Red de regulación de <i>Escherichia coli</i> K-12	68
5.2. Enfoque de descomposición natural	69
5.2.1. La red de regulación de <i>E. coli</i> no es acíclica	69

5.2.2. Identificación de los nodos jerárquicos y modulares	71
5.2.3. Identificación módulos y genes intermodulares	71
5.2.4. Anotación manual y automatizada de los módulos identificados	73
5.3. Reconstituyendo la red regulatoria de <i>E. coli</i>	73
5.3.1. Inferencia de la estructura jerárquica gobernando la red	73
5.3.2. La médula jerárquica de la red es conformada por motivos <i>feedforward</i> . .	74
6. Conclusiones y Perspectivas	76
6.1. Conclusiones	76
6.2. Perspectivas	78
A. Modular analysis of the transcriptional regulatory network of <i>E. coli</i>	80
B. Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in <i>Escherichia coli</i>	86
C. Functional architecture of <i>Escherichia coli</i>: new insights provided by a natural decomposition approach	105
Bibliografía	118

Índice de figuras

1-1. Microfotografía de <i>E. coli</i> obtenida mediante microscopía electrónica	2
1-2. <i>E. coli</i> en la superficie del intestino delgado	3
1-3. <i>E. coli</i> sobre la superficie de la piel humana y un folículo piloso	3
1-4. Ilustración mostrando un corte transversal de una sección de <i>E. coli</i>	6
1-5. Niveles de organización molecular en la célula	8
1-6. El elefante y <i>Escherichia coli</i>	9
2-1. Composición química y organización del ADN	14
2-2. Expresión de los genes y procesamiento para obtener el producto final	16
2-3. Interacciones entre la ARN polimerasa y su promotor	19
2-4. Secuencia de iniciación de la transcripción en los promotores bacterianos	21
2-5. Dibujo original realizado por Jacques Monod, en 1960, ilustrando el operón	23
2-6. Modelo de organización del operón	23
2-7. Modelo de organización del regulón	24
2-8. Modelo de organización del modulón	25
3-1. El Uroboros tal como aparece en el <i>Synosius</i>	29
3-2. El Uroboros como aparece en la <i>Crisopea de Cleopatra</i>	30
3-3. Grabado de 1652 ilustrando la ciudad de Königsberg	31
3-4. Dibujo de los puentes de Königsberg realizado por Euler	32
3-5. Abstracción del mapa de los siete puentes de Königsberg	34
3-6. Tres modelos fundamentales de redes y sus propiedades distintivas	38
3-7. Representación artística del concepto de los seis grados de separación	39

3-8. Todos los distintos subgrafos conexos generados empleando tres nodos	49
3-9. Motivos topológicos encontrados en las redes de <i>E. coli</i> y <i>S. cerevisiae</i>	50
3-10. Los ocho tipos de circuitos <i>feedforward</i>	51
4-1. Ejemplo de red de comunidades	54
4-2. Ejemplo de red jerárquico-modular	55
4-3. Módulos identificados en la red de regulación de <i>E. coli</i>	57
4-4. Módulos de la red de regulación íntegra de <i>E. coli</i>	59
4-5. Módulos identificados en una subred condición-dependiente (WT+Glu/WT)	61
4-6. Estructura jerárquica multicapa propuesta por Ma <i>et al.</i>	65
5-1. Representación gráfica de la red de regulación transcripcional de <i>E. coli</i> recons- truida para este estudio	70
5-2. Identificación de nodos modulares y jerárquicos en la red de <i>E. coli</i>	72
5-3. Mapa de la organización modular y jerárquica de las subrutinas que componen el programa génico de <i>E. coli</i>	75

Índice de tablas

3-1. Principales características de los circuitos de retroalimentación	47
4-1. Jerarquía de la red de regulación de <i>E. coli</i> propuesta por Yu y Gerstein	66

Prólogo

Corría el año 2000; la economía de México superaba las expectativas y en general el país vivía una transición en diversos aspectos. Una fresca noche de aquel año, llegué por primera vez al entonces Centro de Investigación sobre Fijación de Nitrógeno —hoy Centro de Ciencias Genómicas— de la Universidad Nacional Autónoma de México (UNAM); en aquel entonces, ni idea tenía de la existencia de centros de investigación de la UNAM fuera de Ciudad Universitaria. Esa noche acompañaba al Dr. Luis Enrique Sucar, nos dirigíamos al Laboratorio de Biología Computacional para tener una reunión con el Dr. Julio Collado Vides.

En aquellos días, estudiaba una maestría en ciencias de la computación en el Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) bajo la dirección del Dr. Sucar. Mi línea de investigación versaba sobre cómo hacer que una computadora aprendiera a controlar un mecanismo complejo, a partir de analizar la conducta de un humano experto. Un año antes, este enfoque, denominado clonación conductista, había acaparado mi atención durante el seminario donde los investigadores propusieron las posibles líneas de trabajo para tesis. Me entusiasmaba el lograr que una computadora reprodujera una faceta del comportamiento humano.

Cierta noche durante una cena, Patty, mi pareja, me explicaba sobre lo que hacía en su trabajo: estudio de la regulación de los factores de virulencia en *Vibrio cholerae* —el agente causal del cólera—. En su plática, me hablaba sobre biología molecular, células, proteínas, reguladores, toxinas... ¡no alcanzaba a comprender lo que me decía!, le pedía más información con la finalidad de tratar de entender e imaginar los procesos que me explicaba.

La biología nunca fue mi fuerte. Recuerdo bien que, estando en tercero de secundaria, tuve que comprar varios libros sobre genética humana para poder pasar el examen extraordinario de biología. Ya en el bachillerato, el profesor de biología hablaba sobre las células, la membrana, el núcleo, la mitocondria, el ADN, etc. Entendía que esos elementos conformaban a los seres vivos pero, ¿dónde estaban?, ¿cómo eran?, ¿por qué eran tantos?

Cuando niño, me fascinaba la ciencia y poder descubrir el mundo oculto a mis ojos. En secundaria, para poder escribir los reportes del laboratorio de química repetía los experimentos en mi casa —desde entonces me concentraba y rendía mejor en mi soledad—, conectando un mechero de Bunsen a la salida del piloto de la estufa; los reactivos, así como los tubos de ensayo, los obtenía de mis juegos de química *Mi Alegría*; mientras que algunos otros instrumentos como un par de mecheros, matraces de Erlenmeyer y vasos de precipitado eran adquiridos en Médico Dental Arcam, un proveedor de productos médicos y químicos. Mi madre siempre me apoyó en ese aspecto, muchas veces convenciendo a mi padre, quien si bien apoyaba materialmente, nunca lo hizo moralmente.

Recuerdo, como si hubiera pasado ayer, que una navidad, tendría 12 o 13 años, mi madre me regaló un microscopio. Desde que lo vi me fascinó, llevaba ya tiempo anhelando tener uno. Además, no era uno de esos clásicos microscopios *Mi Alegría* tan comunes en aquellos tiempos; éste permitía —o permite, porque aún lo conservo— proyectar las imágenes sobre la pared y tenía una fuente de iluminación propia. ¡Era un nuevo mundo abierto ante mis ojos! Con él pude descubrir muchas cosas; estudiar las partes de un mosquito, una mosca o una araña; pero quería descubrir cosas más pequeñas, aquellas que escapaban a la simple vista. Sabía que las bacterias están por todas partes, que son las responsables de muchas enfermedades. Sin embargo, a pesar de mis múltiples esfuerzos, nunca pude ver una. Recuerdo bien haber leído con emoción, a uno de los costados de la caja del microscopio, una leyenda en inglés que decía: “Un pasatiempo hoy, una profesión mañana.” Y haber pensado: “¡sí!... a eso quiero dedicar mi vida”. El primer día de clases del primer año de secundaria, durante la clase de física el profesor nos pidió que nos presentáramos y dijéramos que queríamos ser cuando grandes. Cuando llegó mi turno, me paré y afirmé: “Quiero estudiar una licenciatura, una maestría y un doctorado.” Instantáneamente se escucharon carcajadas en todo el salón, todos reían incluyendo al profesor; yo no comprendía que pasaba.

En ese entonces, albores de los 80, la llegada de la microcomputadora abría una nueva era. Por primera vez, máquinas que ocupaban varios cuartos, y a veces varios pisos, estaban disponibles a precios asequibles y en dimensiones muy reducidas. Mi primera computadora la tuve a los 12 años. Mientras que mis esfuerzos por ver una célula fracasaban, mis correrías con la computadora iban en ascenso. No pasó mucho tiempo en que me encontré escribiendo mis primeros programas en lenguaje BASIC en una *Commodore 64*¹. Mi éxito con las computadoras desplazó, poco a poco, a otras ciencias, inclinando la balanza hacia las ciencias exactas. Eventualmente, me licencié como ingeniero en sistemas computacionales en el Instituto Tecnológico de Veracruz.

Durante los días subsecuentes a aquella plática con Patty, le pedí que me siguiera explicando; cada vez le solicitaba más detalles. Me trataba de hacer una idea de lo que me platicaba. Tras pensar en todo lo que me había dicho, llegué a la conclusión de que una bacteria, como *V. cholerae* o cualquier otra célula, monitorea sus alrededores detectando cualquier cambio y respondiendo mediante ciertos mecanismos predefinidos.

Repentinamente, todo se armó en mi cabeza; una célula es como un controlador que monitorea ciertas variables y responde a ellas cuando es necesario. ¿Qué pasaría si aplicáramos el enfoque de clonación conductista a la célula? Si registráremos los valores de ciertas variables y la decisión tomada por la célula, para luego alimentar estos datos a un algoritmo de aprendizaje, ¿sería posible predecir la reacción de una célula ante un cierto estímulo?

La idea me entusiasmó mucho, entonces decidí proponersela al Dr. Sucar; como mi director de tesis, pensé, quizá le interesará una nueva aplicación de la clonación conductista. Él escuchó con atención mi propuesta, entonces me comentó que conocía a un investigador, de nombre Julio Collado, que estudiaba cuestiones de biología empleando computadoras, y me dijo que, si así lo deseaba, él podría concertar una cita para ir a visitarlo a su laboratorio. Estuve de acuerdo y él programó la cita.

¹La *Commodore 64* (C64), con tan solo 64 kBytes de RAM y una velocidad de 1.023 MHz, fue la microcomputadora más popular de aquella época. Sólo como comparación, mi *Palm Tungsten E* —ya algo obsoleta en estos días— tiene 512 veces la RAM de la C64 y es 123.2 veces más veloz.

Durante la reunión, le platicué al Dr. Collado la idea de aplicar la clonación conductista para modelar una célula. Él nos escuchó con atención; empero, al terminar, nos expuso una serie de problemas de su interés. Mientras que la idea no fructificó, yo estaba interesado en iniciar un doctorado y se discutió la opción. Se me presentaron dos alternativas: estudiar el doctorado en el ITESM bajo la dirección del Dr. Sucar o estudiar el doctorado en la UNAM bajo la dirección del Dr. Collado. La primera opción sólo brindaba una beca-crédito del ITESM —lo cual incrementaría más mi deuda adquirida al haber estudiado la maestría con el mismo tipo de beca-crédito—; mientras que la segunda una beca Conacyt, pero con el inconveniente que tendría que ingresar en un programa de ciencias biológicas, lo que implicaba que debía dominar la biología, y en particular la biología molecular.

En diciembre de 2000 obtuve el grado de maestro en ciencias por el ITESM. Iniciando enero de 2001, mi madre fue diagnosticada con cáncer de seno, lo que me retuvo ese primer semestre en mi ciudad natal, Veracruz. La empeorante situación económica, más el respaldo de mi madre, me llevó a comunicarme con el Dr. Collado en agosto de ese año, la decisión estaba tomada, deseaba realizar el doctorado en la UNAM.

Los últimos meses de 2001 y los primeros de 2002 estuvieron dirigidos a estudiar biología molecular y empaparme de la parte biológica de las líneas de investigación del Laboratorio de Biología Computacional. Durante ese tiempo tuve la oportunidad de apoyar a la entonces estudiante de doctorado Rosa María Gutiérrez en su proyecto doctoral; basado en un conjunto de reglas propuestas por ella, desarrollé programas para predecir los estados de expresión de genes en *E. coli*; actividad que resultó muy benéfica al permitir compenetrarme, por primera vez, en el análisis de datos biológicos. Además, fui invitado por el Dr. Collado a participar en el desarrollo de los planes de estudio en las áreas de computación y matemáticas para la propuesta de la Licenciatura en Ciencias Genómicas. Paralelamente, con apoyo del Dr. Collado, desarrollé un anteproyecto de cuya evaluación dependía mi ingreso al Programa de Doctorado en Ciencias Biomédicas; en dicho anteproyecto se proponía estudiar y modelar el comportamiento dinámico de la red de regulación de la bacteria *E. coli*.

En junio de 2002, acudí al Instituto de Investigaciones Biomédicas a realizar la defensa de mi anteproyecto. Tras la deliberación del comité evaluador, su dictamen fue contundente: “Es un proyecto muy interesante y se ve que dominas bien la parte teórica; sin embargo, no sabes biología, por lo que consideramos que eres un candidato adecuado para un doctorado en tu área.”

Durante aquel verano, un artículo sobre redes de mundo pequeño [Watts y Strogatz, 1998] me brindó un panorama nuevo sobre las redes complejas, llevándome a cuestionarme si las mismas propiedades se cumplían en la red de regulación de *E. coli*. Durante mis vacaciones, me dedique a estudiar y planificar como evaluar estas propiedades en la red de regulación, ese fue mi primer contacto con el estudio de las propiedades topológicas de la red.

De regreso de las vacaciones, comuniqué la decisión de los evaluadores al Dr. Collado. Él ya tenía información sobre mi desempeño durante la defensa del anteproyecto y me ofertó una segunda, pero última, oportunidad; se intentaría el ingreso por dos flancos: Ciencias Biomédicas y Ciencias Bioquímicas.

Acudí al Instituto de Biotecnología, sede del Programa de Doctorado en Ciencias Bioquímicas, para informarme sobre los requisitos de ingreso. Cuando el coordinador de la unidad de

docencia vio que yo procedía de un área ajena a las ciencias biológicas me informó: “Para tener derecho a presentar el examen de admisión primero deberás cursar y acreditar la materia de Bioquímica que se imparte en el programa.” Dado que el examen de admisión se realizaría en noviembre, antes de que se entregaran las calificaciones finales de Bioquímica, mi aceptación quedaba condicionada a acreditar dicha materia. Ante esta situación, decidí realizar un cambio de estrategia consistente en suspender todas mis demás actividades y concentrarme de tiempo completo, desde agosto de 2002 hasta enero de 2003, en la materia de Bioquímica; algo que, aunque no entusiasmo mucho al Dr. Collado, a la postre mostraría sus frutos.

Presenté el examen de admisión, obteniendo el tercer lugar general y la primer posición de los candidatos a ingresar al doctorado, garantizándome esto obtener una beca de Conacyt con prioridad uno. En Bioquímica obtuve un promedio final de 9.0, quedando así garantizada mi admisión al programa doctoral.

En enero de 2003 realicé mi inscripción al Programa de Doctorado en Ciencias Bioquímicas. El siguiente requisito de permanencia era presentar la candidatura doctoral. Dado que ya había cursado Bioquímica, en el primer semestre de 2003 decidí no tomar materia alguna, el plan era dedicarme de tiempo completo a preparar mi examen de candidatura y en él mostrar ya los primeros resultados de mi trabajo. El Dr. Collado fue invitado a una estancia de tres meses en el Instituto Henri Poincaré en París, lo cual dificultaría mantenernos en contacto. Si bien él me invitó a que lo acompañara, yo no estaba seguro que fuera la mejor decisión, pensé que ya que estando allá seguramente surgirían otras inquietudes que me distraerían de mi objetivo principal; así, decliné su invitación.

Durante cinco meses de trabajo diseñe, desarrollé e implementé un modelo dinámico basado en las reglas desarrolladas por Rosa María Gutiérrez. El 27 de agosto de 2003 presenté mi defensa de candidatura doctoral. En esta ocasión el dictamen de los evaluadores fue muy positivo, sólo sugiriendo invitar al comité tutorial a un investigador experimentalista y acotar el modelo propuesto a uno más interpretativo que sirviera como base para uno predictivo en el futuro; así ese mismo día obtuve la candidatura al grado de doctor en ciencias.

En septiembre de 2003, me involucré en un estudio iniciado por el Dr. Osbaldo Resendis sobre las características topológicas de la red de regulación de *E. coli*. Ya había transcurrido un año desde mi primera incursión estudiando las propiedades topológicas de la red, y consideré que podría aportar elementos útiles al proyecto. Mi experiencia con la teoría de grafos me permitió identificar algunas desviaciones en los resultados del Dr. Resendis y apoyarlo en sus análisis. Posteriormente, por sugerencia del Dr. Collado, realicé un análisis de la red de regulación para identificar su modularidad, el cual enriquecí con una evaluación de la relevancia biológica de los módulos identificados. Durante el primer semestre de 2004, este trabajo se sometió para su publicación, siendo el primer estudio donde se mostró que la red de regulación de *E. coli* posee una organización modular [Resendis-Antonio *et al.*, 2005].

Aún me restaba cursar la materia de Biología Molecular, cursándola de octubre de 2003 a enero de 2004, y obteniendo el tercer lugar de mi clase con un promedio de 9.13. Esto fue una experiencia muy enriquecedora. Cada vez que cursaba una materia, el conocimiento aprendido le daba una nueva perspectiva a mis ideas sobre la regulación transcripcional y cómo abordar el problema de su modelaje.

Lo aprendido en Biología Molecular en conjunto con la experiencia ganada durante la preparación de mi candidatura y mi colaboración con el Dr. Resendis, hicieron que surgieran en

mí un conjunto de inquietudes respecto al modelaje de las redes de regulación. Llegué a la conclusión de que en lugar de tratar de abordar el problema de modelar la red completa —que *per se* podría llegar a ser intratable—, lo mejor era descomponerla en módulos y tratar de modelar cada uno de estos subsistemas en términos de sus entradas y salidas. Esto planteaba un nuevo dilema: ¿cómo particionar la red? Mientras que en el estudio realizado por Resendis *et al.* mostramos que la red posee una organización modular, la metodología empleada para realizar la partición presenta varios inconvenientes (ver capítulo 4).

En diciembre de 2004, le comuniqué al Dr. Collado mis observaciones y le hice saber que si bien el modelaje del comportamiento dinámico de la red me interesaba, desde mi perspectiva, primero tenía que atacar el problema de particionar la red en subsistemas bien definidos, por lo que en mi tesis me enfocaría principalmente en el análisis de la topología de la red que en su dinámica.

En esa dirección, uno de los primeros pasos que di fue definir un marco de trabajo para identificar todos los circuitos de retroalimentación existentes en una red de regulación. Sin embargo, en marzo de 2005, un par de trabajos por Ma *et al.* sacudieron varias preconcepciones que tenía sobre el tema [Ma *et al.*, 2004a,b]. En principio, estos investigadores afirmaban que la red era acíclica, algo que no podía creer dada la importancia de los circuitos de retroalimentación [Thomas y D'Ari, 1990]. Por otra parte, los autores proponían una estructura jerárquica gobernando la red; empero, esta estructura me parecía sospechosa al ubicar en capas superiores de la jerarquía a genes que carecen de efectos globales sobre la célula —algo conocido como pleiotropía—.

Para ese entonces ya había empezado a formar un pequeño grupo de trabajo, liderado por mí, con estudiantes de la Licenciatura en Ciencias Genómicas —en donde he dado clases desde agosto de 2004—, al cual eventualmente se incorporaría el Dr. Luis Gerardo Treviño Quintanilla. Así, decidí reevaluar las afirmaciones de Ma *et al.* respecto a los circuitos de retroalimentación.

De esta forma inició una aventura que me llevaría, eventualmente, a demostrar que la red de regulación de *E. coli* no es acíclica, a proponer un método matemático para la identificación de factores de transcripción jerárquicos y mostrar su correlación con los factores de transcripción globales conocidos, y a crear *de novo* un método objetivo para inferir la organización modular y jerárquica en redes de regulación, identificando elementos nunca antes descritos en ellas, como lo son los genes intermodulares y los genes sólo regulados por jerárquicos.

Esta obra describe dicha aventura y sus consecuencias.

JULIO AUGUSTO FREYRE GONZÁLEZ
Centro de Ciencias Genómicas
Universidad Nacional Autónoma de México
Cuernavaca, Morelos; México
Septiembre de 2008

Disección de la arquitectura funcional de la red de regulación transcripcional de *Escherichia coli*: Un enfoque de descomposición natural

por

Julio Augusto Freyre-González

Resumen

Si bien los proyectos “omicos” han brindado una lista exhaustiva de partes de la célula, ayudando a comprender la diversidad y complejidad de la maquinaria celular, cómo una célula coordina todos sus subsistemas para lograr tomar decisiones sigue siendo un tema difícil de comprender. A lo largo de su vida, las bacterias deben afrontar un ambiente cambiante. En consecuencia, para lograr subsistir requieren de un mecanismo para monitorear su entorno y responder de forma adecuada. Además, las bacterias pueden producir miles de proteínas u otros productos a partir de la información codificada en sus genes, imponiendo así restricciones espaciales y energéticas. ¿Cómo controla la célula la producción de sus productos génicos?

La regulación del inicio de la transcripción es el mecanismo principal que permite a las bacterias responder a un ambiente cambiante. Este mecanismo posibilita a las bacterias para controlar la síntesis de proteínas —las máquinas fundamentales de la célula—. Por más de 20 años se ha reconocido que las redes regulatorias se componen de circuitos complejos con diferentes niveles de control. Esto les permite controlar simultáneamente diferentes subrutinas del programa genético. De hecho, estas interacciones regulatorias dan origen a redes complejas, las cuales obedecen principios de organización que definen su comportamiento dinámico. La comprensión de estos principios es un reto actual. Se ha sugerido que redes que toman decisiones requieren de topologías específicas. Efectivamente, existen fuertes argumentos apoyando la noción de una organización modular en la célula. Un módulo se define como un grupo de elementos que cooperan para una función celular específica. En redes génicas, estos módulos deben comprender genes que responden de forma coordinada bajo la influencia de un estímulo específico.

Recientemente, análisis topológicos globales han sugerido la existencia de modularidad jerárquica en redes de regulación transcripcional. Trabajos previos han propuesto metodologías para inferir esta organización modular y jerárquica. Desgraciadamente, estos enfoques metodológicos han mostrado ser inadecuados para afrontar la presencia de circuitos *feedforward* y de retroalimentación, dos estructuras topológicas relevantes. Además, las conclusiones biológicas obtenidas con estos enfoques son contraintuitivas, dado que ubican, en las capas jerárquicas más altas, a factores de transcripción que reponen a condiciones muy específicas de la célula y los cuales por consecuencia carecen de efectos globales o pleiotrópicos.

En este estudio, la red de regulación transcripcional de *Escherichia coli* es desentrañada mediante separarla en sus componentes claves, revelando así su organización natural. Esta metodología permite la identificación de los elementos principales componiendo a una red de regulación transcripcional: genes jerárquicos, modulares, e intermodulares. Éste último un tipo importante de elemento nunca antes descrito ni identificado. Se propone un criterio matemático, basado en las características topológicas de la red, para clasificar cada gen en la red en uno de dos posibles clases: genes jerárquicos o modulares.

Se encontró que los genes modulares se aglutinan en grupos correlacionados fisiológicamente los cuales fueron validados por un análisis estadístico del enriquecimiento de clases funcionales. Mientras que los genes jerárquicos codifican para factores transcripcionales responsables de coordinar las respuestas de los módulos basándose en señales de interés general para la célula. En efecto, todos los factores transcripcionales globales conocidos en *E. coli* fueron identificados como genes jerárquicos, y se predicen dos nuevos, sugiriendo así el primer posible criterio matemático para identificar factores transcripcionales globales en una célula. Además, los genes intermodulares se definieron como genes estructurales los cuales integran, a nivel promotor, señal provenientes de módulos distintos, y por consecuencia de diferentes respuestas fisiológicas. Por último, empleando el concepto de pleiotropía, se desarrolló un método para reconstruir la jerarquía gobernando la red. Se mostró el papel que juegan los circuitos *feedforward* al interconectar diferentes niveles de organización, moldeando así la médula jerárquica de la red.

Además se desarrolló un algoritmo original para enumerar todos los circuitos de retroalimentación, compuestos por dos o más nodos, existentes en la red. Brindando así la primer enumeración y análisis sistémicos de la presencia y participación global de los circuitos de retroalimentación en la organización funcional de una red de regulación transcripcional.

Este estudio brinda nuevos elementos para comprender los principios de diseño fundamentando la organización de las redes de regulación transcripcional, mostrando una novedosa arquitectura no piramidal compuesta por módulos independientes gobernados globalmente por factores transcripcionales jerárquicos, cuyas respuestas son integradas por los genes intermodulares.

Disección de la arquitectura funcional de la red de regulación transcripcional de *Escherichia coli*: Un enfoque de descomposición natural

by

Julio Augusto Freyre-González

Abstract

While “omic” projects have provided a comprehensive parts list for the cell, shedding light on the diversity and complexity of the cell’s machinery, how a cell coordinates all their subsystems in order to make decisions is still an elusive topic. Through their life, bacteria must deal with a changing environment. As consequence, in order to survive they require a mechanism to sense their surroundings and respond in a proper way. In addition, bacteria can produce thousands of proteins or other products from information encoded in their genes, thus imposing restrictions on space and energy. How does the cell control the production of its genic products?

Regulation of transcription initiation is the main mechanism enabling bacteria to respond to a changing environment. This mechanism enables bacteria to control protein synthesis —the fundamental machines of the cell—. For more than 20 years it has been recognized that regulatory networks comprise complex circuits with different control levels. This makes them able to control different subroutines of the genetic program simultaneously. In fact, these regulatory interactions give rise to complex networks, which obey organizational principles defining their dynamic behavior. The understanding of these principles is currently a challenge. It has been suggested that decision-making networks require specific topologies. Indeed, there are strong arguments supporting the notion of a modular organization in the cell. A module is defined as a group of elements cooperating towards a specific cellular function. In genetic networks, these modules must comprise genes that respond in a coordinated way under the influence of specific stimuli.

Recently, global topological analyses have suggested the existence of hierarchical modularity in transcriptional regulatory networks. Previous works have proposed methodologies to infer this hierarchical modular organization. Unfortunately, these methodological approaches have shown to be inadequate to deal with feedforwards and feedback circuits, two relevant topological structures. In addition, biological conclusions obtained with these approaches are counterintuitive, as they place, in the highest hierarchical layers, transcription factors responding to very specific conditions of the cell and which therefore lack global or pleiotropic effects.

In this study, the transcriptional regulatory network of *Escherichia coli* is unraveled by separating it into its key elements, thus revealing its natural organization. This methodology enables the identification of the principal elements comprising a transcriptional regulatory network: hierarchical, modular, and intermodular genes. The latter an important type of elements never described before nor identified. A mathematical criterion, based on the topological features of the network, is proposed to classify every gene in the network into one of two possible classes: hierarchical or modular genes.

It was found that modular genes are clustered into physiologically correlated groups validated by a statistical analysis of the enrichment of the functional classes. While hierarchical genes encode transcription factors responsible for coordinating module responses based on signals of

general interest for the cell. Actually, all the known global transcription factors of *E. coli* were identified as hierarchical genes, and two new are predicted, thus suggesting the possible first mathematical criterion to identify global transcription factors in a cell. In addition, intermodular genes were defined as structural genes which integrate, at the promoter level, signals coming from different modules, and therefore from different physiological responses. Finally, using the concept of pleiotropy, a method to reconstruct the hierarchy governing the network was developed. It was shown the role of feedforward circuits bridging different organizational levels, thus shaping the hierarchical backbone of the network.

It was also developed a novel algorithm to enumerate all the feedback loops, comprising two or more nodes, existing in the network. Thus providing the first systems-level enumeration and analysis of the global presence and participation of feedback loops in the functional organization of a transcriptional regulatory network.

This study sheds new light on the design principles underpinning the organization of transcriptional regulatory networks, showing a novel nonpyramidal architecture comprised of independent modules globally governed by hierarchical transcription factors, whose responses are integrated by intermodular genes.

Capítulo 1

Prolegómenos

La biología ocupa, entre las ciencias, un lugar a la vez marginal y central. Marginal en cuanto que el mundo viviente no constituye más que una parte ínfima y muy “especial” del universo conocido, de suerte que el estudio de los seres vivos no parece poder lograr jamás la revelación de unas leyes generales, aplicables fuera de la biosfera. Pero si la ambición última de la ciencia entera es fundamentalmente, como creo, dilucidar la relación del hombre con el Universo, entonces es justo reconocer a la biología un lugar central puesto que es, entre todas las disciplinas, la que intenta ir más directamente al centro de los problemas que se deben haber resuelto antes de poder proponer el de la “naturaleza humana”, en unos términos que no sean metafísicos.

— JACQUES MONOD, *El azar y la necesidad* (1970)

Nuestro planeta tierra es habitado por una gran variedad de organismos, los cuales para su estudio se clasifican en tres grandes dominios: *bacteria*, *archaea* y *eukarya*. La célula puede ser vista como una fábrica automatizada y autosostenida, en la cual las máquinas (proteínas) transforman la materia (metabolitos), dependiendo de las circunstancias, ya sea en energía, necesaria para el funcionamiento de la misma fábrica, o en piezas para la construcción de más máquinas. Empero, esta fábrica realmente es un saco de compuestos químicos, cada uno con distintas actividades, entre los que se destacan las enzimas (una de las máquinas centrales para la célula), las cuales son catalizadores que aceleran la velocidad de una reacción química. Cada célula es responsable de producir sus enzimas, teniendo miles distintas. Si en un momento determinado, una célula produjera la mayoría de sus enzimas, entonces tendríamos un caos tanto en términos de espacio como energéticos. Así, una célula debe controlar estricta y eficientemente las decisiones de qué enzimas producir y en qué cantidad. ¿Cómo se organiza esa red de toma de decisiones en la célula? A lo largo de este estudio nos enfocaremos en contestar esta pregunta.

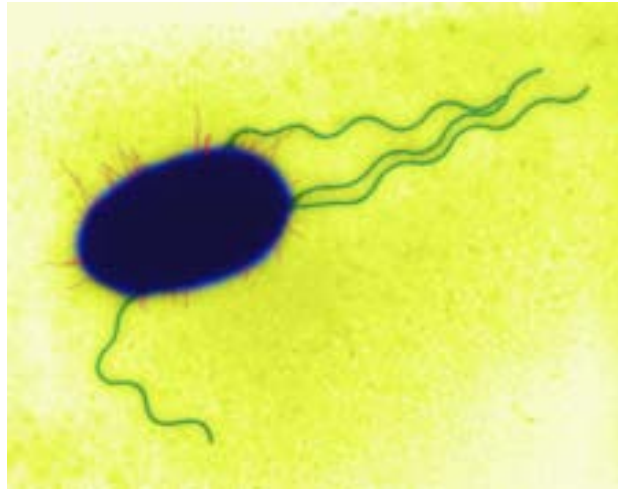


Figura 1-1: Microfotografía de *E. coli* obtenida mediante microscopía electrónica de transmisión con un aumento de 3, 515 X. En verde se observan los flagelos que permiten nadar a la bacteria; mientras en rojo se pueden ver las fimbrias que ayudan a la bacteria a adherirse a las superficies, las unas a otras, o a otras células. (Derechos reservados. Copyright © Dennis Kunkel Microscopy, Inc.)

1.1. ¿Qué es *Escherichia coli*?

Escherichia coli es una bacteria que fue descrita por primera vez en 1885 por el pediatra germano-austríaco Theodor Escherich como *Bacterium coli commune*, la cual fue aislada de las heces de los recién nacidos; ver figura 1-1. En condiciones normales, el tracto intestinal de muchos animales de sangre caliente es colonizado en las primeras horas o días después de nacer. El intestino humano es colonizado a las 40 horas de haber nacido, siendo la bacteria adquirida con los alimentos, el agua o por las personas que entran en contacto con el bebé, ya que ésta puede sobrevivir fuera del organismo; ver figura 1-3. Una vez dentro del organismo, ésta se adhiere a la mucosa del intestino grueso donde puede permanecer como un comensal benigno, conformando el 0.1 % de las bacterias totales en el intestino adulto; ver figura 1-2. *E. coli* juega un papel benéfico para el ser humano al producir vitamina K₂ y evitar la colonización por otras bacterias patogénicas.

No obstante, no todas las *E. coli* son benignas; de hecho, en 1935 se mostró que *E. coli* había causado un brote de diarrea en niños. Actualmente, se sabe que existen cepas o variantes

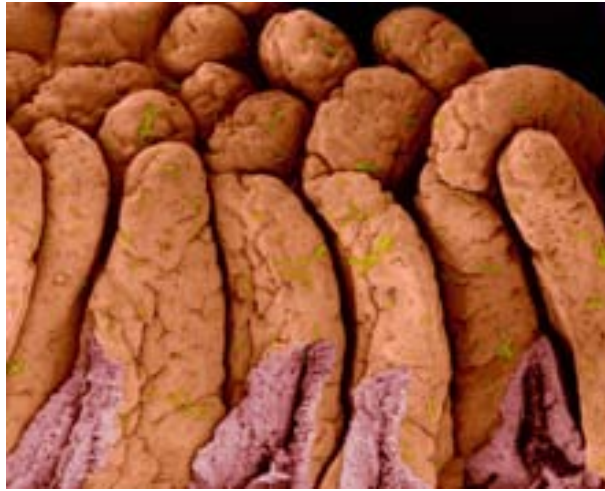


Figura 1-2: *E. coli* en la superficie del intestino delgado. Fotocomposición realizada mediante microscopía electrónica de barrido; aumento de la bacteria 200 X, intestino 49 X. En verde se observa a la bacteria. (Derechos reservados. Copyright © Dennis Kunkel Microscopy, Inc.)

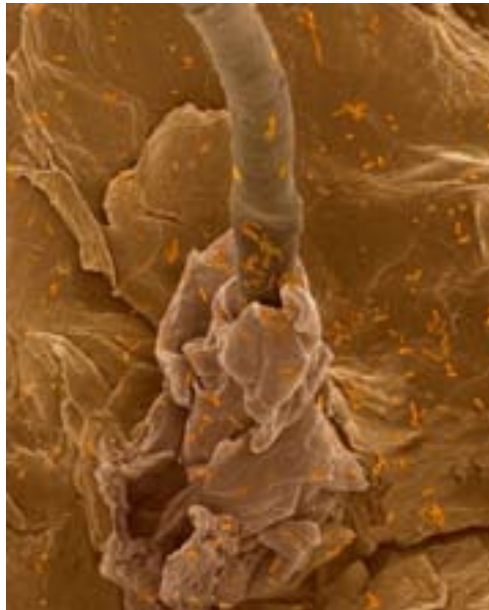


Figura 1-3: *E. coli* sobre la superficie de la piel humana y un folículo piloso. Fotocomposición realizada por microscopía electrónica de barrido; aumento de la bacteria 200 X, piel y folículo 260 X. En naranja se observa a la bacteria. (Derechos reservados. Copyright © Dennis Kunkel Microscopy, Inc.)

virulentas, siendo la O157:H7 la más destacada, capaces de producir gastroenteritis con o sin diarrea sanguinolenta, infecciones del tracto urinario, y meningitis neonatal; entre otros problemas de salud. De hecho, en septiembre de 2006, *E. coli* se hizo infamemente popular entre la población mundial debido a un brote de enfermedad que abarcó 28 estados de los Estados Unidos. Éste fue causado por espinacas contaminadas con *E. coli* O157:H7; dejando un saldo de, al menos, 276 enfermos y 3 muertos. En noviembre y diciembre de ese mismo año se originó un segundo brote en las cadenas de restaurantes Taco Bell y Taco John's; en esta ocasión la causa fue lechuga preempacada contaminada.

La cepa *E. coli* empleada en este estudio es la K-12, la cual es una cepa cultivable y bien adaptada al laboratorio que, a diferencia de la cepa silvestre¹, ha perdido su capacidad de colonizar el intestino. *E. coli* K-12 es una bacteria que ha servido como organismo modelo² para una gran cantidad de estudios en la biología molecular, siendo así la bacteria más estudiada a la fecha.

1.2. ¿Qué es la regulación transcripcional?

El metabolismo de un organismo es una intrincada red de reacciones químicas que se encarga de extraer la energía de los nutrientes y producir los componentes de la célula, permitiéndole así crecer, reproducirse, responder a estímulos, etc. Las enzimas catalizan muchas de las reacciones químicas del metabolismo y son fabricadas por la misma célula siguiendo instrucciones precisas codificadas en los genes. En cada célula existe una gran molécula de ácido desoxirribonucleico (ADN); los genes son fragmentos de esta molécula, y cada organismo puede tener cientos o miles de ellos codificados en su ADN.

Durante finales del siglo XX e inicios del siglo XXI, las investigaciones en biología molecular dieron origen a uno de los proyectos más ambiciosos en la historia del hombre: la secuenciación del genoma humano, que produjo la lista de todos sus genes. Actualmente, también se ha secuenciado el genoma de muchos otros organismos; de hecho, al 15 de julio de 2008, se han terminado de secuenciar 746 organismos de los tres dominios de la vida, mientras que otros 1, 431

¹Forma típica que ocurre en la naturaleza.

²Un organismo modelo es una especie la cual es ampliamente estudiada para comprender fenómenos biológicos particulares, con la confianza de que nos brindará pistas sobre estos fenómenos en otros organismos.

están siendo secuenciados. Sin embargo, una lección aprendida de los proyectos de secuenciación es que tener la lista de genes no es suficiente para comprender como responde un organismo a su ambiente. El genoma de un organismo es sólo una lista de genes, algo así como tener el diccionario de un idioma extranjero que no conocemos. Supongamos que deseamos construir una oración en dicho idioma, ¿será suficiente con sólo tomar las palabras del diccionario y expresarlas en el orden que creemos correcto? La respuesta concreta es: no, no será suficiente; dado que cada idioma, además de un vocabulario, tiene un conjunto de reglas gramaticales que indican el orden y forma en que las palabras se deben de expresar. Análogamente, no es suficiente con sólo tener la lista de genes de un organismo para saber cómo éste va a responder a un estímulo dado, es necesario también conocer las reglas que establecen cómo estos genes deben ser expresados en tiempo y cantidad.

Una célula, como la de la bacteria *E. coli*, consiste de una compleja maraña de moléculas que interactúan unas con otras siguiendo reglas específicas; ver figura 1-4. Dado que cada gen codifica un producto, *e.g.*, una enzima, esta complejidad es también una limitante para la expresión de los genes. Si en un momento determinado, todos los genes o una mayoría, expresaran sus productos, estaríamos ante un caos. Por un lado, sería imposible albergar, en un espacio tan reducido, todas las moléculas que una célula puede producir; por otra parte, dado que producir cada molécula requiere de energía, producirlas sin requerirlas sería un derroche energético. Debido a esto, una célula debe poseer mecanismos que le permitan controlar de forma eficiente qué genes y en qué cantidad deben de expresarse en un instante determinado. Así, una célula produce ciertas proteínas que actúan como sensores, permitiéndole convertir la información del medio en mensajes que la célula puede entender. Estos mensajes son procesados por ciertas proteínas llamadas factores de transcripción, las cuales interactúan unas con otras y con el ADN formando así una red de interruptores que permiten tomar una decisión sobre cómo responder al mensaje. En bacterias, el mecanismo de control de la respuesta celular más importante se denomina regulación transcripcional, el cual se encarga de decidir qué genes deben ser expresados y en qué cantidad en un momento determinado. Este mecanismo fue propuesto por primera vez por François Jacob y Jaques Monod, quienes por su descubrimiento recibieron, en 1965, el Premio Nobel de Medicina.

La preponderante importancia de las redes de regulación en la toma de decisiones de una

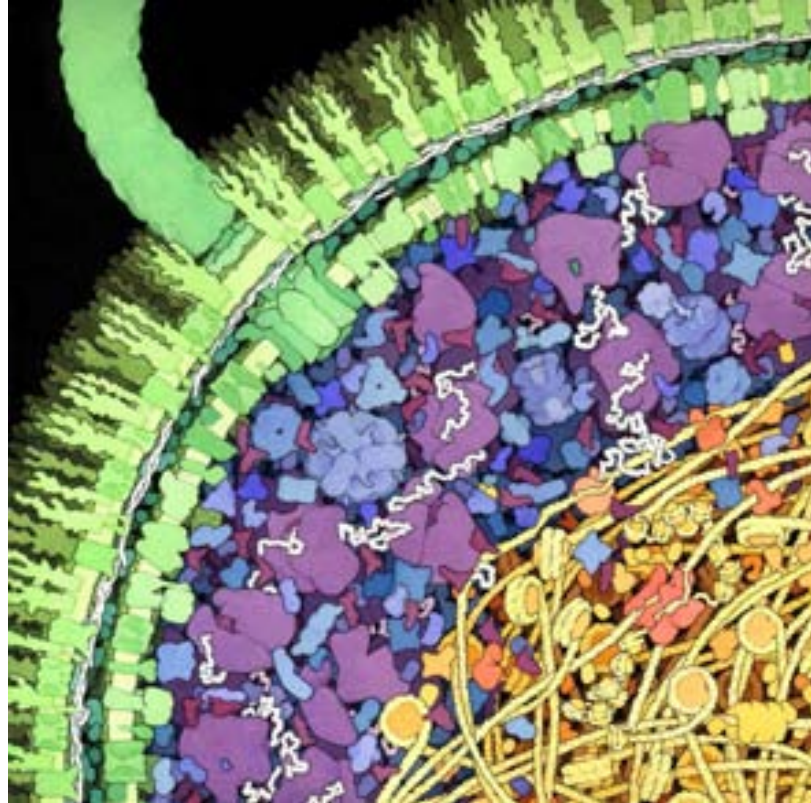


Figura 1-4: Ilustración dibujada a mano mostrando un corte transversal de una sección de *E. coli*. En esta excelente ilustración, una de las más apegadas a la realidad, se observa la extraordinaria complejidad del interior de una célula. En verde se observa la pared celular, con dos membranas concéntricas tachonadas con proteínas transmembranales. En la esquina superior izquierda, se puede observar el flagelo cruzando la pared celular y extendiéndose hacia arriba. En azul y púrpura podemos ver la región del citoplasma. Las moléculas grandes en púrpura son los ribosomas, las moléculas pequeñas en marrón son los ARNt (ácidos ribonucleicos de transferencia) y las hebras blancas son los ARNm (ácidos ribonucleicos mensajeros); las enzimas aparecen en azul. La región del nucleoide se muestra en amarillo y naranja. El ADN aparece en amarillo, enrollado alrededor de proteínas HU (nucleosomas bacterianos). (Técnica mixta tinta acuarela, David S. Goodsell, 1999.)

célula ha ocasionado que éstas sean objeto de estudio por la comunidad científica internacional. Dado que las redes de regulación son los elementos de control principales por medio de los cuales una bacteria patógena, como *Vibrio cholerae* el agente causal del cólera, decide cuándo y en qué cantidad producir las proteínas que le permiten colonizar y desarrollar la infección, su estudio también es de suma importancia para la medicina. La comprensión de la estructura y organización de las redes de regulación nos permitirá, eventualmente, entender mejor cómo y bajo qué condiciones un organismo patógeno desarrolla la infección y la subsecuente enfermedad, posibilitándonos así el diseño de nuevas estrategias de tratamiento. Pero los beneficios del estudio de la regulación transcripcional no se limitan sólo a bacterias y sus enfermedades asociadas; por el contrario, también se han dado pasos importantes que han mostrado el papel que la regulación transcripcional juega, por ejemplo, vía el sistema p53 en la patogénesis de ciertos tipos de cáncer y su combate por el organismo humano [Liu y Chen, 2006; Millau *et al.*, 2008], así como en neuropatías como el Alzheimer [Theuns y Broeckhoven, 2000].

1.3. ¿Por qué un enfoque de descomposición natural?

En una célula, las moléculas interactúan unas con otras siguiendo reglas bien definidas, permitiéndole de esta forma realizar funciones específicas. Como ya mencionamos, en bacterias, el mecanismo de toma de decisiones ocurre mediante la interacción de un conjunto de moléculas. ¿Cómo se organizan las interacciones entre las distintas moléculas tal que una célula pueda tomar decisiones?

Se ha propuesto que, a pesar de exhibir una aparente complejidad, la célula para ser funcional debe poseer una organización no azarosa, en la cual ciertos grupos de moléculas están encargados de funciones fisiológicas específicas; conformando así grupos o módulos de moléculas que se interconectan originando una organización modular [Hartwell *et al.*, 1999]. La comprensión de cómo estos módulos se interconectan es un reto actual de la biología molecular, cuyo estudio nos ha permitido develar algunos de los principios de diseño detrás de la circuitería biológica [Oltvai y Barabási, 2002]; mostrándose que, por ejemplo, estos módulos están sujetos a una organización jerárquica de alto nivel [Ravasz *et al.*, 2002; Ravasz y Barabási, 2003]; ver figura 1-5.

De hecho, estas propiedades topológicas a gran escala parecen estar altamente conservadas

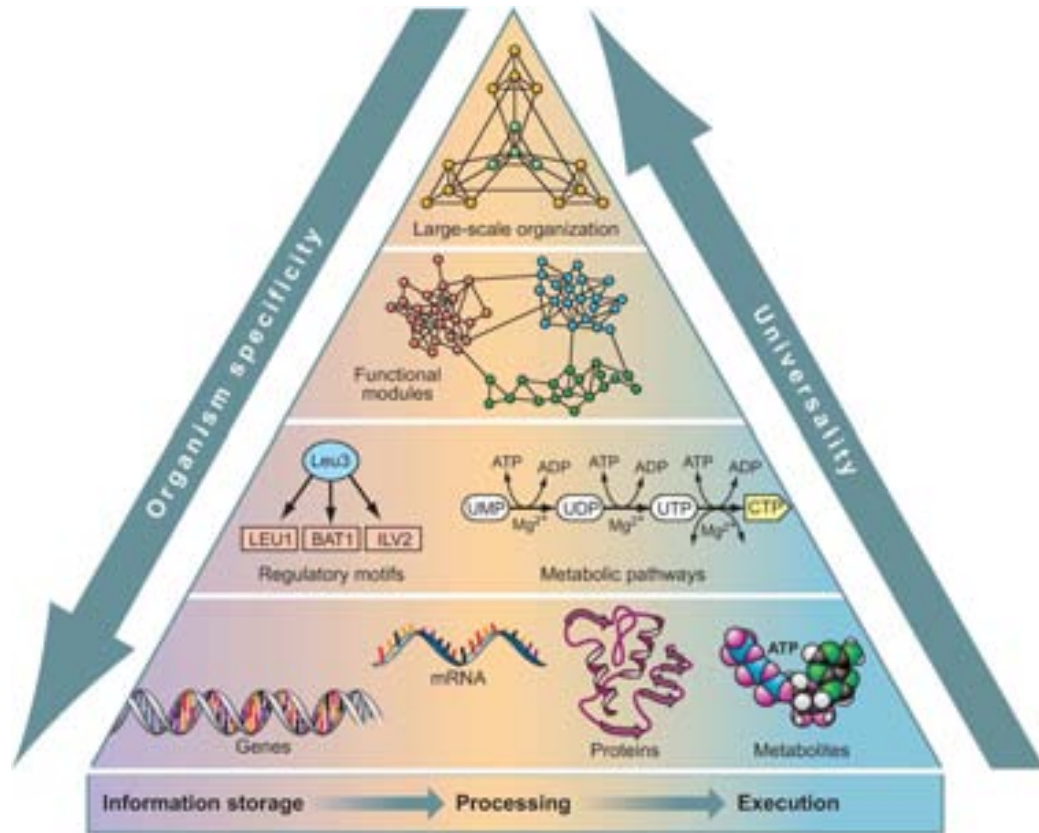


Figura 1-5: Niveles de organización molecular en la célula. En la primer capa de la pirámide se muestran los diferentes tipos de moléculas que están presentes en una célula, conformando, respectivamente, el genoma, transcriptoma, proteoma, y metaboloma; los cuales, además, suelen ser específicos para cada organismo (nivel 1). Estas diferentes moléculas interactúan entre sí para dar origen a subestructuras como las que aparecen en las redes de regulación o en las redes metabólicas (nivel 2). Dichas subestructuras a su vez se agrupan en módulos encargados de funciones fisiológicas particulares (nivel 3). Finalmente, estos módulos se interconectan generando una estructura jerárquica que gobierna a los módulos (nivel 4). Se ha mostrado que, mientras el nivel 1 de la pirámide es específico para cada organismo, las propiedades topológicas de las redes celulares tienden a estar compartidas entre diversos organismos. (Imagen tomada de Oltvai y Barabási, 2002.)



Figura 1-6: El elefante y *Escherichia coli*. Al interior, repetido múltiples veces, se lee la frase “*Tout ce qui est vrai pour le Colibacille est vrai pour l'éléphant*” (“Todo lo que es verdadero para *E. coli* es verdadero para el elefante”). (Aquarela, diciembre de 1972. Archivos del Instituto Pasteur, MON.Bio.20, expediente 24.2.)

en diversas redes biológicas tanto de bacterias como de organismos superiores, llevándonos a recordar aquellas palabras proféticas de Jaques Monod “Todo lo que es verdadero para *E. coli* es verdadero para el elefante”; ver figura 1-6.

Desde 2002, se han desarrollado diversos enfoques para intentar extraer la organización modular y jerárquica de las diversas redes de interacciones moleculares en una célula. Sin embargo, como veremos en el capítulo 4, mientras que cada enfoque posee sus ventajas y desventajas, ninguno de ellos se ha fundamentado en las características intrínsecas que las redes biológicas exhiben para extraer de forma natural su organización; por el contrario, de una u otra forma, las técnicas hasta ahora desarrolladas requieren de métodos artificiales que emplean parámetros imposibles de cuantificar con precisión; ocasionando así que, algunas veces,

sus conclusiones sean irrelevantes o erradas desde una perspectiva biológica.

Por otra parte, el enfoque de descomposición natural propuesto en este estudio explota las características inherentes a una red de regulación transcripcional, para generar un método que permite, mediante la caracterización y desconexión controlada de todas las moléculas participantes, revelar de forma natural la organización modular y jerárquica de la red.

1.4. Descripción de la investigación

El objetivo de esta investigación es el de *disectar y estudiar la arquitectura funcional de la red de regulación transcripcional de la bacteria Escherichia coli K-12 mediante un enfoque de descomposición natural*. Para lograr esto, este estudio se encuentra subdividido en seis capítulos y tres apéndices de la siguiente forma:

Capítulo 1. Introducción. Se introduce al tema del estudio, así como los conceptos básicos de regulación transcripcional y teoría de redes. Además, se exponen las motivaciones que condujeron a esta investigación. Finalmente, se describe la estructura y organización del estudio.

Capítulo 2. Control Transcripcional de la Expresión Genética. Se presentan los conceptos básicos sobre cómo toman decisiones las bacterias para poder afrontar de forma eficiente su ambiente cambiante. Para ello, se introduce al lector en los conceptos elementales de la biología molecular de la célula, explicándole desde qué tipos de moléculas la componen, pasando por los principales mecanismos de control del inicio de la transcripción, hasta los niveles de organización de las redes de regulación. Además, se discute la importancia de los factores de transcripción globales para la organización y la toma de decisiones en las redes de regulación.

Capítulo 3. Teoría de Redes. Se expone el desarrollo de la teoría de redes, desde la antigüedad hasta nuestros días, haciendo especial énfasis en los hitos. Además, se exponen los conceptos básicos tras dicha teoría y se resumen los principales resultados de investigación que han conducido al desarrollo de nuevos paradigmas para comprender la naturaleza como un conglomerado de redes complejas.

Capítulo 4. Modularidad y Jerarquía en Redes Biológicas. Se presentan los conceptos de modularidad y jerarquía como los pilares fundamentales de la organización de las redes biológicas. Además, se exponen las investigaciones fundamentales al respecto, incluyendo algunas en las cuales colaboré, y se analizan las ventajas y desventajas de las técnicas tradicionales empleadas para inferir la organización modular y jerárquica de las redes de regulación.

Capítulo 5. Disectando la Arquitectura Funcional de *E. coli*. Se introduce nuestro objeto de estudio, la red de regulación transcripcional de *Escherichia coli* K-12, dando nociones sobre su complejidad. Se establecen las bases teóricas y biológicas de los métodos desarrollados en esta investigación, así como los resultados de la disección de la arquitectura funcional de la red de regulación de *Escherichia coli* K-12.

Capítulo 6. Conclusiones y Perspectivas a Futuro. Se exponen las conclusiones de la investigación, así como las líneas futuras de trabajo para contestar la nuevas preguntas surgidas de este estudio.

Apéndice A. Modular analysis of the transcriptional regulatory network of *E. coli*.

Primer artículo relacionado con el tema de esta investigación, en el cual colaboré como coautor. En éste se explora la modularidad de la red de regulación de *E. coli*.

Apéndice B. Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. Segundo artículo relacionado con el tema de esta investigación, en el cual colaboré como coautor. En éste se estudia la respuesta global de la red de regulación de *E. coli* a la presencia de glucosa en el medio, identificandose módulos en una red condición-dependiente.

Apéndice C. Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach. Artículo de investigación en el cual se reporta la técnica desarrollada en este estudio y los resultados de la disección de la red de regulación de *E. coli*.

Capítulo 2

Control Transcripcional de la Expresión Genética

El estudio de estos sistemas microscópicos nos revela, en fin, que la complejidad, la riqueza y la potencia de la red cibernética, en los seres vivos, sobrepasan con mucho lo que el estudio de las solas performances globales de los organismos podría jamás dejar entrever. E incluso cuando esos análisis están lejos aún de suministrar una descripción completa del sistema cibernético de la célula más simple, revelan que todas las actividades, sin excepción, que concurren al crecimiento y a la multiplicación de esta célula están, directamente o no, subordinadas unas a otras.
— JACQUES MONOD, *El azar y la necesidad* (1970)

En el genoma de todo organismo se encuentran codificadas las instrucciones necesarias para fabricar la mayoría de los elementos que requiere la célula para sobrevivir. Estos elementos deben ser producidos en el momento correcto y en las cantidades adecuadas para que una bacteria pueda afrontar exitosamente su ambiente cambiante. El principal mecanismo que controla la expresión de los genes en bacterias es la regulación del inicio de la transcripción. En este capítulo, revisaremos las generalidades de este mecanismo en bacterias, partiendo de las interacciones moleculares fundamentales hasta llegar a las propuestas sobre su organización global.

2.1. Los componentes químicos de la célula

Las moléculas esenciales para la célula se pueden agrupar en tres tipos: los *ácidos nucleicos*, las *proteínas*, y los *metabolitos*. Los dos primeros tipos son *polímeros*, y por ende *macromolé-*

culas¹. Para aclarar qué es un polímero imaginemos una fila de elefantes donde cada elefante usa su trompa para agarrar la cola del siguiente; así, cada elefante es lo que se conoce químicamente como monómero, la trompa agarrada de la cola es el tipo de enlace químico que une los monómeros, y la fila de elefantes es el polímero.

2.1.1. Ácidos nucleicos

Los ácidos nucleicos se presentan en la célula en dos formas: *ácido desoxirribonucleico* (ADN) y *ácido ribonucleico* (ARN). Ambos ácidos nucleicos poseen un esqueleto formado por un azúcar de cinco carbonos (una pentosa, que puede ser una desoxirribosa para el ADN o una ribosa para el ARN) y un grupo fosfato, uniéndose a este esqueleto una parte variable, la *base nitrogenada*, formando así lo que se conoce como *nucleótido*; ver figura 2-1. Los ácidos nucleicos son polímeros, unidos por enlaces fosfodiéster, de cuatro posibles nucleótidos: adenina (A), citosina (C), guanina (G), y timina (T) sólo para el ADN o uracilo (U) sólo para el ARN. Debido a esto, la información codificada en un ácido nucleico puede ser leída como una secuencia de letras, aunque el alfabeto que se emplea depende del tipo de ácido nucleico; para el ADN éste es {A, C, G, T}, mientras que para el ARN es {A, C, G, U}.

El ácido desoxirribonucleico

Cada polímero de nucleótidos forma una hebra de ADN; sin embargo, el ADN se forma por un par de hebras que interactúan de forma específica formando así una doble hélice, algo así como una especie de escalera de caracol; ver figura 2-1. Las interacciones específicas que ocurren en el ADN siguen el apareamiento Watson-Crick; esta regla señala que, mediante un tipo débil de enlace químico conocido como puente de hidrógeno, una adenina siempre se enlaza con una timina, mientras una citosina siempre interactúa con una guanina, haciendo que una hebra de ADN sea, de cierta forma, el negativo de la otra; algo conocido como *complementariedad*, lo cual nos garantiza que la información codificada en una hebra del ADN es la misma que la presente en la segunda hebra. Cada célula posee una molécula de ADN; ciertos segmentos de ella son los genes, los cuales codifican la información que le permite a la célula fabricar las proteínas y los ARN que requiere; ver figura 2-2. Al conjunto de genes de un organismo se le denomina

¹Moléculas formadas por un gran número de átomos, y que en consecuencia tienen una masa molecular elevada.

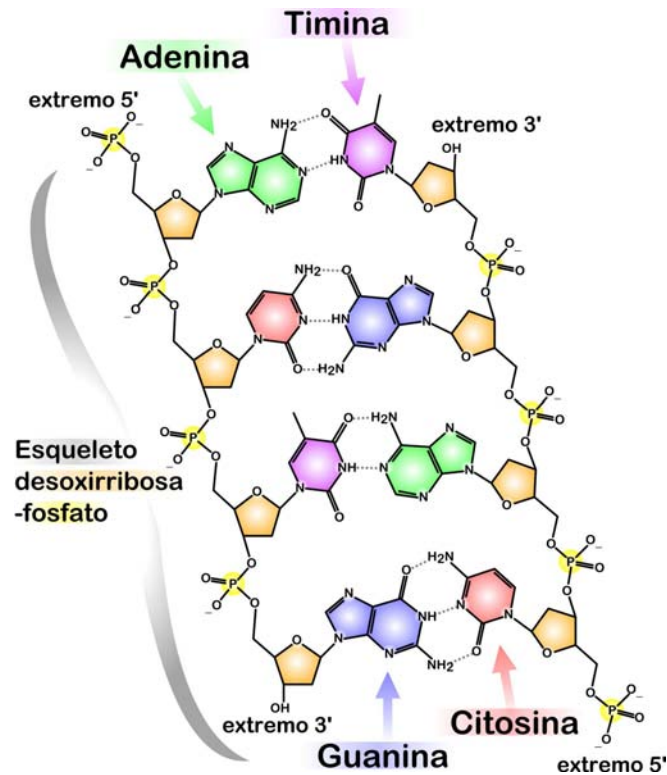


Figura 2-1: Composición química y organización del ADN. El esqueleto, señalado por la llave gris, está compuesto por un grupo fosfato resaltado en amarillo y una desoxirribosa en naranja. En verde, púrpura, azul y rojo se muestran las cuatro bases nitrogenadas; como líneas punteadas grises se muestran los puentes de hidrógeno que enlazan estas bases.

su *genoma*. Finalmente, el ADN puede enrollarse alrededor de ciertas proteínas, tal como un hilo se enrolla alrededor de un carrete, formando así el *cromosoma*. Este *superenrollamiento* del ADN juega dos papeles importantes, uno estructural y otro regulatorio; en su rol estructural permite compactarlo, ocupando así menos espacio en la célula; mientras que su papel regulatorio permite modificar la expresión de un conjunto de genes.

El ácido ribonucleico

En las células bacterianas los ARN son producidos a través de un proceso conocido como *transcripción*, en el cual la información de uno o más genes se copia a una molécula de ARN; este proceso es catalizado por una enzima conocida como *ARN polimerasa*, empleando como patrón o molde a una de las hebras de ADN; ver figuras 2-2 y 2-3. Los ARN se clasifican

de acuerdo a su función en la célula, siendo los principales: el ARN mensajero (ARNm), el ARN de transferencia (ARNt), y el ARN ribosomal (ARNr). Algunos ARN poseen actividades catalíticas como es el caso de las ribozimas, mientras que otros poseen actividades regulatorias como los ARN de interferencia (ARNi).

2.1.2. Proteínas

Las proteínas son polímeros de *aminoácidos* que desempeñan un papel fundamental en la célula, abarcando un amplio rango de funciones: enzimáticas, de transporte, receptoras o sensoras, estructurales, etc. En total existen 20 aminoácidos diferentes, cada uno de los cuales posee una cadena variable llamada residuo y dos extremos: el grupo amino y el grupo carboxilo. El polímero se crea al unir el extremo carboxilo de un aminoácido al extremo amino del siguiente mediante un enlace llamado *peptídico*, ocasionando esto que una proteína tenga dos dominios: el *amino-terminal* y el *carboxilo-terminal*, los cuales corresponden, respectivamente, con el extremo amino del primer aminoácido y el extremo carboxilo del último en la proteína.

El *código genético* establece que tres bases nitrogenadas codifican para un cierto aminoácido, como existen un total de cuatro bases nitrogenadas tenemos $4^3 = 64$ posibles combinaciones, conocidas como *tripletes* o *codones*; dado que sólo existen un total de 20 aminoácidos, necesariamente más de un codón debe codificar para un mismo aminoácido; a esto se le conoce como *degeneración del código genético*.

Una célula es capaz de fabricar las proteínas que requiere a partir de la información codificada en los genes. Primero la información del gen es copiada en una molécula de ARN conocida como *ARNm*². Posteriormente, este ARNm es leído, triplete a triplete, por el ribosoma y, siguiendo el código genético, es traducido en una proteína mediante la polimerización de aminoácidos catalizada por el mismo ribosoma; ver figura 2-2. Al conjunto de todos los ARNm que produce una célula se le conoce como *transcriptoma*, mientras que al conjunto de todas las proteínas que una célula fabrica se le llama *proteoma*.

²Es posible que un ARNm contenga la información transcrita de más de un gen, en cuyo caso estamos ante un ARNm policistrónico; esto ocurre cuando se transcribe un operón, lo que se discutirá más adelante.

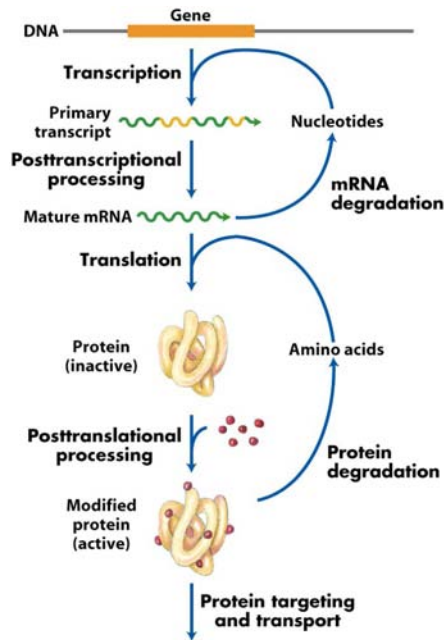


Figura 2-2: Expresión de los genes y procesamiento para obtener el producto final. Siete procesos afectan la concentración en estado estacionario de una proteína, cada uno de dichos procesos puede poseer diversos puntos de regulación. (Imagen tomada de Nelson y Cox, 2000.)

2.1.3. Metabolitos

El metabolismo es una compleja red de reacciones y procesos fisicoquímicos que son la base de la vida a nivel molecular. Esta red se divide en dos procesos íntimamente relacionados: el catabolismo y el anabolismo. El primero, se encarga de extraer la energía capturada en los enlaces químicos mediante la degradación de ciertos compuestos. El segundo, emplea la energía obtenida por el primero para construir moléculas complejas como lo son las proteínas y los ácidos nucleicos. Cualquier molécula producida o consumida por una reacción metabólica es llamada *metabolito*, y pueden ser desde moléculas muy simples y pequeñas como el dióxido de carbono (CO_2), hasta macromoléculas como las proteínas. Algunos metabolitos juegan un papel primordial al ser las señales que permiten activar o desactivar, mediante un mecanismo conocido como *alosterismo*, a ciertas moléculas como las proteínas regulatorias o algunas enzimas, permitiendo de esta forma reconfigurar la respuesta de una célula en función de su estado metabólico; ver figura 2-2 (los metabolitos aparecen en rojo).

2.2. Control del inicio de la transcripción

El elemento central de la transcripción en bacterias es la ARN polimerasa dependiente de ADN, la cual juega un papel indispensable en las cuatro fases de la transcripción: *contacto*, *isomerización*, *iniciación* y *elongación*. El núcleo de esta macromolécula se conforma por varias proteínas o subunidades siguiendo la composición $\alpha_2\beta\beta'\omega$; cada una de estas subunidades posee un papel importante:

- α_2 . Estas subunidades mantienen ensamblada la enzima y sirven para reconocer ciertas secuencias de control de la transcripción denominadas secuencias UP.
- β . Ésta posee el sitio activo con la actividad de ARN polimerasa, siendo activa durante las fases de iniciación y elongación del transcrito.
- β' . Reconoce y se une al ADN de forma no específica.
- ω . Mientras que esta subunidad no tiene un papel directo en la transcripción, funciona como un auxiliar o chaperona ayudando a la subunidad β' a lograr su plegamiento adecuado.

Otro elemento importante para que ocurra la transcripción es la *secuencia promotora* o *promotor*, el cual es una región de ADN justo delante del gen que sirve como secuencia de control indicándole a la ARN polimerasa donde debe iniciar su transcripción. Aun cuando el núcleo de la ARN polimerasa es capaz de reconocer y contactar al ADN de forma no específica, le es imposible iniciar la transcripción ya que para ello requiere poder identificar una región promotora, capacidad que sólo le otorga otra subunidad denominada *factor* σ . Cuando el factor σ se une al núcleo de la ARN polimerasa se forma lo que se conoce como la *holoenzima*, la cual tiene la composición $\alpha_2\beta\beta'\omega\sigma$. Este factor σ incrementa la especificidad de la ARN polimerasa por secuencias promotoras, permitiéndole a la ARN polimerasa dirigirse a transcribir los genes que se requiere transcribir; además, el factor σ posiciona a la holoenzima de la ARN polimerasa en el promotor y facilita el desenrollado de la doble hélice del ADN en el sitio de inicio de la transcripción. Muchas bacterias poseen múltiples factores σ , los cuales reconocen conjuntos distintos de promotores; sin embargo, a excepción de la pequeña familia σ^{54} , todos poseen características comunes.

2.2.1. Arquitectura del promotor bacteriano

El promotor bacteriano se compone de cuatro elementos, algunos de los cuales pueden estar o no presentes; ver figura 2-3. Los dos más importantes son el hexámero³ -10 y el hexámero -35 , conocidos comúnmente como cajas -10 y -35 , los cuales se ubican 10 y 35 pares de bases (pb), respectivamente, río arriba (*upstream*) del sitio de inicio de la transcripción; la caja -10 es reconocida por el segundo dominio del factor σ , mientras que la -35 es reconocida por el cuarto dominio del factor σ . Los otros dos elementos promotores relevantes son la caja -10 extendida y el elemento UP; la caja -10 extendida es una pequeña secuencia de 3-4 pb localizada río arriba de la caja -10 , reconocida por el tercer dominio de la subunidad σ ; mientras el elemento UP es una secuencia de ~ 20 pb ubicada río arriba del hexámero -35 , que es reconocida por los dominios carboxilo terminal de las subunidades α de la ARN polimerasa. Cada uno de los cuatro elementos otorgan especificidad a la unión de la ARN polimerasa al promotor y contribuyen a la formación del *complejo abierto*⁴, aunque sus contribuciones individuales varían de un promotor a otro y las deficiencias en un elemento pueden ser compensadas por otro; de hecho, ningún promotor posee los cuatro elementos perfectos, así cada elemento se describe con una *secuencia consenso* que representa la secuencia ideal que un cierto elemento debería poseer para tener la mayor fuerza de interacción con la ARN polimerasa.

2.2.2. ¿Cómo se inicia la transcripción?

Como ya se mencionó, el primer paso para que se inicie la transcripción de un gen es el reconocimiento del promotor por la ARN polimerasa, así como su adecuado posicionamiento con respecto al inicio de la transcripción; este paso lleva a la formación del *complejo cerrado*. A continuación, el este complejo sufre una isomerización en la cual las dos hebras del ADN se separan desde la posición -10 aproximadamente hasta la posición $+2$, formando así el *complejo abierto* y la *burbuja de transcripción*; ver figuras 2-3 y 2-4. En este punto se polimerizan las primeras bases del ARNm, permitiendo el inicio de la transcripción; sin embargo, hay una tendencia a que la ARN polimerasa se libere del ADN produciendo así un transcrito truncado,

³Un polímero de seis monómeros; en el caso del ADN nos referimos a una secuencia de seis bases nitrogenadas.

⁴Paso inicial de la transcripción donde la ARN polimerasa se encuentra unida al promotor y el ADN se ha desenrollado, formando así la burbuja de transcripción.

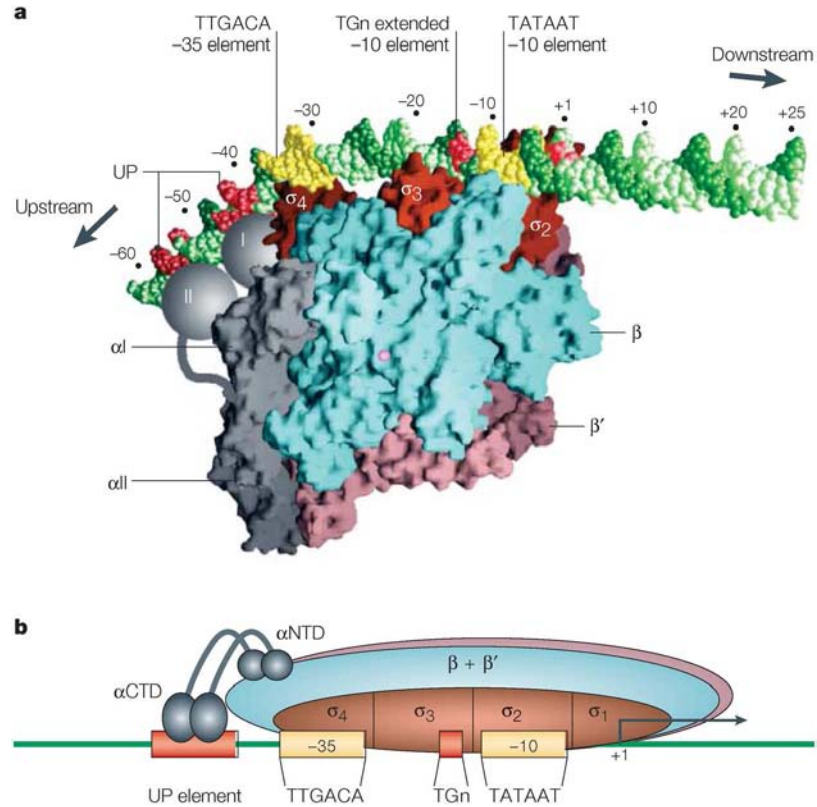


Figura 2-3: Interacciones entre la ARN polimerasa y su promotor. **(a)** Modelo basado en estudios cristalográficos del acoplamiento de la holoenzima de la ARN polimerasa a un promotor. El ADN se observa en verde, con las cajas -10 y -35 en amarillo, y los elementos -10 extendido y UP en rojo. Las subunidades β y β' se muestran, respectivamente, en azul claro y rosa; los dominios α NTD (amino-terminal de la subunidad α) aparecen en gris; mientras los diferentes dominios del factor σ se observan en tonos de rojo. Las esferas grises etiquetadas I y II, representan los dominios del α CTD (carboxilo-terminal de la subunidad α) que contactan al promotor. El sitio activo de la ARN polimerasa es denotado por el ion Mg^{+2} al centro en magenta. **(b)** Una representación del modelo mostrado en la parte **a**, ilustrando las distintas interacciones entre los elementos del promotor y la ARN polimerasa. Además, se muestran las secuencias consenso para las cajas -35 (TTGACA), -10 extendida (TG_n) y -10 (TATAAT). (Imagen tomada de Browning y Busby, 2004.)

lo cual es un fenómeno común y se conoce como *iniciación abortiva*. Una vez que el transcrito alcanza 23 bases aproximadamente, se estabiliza el complejo y se procede a la elongación liberándose la ARN polimerasa del factor σ . Durante la elongación, el ARNm es polimerizado empleando como molde una de las dos hebras del ADN. Eventualmente, la ARN polimerasa llega a un punto de terminación en donde se desestabiliza el complejo y el transcrito es liberado.

2.2.3. Elementos de control de la transcripción

Si bien el factor σ brinda un control selectivo sobre cuales promotores se transcriben, esto no es suficiente para controlar los miles de genes que una célula bacteriana posee. Así, ésta requiere de otros elementos que puedan brindar un control más fino sobre el inicio de la transcripción de los genes. Uno de estos elementos principales son las *proteínas regulatorias*, las cuales reconocen ciertas dianas en el ADN y son capaces de unirse a él alrededor de la zona del promotor para así activar o reprimir la transcripción. Además, como ya mencionamos previamente, estas proteínas vinculan la expresión genética con las señales ambientales a través del alosterismo.

Mientras que algunas proteínas regulatorias funcionan exclusivamente como *activadores* o como *represores*, otras pueden funcionar de una u otra forma dependiendo del promotor. En esencia, los activadores incrementan la afinidad de la ARN polimerasa por el promotor, mientras que los represores pueden ya sea bloquear la interacción entre la ARN polimerasa y el promotor o bien posicionarse delante del promotor impidiendo el avance de la ARN polimerasa. Algunos trabajos [Griffith *et al.*, 2002; Martin *et al.*, 2002; Griffith y Wolf, 2004] han mostrado la existencia de proteínas regulatorias activadoras capaces de interactuar con la ARN polimerasa antes de que ésta se una al promotor, siguiendo un mecanismo que nos recuerda el empleado por los factores σ .

2.3. Niveles de organización de las redes de regulación

Por más de 20 años se ha reconocido que las redes de regulación se componen de circuitos complejos con distintos niveles de control, lo que las capacita para controlar de forma simultánea diferentes subrutinas del programa genético [Gottesman, 1984; Neidhardt y Savageau, 1996]. En bacterias, los genes se encuentran dispuestos secuencialmente en un cromosoma circular,

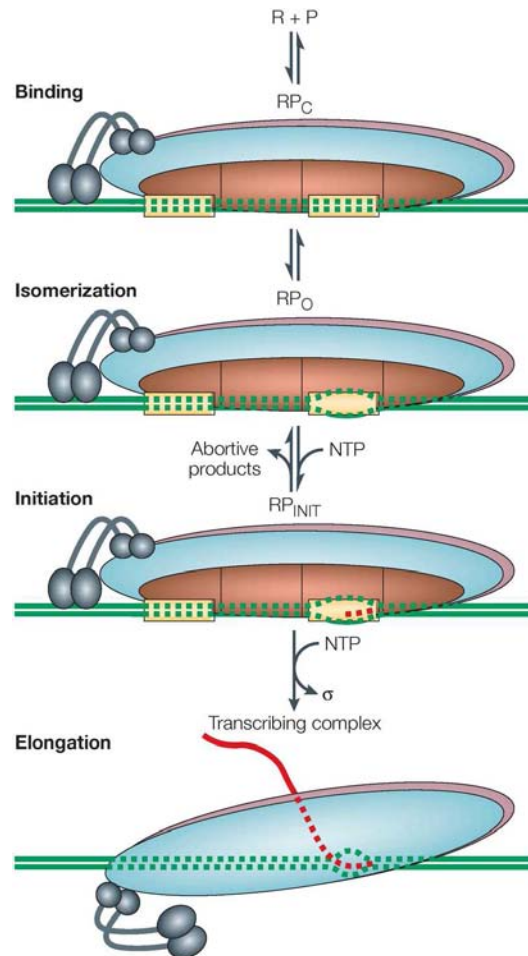


Figura 2-4: Secuencia de iniciación de la transcripción en los promotores bacterianos. **(Contacto)** La ARN polimerasa libre (R) interacciona con el promotor en el ADN (P) para formar el complejo cerrado (RP_C). Las líneas punteadas muestran el promotor que es contactado por la holoenzima de la ARN polimerasa. **(Isomerización)** La doble hélice del ADN es abierta (representado por una *burbuja* en el ADN) para formar el complejo abierto (RP_O). **(Iniciación)** Tras una serie de intentos abortivos, se forma el complejo de iniciación (RP_{INIT}) y comienza la síntesis de la hebra de ARN (mostrada como una línea punteada roja) con la formación de un enlace fosfodiéster entre el primer par de nucleósidos trifosfatos (NTP). **(Elongación)** La elongación es la etapa final en la que se incrementa la longitud de la hebra de ARN (mostrada como una línea roja). (Imagen tomada de Browning y Busby, 2004.)

el cual a su vez puede poseer una cierta estructura tridimensional debido a la introducción de superenrollamiento. Así, estos complejos circuitos de control, encargados del encendido y apagado de los genes, requieren de una bien definida organización genética que facilite a la célula producir, de forma precisa y eficiente, los componentes que requiera. A lo largo de años, se han propuesto algunos esquemas de organización los cuales han ayudado a comprender mejor la complejidad del control celular; aunque como veremos en el capítulo 4, ha sido hasta los albores del siglo XXI que, empleando la teoría de redes, se ha empezado a comprender cómo las redes de regulación se particionan en subcircuitos encargados de funciones fisiológicas bien definidas.

2.3.1. Operón

Hacia 1960, en un estudio sobre el sistema de degradación de lactosa de *E. coli*, Jacob *et al.* definieron el *operón* como un conjunto de genes contiguos en el cromosoma expresados coordinadamente desde un mismo promotor y regulados desde un mismo operador [Jacob *et al.*, 1960]; además, estableciendo que este conjunto de genes contiguos se transcriben en un solo *ARNm policistrónico*; ver figuras 2-5 y 2-6. Ésta fue la primer estructura que mostró la existencia de una organización funcional de los genes, dado que las enzimas codificadas en los genes del operón conforman los pasos iniciales del transporte y degradación del *efector alostérico* que induce la transcripción del mismo operón, formando así un pequeño subsistema de degradación; ver figura 2-5.

Sin embargo, a pesar de su utilidad, la organización en operones tiene sus limitaciones. Por una parte, en la célula existen procesos que pueden requerir de decenas de productos génicos cuya expresión debe ser coordinada de forma precisa, resultando ineficiente acomodar tal cantidad de genes en un solo operón. Por otra parte, muchos procesos celulares requieren de genes que puedan ser controlados tanto de forma independiente como coordinadamente, algo que ocurre comúnmente en procesos de degradación de fuentes de carbono donde los genes deben poderse inducir por la presencia de sus sustratos, pero deben reprimirse por la presencia de un sustrato más importante energéticamente, *e.g.*, la glucosa.

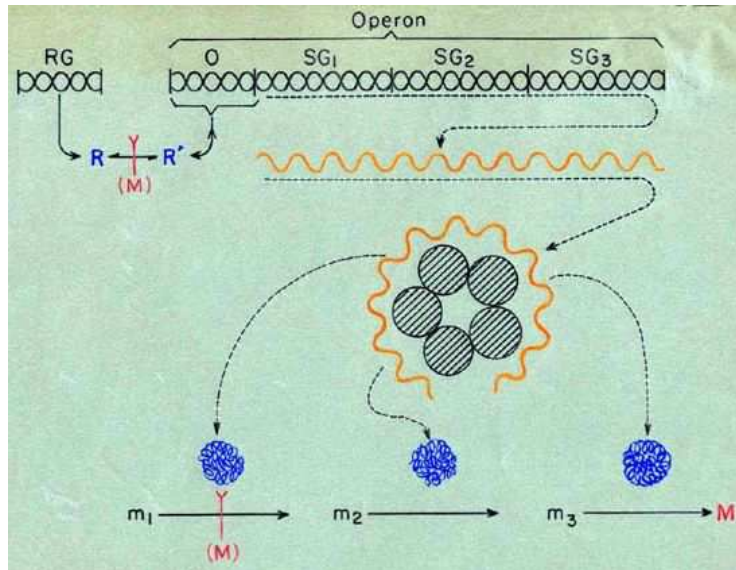


Figura 2-5: Dibujo original realizado por Jacques Monod, en 1960, ilustrando el operón. (Archivos del Instituto Pasteur, MON.Bio.20, expediente 24.11.)

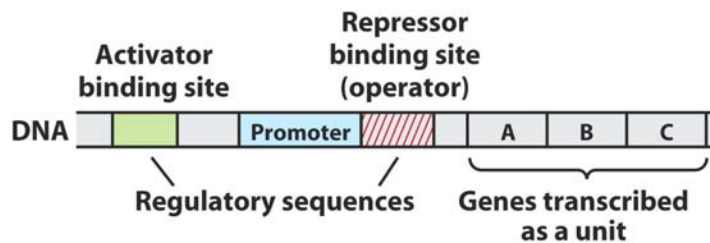


Figura 2-6: Modelo de organización del operón. Los genes A, B y C se transcriben en un solo ARNm conocido como policistrónico. Las secuencias reguladoras típicas incluyen sitios de unión para proteínas que activan o reprimen la transcripción a partir del promotor. (Imagen tomada de Nelson y Cox, 2000.)

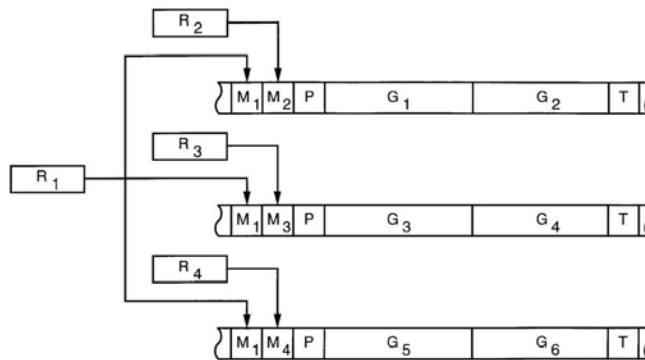


Figura 2-7: Modelo de organización del regulón. Cada operón es mostrado como un segmento de ADN compuesto de un promotor (P) y un terminador (T), así como de sitios moduladores (M_j) sobre los que ciertas proteínas reguladoras (R_i) actúan para alterar la transcripción de los genes (G_k). La proteína reguladora R_1 es la base de este regulón. (Imagen tomada de Neidhardt y Savageau, 1996.)

2.3.2. Regulón

Las limitaciones del operón abrieron las puertas a la búsqueda de otros niveles de control; dado que los genes en un operón son contiguos debía existir una forma de control que permitiera coordinar los diversos operones diseminados a lo largo del cromosoma. En 1964, estudiando la proteína reguladora ArgR de *E. coli*, Maas definió el *regulón* como el conjunto de genes u operones que son regulados por una proteína reguladora específica [Maas, 1964]; estableciendo por primera vez un nivel de control jerárquico capaz de coordinar la respuesta de múltiples genes u operones diseminados en el cromosoma; ver figura 2-7. Hacia 2003, Gutiérrez-Ríos *et al.* extendieron la definición de regulón; en su propuesta, la definición clásica pasó a denominarse *regulón simple*, mientras que definieron como *regulón complejo* al conjunto de genes regulados por exactamente el mismo conjunto de proteínas reguladoras [Gutiérrez-Ríos *et al.*, 2003].

La inducción de los operones individuales que componen a un regulón no es coordinada de forma estricta, permitiendo así variaciones de un operón a otro, tanto en la cantidad de producto como en el tiempo de inducción; estableciendo de esta forma un control jerárquico de la expresión de los genes u operones bajo el control del regulón.

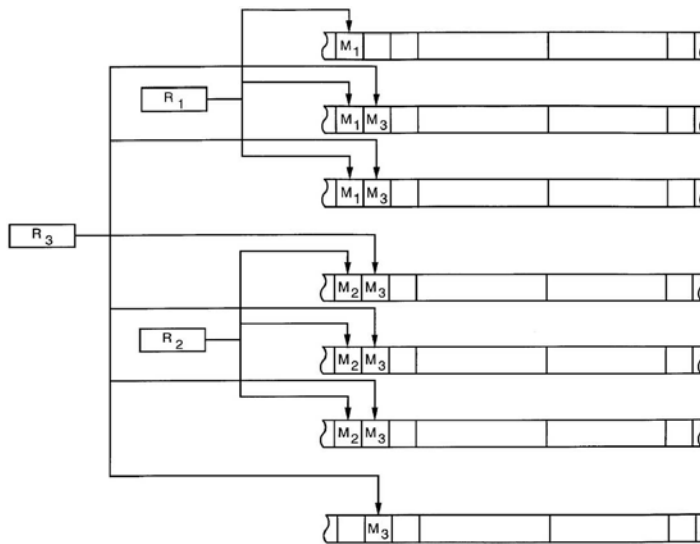


Figura 2-8: Modelo de organización del modulón. En la figura se observa a la proteína regulatoria R_3 controlando a dos regulones (controlados a su vez por los reguladores R_1 y R_2 , respectivamente) y a un operón independiente. (Imagen tomada de Neidhardt y Savageau, 1996.)

2.3.3. Modulón

Por definición, los operones bajo el control de un regulón están relacionados funcionalmente; esto deja abierto el problema de cómo coordinar operones que pueden tener funciones diversas pero los cuales, bajo ciertas circunstancias, *e.g.*, la presencia de un estrés, es necesario que respondan a un objetivo común. Durante 1988, en un estudio sobre la proteína regulatoria ArcA de *E. coli*, Iuchi y Lin propusieron el *modulón* como un conjunto de operones o regulones que son modulados por una *proteína regulatoria pleiotrópica* común, la cual puede ser un factor σ [Iuchi y Lin, 1988]; ver figura 2-8. La *pleiotropía* ocurre cuando un gen influenciaba múltiples características fenotípicas⁵ de un organismo, causando así alteraciones globales en el mismo. Algunas proteínas regulatorias pleiotrópicas o globales son capaces de responder a los niveles de ciertos metabolitos intracelulares, los cuales reflejan la condición de cierta sección del metabolismo o una determinada condición ambiental. Así, los modulones se posicionan como la jerarquía de control global de la célula.

Como se deriva de la definición, uno o más operones pueden ser miembros de un modulón;

⁵El fenotipo es una característica observable de un organismo.

esto hace que esta terminología sea confusa para algunos por dos razones:

1. No es correcto pensar en un modulón como un regulón de regulones.
2. Si el modulón está compuesto sólo por operones, ¿qué lo distingue del regulón?

La clave para romper esta aparente ambigüedad radica en las definiciones originales, ya que mientras los operones de un regulón se encuentran funcionalmente relacionados, los operones o regulones de un modulón no lo están, por lo que el factor transcripcional que los controla posee efectos pleiotrópicos.

2.4. Factores transcripcionales globales

A lo largo de su vida, una bacteria debe hacer frente a una serie de circunstancias o condiciones que, para ser resueltas satisfactoriamente, requieren de cambios dramáticos y eficientes en la expresión de sus genes. Estos cambios deben realizarse en intervalos pequeños de tiempo, de tal suerte que la existencia de una arquitectura de control bien definida es clave para lograr una respuesta eficiente. Como hemos visto en secciones anteriores, se ha reconocido que las redes de regulación integran circuitos regulatorios simples en otros más complejos, formando así una serie de niveles jerárquicos.

2.4.1. Proteínas regulatorias globales

Los modulones ponen en evidencia la importancia que tienen ciertas proteínas regulatorias pleiotrópicas al controlar a una gran cantidad de operones o regulones, permitiendo así reconfigurar a escala global el patrón de expresión de los genes para responder de forma eficiente a un objetivo común. Estas proteínas regulatorias globales y sus circuitos de control global han sido estudiados desde hace más de 20 años. En 1984, Susan Gottesman propuso por primera vez un conjunto de propiedades que un factor transcripcional debe poseer para ser considerado global [Gottesman, 1984]:

- El factor transcripcional regula muchos genes.
- Los genes regulados participan en más de una vía metabólica.

- El factor transcripcional coordina la expresión de un conjunto de genes en respuesta a una necesidad común.

Las propiedades propuestas por Gottesman predijeron, cuatro años antes del trabajo de Iuchi y Lin, aquellas que posee un factor transcripcional pleiotrópico encargado de controlar un modulón. Estudios posteriores han tratado de identificar estos factores de transcripción globales, empleando para ello criterios arbitrarios basados principalmente en el número de genes regulados [Shen-Orr *et al.*, 2002; Babu y Teichmann, 2003; Ma *et al.*, 2004a,b]; lamentablemente, a la fecha no existe un consenso al respecto. En 2003, Martínez-Antonio y Collado-Vides realizaron una revisión de la literatura y analizaron diversas propiedades en búsqueda de criterios diagnósticos para identificar factores de transcripción globales [Martínez-Antonio y Collado-Vides, 2003]; sin embargo, si bien los autores arrojaron luz sobre propiedades pertinentes a los factores de transcripción globales, su trabajo no logró establecer criterios diagnósticos explícitos que puedan, como los mismos autores proponen, ser implementados computacionalmente.

2.4.2. Factores σ

Tradicionalmente, el término *factor de transcripción* se ha empleado para referirse a las proteínas regulatorias, las cuales innegablemente lo son. Sin embargo, la palabra *factor* conlleva un sentido más amplio e implica involucrar no sólo a las proteínas regulatorias, sino a todos aquellos elementos que participan de alguna forma en la decisión de si un gen debe ser transcrito o no, independientemente del mecanismo molecular empleado para ello; precisamente en ese sentido se emplea la frase *factor de transcripción* en el estudio de la regulación transcripcional en eucariotes. Además de las proteínas regulatorias, los factores σ juegan un papel preponderante en la toma de las decisiones acerca de qué conjunto de genes se desea transcribir; al controlar la afinidad de la ARN polimerasa por el promotor, la competencia entre distintos factores σ puede alterar de forma global el patrón de expresión de los genes, mediante redirigir la maquinaria de transcripción hacia nuevas dianas. Por estos motivos, a lo largo de esta investigación también consideraremos a los factores σ como factores de transcripción.

Capítulo 3

Teoría de Redes

La capacidad de simplificar significa eliminar lo innecesario de modo que lo necesario pueda hablar.

— HANS HOFMANN

Personas de distintas edades y estratos sociales han escuchado, al menos una vez, la palabra *red*; su uso se ha arraigado como consecuencia, principalmente, de la popularización, durante la última década del siglo XX, de la llamada red mundial o Internet. Sin embargo, el uso de la palabra *red* data de tiempo atrás; por ejemplo, en la década de los 70, ya se hablaba de la red telefónica o de la red del metro, y fue durante la segunda mitad del siglo XX que en el ámbito científico y tecnológico se popularizó el concepto *red*; empero sus orígenes se remontan a muchos siglos atrás. A lo largo de este capítulo iremos desde sus orígenes conceptuales hasta los más recientes avances en la moderna teoría de redes.

3.1. El Uroboros

La referencia más antigua al concepto *red* la tenemos en el Uroboros¹, uno de los símbolos místicos más antiguos de la humanidad, el cual ha aparecido en las más diversas culturas: egipcia y griega; las mitologías nórdica, azteca, maya y del Medio Oriente; el cristianismo, hinduismo y religiones africanas. Este símbolo muestra una serpiente devorando su propia cola; ver figura 3-1. La primer aparición del Uroboros se registra en la cultura Hongshan (4700 a. de C. al 2900 a. de C.) que floreció en lo que es hoy el noreste de China. Un dibujo del Uroboros,

¹Cuya etimología proviene del griego ουροβόρος que significa “devorador de la cola”.



Figura 3-2: El Uroboros como aparece en la *Crisopea de Cleopatra*; al interior se puede leer en griego el lema “έν τὸ πᾶν” (“Todo es Uno”).

a la humanidad como un ancestro digno de respeto y veneración; aquellos quienes, a lo largo de sus vidas, se mantienen firmes al principio del Ubuntu alcanzarán, al morir, una unidad con aquellos aún vivos.

3.2. El problema de los siete puentes de Königsberg

Aunque desde las culturas primitivas las ideas fundamentales ya habían florecido, el nacimiento de los conceptos matemáticos básicos de la teoría de redes nos lleva varios siglos adelante en la historia hasta el siglo XVIII, y nos ubica en la pintoresca ciudad universitaria de Königsberg², conocida desde 1946 como la provincia rusa de Kaliningrado. Esta ciudad Prusa, fundada en 1255 por los Caballeros Teutones, se transformó en 1554 en un centro cultural germano cuando Albert, duque de Prusia, fundó la universidad de Königsberg, también conocida como universidad Albertina; lo que dió a la ciudad un aire intelectual y cosmopolita. Königsberg fue la residencia de, entre otros, Immanuel Kant, Gustav Kirchhoff, David Hilbert y Christian Goldbach. Un grabado del siglo XVII muestra a Königsberg como una próspera ciudad, donde

²Königsberg significa literalmente “montaña del rey” en prusiano antiguo. La ciudad se encontraba en el reino de Prusia, *Borussia* en latín; de ahí que Euler, en su artículo escrito en latín en 1736, se refiera a Königsberg como *Regiomonti in Borussia*.

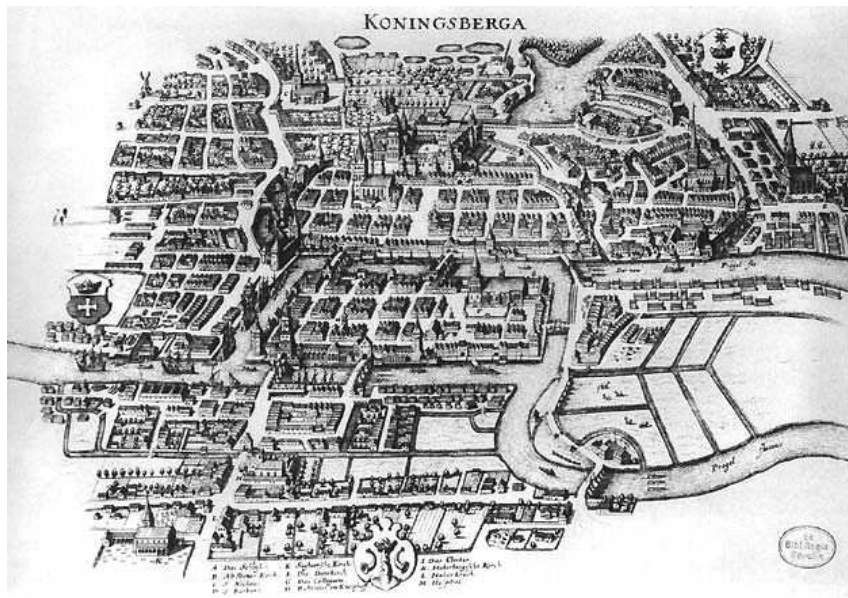


Figura 3-3: Grabado de 1652 ilustrando la ciudad de Königsberg y los siete puentes que entonces existían, algunos de los cuales fueron destruidos por bombardeos durante la segunda guerra mundial; al centro se encuentra el Kneiphof rodeado por el río Pregel.

flotas de embarcaciones cruzan el río Pregel. La sana economía de la ciudad permitió la construcción de siete puentes sobre el río, la mayoría de los cuales conectaban el Kneiphof³; ver figura 3-3.

Se cuenta que los domingos, pasado el medio día, los ciudadanos de Königsberg gustaban de caminar alrededor de su hermosa ciudad. Quizá el ambiente intelectual reinante en la ciudad hizo que uno de sus ciudadanos se preguntara, ¿cómo se puede dar un paseo, de manera que, se cruce cada uno de nuestros siete puentes sólo una vez?; ver figura 3-4. La pregunta se popularizó, y muchos ciudadanos, en aquellos paseos dominicales, intentaron resolver el enigma, siempre sin éxito; pero a pesar de haber fallado nadie había logrado demostrar que la empresa fuera imposible.

En 1679, Gottfried Leibniz tenía ya en mente la idea de lo que posteriormente llamaría *análisis situs*⁴, refiriéndose al análisis de la situación o posición; un concepto que fue ampliamente

³Kneiphof originalmente era *Knypabe* (*Kneip-ape*), que significa “área rodeada por agua (río o arroyo)” en prusiano antiguo.

⁴Originalmente denominado *geometriam situs* en latín.

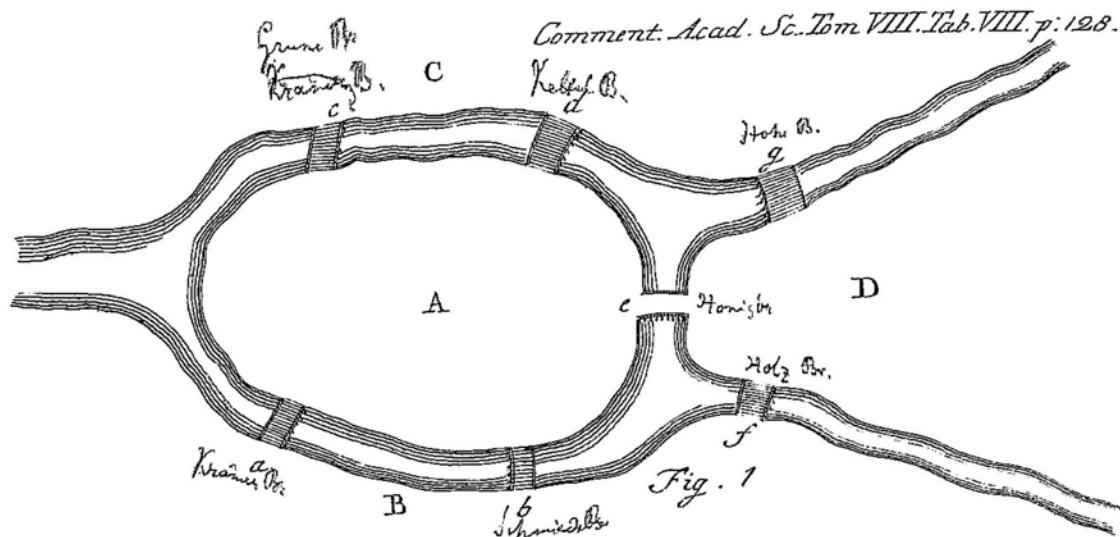


Figura 3-4: Dibujo de los puentes de Königsberg realizado por Euler y publicado en su artículo *Solutio problematis ad geometriam situs pertinentis* [Euler, 1741]. En él se observan cuatro áreas de tierra firme (A, B, C, D) conectadas por siete puentes (a, b, c, d, e, f, g); la zona A es el Kneiphof. (Imagen tomada de Euler, 1741.)

interpretado por sus seguidores como referente a lo que ahora denominamos *topológico*, *i.e.*, de naturaleza geométrica pero sin involucrar ideas métricas como distancia, longitud o ángulo [Hopkins y Wilson, 2004]. Para 1727, Leonhard Euler llega a Rusia y empieza a trabajar para la Academia de Ciencias de San Petersburgo; algún tiempo después, llegaría a sus oídos el problema de los siete puentes de Königsberg. El 26 de agosto de 1735, Euler presenta a sus colegas la solución a “un problema relacionado a la geometría de la posición”, que no era otro que el problema de los puentes de Königsberg, el cual había resuelto sin visitar la ciudad. A pesar de que, inicialmente, el problema le parece trivial, lo intriga; en una carta fechada 13 de marzo de 1736 y dirigida a Giovanni Marinoni⁵, escribe [Hopkins y Wilson, 2004]:

[...] Este problema es muy banal, pero me pareció digno de atención que ni la geometría, ni el álgebra, ni siquiera el arte de contar fue suficiente para resolverlo. En vista de esto, se me ocurrió preguntarme si éste pertenecía a la geometría de la posición (*geometriam situs*), que Leibniz una vez tanto había deseado. [...]

⁵Matemático e ingeniero italiano quien vivió en Viena y fue el astrónomo de la corte del kaiser Leopoldo I.

Euler no estaba equivocado, su trabajo realizado sobre los puentes de Königsberg sentó las bases de lo que en matemáticas se conocería posteriormente como teoría de grafos. En ese mismo año, 1736, Euler escribe su artículo *Solutio problematis ad geometriam situs pertinentis* [Euler, 1741], describiendo el problema de los puentes de Königsberg y demostrando que no existe una solución al mismo; empero, Euler va más allá, generaliza el problema y establece un conjunto de reglas que permiten determinar si existe o no un paseo, que cruce por cada puente sólo una vez, dado un conjunto arbitrario de puentes y áreas de tierra firme. Las reglas establecidas por Euler son las siguientes [Euler, 1741]:

- Si existen más de dos áreas de tierra firme a las cuales llegan un número impar de puentes, entonces el paseo es imposible.
- Sin embargo, si el número de puentes es impar para dos áreas de tierra exactamente, entonces el paseo es posible si se inicia desde alguna de estas dos áreas.
- Finalmente, si no existen áreas a las cuales lleguen un número impar de puentes, entonces el paseo puede ser logrado empezando desde cualquier área.

3.3. Conceptos básicos de teoría de grafos

Hagamos un experimento mental, tomemos el mapa de Königsberg hecho por Euler y hagamos lo que los matemáticos hacen, abstraer de la realidad los elementos relevantes al problema; al hacerlo veremos que no es relevante la forma de los puentes, ni la forma de las áreas de tierra, etc.; lo relevante son tres cosas: ¿cuántas áreas de tierra firme hay?, ¿cuántos puentes hay?, y ¿qué áreas de tierra une cada puente? De esta forma hemos eliminado muchos distractores que no contribuyen a la solución del problema, y nos hemos quedado sólo con lo necesario para resolverlo. Ahora, tomemos el mapa hecho por Euler y representemos cada área de tierra como un punto y cada puente como una línea uniendo dos puntos; ver figura 3-5.

En matemáticas, el dibujo que hemos obtenido se denomina *grafo* o *grafo no dirigido*, y es una herramienta muy útil para estudiar problemas que involucran un conjunto de elementos y cómo estos se interrelacionan. Formalmente, un grafo G se define como un par ordenado $G = (\mathcal{V}, \mathcal{E})$, donde \mathcal{V} es un conjunto de elementos llamados *vértices* o *nodos*, y \mathcal{E} es un conjunto

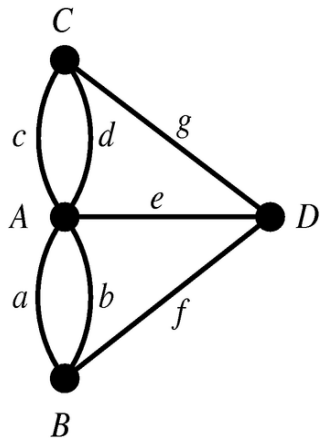


Figura 3-5: Dibujo obtenido tras abstraer los elementos relevantes del mapa de los siete puentes de Königsberg.

de pares no ordenados de vértices llamados *aristas*. Si pensáramos en los puentes como calles de una sola dirección, entonces tendríamos que recurrir a un *grafo dirigido* o *digrafo*, el cual se define como un par ordenado $G = (\mathcal{V}, \mathcal{A})$, donde \mathcal{V} es un conjunto de elementos llamados vértices o nodos, y \mathcal{A} es un conjunto de pares ordenados de vértices llamados *aristas dirigidas* o *arcos*.

El área de las matemáticas y las ciencias computacionales encargada del estudio de los grafos se conoce como *teoría de grafos*, la cual coincide con un amplio tópico de estudio de las matemáticas aplicadas y la física denominado *teoría de redes*, con aplicaciones en un diverso rango de disciplinas incluyendo ciencias computacionales, biología, economía y sociología. Así, actualmente, los grafos también pueden recibir, indistintamente, el nombre de redes.

En biología, tanto los grafos no dirigidos, como los digrafos, son útiles. Las redes de regulación transcripcional o las redes metabólicas suelen ser representadas por digrafos ya que en ellas la dirección es importante, *e.g.*, en una red metabólica la dirección nos indica la ruta que sigue una reacción química. Por otra parte, las redes de interacciones proteína-proteína describen los contactos físicos directos entre las proteínas de un organismo; en consecuencia en ellas la dirección no es importante y se describen empleando grafos no dirigidos, en donde los nodos representan proteínas y las aristas sus interacciones. A continuación algunos de los conceptos básicos de la teoría de grafos:

Adyacencia. Se dice que dos nodos i y j son *adyacentes* o *vecinos* si existe una arista o un arco que los una.

Matriz de adyacencia. Mientras que una representación gráfica puede ser útil en términos visuales, el análisis de redes requiere de una representación más formal conocida como *matriz de adyacencia*. Sea G un grafo o digrafo con $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$ como el conjunto de todos sus vértices y \mathcal{E} el conjunto de sus aristas o arcos, entonces la matriz de adyacencia es una matriz cuadrada \mathbf{A} de $n \times n$ donde cada elemento esta dado por

$$a_{i,j} = \begin{cases} 1 & \text{si } v_i v_j \in \mathcal{E}, \\ 0 & \text{si } v_i v_j \notin \mathcal{E}. \end{cases} \quad (3-1)$$

Grado de un nodo. En grafos no dirigidos, el *grado*, *valencia* o *conectividad* (k) de un nodo es el número de aristas que convergen en él. En grafos dirigidos, existen dos grados posibles, el grado de entrada (k_e) y el grado de salida (k_s); siendo cada uno, respectivamente, el número de arcos que llegan al nodo y el número de arcos que salen del nodo. Es importante hacer notar que si transformamos un digrafo en un grafo no dirigido, mediante eliminar la dirección de los arcos, entonces no necesariamente se cumple que $k = k_e + k_s$. En particular, por cada nodo que participa en un ciclo de dos elementos⁶ ocurre que $k \neq k_e + k_s$; esto debido al hecho de que al eliminar la dirección de los dos arcos queda sólo una arista resultante.

Conectividad máxima. La *conectividad máxima* ($k_{\text{máx}}$) de una red es el valor máximo de conectividad presente en dicha red.

Camino. Una arista conecta de forma directa dos nodos; de forma similar, una sucesión de aristas, con vértices en común, nos permite trazar un *camino* que conecta de forma indirecta dos nodos.

Longitud del camino. El número de aristas necesarias para llegar desde un nodo a otro en un camino, es denominado *longitud del camino*.

⁶Un ciclo de dos elementos es una estructura en la cual dos elementos a y b se interrelacionan, *i.e.*, $a \rightleftarrows b$.

Camino mínimo. Es el camino más corto entre un cierto par de nodos. En ocasiones este camino puede ser igual a sólo una arista, pero no necesariamente, *e.g.*, en la figura 3-5 el *camino mínimo* que lleva de *B* a *C* es de 2, ya sea que el camino cruce por *A* o por *D*.

Camino mínimo promedio. También llamado *camino característico*, es el camino mínimo promediado para todos los pares de vértices.

Diámetro. Si calculamos los caminos mínimos entre todos los pares de nodos de un grafo, el *diámetro* del grafo será igual al camino mínimo de mayor longitud.

Grafo conexo. Es aquél en el cual existe al menos un camino entre cualquier par de nodos del grafo.

Grafo inconexo. Dícese de aquél en el cual no existe un camino entre al menos un par de nodos del grafo.

Grafos isomorfos. Son aquellos grafos que contienen el mismo número de nodos conectados de la misma forma, tal que es posible encontrar una reorganización espacial de los nodos y un mapeo de sus etiquetas que haga a los grafos visualmente idénticos.

Componente gigante. Dado un grafo inconexo, el componente gigante será aquel subgrafo conexo que contiene la mayoría de los nodos del grafo total, *i.e.*, el subgrafo conexo más grande en términos del número de nodos que contiene.

Distribución de conectividad. Una de las primeras herramientas para analizar la estructura topológica global de una red es la *distribución de conectividad*, $P(k)$, la cual mide la proporción de nodos con grado k en una red. Formalmente,

$$P(k) = \frac{n_k}{n}, \quad (3-2)$$

donde n_k es el número de nodos con conectividad k y n es el número total de nodos en la red.

3.4. ¿Unos cuantos principios gobiernan el todo?

Con estos conceptos básicos sigamos nuestro recorrido. Después de la publicación del artículo de Euler en 1741, gradualmente los matemáticos empezaron a interesarse en esta nueva área de estudio, se hicieron descubrimientos importantes, pero fue hasta 1959 que se dio el siguiente hito en la teoría de redes. En ese año Paul Erdős y Alfréd Rényi publicaron su artículo titulado *On random graphs*, en el cual definieron por primera vez las redes aleatorias [Erdős y Rényi, 1959].

En esencia una red aleatoria es aquella en la cual cada nodo tiene una cierta probabilidad p de conectarse con algún otro nodo. En el modelo Erdős-Rényi [Erdős y Rényi, 1961], p es un parámetro dado de antemano que establece la probabilidad de que exista una arista entre un cierto par de nodos, independientemente del resto de las aristas; dando así origen a grafos en donde muchos nodos tienen un valor de conectividad muy cercano o igual al valor promedio de conectividad, $\langle k \rangle$, siguiendo una distribución de Poisson; ver figura 3-6Ab; donde nodos con conectividades muy inferiores o superiores a $\langle k \rangle$ son muy raros o no existen [Erdős y Rényi, 1959, 1961]. Así, se puede decir que el valor de conectividad $\langle k \rangle$ es típico de la red, la mayoría de los nodos se aglomeran alrededor de este valor y la distribución de conectividad decae exponencialmente conforme se incrementa el valor de $|k - \langle k \rangle|$.

El aspecto más importante de las redes aleatorias es brindar tanto un método probabilístico para probar la existencia de grafos que satisfacen distintas propiedades, como una definición rigurosa de lo qué significa que una propiedad se cumpla para “casi todos” los grafos. Finalmente, el descubrimiento de que unas cuantas reglas permiten generar redes con arquitecturas específicas, incentivó el estudio de la teoría de redes.

3.5. ¡El mundo es un pañuelo!

En 1909, Guglielmo Marconi recibió el Premio Nobel de Física en reconocimiento a sus contribuciones para el desarrollo de la telegrafía inalámbrica. Durante su conferencia Nobel afirmó que era posible, mediante el emplazamiento de seis estaciones permanentes, transmitir mensajes sobre el océano Atlántico [Marconi, 1967], acortando así los lazos de comunicación entre América y Europa. Se ha especulado que esta afirmación incentivó la imaginación del

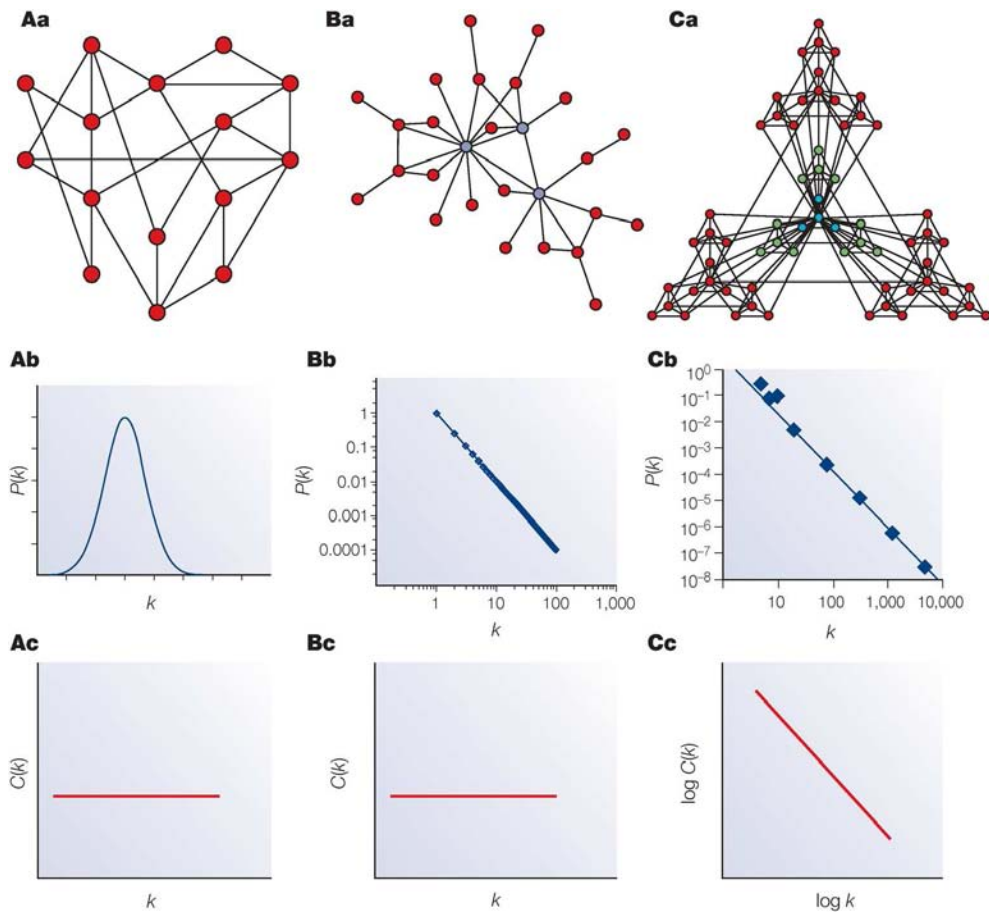


Figura 3-6: Tres modelos fundamentales de redes y sus propiedades distintivas, los cuales han contribuido a comprender fenómenos en muy diversas áreas del conocimiento. **(A)** Redes aleatorias. **(B)** Redes libres de escala (*scale-free*). **(C)** Redes jerárquico-modulares. Los gráficos asociados corresponden con **(b)** las distribuciones de conectividad, $P(k)$, y **(c)** del coeficiente de *clustering*, $C(k)$; ver el texto principal para las definiciones de los conceptos. (Imagen tomada de Barabási y Oltvai, 2004.)

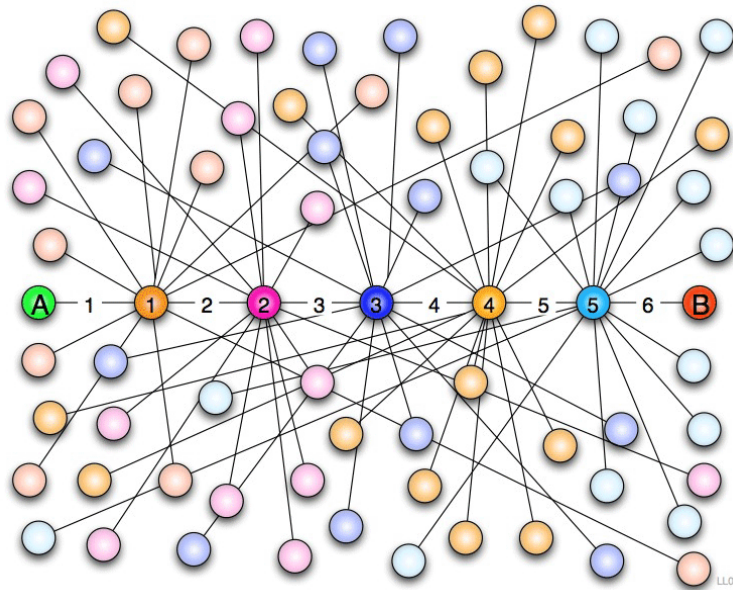


Figura 3-7: Representación artística del concepto de los seis grados de separación. Cada nodo representa a una persona; A y B son dos personas cualesquiera vinculadas por un camino de cinco conocidos.

escritor húngaro Frigyes Karinthy [Barabási, 2002], quien, en 1929, en su libro *Todo es diferente* publicó un cuento corto titulado *Cadenas (Láncszemek)*. En *Cadenas*, el autor investiga en términos abstractos, conceptuales y ficticios muchos de los problemas que podrían cautivar a las futuras generaciones de matemáticos, sociólogos y físicos dentro del campo de la teoría de redes. El autor propone que dados los avances tecnológicos en el transporte y las telecomunicaciones, las redes de amistad serán más grandes y abarcarán mayores distancias. Karinthy creía que el mundo moderno se estaba “encogiendo” debido a la cada vez mayor interconexión entre los seres humanos. En consecuencia, afirmó que, a pesar de las grandes distancias geográficas entre los individuos en el planeta, la densidad de las redes humanas harían la distancia social mucho menor. Los personajes de *Cadenas* creen que cualesquiera dos individuos podrían vincularse por a lo más cinco conocidos. Su análisis de estos conceptos hizo que se considere a Karinthy el padre de la noción de los *seis grados de separación*; ver figura 3-7.

En 1967, Stanley Milgram realizó un conjunto de experimentos para investigar el *problema del mundo pequeño* y la noción de los seis grados de separación, publicándolos en la revis-

ta de ciencia popular *Psychology Today* [Milgram, 1967]; una versión más rigurosa de estos experimentos fue publicada en *Sociometry* dos años después [Travers y Milgram, 1969]. El experimento de Milgram partió del deseo de saber más sobre cuál es la probabilidad de que dos personas elegidas al azar se conozcan mutuamente. Una forma alterna de ver este problema es imaginar a la población como una red social e intentar encontrar la longitud del camino promedio entre cualquier par de nodos; el experimento de Milgram fue diseñado para medir las longitudes de estos caminos. El procedimiento básico fue el siguiente [Milgram, 1967; Travers y Milgram, 1969]:

1. Milgram eligió las ciudades de Omaha, Nebraska y Wichita, Kansas para ser los puntos de inicio y a Boston, Massachusetts para ser el punto final de una cadena de correspondencia. Estas ciudades fueron seleccionadas debido a que representaban una gran distancia tanto social como geográficamente.
2. Se enviaron sobres con información a un conjunto de individuos elegidos aleatoriamente en Omaha y Wichita. Estos sobres incluían una carta de invitación la cual detallaba el objetivo del estudio, una lista de nombres, tarjetas postales con la dirección de los investigadores en Harvard, e información básica sobre el destinatario final en Boston.
3. En la invitación se solicitaba que, si el participante conocía personalmente al destinatario final, reenviara el sobre directamente a esa persona.
4. En el caso más probable de que el participante no conociera personalmente al destinatario final, entonces debía pensar en un amigo o pariente, al que conociera personalmente, quien fuera más probable que conociera al destinatario final. El participante debía registrar su nombre en la lista y reenviar el sobre a la persona elegida; además, debía enviar una tarjeta postal a los investigadores en Harvard tal que ellos pudieran seguir el avance hacia el destinatario final.
5. Cuando el sobre eventualmente llegaba a su destino final en Boston, los investigadores examinaban la lista de nombres para contar el número de veces que había sido reenviado de una persona a otra. En el caso de los paquetes que nunca llegaron a su destino, las tarjetas postales ayudaron a identificar el punto donde la cadena se había roto.

Poco después de que el experimento inició las cartas empezaron a llegar a su destino final; a veces en sólo dos reenvíos y en otros casos a través de cadenas de nueve a diez reenvíos. Tras analizar el total de los sobres que llegaron a su destino, encontraron que el número promedio de reenvíos fue de alrededor de 5.5 [Milgram, 1967; Travers y Milgram, 1969]; por lo que los investigadores concluyeron que las personas en los Estados Unidos se encontraban separadas por seis personas en promedio. A pesar de que Milgram nunca usó la frase “seis grados de separación” sus descubrimientos posiblemente contribuyeron a su amplia aceptación [Barabási, 2002].

En 1998, Duncan J. Watts y Steven H. Strogatz, ambos del Departamento de Mecánica Teórica y Aplicada de la Universidad de Cornell, publicaron el primer modelo de redes basado en el fenómeno de mundo pequeño [Watts y Strogatz, 1998], mostrando que redes sociales, naturales y hechas por el hombre exhiben propiedades particulares de este fenómeno. Las redes estudiadas fueron: la red de colaboración de actores de cine con datos obtenidos de la *Internet Movie Database*; la red neuronal del nemátodo *Caenorhabditis elegans*; y la red de energía eléctrica del occidente de los Estados Unidos. Además, Watts y Strogatz mostraron que, iniciando con una red regular, la reconexión aleatoria de una cierta fracción de sus enlaces reduce el camino característico y el *coeficiente de clustering* (*coeficiente de agrupamiento*) en función de la probabilidad p de que un enlace sea reconectado. Mediante este proceso de reconexión de la red identificaron que existe un punto intermedio, entre una red regular y una red aleatoria, en el que surge una *red de mundo pequeño* (*small-world*), la cual posee un camino característico pequeño y un coeficiente de *clustering* alto. De hecho, ellos fueron los primeros en definir el coeficiente de *clustering*, que es una medida de exclusividad en un grupo de amistades. Sea i el nodo que se desea analizar, k_i el número de vecinos y n_i el número de conexiones entre sus vecinos, entonces el coeficiente de *clustering* C_i de dicho nodo está dado por

$$C_i = \frac{2n_i}{k_i(k_i - 1)}. \quad (3-3)$$

Si cada uno de los k_i vecinos de i puede conectarse con los demás $k_i - 1$ vecinos, entonces el número total posible de conexiones que puede ocurrir ésta dado por $\frac{k_i(k_i-1)}{2}$. Así la ecuación 3-3 mide la razón entre el número real de conexiones entre los vecinos de i (n_i) con respecto al

número total posible.

Hacia 2003, se publicó un experimento que recreó el experimento de Milgram, empleando en esta ocasión correo electrónico [Dodds *et al.*, 2003]; en el cual más de 60,000 usuarios intentaron alcanzar uno de los 18 destinatarios finales en 13 países mediante reenviar mensajes a sus conocidos. Los resultados arrojaron que, dependiendo de la separación geográfica entre el origen y el destino, la longitud promedio de la cadena fue de entre cinco y siete reenvíos. En 2008, un estudio de *Microsoft* reconstruyó la red formada por 30 billones de conversaciones entre 240 millones de personas de todo el mundo, capturadas durante un mes de uso del sistema de mensajería instantánea *Microsoft Messenger*. La red analizada mostró que la longitud del camino promedio entre los usuarios del *Microsoft Messenger* es de 6.6 [Leskovec y Horvitz, 2008].

Todos estos estudios han reforzado la hipótesis de Karinthy⁷ sobre la noción de los seis grados de separación, mostrando además que el fenómeno de mundo pequeño va más allá de las redes sociales, invadiendo las redes naturales y las hechas por el hombre, revolucionando nuestra perspectiva del mundo. De cierta forma el camino característico de una red nos indica con cuanta facilidad se puede transmitir información o materia a través de ella. La aparición de este fenómeno en redes biológicas sugiere que éstas son eficientes en términos de la transferencia de información y materia.

3.6. La importancia de la popularidad

Durante 1999, Albert-László Barabási y sus colegas de la Universidad de Notre Dame mapearon, empleando un robot *Web* (o *Web crawler*), la red de Internet a nivel de páginas *Web*. Al analizar la topología global de la red encontraron que, para su sorpresa y contrario a lo que esperaban, ésta no seguía una distribución de Poisson o gaussiana como lo predecía el modelo Erdős-Rényi. En lugar de ésta, la distribución de conectividad encontrada mostraba una larga cola indicando la existencia de unos cuantos nodos altamente conectados, los cuales fueron denominados *hubs*⁸; mientras la mayoría de los nodos mostraban bajas conectividades. Se reportó

⁷ Así como la de aquellos de nosotros quienes, al menos una vez en la vida, hemos afirmado: ¡el mundo es un pañuelo!

⁸ El término *hub* fue tomado de las redes de telecomunicaciones, en donde éste se refiere a un enrutador, conmutador o concentrador que se encarga de redirigir los paquetes de información hacia su destino.

[Albert *et al.*, 1999; Barabási y Albert, 1999] que esta distribución de conectividad se ajustaba a una ley de potencia de la forma

$$P(k) \sim k^{-\gamma}, 2 < \gamma < 3, \quad (3-4)$$

la cual en un espacio log-log se observa como una línea recta con pendiente $-\gamma$; ver figura 3-6Bb. A estas redes se les denominó *libres de escala* (*scale-free*) debido a que, a diferencia de las redes aleatorias, carecen de un nodo característico, *i.e.*, es imposible definir un nodo cuyas propiedades reflejen el promedio del resto de los nodos de la red. El valor de γ determina muchas propiedades del sistema; de hecho, entre menor es el valor de γ más importantes es el papel de los *hubs* en la red. Para $\gamma > 3$ los *hubs* no son relevantes, al grado que las redes se asemejan a una red aleatoria; cuando $2 < \gamma < 3$ se hace patente una jerarquía de *hubs*, y para $\gamma = 2$ emerge un tipo de red en estrella⁹ con el *hub* más conectado interactuando con una significativa fracción de todos los nodos. Además, las redes libres de escala con exponente $2 < \gamma < 3$ son ultra pequeñas, dado que su camino característico cumple $l \sim \log \log n$ [Chung y Lu, 2002; Cohen y Havlin, 2003], donde n es el número total de nodos de la red; lo cual es significativamente menor que el valor $\log n$ que caracteriza a las redes aleatorias.

Otra particularidad que hace a los *hubs* importantes y distintivos de las redes libres de escala es la robustez que brindan antes fallas aleatorias. Dado que la cantidad de nodos altamente conectados es muy pequeña, la probabilidad de elegir al azar un nodo altamente conectado es muy baja; por lo tanto si nos ponemos a remover nodos de la red al azar, la probabilidad de lograr desconectarla es muy pobre. Por el contrario, el talón de Aquiles de las redes libres de escala es el ataque dirigido, tal que si los *hubs* son removidos en orden decreciente de conectividad es posible desintegrar la red [Albert *et al.*, 2000]. Curiosamente, a mitad de los 60, el Departamento de Defensa de los Estados Unidos encargó a su brazo de investigación, la ARPA (*Advanced Research Projects Agency*)¹⁰, el diseño de la ARPANET (*Advanced Research Projects Agency Network*), el predecesor de lo que hoy es Internet. Se dice que uno de sus objetivos de

⁹Una red en estrella es aquella en la que todos los nodos se conectan directamente a un nodo central, comúnmente llamado concentrador (*hub*); en consecuencia, todas las comunicaciones fluyen a través de éste.

¹⁰La Agencia de Proyectos de Investigación Avanzada (ARPA por sus siglas en inglés), cuya misión era hacer avanzar la tecnología que pudiera ser útil para la milicia, se creó en respuesta al lanzamiento del Sputnik, en 1957, por parte de la entonces Unión Soviética.

diseño era lograr crear una red militar de comunicaciones que fuera robusta ante un ataque nuclear; de tal suerte que en caso de que una sección de la red quedara destruida, el resto podría seguir funcionando. Mientras que *RAND Corporation*, un *think tank*¹¹ estadounidense sobre política global, incluyó en uno de sus estudios sobre comunicación segura por voz el escenario de una guerra nuclear, Charles Herzfeld, director de la ARPA de 1965 a 1967, ha afirmado que si bien “construir tal sistema era claramente una importante necesidad militar, la misión de la ARPA no fue la de hacerlo”, sino que “la ARPANET resultó de nuestra frustración de que sólo existieran un número limitado de poderosas computadoras de investigación en el país, y que muchos investigadores que deberían tener acceso a las mismas estaban separados geográficamente de ellas”. Verdad o mentira, la realidad es que, como lo han demostrado los resultados de Barabási y colegas, Internet es una red muy robusta capaz de soportar las fallas aleatorias de sus nodos y aún seguir funcionando; a costa de tener un talón de Aquiles, el ataque dirigido a sus *hubs*.

Estudios posteriores analizaron, entre otros tipos de redes [Albert y Barabási, 2002], diversas redes biológicas en una variedad de organismos, encontrando que éstas también mostraban un comportamiento libre de escala [Barabási y Oltvai, 2004; Albert, 2005]; *e.g.*, en un estudio publicado en el 2000 se analizaron las redes metabólicas de 43 organismos abarcando los tres dominios de la vida: *bacteria*, *archaea* y *eukarya*. Los resultados de este trabajo indicaron que, para los 43 organismos estudiados, las distribuciones de conectividad de entrada, $P_e(k_e)$, y de salida, $P_s(k_s)$, seguían una ley de potencia de la forma de la ecuación 3-4 [Jeong *et al.*, 2000]. Por otra parte, análisis de redes regulatorias mostraron que mientras la distribución de la conectividad de salida sigue una ley de potencia, la distribución de la conectividad de entrada se ajusta mejor a un comportamiento exponencial de la forma $P_e(k_e) \sim e^{-\beta k_e}$ [Guelzim *et al.*, 2002; Barabási y Oltvai, 2004].

¹¹También conocido como “fábrica de ideas” es una organización, instituto, corporación o grupo que realiza investigación y se encarga de apoyar áreas tales como política social, estrategia política, economía, asuntos de ciencia o tecnología, políticas industriales o de negocios, o asesoramiento militar. La frase *think tank* en el argot americano de tiempos de guerra se refiere a los cuartos en donde los estrategas discuten el plan de batalla.

3.7. La tragedia de los cosmopolitas

Hacia 2002, un estudio publicado por Dorogovtsev *et al.* realizó una interesante observación; al analizar ciertas redes libres de escala deterministas, *i.e.*, crecidas siguiendo reglas específicas, notaron la existencia de una correspondencia entre el coeficiente de *clustering* y el grado de un nodo, $C = 2/k$ [Dorogovtsev *et al.*, 2002]. Esta observación, en conjunto con la sugerencia de que la organización celular debe ser modular [Hartwell *et al.*, 1999], llevó a Ravasz *et al.* a analizar el coeficiente de *clustering* promedio, $\langle C \rangle$, y la distribución del coeficiente de *clustering*, $C(k)$, de las redes metabólicas de 43 organismos¹². Formalmente la distribución del coeficiente de *clustering* se define como

$$C(k) = \frac{\langle C \rangle_k}{n_k}, \quad (3-5)$$

donde $\langle C \rangle_k$ es el coeficiente de *clustering* promedio para todos los nodos con conectividad k y n_k representa el número total de nodos con grado k . Los resultados de Ravasz *et al.* mostraron que los coeficientes de *clustering* promedio de las 43 redes analizadas eran, al menos, un orden de magnitud mayor a los coeficientes de *clustering* promedio de redes aleatorias generadas empleando el modelo Barabási-Albert de redes libres de escala; además, mostraron que para las 43 redes analizadas las distribuciones del coeficiente de *clustering* seguían una ley de potencia de la forma

$$C(k) \sim k^{-1} \quad (3-6)$$

[Ravasz *et al.*, 2002]; ver figura 3-6Cc; contrario a la independencia del coeficiente de *clustering* con respecto a la conectividad observada en redes que siguen el modelo Erdős-Rényi de redes aleatorias o el Barabási-Albert; ver figuras 3-6Ca y Cb. En consecuencia, los autores de dicho estudio sugirieron una estructura jerárquico-modular para las redes metabólicas, en la cual los módulos individuales se componen de nodos densamente agrupados y con baja conectividad, mientras los distintos módulos son interconectados por *hubs*¹³. Estudios posteriores brindaron evidencia apoyando la idea de que estructuras similares se encuentran presentes en diversas redes libres de escala, como en redes sociales, semánticas e Internet [Ravasz y Barabási, 2003],

¹²El conjunto de datos empleado por Ravasz *et al.* fue el mismo utilizado por Jeong *et al.*

¹³El título de esta sección alude a este fenómeno. La tragedia de los cosmopolitas es no pertenecer a lugar alguno, fieles a la filosofía de Facundo Cabral, “no soy de aquí, ni soy de allá”; así, su papel es el de servir de puentes de enlace entre los diversos grupos o módulos.

así como en varias redes biológicas [Dobrin *et al.*, 2004; Resendis-Antonio *et al.*, 2005; Barabási y Oltvai, 2004], lo que llevó eventualmente a denominarlas *redes jerárquico-modulares*.

Diferentes técnicas se han empleado en un esfuerzo por extraer la estructura modular y jerárquica de las redes, pero este tema se tratará a profundidad en el capítulo 4.

3.8. El nivel mesoscópico

En el nivel macroscópico reside todo aquello que podemos ver a simple vista, *i.e.*, los objetos que nos rodean; mientras que al reino de lo microscópico pertenecen los elementos invisibles a nuestros ojos, pero que pueden ser visualizados o cuantificados por diversos instrumentos. Análogamente, es posible estudiar una red a nivel macroscópico analizando sus propiedades globales (*e.g.*, $P(k)$, $C(k)$, $\langle C \rangle$), quedando confinados al nivel microscópico los elementos atómicos de la red: sus nodos y aristas. Empero, existe un nivel intermedio llamado mesoscópico, al cual pertenecen las moléculas y sus interacciones, donde uno puede discutir propiedades de la materia independientemente del comportamiento de los átomos; similarmente, es posible analizar una red a escala mesoscópica, buscando comprender las subestructuras que la componen, así como la función de las mismas. A esta escala existen dos subestructuras de interés: los *circuitos de retroalimentación (feedback)* y los *motivos topológicos*. Mientras que los módulos también podrían considerarse aquí, estos se tratarán más a detalle en el capítulo 4.

3.8.1. Circuitos de retroalimentación

Norbert Wiener, en su libro *Cibernética o el control y comunicación en animales y máquinas*, argumentó que los circuitos de retroalimentación fundamentan el comportamiento teleológico de organismos vivos y máquinas. Estos circuitos se componen de uno o más nodos interconectados mediante un conjunto de interacciones circulares. Si el circuito posee sólo un nodo i entonces tenemos una autoregulación, *i.e.*, el nodo i se controla a sí mismo, \widehat{i} . Cuando el circuito posee dos nodos, i y j , entonces tenemos que i puede afectar a j y viceversa, *i.e.*, $i \rightleftarrows j$.

Trabajo teórico pionero de René Thomas [Thomas y D'Ari, 1990; Thomas, 1998; Thieffry y Thomas, 1998; Thomas y Kaufman, 2001] y trabajos experimentales [Kaern *et al.*, 2005; Smits *et al.*, 2006] han subrayado la relevancia dinámica y biológica de los circuitos de retroalimenta-

Característica	Circuito negativo	Circuito positivo
Número de interacciones negativas	Impar	Par
Propiedad dinámica	Periodicidad	Multiestabilidad
Fenómeno biológico	Homeóstasis	Variabilidad fenotípica y diferenciación

Tabla 3-1: Principales características de los circuitos de retroalimentación.

ción. Se ha mostrado que el comportamiento dinámico de estos circuitos está determinado por el número de interacciones negativas presentes, dando origen al concepto de *signo del circuito* [Thomas y D'Ari, 1990]; así como que la presencia de, al menos, un circuito es una condición necesaria para que se manifieste su dinámica asociada; ver tabla 3-1.

Imaginemos que i y j son dos colaboradores, si i anima a j a dar lo mejor de sí y j anima a i a hacer lo mismo, entonces tenemos lo que comúnmente se llama un círculo virtuoso, que desembocará en el mejoramiento continuo del trabajo realizado por i y j . Por el contrario, si i desanima constantemente a j y viceversa, entonces el trabajo que realicen será pobre o en el peor de los casos no se realizará, estamos ante un círculo vicioso. Así, podemos ver que pueden existir dos posibles efectos de i sobre j , uno positivo y otro negativo. Ambos ejemplos anteriores, el del círculo virtuoso y el del vicioso, son ejemplos de dinámicas multiestables, las cuales dependiendo de las condiciones iniciales y de los parámetros del sistema (*e.g.*, intensidad inicial de la presión sobre el otro, resistencia a la opinión del otro, autoestima, tolerancia a la frustración) evolucionarán hacia uno de los dos o más estados finales posibles. Por otra parte, imaginemos que i es un sujeto optimista, mientras j es un sujeto pesimista; i constantemente está animando a j a mejorar en su trabajo, pero j siempre desanima a i ; el rendimiento neto de ambos será oscilatorio con picos y caídas en su productividad, *i.e.*, mostrarán un comportamiento dinámico periódico. En redes regulatorias, estos comportamientos de los circuitos de retroalimentación se han asociado a fenómenos biológicos tales como homeóstasis, variabilidad fenotípica y diferenciación [Thomas, 1998; Thomas y Kaufman, 2001; Kaern *et al.*, 2005; Smits *et al.*, 2006]; ver tabla 3-1. Además, estudios previos han sugerido que los circuitos de retroalimentación juegan un papel importante en la modularidad de las redes regulatorias [Thieffry y Romero, 1999].

Durante 1998, un estudio realizado por Thieffry *et al.* reveló la existencia de circuitos de retroalimentación de sólo un elemento, en su mayoría negativos, en la red de regulación de *E. coli*

[Thieffry *et al.*, 1998]; empero, no se realizó un análisis sistemático para identificar circuitos de retroalimentación de más de dos elementos. En 2004, Ma *et al.* analizaron la red de regulación de *E. coli* a nivel de operones y encontraron que ésta no contenía circuitos de retroalimentación de más de un elemento, afirmando así que la red era acíclica [Ma *et al.*, 2004a]. Ese mismo año, en un estudio posterior empleando una reconstrucción extendida de la red de regulación de *E. coli*, que incluyó interacciones con factores σ , Ma *et al.* reportaron haber encontrado siete circuitos de retroalimentación compuestos por dos nodos, dejando sin esclarecer si habían buscado circuitos de más de dos nodos [Ma *et al.*, 2004b]. Empero, argumentaron que cada uno de los circuitos identificados estaban compuestos de genes que se encontraban en un mismo operón, lo que los llevó a concluir que los circuitos existentes no eran relevantes para la organización y la dinámica de la red de regulación.

3.8.2. Motivos topológicos

Durante 2002, dos trabajos realizados por Uri Alon y sus colegas del Instituto Weizmann llamaron la atención hacia un nuevo conjunto de estructuras topológicas que ellos denominaron motivos topológicos o motivos de red. En bioinformática, un motivo es una secuencia de nucleótidos o aminoácidos que está sobrerrepresentada estadísticamente con respecto al azar, *i.e.*, una secuencia que dada su composición es poco probable que ocurra por puro azar. De forma análoga, un motivo topológico en una red se define como una subestructura que aparece en una cantidad estadísticamente mayor a lo esperado al azar. El trabajo realizado por Shen-Orr *et al.* fue el primero en tomar prestado el concepto de motivo para definir e identificar los motivos topológicos presentes en la red de regulación de *E. coli* [Shen-Orr *et al.*, 2002]. Posteriormente, un estudio realizado por Milo *et al.* extendió el análisis de motivos incluyendo la red de regulación de *Saccharomyces cerevisiae*, redes neuronales, tróficas o alimentarias, circuitos electrónicos e Internet [Milo *et al.*, 2002]. El enfoque para la identificación de motivos topológicos consiste en lo siguiente [Milo *et al.*, 2002]:

1. Dado un cierto número de nodos, generar todos los posibles subgrafos no isomorfos; ver figura 3-8.
2. Contar el número de ocurrencias de cada subgrafo en la red bajo estudio.

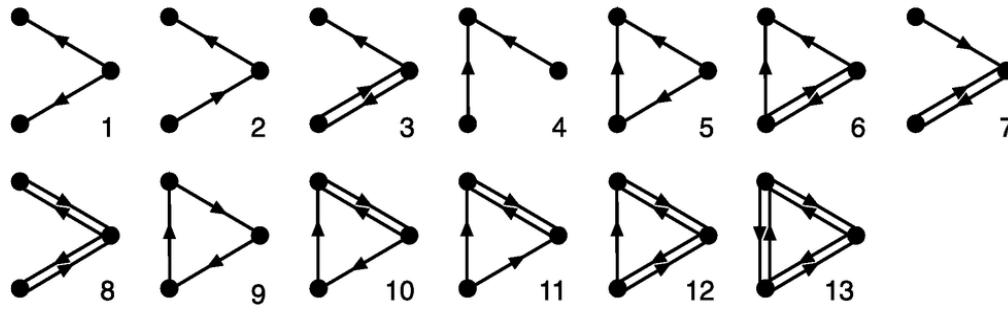


Figura 3-8: Todos los distintos subgrafos conexos, *i.e.*, eliminando isomorfismos, generados empleando tres nodos. (Imagen tomada de Milo *et al.*, 2002.)

3. Generar un número estadísticamente significativo de redes aleatorias a través de aleatorizar la red bajo estudio reconectando sus nodos, pero conservando el grado de entrada y de salida de cada nodo.
4. Por cada red aleatoria, contar el número de ocurrencias de cada subgrafo.
5. Por cada subgrafo, evaluar estadísticamente la probabilidad de que la frecuencia con la que aparece en la red bajo estudio sea la misma que se esperaría al azar.
6. Si un subgrafo muestra una frecuencia cuya probabilidad de ocurrencia al azar es muy pequeña, entonces dicho subgrafo es considerado un motivo topológico.

Milo *et al.* identificaron dos motivos comunes en las redes de regulación de *E. coli* y *S. cerevisiae*, denominándolos *feedforward*¹⁴ y *bi-fan*; ver figura 3-9. Un análisis posterior reveló que la gran mayoría de estos motivos se traslapan, agregándose así en *grupos de motivos homólogos*; los cuales no están del todo aislados sino que se interconectan formando un *supergrupo de motivos* [Dobrin *et al.*, 2004].

Otros estudios definieron motivos alternativos en función de estructuras que son muy comunes en redes biológicas como la autoregulación, los *feedforward* multisalida¹⁵, los módulos

¹⁴ Aunque suele también ser llamado *feedforward loop*, este término es incorrecto. Topológicamente la palabra *loop* (*bucle*) se refiere a un camino que termina en el mismo punto que inició, pero al considerar la dirección de los arcos del *feedforward* fácilmente podemos ver que esto no se cumple.

¹⁵ *Feedforward* en el cual los mismos factores de transcripción X y Y regulan múltiples genes, *i.e.*, *feedforwards* traslapados que comparten los nodos regulatorios X y Y.

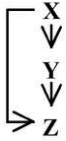

Network	Nodes	Edges	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score	N_{real}	$N_{\text{rand}} \pm \text{SD}$	Z score
Gene regulation (transcription)				Feed- forward loop			Bi-fan	
<i>E. coli</i>	424	519	40	7 ± 3	10	203	47 ± 12	13
<i>S. cerevisiae</i> *	685	1,052	70	11 ± 4	14	1812	300 ± 40	41

Figura 3-9: Motivos topológicos encontrados en las redes de regulación de *E. coli* y *S. cerevisiae*. Ésta última presentó un motivo de cuatro nodos adicional: ($X \rightarrow Z, W$; $Y \rightarrow Z, W$; $Z \rightarrow W$), $N_{\text{real}} = 150$, $N_{\text{rand}} = 85 \pm 15$, $Z_{\text{score}} = 4$. (Imagen tomada de Milo *et al.*, 2002.)

de una entrada¹⁶ y el regulón dénsamente traslapado¹⁷ [Shen-Orr *et al.*, 2002; Alon, 2007].

La presencia de interacciones positivas y negativas en una red de regulación permite la existencia de ocho tipos distintos de *feedforwards*, cuatro de ellos llamados *coherentes* y cuatro *incoherentes*, donde la coherencia de un *feedforward* está en función del efecto neto sobre el gen regulado; ver figura 3-10. Análisis de las redes de regulación de *E. coli* y *S. cerevisiae* han mostrado que dos de los ocho tipos de *feedforwards* ocurren con mucha mayor frecuencia, éstos son el coherente tipo 1 y el incoherente tipo 1 [Ma *et al.*, 2004b; Alon, 2007]. Se ha estudiado el comportamiento dinámico de estos *feedforwards* más abundantes, encontrando que cada uno de ellos exhibe respuestas dinámicas específicas [Alon, 2007].

En términos evolutivos, se ha sugerido que los *feedforwards* evolucionan de forma convergente, tal que la naturaleza constantemente está redescubriéndolos debido a sus relevantes propiedades dinámicas [Alon, 2007]. Sin embargo, varios estudios han cuestionado tanto el valor adaptativo como la relación estructura-función de los motivos [Mazurie *et al.*, 2005; Solé y Valverde, 2006], mostrando que algunos de ellos pueden exhibir comportamientos dinámicos muy diversos [Voigt *et al.*, 2005; Ingram *et al.*, 2006], y brindando evidencia que apunta a que estos son sólo una consecuencia secundaria de la evolución del genoma y la organización de la red [Cordero y Hogeweg, 2006; Kuo *et al.*, 2006; Solé y Valverde, 2006].

¹⁶Los *single-input modules* (SIMs) son motivos en los cuales un regulador X, el cual puede o no autoregularse, regula a un conjunto de genes.

¹⁷En el *dense overlapping regulon* (DOR) muchas entradas regulan a muchas salidas.

Coherent FFL

Coherent
type 1



Coherent
type 2



Coherent
type 3



Coherent
type 4



Incoherent FFL

Incoherent
type 1



Incoherent
type 2



Incoherent
type 3



Incoherent
type 4



Figura 3-10: Los ocho tipos de circuitos *feedforward* (FFL en la figura). En los *feedforward* coherentes el signo del camino directo desde el factor transcripcional X hasta el gen regulado Z es el mismo que el signo del camino indirecto a través del factor transcripcional Y. Los *feedforwards* incoherentes tienen signos opuestos para los dos caminos. (Imagen tomada de Alon, 2007.)

Capítulo 4

Modularidad y Jerarquía en Redes Biológicas

Había una vez dos relojeros, llamados Hora y Tempus, quienes hacían relojes muy finos. Los teléfonos de sus talleres timbraban con frecuencia; nuevos clientes constantemente los llamaban. Sin embargo, Hora prosperó mientras que Tempus se volvió cada vez más pobre. Al final, Tempus perdió su tienda. ¿Cuál fue la razón detrás de esto? Los relojes se componían de alrededor de 1000 piezas cada uno. Los relojes que Tempus hizo estaban diseñados de forma que, cuando él tenía que dejar un reloj parcialmente ensamblado (por ejemplo, para contestar el teléfono), éste inmediatamente se desarmaba y tenía que ser reensamblado por completo. Hora había diseñado sus relojes tal que él podía armar subensamblados de alrededor de diez componentes cada uno. Diez de estos subensamblados podían ser armados para fabricar un subensamblado mayor. Finalmente, diez de los subensamblados mayores constituían el reloj completo. Cada uno de los subensamblados podía ser dejado sin que se desarmara.

— HERBERT A. SIMON

El estudio de las propiedades topológicas globales de las redes biológicas ha permitido la identificación de principios generales de organización; sin embargo, estos estudios de nivel macroscópico arrojan poca luz sobre la relación estructura-función. En esa dirección se ha mostrado que, por ejemplo, redes que toman decisiones requieren de topologías específicas [Oosawa y Savageau, 2002]; además, se ha subrayado la importancia que la organización tiene sobre la dinámica del sistema y viceversa [Variano *et al.*, 2004]. Pero cabe preguntarnos, ¿cómo ocurre la división de trabajo y el control al interior de una red de regulación? En este capítulo sentaremos las bases de la visión jerárquico-modular que actualmente se tiene de la organización celular, así como revisaremos las principales metodologías empleadas para la identificación de módulos y

la inferencia de la estructura jerárquica gobernándolos, haciendo especial énfasis en las técnicas empleadas para redes de regulación; además, veremos que si bien estas metodologías han permitido avanzar en la comprensión de la estructura jerárquico-modular de las redes de regulación, poseen también inconvenientes que han llevado, en ocasiones, a conclusiones inadecuadas al ser analizadas desde un enfoque biológico.

4.1. ¿Cómo ensamblar un reloj sin morir en el intento?

En 1974, tomando prestadas ideas propuestas originalmente por Arthur Koestler, François Jacob en su libro *La lógica de lo viviente* afirmó que los organismos se construyen por una “jerarquía de integraciones”, implicando que los sistemas a los que denominó *integrones* se organizan en sistemas más grandes, siendo así a su vez parte y todo. Un *módulo* se define como un conjunto de elementos que cooperan para lograr una función específica. En años recientes se ha propuesto que la *modularidad* es un nivel crítico de organización biológica en la célula [Hartwell *et al.*, 1999]; en consecuencia se ha establecido que este tipo de organización modular también juega un papel central en las redes biológicas [Alon, 2003], incluyendo las redes de regulación [Wolf y Arkin, 2003]. Como vimos en el capítulo 2, a lo largo de los años se han propuesto diferentes niveles de organización genética; sin embargo, el solo estudio de estos niveles nunca antes permitió particionar alguna red de regulación en sus componentes modulares.

La búsqueda de elementos que interactúan unos con otros tendiendo a formar conglomerados es un problema que ha interesado a la sociología desde hace ya algún tiempo, debido a que los seres humanos tienden a reunirse e interactuar en la búsqueda de objetivos comunes. En redes sociales, estas comunidades de agentes que interactúan permiten identificar grupos de individuos compartiendo gustos u objetivos, permitiendo así poner en relieve las tendencias de los grupos sociales [Newman y Girvan, 2004]. En las *redes comunitarias*, los nodos al interior de una comunidad tienen a interactuar preferentemente con otros miembros de la misma comunidad que con miembros de otras comunidades; ver figura 4-1.

El advenimiento en el 2002 de las redes jerárquico-modulares marcó un parteaguas en el estudio de la modularidad. Mientras que una *red jerárquico-modular* exhibe conglomerados, tal

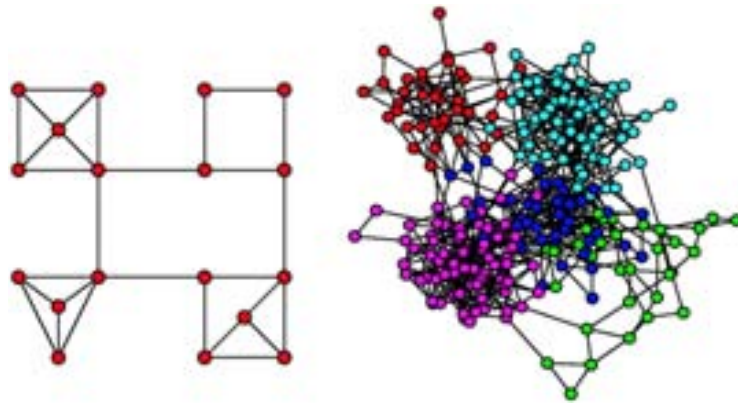


Figura 4-1: Ejemplo de red de comunidades, en la cual se observa cómo los miembros de una comunidad tienden a interactuar preferentemente con otros miembros de la misma comunidad que con los de comunidades alternas. (Imagen tomada de Ravasz *et al.*, 2002.)

como lo hace una red comunitaria, sus propiedades libres de escala la hacen presentar también *hubs* los cuales interconectan a los módulos, complicando su identificación; ver figura 4-2. Se han propuesto diferentes metodologías para identificar los módulos que componen una red jerárquico-modular; sin embargo, la mayoría de estos métodos se basan en un conjunto de herramientas comunes: técnicas de agrupamiento y de optimización.

4.1.1. Agrupamiento jerárquico aglomerativo

Esta es una técnica de agrupamiento de datos que, a diferencia de otras más rudimentarias como *k-means*, no requiere que se le indique *a priori* el número de grupos que se desean; por el contrario, el algoritmo va definiendo parejas de grupos en función de una métrica estadística de disimilitud, los cuales luego son comparados contra otros grupos resultando así un árbol o dendograma de distancias el cual, conforme se acerca a la raíz, indica como disminuye la similitud entre los grupos. Para calcular las medidas de disimilitud entre dos grupos necesitamos definir primero qué representan los valores que deseamos comparar, los cuales generalmente son alguna métrica topológica de la red.

La primer aplicación de esta metodología a redes jerárquico-modulares la realizaron Ravasz *et al.*, quienes emplearon una métrica topológica que denominaron traslape topológico, la cual dados un par de nodos cuantifica el número de enlaces a intermediarios comunes entre

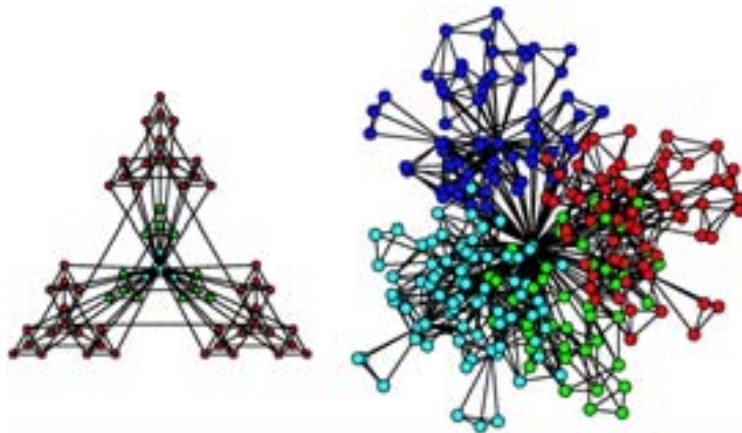


Figura 4-2: Ejemplo de red jerárquico-modular, en la cual se observa cómo los módulos que la componen son interconectados de forma jerárquica por los *hubs*. (Imagen tomada de Ravasz *et al.*, 2002.)

ambos; ésta fue aplicada a la red metabólica de *E. coli* obteniendo una matriz de traslape topológico cuyas columnas y renglones fueron reordenados empleando un agrupamiento jerárquico aglomerativo, permitiéndoles así identificar un conjunto de módulos en la red [Ravasz *et al.*, 2002]. Sin embargo, los autores de este estudio no analizaron la red tal y como es, sino que realizaron un preprocesamiento de ella modificando su estructura general, el cual incluyó la eliminación arbitraria de ciertos metabolitos altamente conectados, sin definir de forma concreta cuales fueron eliminados y cómo fueron elegidos.

Posteriormente, en el 2003 tuve la oportunidad de participar en uno de los primeros estudios dirigidos a comprender la organización modular de las redes de regulación. Inspirados por un estudio realizado en redes proteína-proteína por Rives y Galitski, en el cual se definió una nueva métrica topológica basada en el camino mínimo entre dos nodos, denominada asociación topológica [Rives y Galitski, 2003], Resendis-Antonio *et al.* nos dimos a la tarea de identificar los módulos presentes en la red de regulación de *E. coli*, utilizando para ello la red de genes que sólo codifican proteínas regulatorias¹. Mi papel en dicho estudio estuvo concentrado, principalmente, en el análisis de las propiedades topológicas de la red, así como en la identificación y validación de los módulos que la componen. A partir de la red de regulación generamos una matriz de

¹En el apéndice A, el lector encontrará la publicación donde se reportaron los resultados de este estudio.

asociación \mathbf{M} cuyos elementos fueron calculados empleando la fórmula $1/d_{i,j}^2$, siendo $d_{i,j}$ la longitud del camino mínimo entre el par de nodos i y j ; esta métrica permite maximizar los valores pequeños de distancia minimizando los grandes. El siguiente paso fue reordenar los renglones y columnas de la matriz de asociación mediante aplicar un agrupamiento jerárquico aglomerativo empleando como métrica de disimilitud el coeficiente τ de Kendall, lo cual reveló un conjunto de módulos funcionalmente relevantes, ver figura 4-3; lo que complementado con otros análisis mostró que los motivos *feedforward* tienden a encontrarse dentro de los módulos, mientras los motivos *bi-fan* en su mayoría interconectan módulos [Resendis-Antonio *et al.*, 2005].

Mientras que los análisis realizados nos brindaron una perspectiva global de la presencia de módulos y la distribución de los motivos topológicos en la red, también hicieron emerger en mí una serie de observaciones e inquietudes respecto al método empleado:

- Realicé una serie de experimentos independientes los cuales mostraron que el resultado de un agrupamiento jerárquico aglomerativo está en función de los parámetros elegidos: métrica topológica y métrica de disimilitud; tal que al variar alguno de ellos, los módulos obtenidos son distintos. ¿Cuál es la combinación de métricas más adecuada para identificar módulos en una red? ¿Cómo puede evaluarse esto?
- El dendograma obtenido por Resendis-Antonio *et al.* nos mostró una clara división en módulos embebidos unos dentro de otros; sin embargo, si la partición no fuera tan clara, ¿cuál sería la forma más adecuada para definir un corte en el dendograma?
- A diferencia del trabajo realizado por Ravasz *et al.*, Resendis-Antonio *et al.* decidimos no eliminar ningún nodo de la red conservando aun los reguladores globales; así, en la figura 4-3 es posible observar cómo estos reguladores globales se conectan no sólo con los genes del módulo donde se encuentran, sino que además establecen una serie de entrecruzamientos al conectarse con genes en otros módulos. Si un regulador global, por definición, es pleiotrópico, ¿es correcto que estos posean pertenencia a algún módulo, o debieran, por el contrario, carecer de pertenencia a ellos? Empero, si por el contrario, debemos removerlos, ¿cómo identificar los globales?
- La red empleada por Resendis-Antonio *et al.* sólo contenía genes que codifican proteínas regulatorias, cuyo componente gigante, en ese entonces, poseía 55 nodos; empero la red

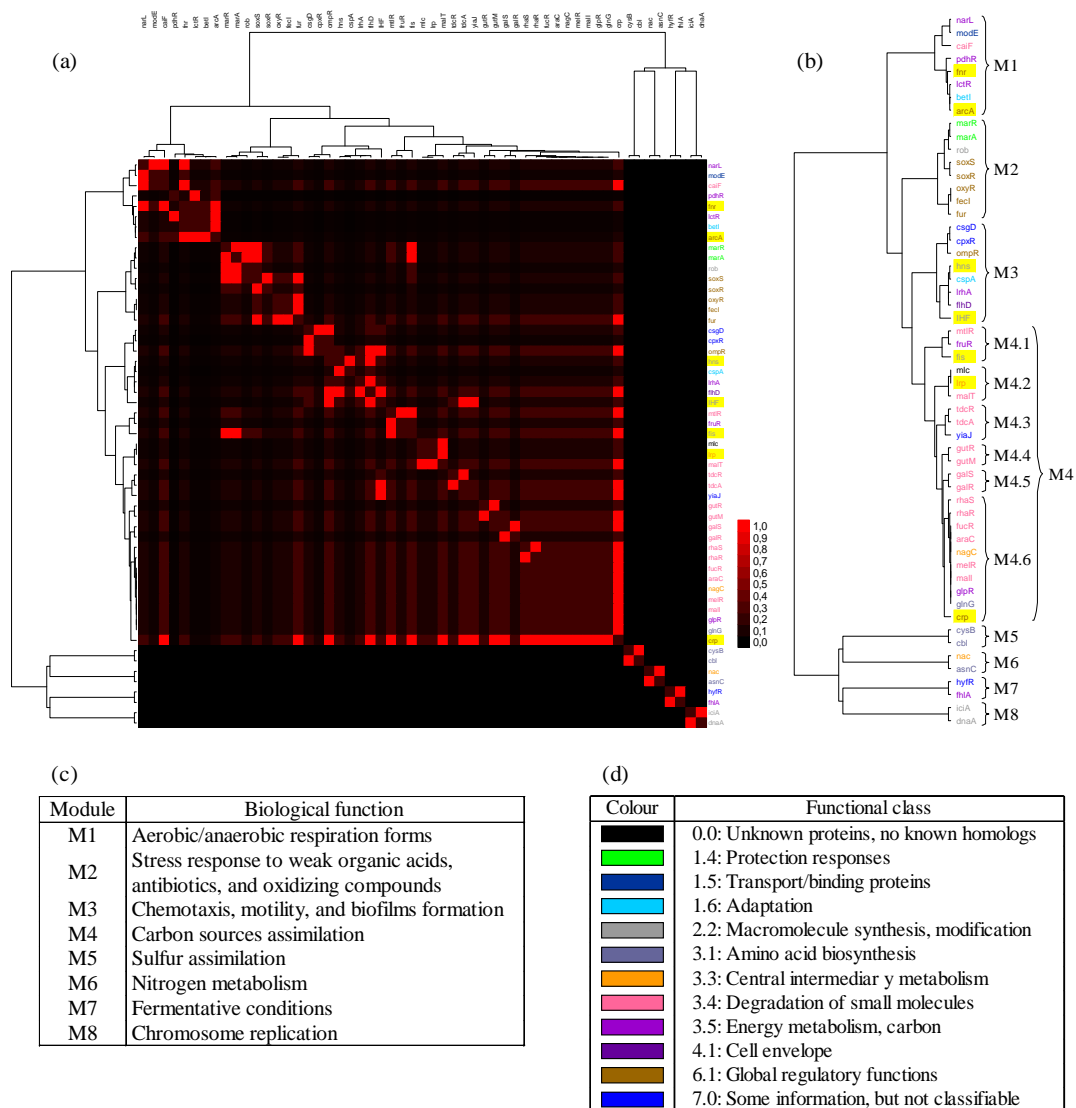


Figura 4-3: Módulos identificados en la red de regulación de *E. coli*. (a) Agrupamiento jerárquico aglomerativo donde la intensidad de cada punto representa la longitud del camino mínimo entre un par de nodos; el valor de intensidad puede variar entre 0 y 1, un valor de 1 implica que dos genes tiene una conexión directa mientras que un valor de 0 nos dice que una pareja de genes no se conecta en la red. Resaltados en amarillo se aprecian los reguladores globales (*hubs*). (b) Dendograma mostrando el conjunto de genes que componen cada módulo. (c) Funciones biológicas de cada uno de los ocho módulos identificados. (d) Código de color empleado para anotar cada gen con su correspondiente clase funcional. (Imagen tomada de Resendis-Antonio *et al.*, 2005.)

de regulación completa es varias veces más grande y mucho más compleja, ¿qué pasaría si se añadieran todos los genes estructurales y sus interacciones?

Para analizar este último punto extraje la red completa que incluía genes que codifican proteínas reguladoras, genes estructurales y las interacciones regulatorias entre ellos, procediendo a repetir la metodología empleada por Resendis-Antonio *et al.*; los resultados de este análisis mostraron lo siguiente:

- El carácter pleiotrópico de los reguladores globales y la restricción de que éstos sólo pueden poseer pertenencia a un módulo confundió al algoritmo de agrupamiento, impidiendo en consecuencia la identificación de una correcta partición de la red; ver figura 4-4.
- Se presentó una gran cantidad de traslape entre módulos, evidente en las regiones lejanas de la diagonal de la matriz del agrupamiento; ver figura 4-4. Mientras que en el agrupamiento realizado por Resendis-Antonio *et al.* el traslape se originó principalmente por proteínas regulatorias globales interactuando con genes en otros módulos, en este análisis con la red íntegra, el traslape además se debió a la presencia de genes coregulados por dos o más módulos.
- La anotación de los genes con sus clases funcionales reveló que los módulos identificados de esta forma no tienden a estar enriquecidos funcionalmente con una clase funcional en particular, sino que, por el contrario, estas se distribuyen de forma muy heterogénea; ver figura 4-4.

Posteriormente, recibí una invitación para participar en un interesante estudio realizado por Gutiérrez-Ríos *et al.* sobre cómo la red de regulación de *E. coli* responde a la presencia de glucosa en el medio². La disponibilidad de estudios de microarreglo en condiciones con y sin glucosa me permitió proponer la reconstrucción de una red condición-dependiente, *i.e.*, una red en la que sólo estuvieran implicados los genes que sufrieron un cambio estadísticamente significativo en su nivel de expresión (expresión atípica), así como las interacciones regulatorias entre ellos. Sin embargo, un problema común con los microarreglos es la baja sensibilidad que tienen para

²En el apéndice B, el lector encontrará la publicación donde se reportaron los resultados de este estudio.

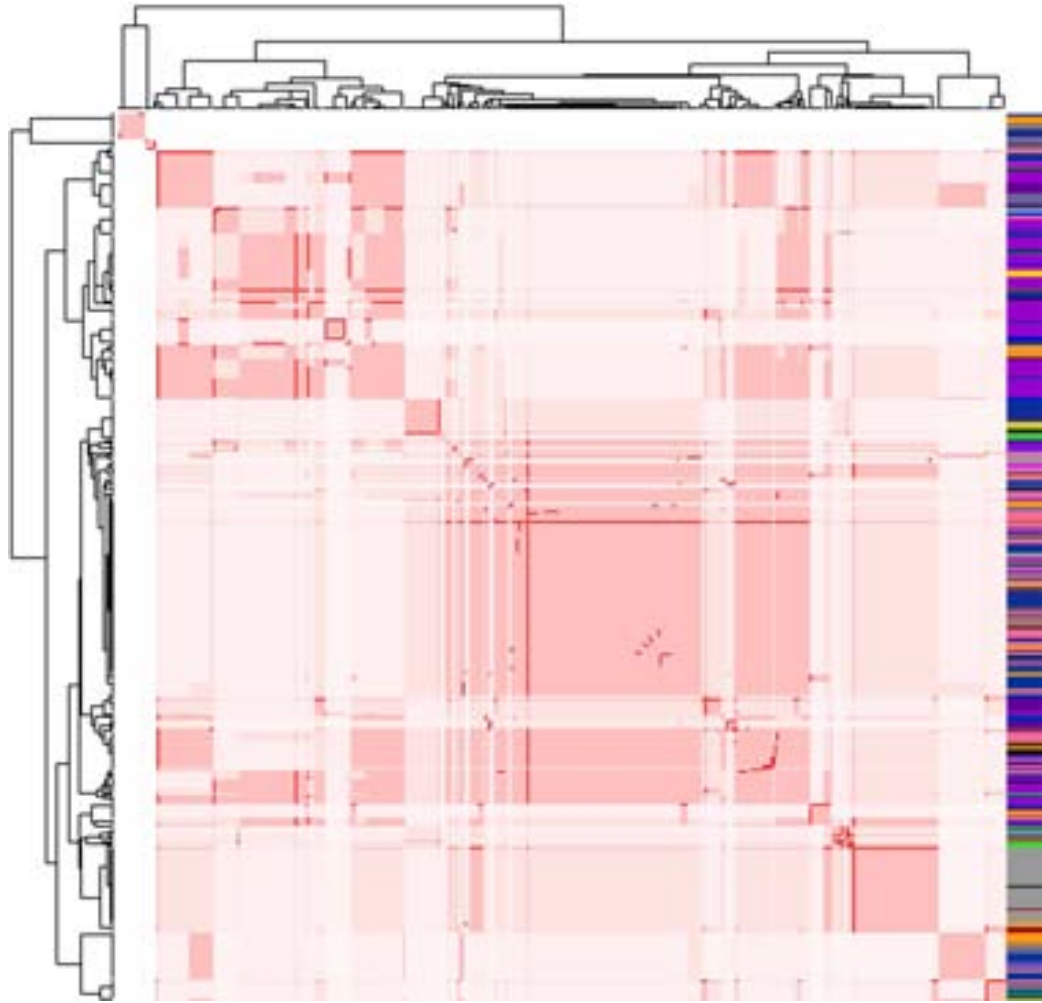


Figura 4-4: Módulos de la red de regulación íntegra de *E. coli* obtenidos empleando los mismos parámetros descritos por Resendis-Antonio *et al.*; en la columna de la izquierda se muestran las clases funcionales codificadas siguiendo el mismo esquema de la figura 4-3.

detectar los cambios de expresión de la mayoría de los factores transcripcionales debido a su baja concentración, lo cual impacta la calidad de la red obtenida al estar ausentes muchos factores transcripcionales. Para sortear este problema propuse incluir a todos aquellos factores de transcripción que regularan de forma directa a los genes identificados con una expresión atípica en el microarreglo. A esta red condición-dependiente se le aplicó la metodología descrita por Resendis-Antonio *et al.*, obteniéndose ocho módulos y ocho minimódulos, siete de estos últimos desconectados del componente gigante, compuestos por genes que responden directa o indirectamente a la presencia de glucosa en el medio [Gutierrez-Ríos *et al.*, 2007]; ver figura 4-5.

La reconstrucción de una red condición-dependiente permitió simplificar la red de regulación e identificar módulos de respuesta a una condición ambiental precisa. A pesar de esto, el traslape entre módulos siguió estando patente en las regiones alejadas de la diagonal de la matriz del agrupamiento; por una parte debido a los factores de transcripción globales, y por otra a la corregulación de genes por módulos distintos.

4.1.2. Maximizando la modularidad

En 2005, Guimerà y Amaral propusieron un método para identificar módulos en redes metabólicas, el cual transforma el problema de identificación de módulos en un problema de optimización [Guimerà y Amaral, 2005]; para lograr esto, los autores proponen que dada una red existe un conjunto finito \mathcal{P} de particiones, o agrupamientos de nodos en módulos, y que la modularidad de cada partición $P \in \mathcal{P}$ puede ser cuantificada como

$$M(P) \equiv \sum_{i=1}^m \left[\frac{l_i}{L} - \left(\frac{d_i}{2L} \right)^2 \right], \quad (4-1)$$

donde m es el número de módulos en la partición P , L es el número total de enlaces en la red, l_i es el número de enlaces dentro del módulo i , y d_i es la suma de los grados de todos los nodos del módulo i . De acuerdo a los autores, el razonamiento detrás de esta medida es que una buena partición en módulos debe presentar muchos enlaces intramódulos pero pocos intermódulos; sin embargo, si sólo se maximiza el número de enlaces intramódulos la partición óptima es aquella compuesta de un solo módulo; por lo que la ecuación 4-1 impone que $M(P) = 0$ si los nodos son puestos en módulos al azar o si todos los nodos están en el mismo grupo.

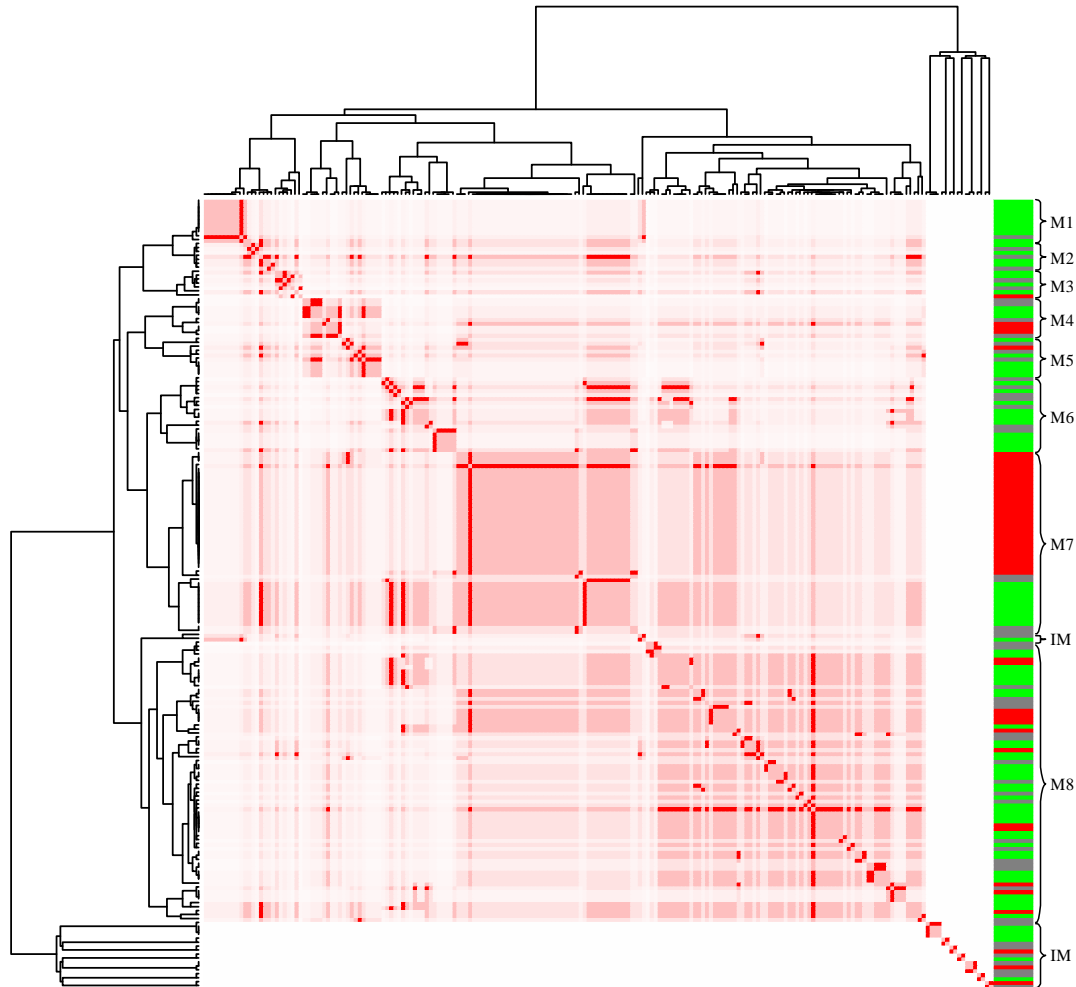


Figura 4-5: Módulos identificados en una subred condición-dependiente (WT+Glu/WT), en la cual se identificaron ocho módulos interconectados y compuestos de genes con funciones relacionadas, más ocho minimódulos, siete de los cuales aparecen desconectados del componente gigante. En el lado derecho cada gen fue etiquetado con un color en función del nivel relativo de ARNm identificado en el microarreglo, rojo si incrementó, verde si decrementó, y gris si el gen no se identificó en el microarreglo. (Imagen tomada de Gutierrez-Ríos *et al.*, 2007.)

Con esta medida de modularidad en mano, los autores aplicaron la técnica de recocido simulado (*simulated annealing*) para encontrar la partición que maximiza el número de enlaces intramódulos. Sin embargo, un problema que emerge de este enfoque es que la técnica de recocido simulado es un algoritmo estocástico, y este tipo de algoritmos no garantizan encontrar la mejor solución pero, si encuentran una, la solución es muy buena, *i.e.*, es un máximo local. Esto implica que cada ejecución del algoritmo producirá distintas particiones cuasióptimas, lo que deja abierta la pregunta, ¿cómo determinar cuál de todas estas particiones es la mejor?

Por otra parte, el método propuesto por Guimerà y Amaral se basa en una premisa fundamental de las redes comunitarias, la existencia de comunidades cuyos nodos tienden preferentemente a interactuar con otros nodos de la misma comunidad que con nodos de otras comunidades. Empero, como ya comentamos, en una red jerárquico-modular existen estos nodos denominados *hubs*, los cuales por definición interconectan una gran cantidad de nodos en la red. Bajo estas condiciones, asignar cualquier *hub* a un módulo tendrá el efecto de elevar el número de conexiones intermódulos, por lo que el problema de maximizar el número de conexiones intramódulos se torna más difícil de resolver. Este problema con los *hubs* es el mismo que, como subrayé previamente, se observa en el análisis de Resendis-Antonio *et al.* respecto al traslape generado por los reguladores globales.

En un estudio posterior, Sales-Pardo *et al.* afirmaron que el hecho de que el coeficiente de *clustering* siga una ley de potencia de la forma $C(k) \sim k^{-1}$ no es una condición necesaria ni suficiente para que una red posea modularidad jerárquica [Sales-Pardo *et al.*, 2007]. En este punto y como sustento de su afirmación, su artículo cita un estudio realizado por Soffer y Vázquez; sin embargo, en dicho estudio lo establecido por los autores es que la existencia de modularidad jerárquica se podía predecir a partir del análisis de las correlaciones entre los grados de los nodos, afirmando [Soffer y Vázquez, 2005]:

[...] El presente trabajo tiende un puente entre estos dos enfoques distintos [correlación de grados y ley de potencia de la $C(k)$] para cuantificar la estructura jerárquica de la Internet, mostrando que las variaciones en el coeficiente de *clustering* con respecto al grado del vértice, como es medido con la definición acostumbrada [$C(k) \sim k^{-1}$], sólo están reflejando la existencia de correlaciones entre los grados. [...]

4.1.3. Técnicas actuales son inadecuadas para redes jerárquico-modulares

Como hemos visto hasta ahora, las principales técnicas empleadas para la identificación de módulos presentan inconvenientes al ser aplicadas a redes jerárquico-modulares debido a las propiedades intrínsecas de éstas. ¿Cuál es la combinación de métricas, topológica y de disimilitud, más adecuada para identificar módulos en una red? ¿Cuál sería la forma más adecuada para definir un corte en el dendograma?

Por otra parte, un problema más de fondo que impacta a ambas de las técnicas aquí revisadas es el de los *hubs*. Este tipo de nodos son una característica inherente a las redes libres de escala, y por extensión a las jerárquico-modulares. Los *hubs* están íntimamente vinculados con la organización jerárquica que gobierna a una red jerárquico-modular, como lo revela una análisis minucioso de la figura 4-2. En redes de regulación transcripcional, estos *hubs* deben corresponder con los factores de transcripción globales, los cuales al ser pleiotrópicos deben interconectar distintos módulos estableciendo así una jerarquía en la toma de decisiones de la bacteria. Estos razonamientos nos llevan a la conclusión de que los *hubs* no deben poseer pertenencia a los módulos, sino por el contrario, estos deben interconectarlos estableciendo así una jerarquía en la red.

4.2. ¿Cómo coordinar las partes para darle sentido al todo?

Si bien muchos de estos estudios dejaban en claro la existencia de una estructura modular en la red de regulación de una célula, la existencia de una *jerarquía* gobernando a dichos módulos aún no era evidente. El concepto de organización jerárquico-modular sugiere la existencia de módulos tipo integrones, *i.e.*, módulos que son a su vez parte y todo; por lo que el siguiente paso fue dedicar esfuerzos a dilucidar la organización jerárquica de una red.

4.2.1. Agregación de motivos topológicos

Uno de los primeros pasos en esta dirección fue dado en 2004 por Dobrin *et al.*, quienes estudiando la organización de los motivos topológicos de la red de regulación de *E. coli* descubrieron que éstos no se encuentran aislados sino que se agregan formando grupos homólogos de motivos los cuales en gran parte se traslapan fusionándose en un supergrupo, formando así la médula

de la red de regulación transcripcional. En consecuencia, los autores sugieren que los motivos no están aislados sino embebidos en una jerarquía multinivel de interacciones regulatorias. Sin embargo, si bien este trabajo arrojó luz por primera vez sobre una organización de más alto nivel en la red de regulación, aún había preguntas abiertas. ¿Quiénes son los actores de dicha jerarquía y cómo se organizan? ¿Cuáles son los niveles de la misma? ¿Cómo se relaciona la jerarquía con los módulos?

4.2.2. Estructuras jerárquicas piramidales

Tradicionalmente, por arraigo matemático, la idea de jerarquía viene a la mente en forma de *estructuras piramidales* las cuales presentan en sus capas superiores a los actores con más alta jerarquía. Siguiendo esta idea en el 2004, Ma *et al.* realizaron un par de estudios; en uno analizaron la red de regulación transcripcional de *E. coli* a nivel de operones [Ma *et al.*, 2004a] y en otro al nivel de genes [Ma *et al.*, 2004b]; incluyendo en ambos análisis a los genes que codifican factores σ y las interacciones con sus genes transcritos.

En su primer estudio, los autores establecieron que la red analizada no poseía ciclos o circuitos de retroalimentación, así que propusieron un enfoque descendente con el que infirieron una estructura jerárquica de cinco capas [Ma *et al.*, 2004a]. El enfoque propuesto por ellos es una variante del algoritmo conocido en ciencias de la computación como *ordenación topológica* u *ordenación ancestral* [Lipschutz, 1986], el cual tiene el objetivo de, a partir de un grafo, generar un ordenamiento de los nodos tal que el ancestro de cada nodo en el grafo preceda a ese nodo en el ordenamiento. Un problema con este algoritmo es que no puede ser aplicado a redes que contengan al menos un ciclo, ya que en esos casos el algoritmo no sabe que hacer al surgir una paradoja; supongamos, sin pérdida de generalidad, que tenemos un ciclo de dos elementos, ¿cuál de los dos elementos debe preceder al otro? Sin mayor información, no existe respuesta a esta paradoja, por lo que el algoritmo no puede continuar.

En su segundo estudio, Ma *et al.* descubrieron que, a diferencia de la red a nivel de operones, la red analizada a nivel de genes sí poseía ciclos, y afirmaron haber identificado siete de ellos de dos elementos, *i.e.*, de la forma $a \rightleftharpoons b$. Sin embargo, sin nunca enumerar exhaustivamente todos los ciclos identificados, aseveraron que todos los genes que participaban en un ciclo estaban en el mismo operón, motivo por el cual estos ciclos no habían sido identificados en su análisis de

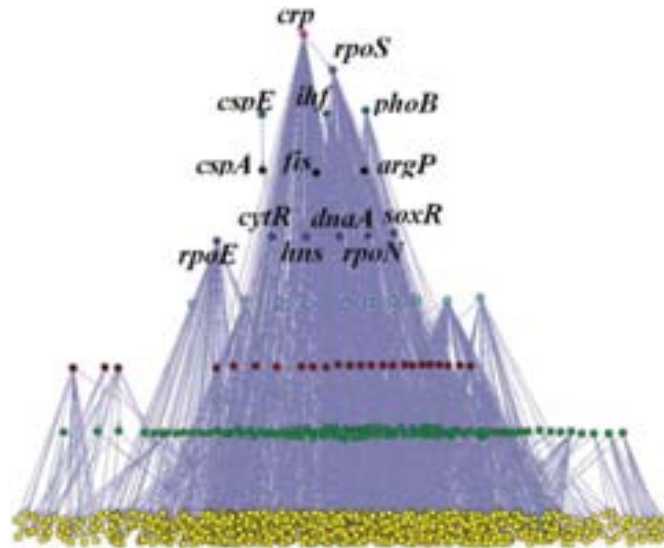


Figura 4-6: Estructura jerárquica multicapa propuesta por Ma *et al.*; quienes argumentan que los genes que aparecen en las cinco capas superiores tienden a ser reguladores globales, y que la red carece de ciclos de retroalimentación relevantes para su organización. (Imagen tomada de Ma *et al.*, 2004b.)

la red a nivel de operones. Debido a esto, propusieron que los genes en ciclos se encontraban al mismo nivel de jerarquía y en consecuencia los ciclos eran irrelevantes para la organización y dinámica de la red; lo que los ayudó a eliminar el problema que los ciclos planteaban para la aplicación de su metodología. Empleando una vez más su enfoque descendente, identificaron una estructura jerárquica de nueve capas y se percataron que el 71 % de los genes en las cinco capas superiores habían sido reportados, al menos una vez, como reguladores globales; ver figura 4-6. Lo que los autores no discutieron en su estudio es lo incongruente, en términos biológicos, del hecho que reguladores no pleiotrópicos como CspE, CspA o ArgP aparezcan en la segunda y tercera capa de la jerarquía, mientras otros reguladores reconocidos como globales como ArcA, FNR y Lrp no aparezcan en ninguna de las cinco capas superiores y en consecuencia no sean identificados por esta metodología como reguladores globales.

Durante el 2006, Yu y Gerstein publicaron un estudio siguiendo la misma línea basada en estructuras jerárquicas piramidales, en el cual proponen un enfoque ascendente para inferir la jerarquía [Yu y Gerstein, 2006]. Su enfoque consiste en primero identificar todos los factores de transcripción en el nivel inferior (nivel 1), *i.e.*, aquellos que no regulan a otros factores de

Nivel	Genes
4	<i>yhiW gntR soxR cspE</i>
3	<i>oxyR lrhA cspA yhiX rob marR soxS exuR</i>
2	<i>yhiE fur crp lrp metJ cytR tdcR flhC rhaR gutM narL himA rpoS feaB cysB fis B2087 cpxR flhD rcsB rpoN fruR flhA glnG marA nac fnr srlR dnaA rpoE uxuR modE himD hns ompR galR arcA mlc feaR lysR rhaS <i>phoP pdhR fadR</i></i>
1	<i>metR appY trpR tyrR argR glcC xylR purR rpiR gals lldR mtlR malT atoC malI hydG emrR hycA cadC asnC yeiL idnR ilvY hupB betI uidR lexA rpoH gcvA fucR hcaR B2531 ada melR yiaJ glpR rcsA fliA cynR putA cbl dsdC treR arsR nagC csgD tdcA rtcR farR phoB araC hupA hipB yhhG fecI iclR B2090 torR caiF sdiA uhpA yjdG xapR evgA nadR adiY narP B1399 deoR gcvR acrR leuO ygaE envY alpA pspF ylcA hyfR yjbK ebgR kdpE yhdM slyA ygaA lacI rbsR nhaR mhpR birA</i>

Tabla 4-1: Jerarquía de la red de regulación de *E. coli* propuesta por Yu y Gerstein; en ella se observa que los reguladores globales conocidos (en negritas), así como algunos factores σ (RpoS y RpoN), se localizan en la tercera capa (nivel 2) de jerarquía, mientras que los genes en las capas superiores (niveles 3 y 4) carecen de efectos pleiotrópicos, lo cual, por definición, es ilógico desde una perspectiva biológica. (Tabla tomada de Yu y Gerstein, 2006.)

transcripción, excepto a sí mismos. Enseguida, empezando desde cada factor de transcripción en el nivel inferior, se realiza una búsqueda en anchura para transformar la red en un árbol de búsqueda en anchura. Esto es equivalente a definir el nivel jerárquico de un cierto factor de transcripción como su distancia más corta desde el factor de transcripción en el nivel inferior. Sin embargo, como los mismos autores mencionan, un problema con este enfoque es su incapacidad para resolver satisfactoriamente estructuras de tipo *feedforward*, las cuales, como se mencionó en el capítulo 3, son muy comunes en redes de regulación. Por otra parte, los autores afirman que los cuellos de botella en el flujo de la información se encuentran en la parte media de la jerarquía; ellos llegan a esta conclusión dado que los factores de transcripción pleiotrópicos son ubicados por su algoritmo en los niveles medios, lo cual es incongruente desde una perspectiva biológica; ver tabla 4-1.

Finalmente, los autores argumentan que su método basado en búsqueda en anchura asigna el nivel jerárquico más bajo posible, debido a que está basado en caminos mínimos. Así, proponen como alternativa calcular el camino más largo desde un factor de transcripción hasta el nodo en la capa inferior y asignar este valor como nivel jerárquico. Este segundo método trae a colación dos problemas:

1. El problema de encontrar el camino más largo entre un par de nodos es intratable computacionalmente, *i.e.*, es NP-completo, lo cual implica que el tiempo de ejecución del algoritmo crece exponencialmente conforme crece linealmente el número de nodos que componen la red.
2. Si existen ciclos en la red, estos funcionarán como atractores causando que el algoritmo de búsqueda del camino más largo nunca se detenga.

Capítulo 5

Disectando la Arquitectura Funcional de *E. coli*

La simplicidad es la sofisticación suprema.

— LEONARDO DA VINCI

Sin algún elemento de gobierno desde la cima, el control ascendente se congelará cuando las opciones sean muchas. Sin algún elemento de liderazgo, los muchos en la parte inferior se paralizarán con las alternativas.

— KEVIN KELLY, *New rules for the new economy* (1999)

Como vimos en el capítulo 4, diversos enfoques se han aplicado en un esfuerzo por identificar la organización modular y jerárquica de las redes de regulación transcripcional. Sin embargo, estos enfoques no son naturales dado que dependen de parámetros los cuales es difícil evaluar con precisión, agrupan genes altamente conectados en módulos, y ubican genes que se sabe poseen efectos pleiotrópicos en capas inferiores de la jerarquía. Estos motivos hacen evidente la necesidad de desarrollar un nuevo enfoque para revelar la organización jerárquico-modular gobernando las redes de regulación. En este capítulo presentamos las bases del método de descomposición natural desarrollado en esta investigación y los resultados obtenidos al disectar la red de regulación de *E. coli*.

5.1. Red de regulación de *Escherichia coli* K-12

La red de regulación transcripcional de la bacteria *Escherichia coli* es la más completa y mejor caracterizada de un organismo procarionte. La red empleada en este estudio se reconstru-

yó a partir de datos obtenidos de la base de datos RegulonDB versión 5.0 [Salgado *et al.*, 2006], los cuales se complementaron con 81 nuevas interacciones regulatorias encontradas mediante una revisión de la literatura sobre los promotores transcritos por los siete factores σ conocidos de *E. coli*¹; ver figura 5-1.

Como se mencionó en el capítulo 2, algunos trabajos [Griffith *et al.*, 2002; Martin *et al.*, 2002; Griffith y Wolf, 2004] han mostrado la existencia de proteínas regulatorias capaces de interactuar con la ARN polimerasa antes de que ésta se una al promotor, en un mecanismo que nos recuerda el empleado por los factores σ ; además en dicho capítulo se argumentó que los factores σ también son factores de transcripción al participar de forma activa en la toma de decisiones sobre qué genes deben de ser expresados. Por estos motivos, se decidió incorporar a la red de regulación transcripcional estudiada aquí los factores σ como factores de transcripción activadores, ya que su presencia es una condición necesaria para que la transcripción ocurra al promover la expresión de los genes sujetos a sus promotores particulares.

5.2. Enfoque de descomposición natural

El enfoque de descomposición natural desarrollado en esta investigación está fundamentado en propiedades intrínsecas de las redes jerárquico-modulares [Freyre-González *et al.*, 2008]. A continuación, un breve resumen de los puntos más relevantes de este enfoque y sus resultados al disectar la red de regulación de *E. coli*².

5.2.1. La red de regulación de *E. coli* no es acíclica

El primer paso para desarrollar esta metodología fue demostrar mediante la enumeración sistemática de todos los circuitos de retroalimentación presentes en la red de regulación de *E. coli* que ésta no es acíclica, mostrando que, por el contrario, ésta posee circuitos de retroalimentación que tienden principalmente a interconectar factores de transcripción globales y locales (en este punto de la investigación empleamos las definiciones de factores de transcripción globales y

¹Los datos de esta revisión de la literatura fueron amablemente facilitados por el Dr. Luis Gerardo Treviño Quintanilla, quien realizó la curación de éstos siguiendo los mismos estándares empleados por el equipo de curadores de RegulonDB. Todas estas nuevas interacciones se encuentran disponibles en RegulonDB versión 6.1.

²En el apéndice C, el lector encontrará la publicación donde se reportaron tanto la metodología como los resultados de este estudio.

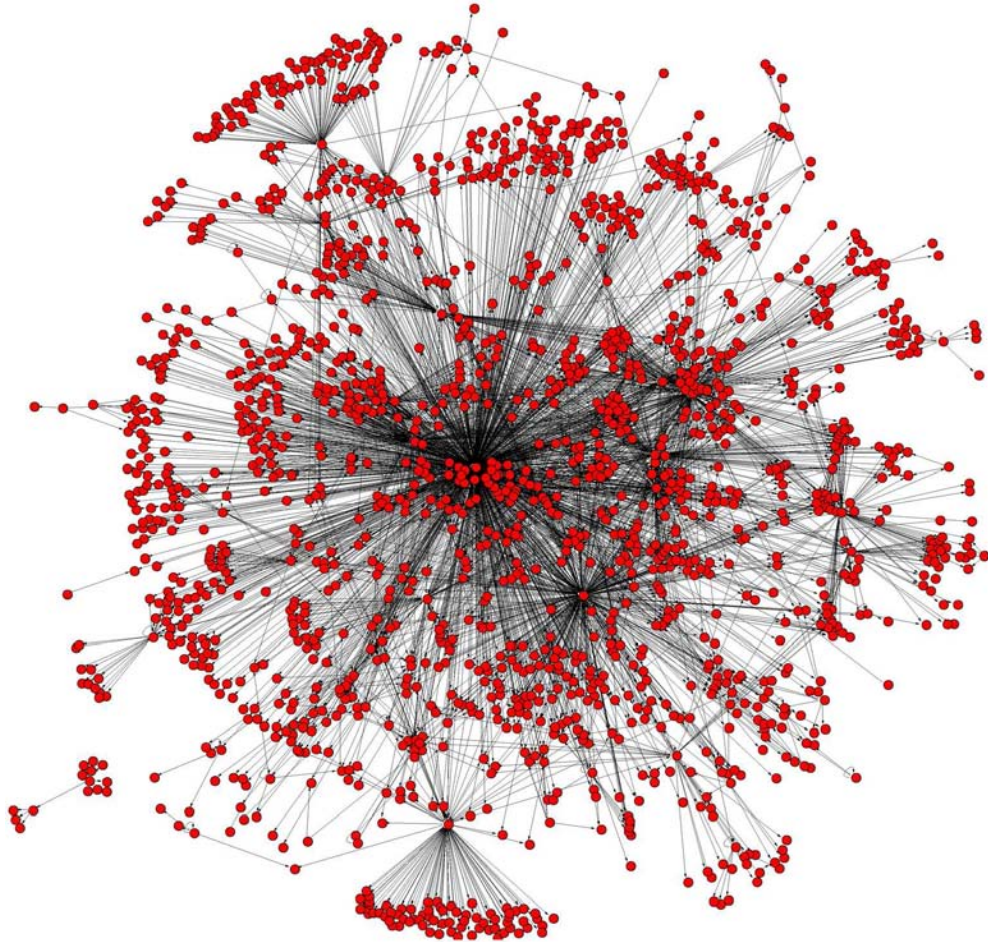


Figura 5-1: Representación gráfica de la red de regulación transcripcional de *E. coli* reconstruida para este estudio; cada nodo en rojo representa un gen, y cada arista entre ellos una interacción regulatoria.

locales dada por Martínez-Antonio y Collado-Vides [Martínez-Antonio y Collado-Vides, 2003]). Esto se logró mediante el desarrollo de un algoritmo para la identificación de circuitos de retroalimentación optimizado para redes de regulación.

5.2.2. Identificación de los nodos jerárquicos y modulares

Mientras que las redes libres de escala son altamente robustas a ataques aleatorios, su talón de Aquiles es el ataque dirigido a los *hubs* [Albert *et al.*, 2000]. Así, el primer paso del enfoque de descomposición natural se basa en la identificación de los *hubs* y su subsecuente remoción de la red. En la $C(k)$ se observan dos comportamientos aparentemente contradictorios; por un lado nodos con alta conectividad poseen muy bajo coeficiente de *clustering*, mientras por el otro nodos con conectividad baja muestran altos coeficientes de *clustering*. Se ha argumentado que lo primero se debe a la presencia de *hubs* en la red, mientras lo segundo a la modularidad de la misma [Ravasz *et al.*, 2002]. En consecuencia, para identificar los *hubs* se definió un punto de equilibrio en la $C(k)$ en el cual se cumple que $\frac{dC(k)}{dk} = -1$; la solución de esta ecuación produce el valor de conectividad κ donde se alcanza dicho equilibrio; ver figura 5-2b. De esta forma, los nodos con conectividad mayor que κ son *nodos jerárquicos*, mientras que aquellos con conectividad menor que κ son *nodos modulares*. La remoción acumulativa, en orden decreciente de conectividad, de todos los nodos jerárquicos y algunos modulares mostró que la red se desintegra al remover los nodos jerárquicos; ver figura 5-2c.

Los 15 *hubs* así identificados correlacionaron altamente con los factores transcripcionales globales conocidos [Martínez-Antonio y Collado-Vides, 2003; Browning y Busby, 2004] recuperándolos completamente y prediciendo dos nuevos: Fur y FlhDC; uno de los cuales, Fur, ha sido sugerido previamente como factor de transcripción global [Babu y Teichmann, 2003].

5.2.3. Identificación de módulos y genes intermodulares

La remoción de los factores de transcripción jerárquicos reveló la existencia de 61 módulos más un megamódulo. Un estudio previo realizado con la red de genes que sólo codifican factores de transcripción no había revelado la existencia de un megamódulo [Freyre-González *et al.*, 2005], por lo que el megamódulo debía interconectarse por la correulación de genes estructurales. En consecuencia se definió un gen intermodular como un gen estructural correulado por

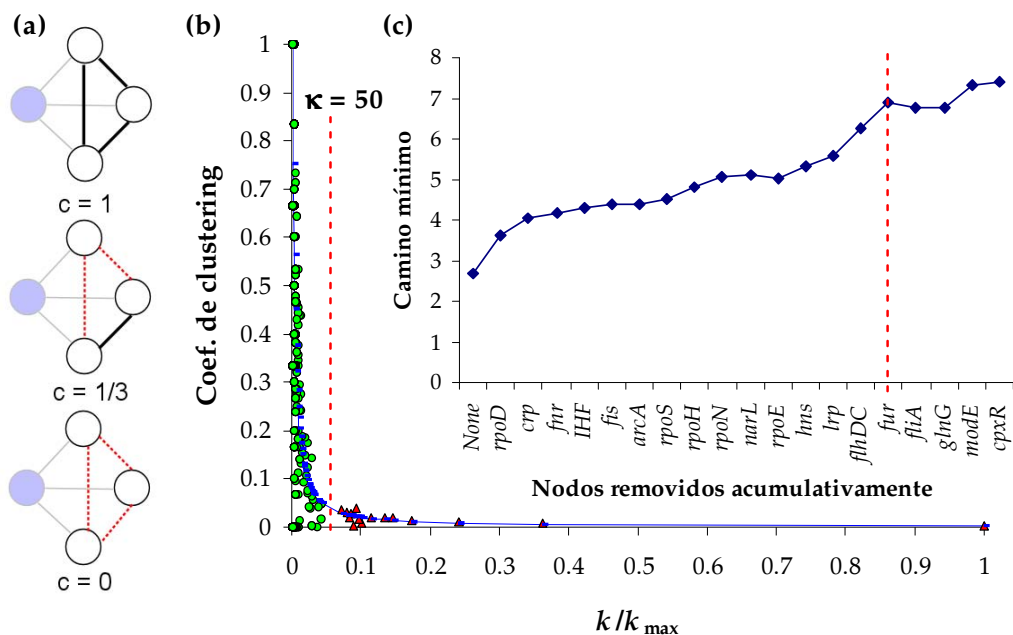


Figura 5-2: Identificación de nodos modulares y jerárquicos en la red de *E. coli*. (a) Ejemplo de coeficientes de *clustering* para el nodo i sombreado en azul. Las líneas negras son aristas existentes entre los vecinos de i , mientras las punteadas rojas son aristas potenciales. (b) Distribución del coeficiente de *clustering*, $C(k)$, y el valor κ calculado para la red. La curva azul representa la ley de potencia de la $C(k)$. La línea punteada roja indica el valor κ obtenido para esta distribución. Los triángulos rojos representan los nodos jerárquicos ($k > \kappa$), mientras que los círculos verdes indican los nodos modulares ($k < \kappa$). (c) Variación en la longitud del camino característico tras la remoción acumulativa de todos los nodos jerárquicos y algunos modulares. La línea punteada roja indica el cambio repentino en la tendencia creciente original cuando los últimos factores de transcripción jerárquicos (FlhDC y Fur) fueron removidos. Esto sugiere que la remoción de los nodos jerárquicos rompió las conexiones entre los módulos, desintegrando así la red de regulación. (Imagen tomada y adaptada de Freyre-González *et al.*, 2008.)

dos o más módulos. Estos genes integran a nivel promotor las señales fisiológicas provenientes de módulos distintos, actuando como un multiplexor molecular al permitir reutilizar ciertos componentes en más de una condición fisiológica o integrar señales fisiológicas distintas para lograr decisiones complejas. Así, se mostró que el megamódulo está compuesto por 39 submódulos interconectados por la coregulación de 136 genes intermodulares. Además, se identificaron 691 genes sólo regulados por factores transcripcionales jerárquicos; un análisis de sus funciones biológicas reveló una tendencia a ser genes que codifican la maquinaria basal de la célula, aunque sin descartar la posibilidad de que algunos de ellos estén fuera de módulos debido a la incompletez de la red.

5.2.4. Anotación manual y automatizada de los módulos identificados

Para identificar la función fisiológica de cada módulo cada gen se anotó con su clase funcional de acuerdo al sistema MultiFun [Serres *et al.*, 2004] y se realizaron dos análisis independientes. Por una parte, se hizo una curación y anotación manual de cada módulo empleando la información disponible en las bases de datos EcoCyc [Keseler *et al.*, 2005] y RegulonDB [Salgado *et al.*, 2006]. Por la otra, se realizó una anotación automatizada y ciega basada en la clase funcional que mostró un enriquecimiento estadísticamente significativo ($p\text{-value} < 0.05$). Mientras que ambos análisis correlacionaron altamente, el análisis manual añadió detalles finos que escaparon al automatizado debido a incompletez en la anotación del sistema MultiFun.

5.3. Reconstituyendo la red regulatoria de *E. coli*

5.3.1. Inferencia de la estructura jerárquica gobernando la red

Basado en las definiciones originales de Gottesman sobre factores de transcripción globales [Gottesman, 1984], y empleando el concepto de pleiotropía, se definió un enfoque ascendente para inferir la estructura jerárquica gobernando la red de regulación de *E. coli*, en el cual los nodos que pertenecen al mismo módulo se redujeron a un solo nodo. Este enfoque ubica cada factor de transcripción jerárquico en una capa específica dependiendo de dos factores:

1. Pleiotropía teórica, definida como el número de módulos y factores de transcripción jerárquicos regulados.

2. La presencia de regulación directa sobre factores transcripcionales jerárquicos ubicados en la capa inferior inmediata.

El segundo factor se toma en cuenta debido a que un factor de transcripción jerárquico puede propagar indirectamente su control a otros módulos mediante alterar el patrón de expresión de un segundo factor transcripcional jerárquico que los controle directamente. Dado que cada módulo está a cargo de una respuesta fisiológica distinta, este enfoque está basado en pleiotropía. La jerarquía obtenida con este enfoque muestra la presencia de cinco cadenas de mando globales, cada una de las cuales está a cargo de enviar señales de interés general para un gran número de genes en la célula; ver figura 5-3.

5.3.2. La médula jerárquica de la red es conformada por motivos *feedforward*

La remoción uno a uno, en orden decreciente de conectividad, de todos los factores transcripcionales jerárquicos mostró que el número de motivos *feedforward* decreció en 96.5%, implicando que la médula jerárquica de la red regulatoria se conforma por *feedforwards* los cuales interconectan módulos a través de los *hubs*.

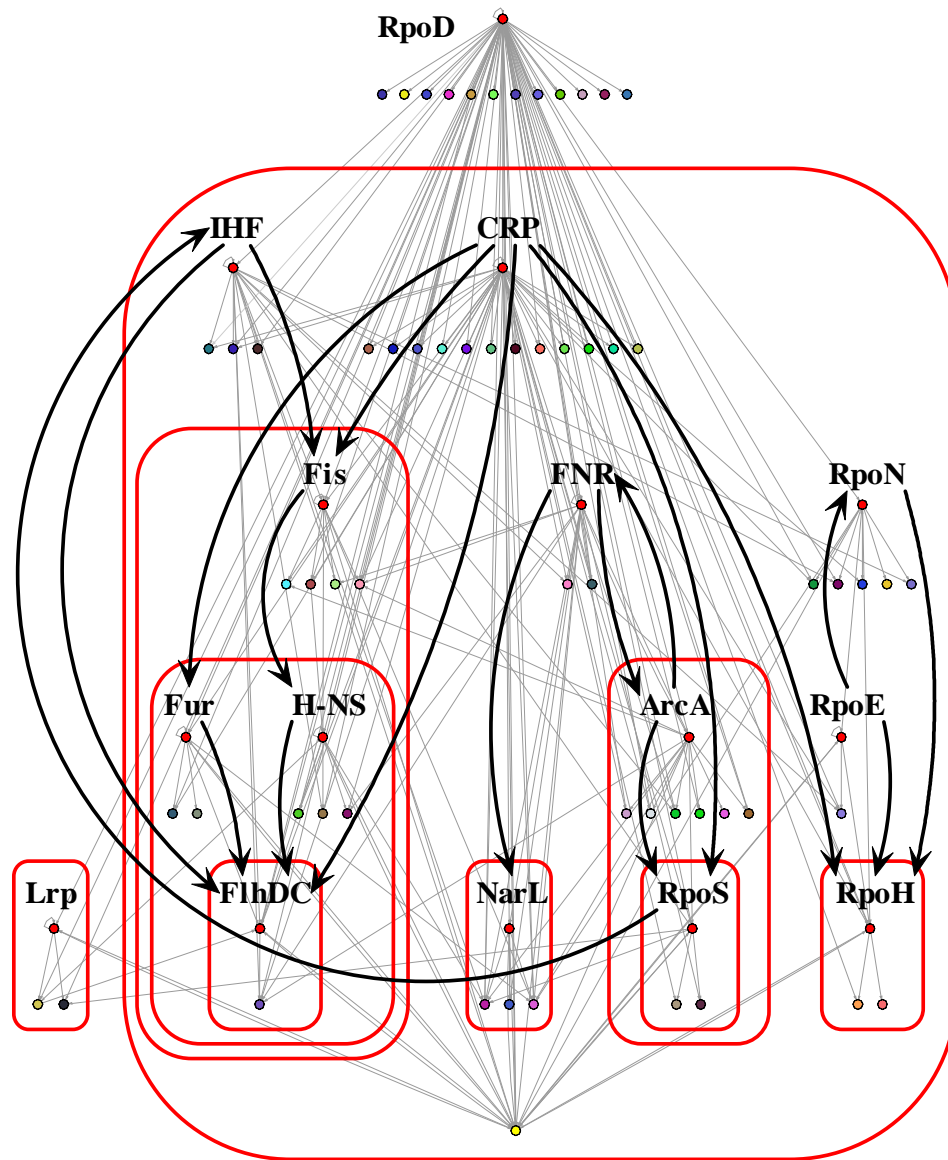


Figura 5-3: Mapa de la organización modular y jerárquica de las subrutinas que componen el programa génico de *E. coli*. Cada color representa un módulo, los factores de transcripción jerárquicos se muestran en rojo. Las flechas negras indican las interacciones regulatorias entre los factores de transcripción jerárquicos. En favor de la claridad, las interacciones de RpoD no se muestran y el megamódulo es mostrado como un solo nodo amarillo en la parte inferior. Los rectángulos redondeados rojos delimitan las capas jerárquicas. Se observa la presencia de cinco cadenas de mando globales: sensor huésped/vida libre y fimbria tipo 1 (Lrp); replicación, recombinación, pili y elementos extracitoplasmáticos (Fis, Fur, H-NS, FlhDC); formas de respiración (NarL); estrés por inanición (ArcA, RpoS); y estrés por choque térmico (RpoH). (Imagen tomada de Freyre-González *et al.*, 2008.)

Capítulo 6

Conclusiones y Perspectivas

*For what is a man, what has he got?
If not himself, then he has naught.
To say the things he truly feels,
and not the words of one who kneels.
The record shows I took the blows —
and did it my way!*
— PAUL ANKA, *My way* (1969)

Las bacterias requieren ser robustas para contender con su ambiente cambiante, implicando que deben ser capaces de monitorear el medio y responder de forma eficiente a las perturbaciones. Pero, ¿cómo se organizan las interacciones entre los componentes moleculares de la célula para realizar una eficiente toma de decisiones? A continuación, resumiremos las conclusiones y perspectivas de esta investigación.

6.1. Conclusiones

El método de descomposición natural desarrollado en este trabajo permitió diseccionar la red de regulación de *E. coli* identificando así sus componentes clave. Este estudio reveló que, a pesar de su aparente complejidad, la red de regulación exhibe una elegante organización no piramidal donde sólo un puñado de factores de transcripción globales coordinan a módulos causalmente independientes, cuyas respuestas se integran a nivel promotor vía los genes intermodulares. Así, este método permitió lograr una mayor comprensión de la arquitectura funcional de la red de regulación y generar un extenso conjunto de hipótesis validables experimentalmente. Las conclusiones de este trabajo pueden resumirse en los siguientes puntos:

- Contrario a lo reportado previamente [Ma *et al.*, 2004a,b], la red de regulación de *E. coli* contiene circuitos de retroalimentación involucrando diferentes capas jerárquicas; esto implica que la expresión de algunos factores de transcripción jerárquicos puede depender también de los factores de transcripción modulares, permitiendo así la reconfiguración de la maquinaria regulatoria en respuesta a cambios ambientales precisos detectados por los factores de transcripción modulares a través del alosterismo.
- El punto de equilibrio κ de la $C(k)$ mostró funcionar como un excelente predictor de los *hubs* al identificar como factores de transcripción jerárquicos a todos los factores transcripcionales globales conocidos [Martínez-Antonio y Collado-Vides, 2003; Browning y Busby, 2004] y predecir dos nuevos: Fur y FlhDC. Recientemente, un análisis de la red de regulación de *Bacillus subtilis* confirmó la capacidad predictiva de este método [Freyre-González *et al.*, 2007], ofreciendo así el primer posible criterio objetivo para identificar factores de transcripción globales en una célula.
- Los 15 factores de transcripción jerárquicos identificados gobiernan las respuestas de la célula, permitiendo a los módulos prepararse para reaccionar en respuesta a estímulos de interés general, mientras los módulos retienen su independencia para responder a estímulos de interés local.
- Se identificaron 691 genes sólo regulados por factores transcripcionales jerárquicos, los cuales principalmente codifican elementos que conforman la maquinaria basal de la célula: ADN y ARN polimerasas, varios tARN, aminoacil-tARN sintasas, proteínas ribosomales, varios ARN, enzimas del ciclo de ácidos tricarbóxicos y cadena respiratoria, enzimas para metilación del ADN, etc. Esto sugiere que la célula mantiene algunos elementos sólo bajo control global para ser capaz de reconfigurar de forma eficiente la maquinaria transcripcional bajo ciertas condiciones.
- Se encontraron 61 módulos más un megamódulo, éste último compuesto por 39 submódulos interconectados por la corregulación de los genes intermodulares, mostrando así la alta granularidad que el método desarrollado tiene para identificar módulos. Además, los análisis mostraron que el 97% de los módulos presentan un enriquecimiento funcional estadísticamente significativo.

- Se hallaron 136 genes intermodulares los cuales integran a nivel del promotor las señales fisiológicas provenientes de módulos distintos. Estos actúan como un multiplexor molecular, permitiendo reutilizar ciertos componentes en más de una condición fisiológica o integrar señales fisiológicas distintas para lograr decisiones complejas.
- La metodología desarrollada en esta investigación mostró que, a pesar de su aparente complejidad, la red de regulación de *E. coli* exhibe una singular elegancia en la organización de su programa génico, mostrando una organización jerárquica no piramidal donde sólo 15 factores de transcripción jerárquicos gobiernan a 100 módulos causalmente independientes, cuyas respuestas se integran a nivel de los promotores de los 136 genes intermodulares.
- Los motivos *feedforward* son la médula de la organización jerárquica de la red de regulación de *E. coli*, como lo evidenció, la remoción de todos los nodos jerárquicos y la consecuente disminución del número total de *feedforwards* en 96.5%. Esto mostró que los *feedforward* actúan como puentes que interconectan genes con diversas funciones fisiológicas, sugiriendo que son consecuencia de la organización de la red y no se encuentran involucrados en funciones fisiológicas particulares.

6.2. Perspectivas

Los resultados obtenidos en este estudio abren nuevas preguntas y planteamientos cuyas respuestas, eventualmente, contribuirán a ayudarnos a comprender mejor la organización de la toma de decisiones en bacterias, y cómo ésta ha evolucionado hasta convertirse en un sistema complejo con características bien definidas. Estas nuevas líneas de trabajo son:

- Evaluar la significancia estadística de la distribución de los circuitos de retroalimentación.
- Caracterizar experimentalmente el comportamiento dinámico de los circuitos de retroalimentación identificados en esta investigación.
- Analizar minuciosamente la biología de los módulos y los genes intermodulares identificados.

- Aplicar la metodología desarrollada en esta investigación a las redes de regulación de otros organismos: *Bacillus subtilis*, *Corynebacterium glutamicum*, *Saccharomyces cerevisiae*. ¿Qué tan universales son las propiedades identificadas en la organización de la red regulatoria de *E. coli*?
- Estudiar cómo la evolución puede moldear la organización de estas estructuras de control.
- Proponer un modelo dinámico que aproveche la partición de la red en los elementos clave descritos en esta investigación para así simplificar el estudio de la dinámica celular.

Apéndice A

Modular analysis of the transcriptional regulatory network of *E. coli*

Referencia:

RESENDIS-ANTONIO, O., FREYRE-GONZÁLEZ, J.A., MENCHACA-MÉNDEZ, R.,
GUTIÉRREZ-RÍOS, R.M., MARTÍNEZ-ANTONIO, A., AVILA-SÁNCHEZ, C., Y COLLADO-
VIDES, J. Modular analysis of the transcriptional regulatory network of *E. coli*.
Trends Genet **21**(1):16–20 (2005)

- 19 Kelly, D.P. and Scarpulla, R.C. (2004) Transcriptional regulatory circuits controlling mitochondrial biogenesis and function. *Genes Dev.* 18, 357–368
- 20 Li, R. *et al.* (1996) Expression of the human cytochrome c1 gene is controlled through multiple Sp1-binding sites and an initiator region. *Eur. J. Biochem.* 241, 649–656
- 21 Gulick, T. *et al.* (1994) The peroxisome proliferator-activated receptor regulates mitochondrial fatty acid oxidative enzyme gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 91, 11012–11016
- 22 Arbeitman, M.N. *et al.* (2002) Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297, 2270–2275
- 23 Jegga, A.G. *et al.* (2002) Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res.* 12, 1408–1417
- 24 Hertz, G.Z. and Stormo, G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15, 563–577
- 25 Pavesi, G. *et al.* (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 17(Suppl. 1), S207–S214
- 26 Pavesi, G. *et al.* (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* 32, W199–W203
- 27 Grillo, G. *et al.* (2003) PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.* 31, 3608–3612

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.11.009

Modular analysis of the transcriptional regulatory network of *E. coli*

Osbaldo Resendis-Antonio, Julio A. Freyre-González, Ricardo Menchaca-Méndez, Rosa M. Gutiérrez-Ríos, Agustino Martínez-Antonio, Cristhian Ávila-Sánchez and Julio Collado-Vides

Program of Computational Genomics, Nitrogen Fixation Research Center, Universidad Nacional Autónoma de México, Ave Universidad s/n, Col Chamilpa, Cuernavaca, Morelos 62100 México

The transcriptional network of *Escherichia coli* is currently the best-understood regulatory network of a single cell. Motivated by statistical evidence, suggesting a hierarchical modular architecture in this network, we identified eight modules with well-defined physiological functions. These modules were identified by a clustering approach, using the shortest path to trace regulatory relationships across genes in the network. We report the type (feed forward and bifan) and distribution of motifs between and within modules. Feed-forward motifs tend to be embedded within modules, whereas bi-fan motifs tend to link modules, supporting the notion of a hierarchical network with defined functional modules.

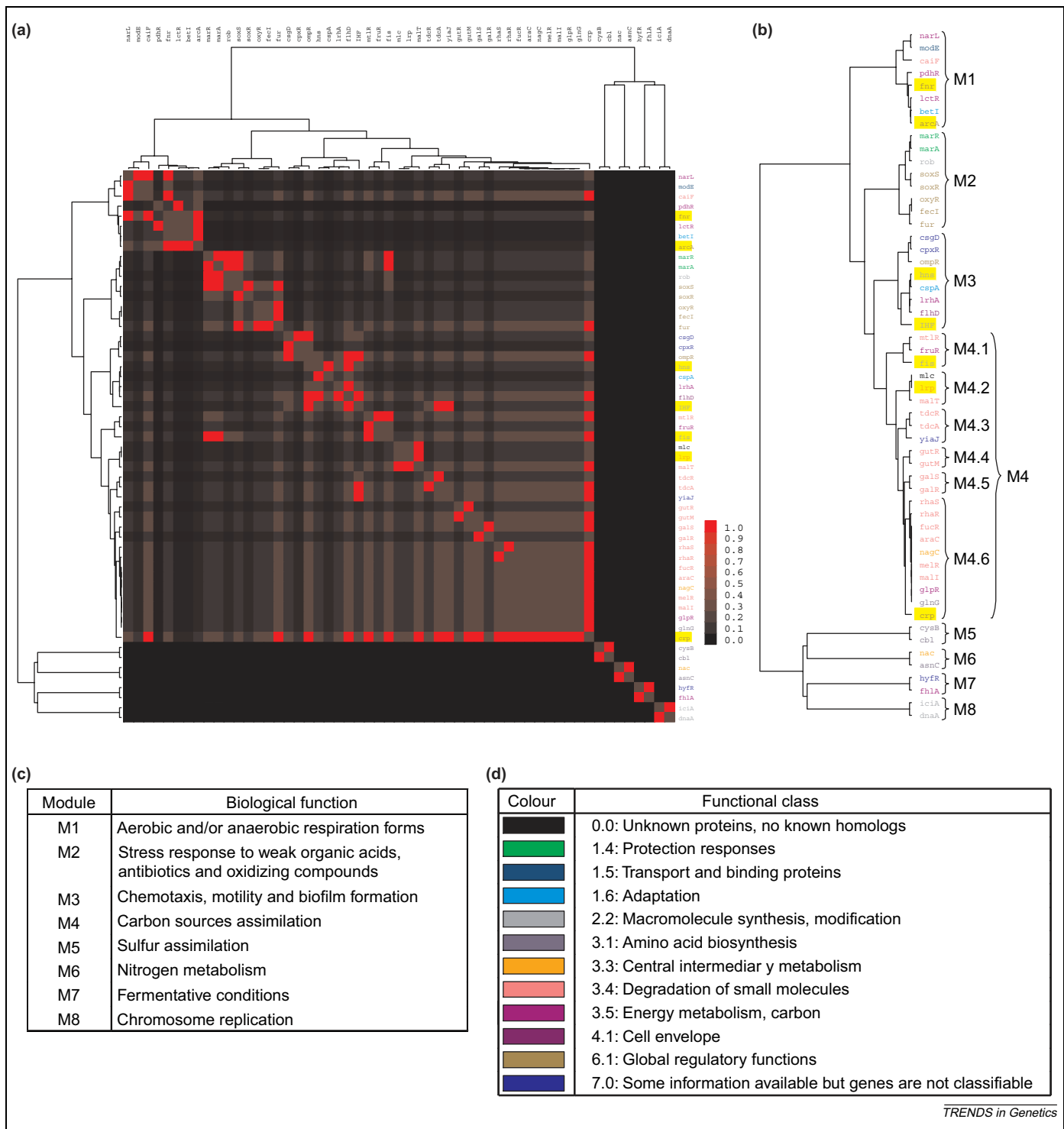
There is experimental evidence suggesting that, at certain times, different fractions of the complete set of transcriptional factors (TFs) are used depending on the growth conditions within the cell [1,2]. The global topological analysis of transcriptional networks supports the notion of their organization into modules or large groups of genes that, in many cases, respond to external or internal stimuli [2,3]. A network with a scale-free property is expected to be robust to failure of individual components, and could contribute to the ability of the cell to respond to changes in environmental and evolutionary conditions [4–9]. At a finer topological scale, over-represented topological units called network motifs, consisting of three or four genes, contribute to the local dynamic behavior of transcriptional regulation [10,11]. Despite studies that are

focused on motifs and modules, little is known about how these motifs can integrate to construct modular structures and whether their components correspond to subsets of genes that regulate integrated cellular responses to external conditions.

It was recently reported that the transcriptional regulatory network of *Escherichia coli* is a scale-free network consisting of a hierarchy of modules [12]. In this article, we describe our reconstruction of eight modules (Figure 1a–d) with clearly defined physiological functions. In addition, using the fraction of feed forward (FF) and bi-fan (BF) motifs [11] that are shared between two modules, we quantified the overlap between them (Figure 2a–f). We found that the largest module that is involved in carbon sources has the greatest number of connection via motifs with other modules.

We analyzed the network of known transcriptional interactions of *E. coli* K-12 that were organized in RegulonDB [13] (http://www.cifn.unam.mx/Computational_Genomics/regulondb/). Neglecting genes without experimental evidence, indicating that they encode TFs (supplementary material online), and ignoring autoregulation, the total number of TFs that regulate the expression of other TFs is 55, all of which control the expression of 747 genes. Based on the number of genes in the genome, and the total number of estimated TFs (~320), we estimate that this fraction represents ~18% of the transcriptional network in *E. coli*. In our graphical representation, vertices represent genes and the transcriptional interactions between them are represented by edges.

Corresponding author: Collado-Vides, J. (collado@cifn.unam.mx).
Available online 25 November 2004



TRENDS in Genetics

Figure 1. Identifying modules in *Escherichia coli*. **(a)** Clustering analysis where the intensity of each spot represents the shortest path length between pairs of genes. The intensity value can be 1.0 or 0.0, where a value of 1.0 means that two genes have a direct connection and a value of 0.0 means that one pair of genes is not connected in the network. Global regulators described in Ref. [16] are highlighted in yellow (Box 2). **(b)** This hierarchical structure shows the sets of genes that conform each module (denoted by M). **(c)** The biological function for each of the eight modules. **(d)** The color code used to annotate each gene with its corresponding functional class [15].

Identifying modules

The clustering analysis was performed using only TF-encoding genes. We obtained the shortest path length between every pair of genes (d_{ij} is the shortest path length between gene i and gene j). Next, we calculated the association function ($1/d_{ij}^2$) of these shortest path lengths [14]. This function gives a measure of the closeness among genes, amplifying close relationships and

minimizing remote distances. The data were used as input to perform a hierarchical agglomerative average-linkage clustering [2] (Figure 2a and supplementary material online). Modules identified by this algorithm and the functional classes of the genes that compose them are shown in Figure 1.

An analysis of sensitivity and specificity for each set of genes inside a module indicates that topological modules

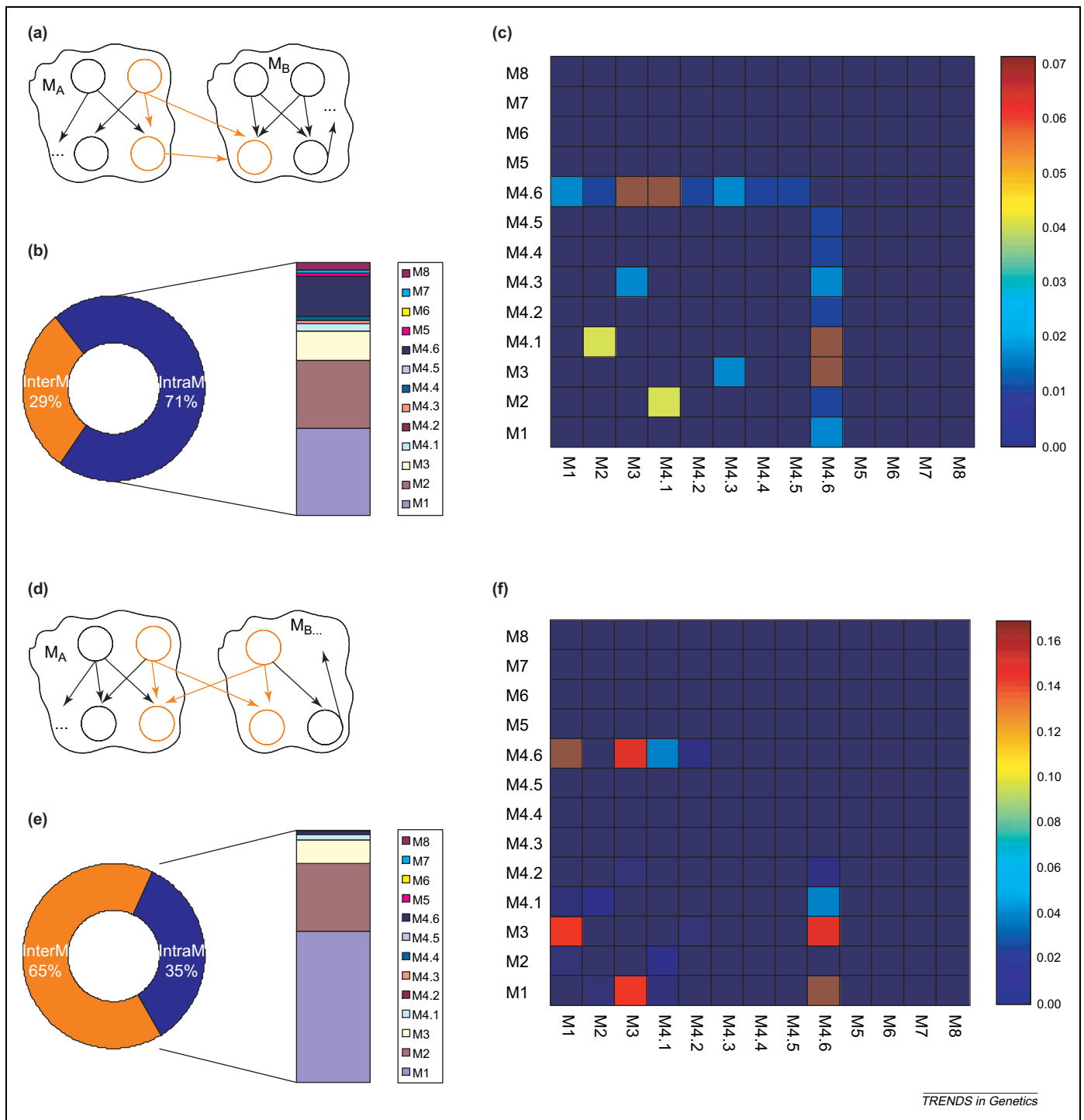


Figure 2. The relationship between modules and motifs. **(a)** An example of a feed forward (FF) motif formed by the interaction of two different modules M_A and M_B . **(b)** The percentage of FF motifs formed by the interaction of pairs of modules (InterM) and inside (IntraM) the modules previously identified (M1–M8). The bar shows the distribution of FF in each module. **(c)** A matrix representation of the percentage of FF motifs formed by the interaction of every specific pair of modules. **(d)** An example of a bi-fan (BF) motif formed by the interaction of two different modules M_A and M_B . **(e)** and **(f)** show the same analysis as **(b)** and **(c)** respectively, but focus on the bi-fan motif. The network has a total of 325 FF motifs and 3918 BF motifs.

are composed of functionally related genes (Box 1 and Table 5 in the supplementary material online). To perform this analysis, we defined a control set using the available information in RegulonDB about the expression of the TFs and their regulated genes in a particular condition. In addition, we used information about the functional classes, described by Riley [15], as stored in RegulonDB. Then, we tested how well the modules, shown

in Figure 1a, were integrated by functionally correlated genes (Box 1).

Module 1 (M1) contains genes involved in respiration, M2 contains genes involved in stress response and M3 contains genes involved in chemotaxis, motility and biofilm formation. The module with the greatest number of TFs (23) is involved in the regulation of the different carbon sources (M4). This large module is devoted to

Box 1. Specificity and sensitivity analysis of modules

To quantify how clustered genes and modules are functional correlated, we performed a sensitivity and specificity analysis. First, we identified a set of functionally correlated groups using information from RegulonDB [13]. This information was used as the control set to test how well the hierarchical clustering method grouped genes into functionally correlated groups. Then, we perform the sensitivity and specificity analysis for each module.

Sensitivity was defined as:

$$\text{Sensitivity} = \frac{\text{pres}^+}{\text{pres}^+ + \text{pres}^-}$$

Specificity was defined as:

$$\text{Specificity} = \frac{\text{abs}^-}{\text{abs}^- + \text{abs}^+}$$

where:

pres^+ is the number of genes that belong to the module and that were clustered into that module (true positives).

pres^- is the number of genes that belong to the module but were not clustered into the module (false negatives).

abs^+ is the number of genes that do not belong to the module but were clustered into the module (false positives).

abs^- is the number of genes that do not belong to the module, and that were not clustered into that module (true negatives).

the use of preferential carbon sources and can be divided into smaller sub-modules related to alternative carbon sources. Furthermore, four smaller modules (M5, M6, M7 and M8) of interacting TFs are fully disconnected from the largest connected network component. These involve genes regulating relevant cellular responses such as sulfur assimilation, metabolism of nitrogen sources, fermentative metabolism and chromosome replication.

Global TFs (Box 2) [16] are evenly distributed within major modules. Cyclic AMP receptor protein (CRP) is positioned clearly within the module of carbon metabolism, together with factor for inversion stimulation (FIS) and leucine-responsive regulatory protein (Lrp). Aerobic respiration regulatory protein (ArcA) and Fumarate and nitrate reductase regulatory protein (FNR) belong to the respiration-response module. Histone-like protein or nucleoid-associated protein (Hns) and integration host factor (IHF) belong to the module involved in chemotaxis, motility and biofilm formation (Figure 1b,c).

Relationship between modules and regulons

We distinguish between simple regulons (genes that are regulated by only one TF) and complex regulons (genes that are regulated by the same set of two or more TFs) as defined in Ref. [17]. We define the degree of participation as the largest fraction of TFs and regulated genes of a regulon that belong to one module, comparing both strict complex regulons (SCR) and the non-strict complex

regulons (NSCR), where strict means that a regulator has the same effect (either positive or negative) on every gene in the regulon. The SCRs and NSCRs have an average degree of participation of 0.29 ± 0.16 and 0.24 ± 0.16 , respectively (supplementary material online). On average, complex regulons involve 2.42 different TFs, implying that the probability of each TF of a SCR belonging to a different module is almost 0.75. In this sense regulons are multi-functional. The interesting conclusion is that a single promoter subject to regulation by several proteins, can function as an integrator of TFs that belong to different modules, suggesting that the promoter senses the status of different physiologically relevant cellular entities or modules.

The interrelationship between motifs and modules

Network motifs are statistically significant interconnection patterns with three-to-four genes that characterize the local topology of the regulatory network [11]. It has been shown that these network motifs were not formed by gene duplication but were formed as a result of convergent evolution [18]. When analyzing their distribution within the larger modules, we found that 71% of the FF motifs are inside a single module, whereas this is true for only 35% of the BF motifs (Figure 2b,e, and supplementary material online). This is not surprising because BFs are complex regulons, regulating TFs that belong to distinct modules. As described in the supplementary material online, we identified the fraction of motifs that are shared between every pair of modules (Figure 2c,f).

Module M4.6 has the highest degree of overlap with other modules, via BF motifs mainly involving CRP. It also has a high degree of overlap with modules M1 (related to respiration) and M3 (related with chemotaxis, motility and biofilm formation). The connection between M4.6 and M1 makes sense because carbon metabolism is closely related to respiration through the tricarboxylic acids cycle. Modules M4.6 and M3 are highly related, being both involved in nutritional mechanisms. However, the same modules have a lower degree of overlap when they are related via FF motifs. Respiration (M1) is connected with the module of chemotaxis, motility and biofilm formation (M3) via BF motifs (Figure 2f). Certainly, *E. coli* is a facultative, anaerobic organism that uses its motility capability to change environments, seeking optimal oxygen concentrations, a mechanism known as aerotaxis [19]. The genes involved in this mechanism have been identified in *E. coli* [20,21]. These specific cases indicate that the BF motif is the main motif connecting modules, whereas FF motifs are mainly located inside modules.

Concluding remarks

This study was initially motivated by the mathematical suggestion, based on the scale-free distribution of the clustering coefficient, that transcriptional regulation is a hierarchical modular network [12]. We wanted to provide biological content and, at the same time, enhance our detailed understanding of the internal organization of this network. Physiological functions are usually performed by the activity of several proteins. Depending on the particular conditions and cellular responses, genes function coordinately to perform a given task.

Box 2. Definition of global regulators

We used an operational definition of 'global' TFs based on a collection of diagnostic criteria: (i) they regulate many genes; (ii) they regulate several genes encoding TFs; (iii) they cooperate with numerous TFs and together regulate other genes; (iv) they directly affect the expression of a variety of promoters that use different sigma factors; and (v) their regulated genes belong to different functional classes. The analysis supporting these criteria is described in detail in Ref. [16].

We know that genes in bacteria are organized in operons and regulons. More recently, it has been shown that small sets of genes form motifs that convey specific dynamical advantages to cellular responses and coordinated expression [22,23]. We show here the interplay of such motifs with larger groups of physiologically related genes or modules that reflect a higher level of organization of gene content of the cell. The hierarchical organization is demonstrated by FF motifs being confined to individual modules that are devoted to particular cellular processes. Larger modules are not strictly separated because a significant number of BF motifs overlap and, thus, interconnect. This interconnection is also reflected by complex regulons, which are mostly regulated by TFs that belong to different modules. Thus, we have begun to understand in detail the precise structure of the biological complexity of the regulatory network of a single cell.

It came as a surprise that blind computational processing of undirected shortest path distances between regulatory genes (local topological properties of the network) permitted a highly accurate reconstruction of groups of functionally related genes. This is an interesting result validated in *E. coli* and has potentially useful applications in describing the regulatory networks of others organisms.

Acknowledgements

J.F-G. is supported by Ph.D. fellowship 176341 from CONACyT-Mexico. This work is supported by grant GM 62205-03 from NIH to J.C-V. We thank Verónica Jiménez for help extracting the dataset from RegulonDB. We also acknowledge Fabiola Sánchez for her computing support. We acknowledge the Editor for helping to improve this article.

Supplementary data

Supplementary data associated with this article can be found at [doi:10.1016/j.tig.2004.11.010](https://doi.org/10.1016/j.tig.2004.11.010)

References

- 1 Segal, E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176
- 2 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
- 3 Hartwell, H. *et al.* (1999) From molecular to modular cell biology. *Nature* 402, C47–C52
- 4 Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* 286, 509–512
- 5 Ravasz, E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555
- 6 Albert, R. and Barabasi, A.L. (2002) Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97
- 7 Ravasz, E. and Barabasi, A.L. (2003) Hierarchical organization in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 67, 026112(<http://link.aps.org/abstract/PRE/v67/e026112>)
- 8 Spiro, S. and Guest, J.R. (1991) Adaptive responses to oxygen limitation in *Escherichia coli*. *Trends Biochem. Sci.* 16, 310–314
- 9 Babu, M. and Teichmann, S.A. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* 31, 1234–1244
- 10 Shen-Orr, S.S. *et al.* (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68
- 11 Milo, R. *et al.* (2002) Network motifs: simple building blocks of complex networks. *Science* 298, 824–827
- 12 Dobrin, R. *et al.* (2004) Aggregation of topological motifs in the *Escherichia coli* regulatory network. *BMC Bioinformatics* 5, 10
- 13 Salgado, H. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* 32, D303–D306
- 14 Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100, 1128–1133
- 15 Riley, M. (1997) Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucleic Acids Res.* 25, 51–52
- 16 Martínez-Antonio, A. and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* 6, 482–489
- 17 Gutiérrez-Ríos, R.M. *et al.* (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.* 13, 2435–2443
- 18 Conant, G.C. and Wagner, A. (2003) Convergent evolution of gene circuits. *Nat. Genet.* 34, 264–266
- 19 Adler, J. (1966) Effect of amino acids and oxygen on chemotaxis in *Escherichia coli*. *J. Bacteriol.* 92, 121–129
- 20 Rebbapragada, A. *et al.* (1997) The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior. *Proc. Natl. Acad. Sci. U. S. A.* 94, 10541–10546
- 21 Bibikov, S.I. *et al.* (1997) A signal transducer for aerotaxis in *Escherichia coli*. *J. Bacteriol.* 179, 4075–4079
- 22 Rosenfeld, N. and Alon, U. (2003) Response delays and the structure of transcription networks. *J. Mol. Biol.* 329, 645–654
- 23 Mangan, S. *et al.* (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.* 334, 197–204

0168-9525/\$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.11.010

Apéndice B

Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*

Referencia:

GUTIERREZ-RÍOS, R.M., FREYRE-GONZÁLEZ, J.A., RESENDIS, O., COLLADO-VIDES, J., SAIER, M., Y GOSSET, G. Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. *BMC Microbiol* **7**:53 (2007)

Research article

Open Access

Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*

Rosa María Gutierrez-Ríos¹, Julio A Freyre-Gonzalez³, Osbaldo Resendis³, Julio Collado-Vides³, Milton Saier⁴ and Guillermo Gosset*²

Address: ¹Departamentos de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Apdo. Postal 510-3, Cuernavaca, Morelos 62250, México, ²Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Apdo. Postal 510-3, Cuernavaca, Morelos 62250, México, ³Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México and ⁴Division of Biological Sciences, University of California at San Diego, La Jolla, CA, 92093-0116, USA

Email: Rosa María Gutierrez-Ríos - rmaria@ibt.unam.mx; Julio A Freyre-Gonzalez - jfreyre@ccg.unam.mx; Osbaldo Resendis - resendis@ccg.unam.mx; Julio Collado-Vides - collado@ccg.unam.mx; Milton Saier - saier@biomail.ucsd.edu; Guillermo Gosset* - gosset@ibt.unam.mx

* Corresponding author

Published: 8 June 2007

Received: 15 January 2007

BMC Microbiology 2007, **7**:53 doi:10.1186/1471-2180-7-53

Accepted: 8 June 2007

This article is available from: <http://www.biomedcentral.com/1471-2180/7/53>

© 2007 Gutierrez-Ríos et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Glucose is the preferred carbon and energy source for *Escherichia coli*. A complex regulatory network coordinates gene expression, transport and enzyme activities in response to the presence of this sugar. To determine the extent of the cellular response to glucose, we applied an approach combining global transcriptome and regulatory network analyses.

Results: Transcriptome data from isogenic wild type and *crp*⁻ strains grown in Luria-Bertani medium (LB) or LB + 4 g/L glucose (LB+G) were analyzed to identify differentially transcribed genes. We detected 180 and 200 genes displaying increased and reduced relative transcript levels in the presence of glucose, respectively. The observed expression pattern in LB was consistent with a gluconeogenic metabolic state including active transport and interconversion of small molecules and macromolecules, induction of protease-encoding genes and a partial heat shock response. In LB+G, catabolic repression was detected for transport and metabolic interconversion activities. We also detected an increased capacity for *de novo* synthesis of nucleotides, amino acids and proteins. Cluster analysis of a subset of genes revealed that CRP mediates catabolite repression for most of the genes displaying reduced transcript levels in LB+G, whereas Fis participates in the upregulation of genes under this condition. An analysis of the regulatory network, in terms of topological functional units, revealed 8 interconnected modules which again exposed the importance of Fis and CRP as directly responsible for the coordinated response of the cell. This effect was also seen with other not extensively connected transcription factors such as FruR and PdhR, which showed a consistent response considering media composition.

Conclusion: This work allowed the identification of eight interconnected regulatory network modules that includes CRP, Fis and other transcriptional factors that respond directly or indirectly to the presence of glucose. In most cases, each of these modules includes genes encoding physiologically related functions, thus indicating a connection between regulatory network topology and related cellular functions involved in nutrient sensing and metabolism.

Background

In their natural environments, bacteria must adapt to changing physicochemical conditions. Adaptation responses are controlled by a complex network of sensory and regulatory proteins that modulate cellular functions at the transcriptional and posttranscriptional levels. Nutrient availability, ranging from sufficiency to total deprivation, is one of the environmental variables the cell is constantly sensing. Among nutrients, carbohydrates are particularly important to the cell since they are utilized as both carbon and energy sources. Glucose is the most abundant aldose in nature, being present mostly in polymeric states as starch and cellulose [1]. This sugar is the preferred carbon and energy source for the gram-negative bacterium *Escherichia coli* (*E. coli*) [2]. Specialized protein systems are present in *E. coli* to sense, select and transport glucose. This sugar is internalized and phosphorylated by the phosphoenolpyruvate:sugar phosphotransferase system (PTS). This system catalyzes group translocation, a process that couples transport of sugars to their phosphorylation. The PTS is widespread in bacteria but absent in Archaea and eukaryotic organisms [3,4]. It is composed of soluble non sugar-specific protein components, Enzyme I (EI) and the phosphohistidine carrier protein (HPr) which relay a phosphoryl group from the glycolytic intermediate, phosphoenolpyruvate (PEP), to any of the different sugar-specific enzyme II complexes. Glucose is imported by the IIGlc complex, composed of the soluble IIA^{Glc} enzyme and the integral membrane permease IICB^{Glc} [5].

The preferred nutritional status of glucose for *E. coli* is evidenced by the observed repression and inhibition exerted by this sugar on gene expression and the activities of enzymes and transporters related to the consumption of other carbon sources. This example of global regulation is called carbon catabolite repression (CCR) [2]. As a sensor of the presence of glucose in the external medium, the PTS plays a central role in CCR. When glucose is present in the medium and it is being transported by the PTS, the IIA^{Glc} protein is non-phosphorylated, and in this state, it binds to various non-PTS permeases inhibiting uptake of other carbon sources. This form of IIA^{Glc} also binds to the enzyme glycerol kinase (GK), inhibiting its activity. When glucose is absent from the culture medium, IIA^{Glc} is mainly in its phosphorylated state. In this condition, IIA^{Glc}~P binds to the enzyme adenylate cyclase (AC), activating its cyclic AMP (cAMP) biosynthetic capacity. Therefore, cAMP concentrations increase in the cell. Then cAMP binds to the cAMP receptor protein (CRP) and promotes the induction of catabolite-repressed genes [2].

The global transcriptional response of *E. coli* to different nutrient/environmental conditions has been studied using microarray technology. These studies have revealed

complex genome-wide expression patterns that reflect the roles of different cellular regulators on cell adaptability and survival. Some of these works have focused on analyzing the effects on global transcription patterns of growing *E. coli* in minimal or complex media with different glucose concentrations [6-9]. These studies have enabled the identification of genes whose transcript levels change in response to each specific condition. In order to characterize the cellular response to glucose, conditions must be chosen that represent sufficiency and the complete lack of this nutrient. A comparison of genome-wide transcriptome patterns between strains grown under these conditions should be adequate for identifying the group of genes displaying a transcriptional response to glucose which we term, the "glucose stimulon". In this work, we use transcriptome data, collected under conditions of glucose absence or excess in a complex medium. Analyses of the data set were used to identify the genes encoding cellular functions that respond to this stimulus and enable the cell to adapt to nutrient availability. Topological analysis of the regulatory network involved in this response revealed modular organization where global and local transcriptional factors integrate different signals to detect and respond to the presence of glucose.

Results and Discussion

Global transcriptome response to the presence of glucose in complex medium

Transcriptome data was obtained from previously reported experiments performed with *E. coli* strain BW25113 and an isogenic *crp* mutant (LJ3017) [10]. These strains were grown in LB medium with (LB+G) or without (LB) 0.4% glucose. Total RNA was extracted from each condition, processed and hybridized to the Affymetrix *E. coli* array which includes 4327 genes [11]. Three data sets were obtained for each of three experimental conditions: wild type grown in LB medium (WT), wild type grown in LB medium + glucose (WTg) and a *crp* mutant grown in LB medium (CRP). Starting with these data, differentially transcribed genes were selected using an outlier iteration method [12-14]. Analysis of the data from the WTg/WT log ratios, allowed the identification of genes having a significant change in transcript level (Table 1) [15]. 180 genes showed increased and 200 reduced relative transcript levels. Of these 380 genes, 87 belong to the hypothetical, unknown class.

Figure 1 shows an integrated view of the transcriptional responses and their roles in cell physiology of the 293 genes having a known function and found to respond to glucose. As can be seen, the presence of glucose induced the expression of genes encoding the general PTS protein Hpr and the glucose-specific IICB^{Glc} permease. This response is expected to increase cellular glucose transport capacity [16]. Glucose-dependent induction was also

detected for genes encoding proteins involved in the import of polyamines, inorganic phosphate and magnesium ions, thus suggesting that these nutrients are required to sustain the higher growth rate observed in the LB+G medium. In contrast, glucose had a repressive effect on genes encoding transporters and periplasmic receptor proteins related to the import of alternative carbon and carbon-nitrogen sources. These included: amino acids, carbohydrates, lactate, glycerol, peptides, dipeptides and nucleosides. Furthermore, a reduction in transcript levels was observed for genes encoding proteins involved in the catabolism of several sugars and amino acids. This transcriptome pattern is the expected result of carbon catabolite repression exerted by glucose [2].

The presence of glucose had a significant effect on transcript levels of genes encoding enzymes of central metabolism. Upregulation with glucose was detected for the genes encoding the E1 and the lipoate acetyltransferase/dihydrolipoamide acetyltransferase subunits of the pyruvate dehydrogenase multienzyme complex (Pdh) as well as the genes encoding phosphotransacetylase and acetate kinase, that constitute an acetate synthesis pathway. On the other hand, downregulation was observed for genes encoding nearly all enzymes involved in gluconeogenesis, the TCA cycle and the glyoxylate bypass [17]. The observed responses are consistent with the expected glycolytic metabolism induced by exogenous glucose.

Functions related to nucleotide biosynthesis and salvage pathways of purines and pyrimidines were found to change in response to glucose. Growth in LB+G medium reduced transcript levels of genes encoding proteins involved in (deoxy)ribose phosphate degradation, as well as the salvage pathways for both adenine, hypoxanthine, and their nucleosides and pyrimidine ribonucleotides and pyrimidine deoxyribonucleotides. By contrast, transcript levels for genes encoding enzymes that participate in the *de novo* biosynthesis of purine and pyrimidine ribonucleotides were increased. These results suggest that the cell exists in a metabolic state where it is importing and interconverting ribo and deoxyribonucleotides present in the LB medium, but addition of glucose induces another state where *de novo* synthesis capacity is increased.

For genes encoding enzymes of amino acid metabolism, different effects of glucose were observed. Downregulation in LB+G medium relative to LB medium was detected for genes involved in biosynthetic pathways for aromatic amino acids, aspartate, cysteine, isoleucine-valine, phenylalanine and threonine. Interestingly, downregulation was also observed for genes encoding activities involved in the degradation of aspartate, cysteine, glycine and threonine. In addition, as mentioned above, a decrease in transcript level was detected for genes encoding importers

for alanine, glutamine, glycine, histidine, proline and serine. These results indicate a reduction in import and degradation capacity for several amino acids when growing in LB+G medium. This can be explained considering that in this condition amino acids utilization as carbon sources should be significantly reduced. The apparent reduction in the demand for external amino acids to be used as alternative carbon sources or building blocks could also be a consequence of increased capacity for the *de novo* synthesis of amino acids once glucose is available. However, as noted above, induction of genes of amino acid synthesis pathways was never observed, and, in fact, repression was observed for several pathways. Therefore, the effects of glucose on degradative and biosynthetic capacities do not seem to be global but amino acid-specific.

A general trend of upregulation in LB medium was detected for genes encoding proteases, indicating higher proteolytic activity under this condition when compared to growth in LB+G medium. This makes sense since peptide degradation and protein turnover can provide carbon and energy for biosynthetic purposes in the absence of glucose. A similar pattern was observed for heat shock proteins and chaperones. These results suggest a higher protein turnover rate in the absence of glucose. The possible presence of partially degraded or misfolded proteins when the cells are growing in LB medium could cause the induction of heat shock proteins and chaperones, as has been previously reported [18,19]. It should be pointed out that several of the induced proteases are involved in regulatory processes (Fig. 1). The overall regulatory effects of such a response remain to be determined. A decrease in transcript level for heat shock proteins and chaperones upon growth in LB+G medium indicates that functions related to protein turnover are reduced by the presence of glucose, suggesting a lower capacity or need to use of amino acids derived from proteins as sources of carbon or protein constituents, or that proteins are more stable in an energized cell.

Medium composition had an important effect on genes encoding proteins involved in translation. Increased transcript levels were observed for genes encoding 20 of the 30 ribosomal protein components of the 50S subunit and 16 of the 22 ribosomal proteins of the 30S subunit. Also increased were transcript levels for 46 tRNA genes, grouped in 14 transcriptional units (TUs). Two of these TUs include genes *rrnA* and *rrnD*, encoding two of the seven 16S ribosomal RNAs. These genes are known to be subject to growth rate-dependent regulation [20]. In cultures used to obtain the RNA to generate the transcriptome data, a 5% higher growth rate was observed when comparing LB+G to LB conditions [10,21]. Therefore, induction of genes encoding ribosomal proteins, tRNAs

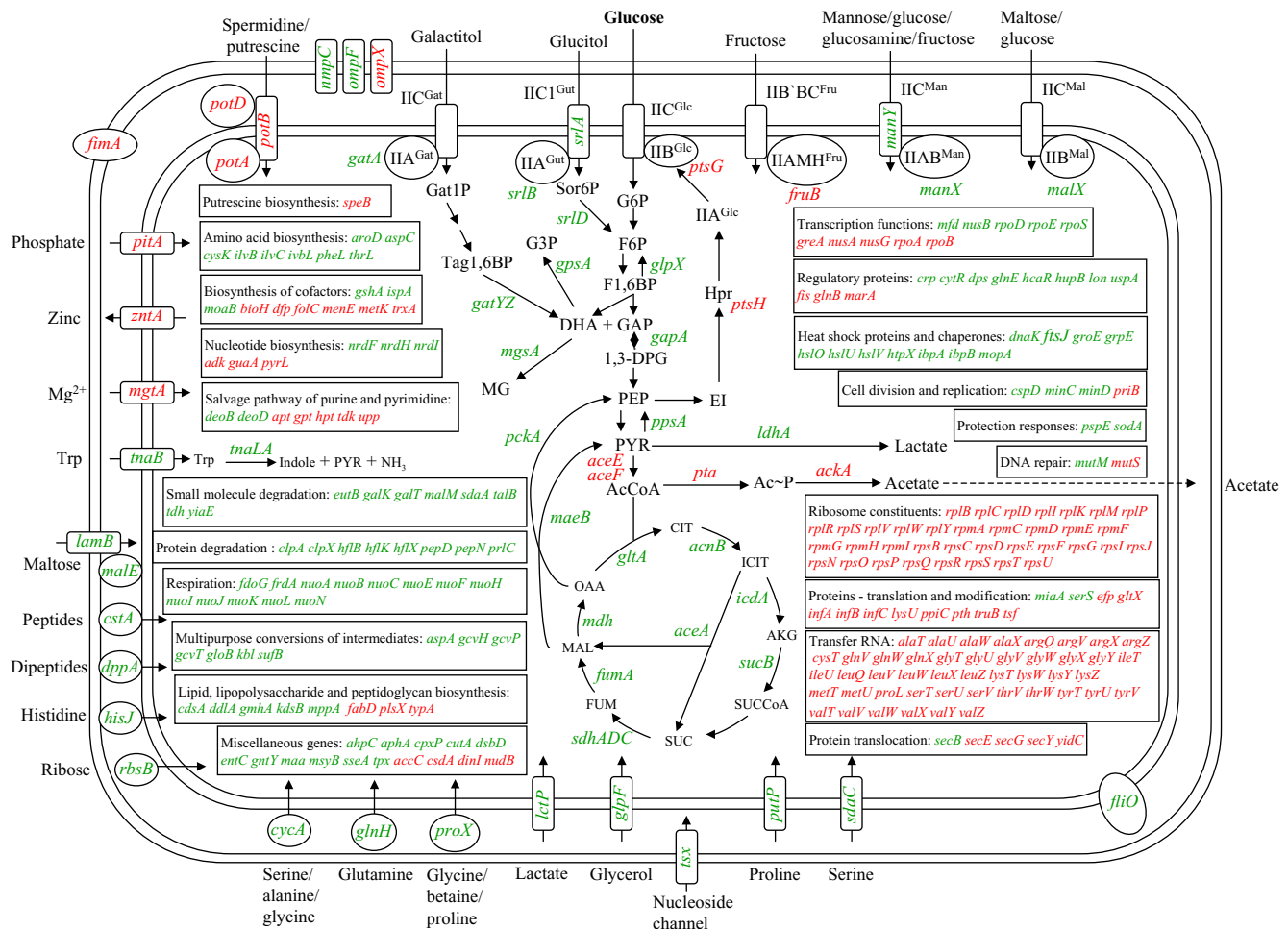


Figure 1
Transcriptome profile of *E. coli* comparing growth in LB+G to that in LB. Genes displaying higher and lower transcript levels due to the presence of glucose are shown in red and green, respectively. Abbreviations: 1,3-DPG, 1,3-diphosphateglycerate; AcCoA, acetyl coenzyme-A; Ac~P, acetyl phosphate; AKG, α -ketoglutarate; CIT, citrate; DHA, dihydroxy-acetone-phosphate; F1,6BP, fructose-1,6-bisphosphate; F6P, fructose-6-phosphate; FUM, fumarate; G3P, glycerol-3-phosphate; G6P, glucose-6-phosphate; GAP, glyceraldehyde-3-phosphate; GAT1P, galactitol-1-phosphate; ICIT, isocitrate; MAL, malate; MG, methylglyoxal; OAA, oxaloacetate; PEP, 3-phosphoenolpyruvate; PYR, pyruvate; Sor6P, sorbitol-6-phosphate; SUC, succinate; SUCCoA, succinyl-CoA; Tag1,6BP, tagatose-1,6-bisphosphate; Trp, L-tryptophan.

and rRNAs is an expected response to the higher growth rate in LB+G medium.

Cell division and replication functions were found to respond to medium composition. Glucose lowered transcript levels for genes encoding DNA replication inhibitor protein CspD and the cell division inhibitor and membrane ATPase MinCD of the MinC-MinD-MinE and DicB-MinC systems. The *cspD* gene is known to be induced upon glucose starvation [22,23]. An increase was observed in transcript level for the gene encoding PriB protein that is a component of the multiprotein complex called primosome. This complex is believed to be

involved in the restart of stalled DNA replication forks. The concerted down regulation of inhibiting and up regulation of activating chromosomal replication and cell division functions is consistent with a cellular response to favorable growth conditions afforded by the presence of glucose.

We found several transcription-related functions to be induced by glucose, these included the α and β subunits of the RNA polymerase core enzyme, as well as the elongation and antitermination factors GreA, NusA and NusG. Under the same growth condition, repression was observed for genes encoding the transcriptional termina-

tion factors NusB and Mfd. Thus, the observed responses for these functions are consistent with an expected increase in the transcriptional rate and efficiency caused by the increased biosynthetic demand of the higher growth rate in the presence of glucose. However, we also detected a reduction in transcript levels for genes encoding sigma 70, sigma E and sigma 38. It remains to be determined what the net consequences on the transcriptional capacity and RNA polymerase promoter selectivity would be, resulting from the observed expression changes.

Increased transcript levels were detected for the gene encoding agmatine ureohydrolase (*speB*), an enzyme involved in the putrescine biosynthetic pathway. Genes encoding the integral membrane component of the flagellar export apparatus FliO (*fliO*) displayed a decrease in transcript levels. Putrescine synthesis in *E. coli* can proceed from the decarboxylation of arginine to agmatine and its subsequent hydrolysis to putrescine, reactions catalyzed by the products of genes *speA* and *speB*, respectively[24]. The higher transcript levels when growing in LB+G medium for *potABD* and *speB* encoding components of the spermidine/putrescine ATP-dependent importer and an enzyme of the putrescine biosynthetic pathway, respectively, are indicative of an increased demand for polyamines when conditions favor a higher growth rate for *E. coli*. Growth in medium containing glucose is known to repress flagellum synthesis[25]. Gene *fliO* is a member of the flagellar class II operon *fliLMNOPQR*, encoding proteins of the export apparatus and the motor/switch complex for flagellar function. This operon can be transcribed by either sigma 70 or the flagellum-specific sigma 28[26].

This analysis enabled us to demonstrate that glucose causes a change in transcript levels of 380 genes, grouped in 142 TUs, corresponding to 9% of the *E. coli* genome. If it is assumed that complete operons are induced when at least one gene member is detected in the microarray, then, this number would increase to 492 genes, corresponding to 11% of the *E. coli* genome. The comparison of the observed transcriptome pattern under the two nutritional conditions studied revealed global responses that involve functions not limited to nutrition/metabolism. Although *E. coli* displays high and similar growth rates in LB and LB+G media, this analysis reveals different transcriptome patterns that are consistent with distinct physiological states under these two conditions.

Transcriptional regulatory elements and mechanisms involved in glucose responses in *E. coli*

In recent years, many groups have concentrated on the study of the transcriptional responses of genes that integrate the regulatory network (RN) of some model organ-

isms such as *S. cerevisiae* and *E. coli* [27,28]. Some of these studies have analyzed the connectivities between the genes and TFs to understand topological properties of the RN [28,29] and infer modules that reflect a correlation between physiological and genetic responses. External stimuli provoke changes in the RN that help the cell to contend with a changing environment. The development of microarray technologies, gives us the opportunity to study globally the expression of genes in response to a given stimulus and try to detect the part of the RN (sub-network) responsible for the adaptative response.

The second part of this study consisted on the identification of the transcriptional RN involved in the observed glucose responses. This analysis represents an approach to understand at a systems level the behavior of the RN. The complete RN in the current version of the RegulonDB data base [30] represents 693 interactions involving 402 genes and 89 TFs. From the 380 regulated genes identified in the WTg/WT experiment, 142 possess a known regulatory interaction in RegulonDB. For these genes, we extracted from RegulonDB, the known information about TFs involved in their regulation. With this information, the RN was defined. We organized the regulatory interactions (RI) in strict simple and complex regulons (as previously described [31]). This data organization enabled us to analyze the interplay of the TFs involved in the regulatory changes of expression shown in the microarray data. We observed that 114 of these genes are regulated or coregulated by a global TF [32] (CRP, FNR, IHF, Fis, ArcA, NarL or Lrp), and only 28 of them don't interact with a global regulator (*zntA*, *mtgA*, *mgrB*, *metK*, *sufB*, *lon*, *cysK*, *uspA*, *fliO*, *fruB*, *pps*, *pckA*, *entC*, *nrdF*, *nrdH*, *nrdI*, *gatY*, *gatZ*, *gatA*, *ilvC*, *rpoD*, *rpsU*, *ahpC*, *hisJ*, *sufB*, *glnB*, *speB*, *proX*). The TFs involved in the regulation of these 28 genes are GadX, CysB, FadR, FhlD, FruR, Fur, GatR, LexA, OxyR, IlvY, MetJ, PhoB, PurR and NtrC.

Our data revealed a very small number of genes encoding TFs (*hupB*, *crp*, *fis*, *marA*, *cytR*, *yagA* and *hcaR*) that responded to the conditions studied (presence of glucose or loss of *crp* function). Although this will be explained in detail below, it should be pointed out that several of these TFs are involved in the regulation of a large number of the genes displaying a significant response to glucose.

As previously reported, we found that glucose responses are highly dependent on the TF, CRP [2], which is a global dual regulator, that governs the expression of at least 140 genes and corregulates gene expression with 75 other TFs [2]. In *E. coli*, CCR is mainly mediated by the PTS. When glucose is present in the culture medium, protein IIA^{Glc} lacks the capacity to activate adenylate cyclase; therefore, cAMP is present at relatively low levels. Lacking cAMP, the CRP protein cannot bind DNA and activate catabolite-

repressed genes [3]. Therefore, in the presence of glucose, CRP is unable to exert its usually positive effect on its regulated genes. The microarray and RegulonDB data revealed that of the 142 genes with known regulatory interactions, 50 are CRP regulated. Seven of these genes (*crp*, *cstA*, *ivbL*, *ilvB*, *putP*, *spf* and *trxA*), are regulated only by CRP. The other 42 genes are coregulated by CRP and one or more of 26 other TFs. From the 50 CRP affected genes, RegulonDB data indicates that 34 of them are activated by CRP and other TFs, 7 of them are exclusively activated by CRP, 6 are dual regulated and 3 genes present two CRP sites with opposite functions (Table 1). Except for the gene *putP* the seven genes that are solely regulated by a negative CRP binding site are induced in our experiment as expected. In the cases of *truB*, *infB* *nusA* and *rpsO*, the effect of Fis seems to enhance the expression of these genes, suggesting that the repression of *putB* could occur because of the presence of another TF, alternative regulatory mechanisms or additional CRP binding sites acting as positive regulators.

Transcriptome data showed that some of the genes positively regulated by CRP were down-regulated, in spite of the presence of other positive TFs like MalT, TorR and FNR. This effect had been previously described for the *melAB* and *malM* promoters [33,34], where CRP acts as a coactivator with a second TF. In our data, we found this response for the *malE* and *malM* genes, in which CRP triggers the repositioning of MalT to an appropriate activating position, causing the genes to be expressed [34]. The rest of the CRP regulated genes that do not appear repressed by glucose, are exclusively negatively regulated by CRP (*trxA*), or have one or more regulators that may counteract the effect of CRP (Table 1).

We found an important number of genes to be under the influence of Fis. RegulonDB reports 94 genes regulated by Fis. Our RN data showed 52 genes affected in the presence of glucose by Fis, grouped in 21 transcription units, out of which 48% belong to the Fis simple regulon, sharing some interesting characteristics: a) All are positively regulated by Fis, b) all are tRNA genes and c) when a binding site was reported, the central position varies between -66 and -75. Other members of the group, like *tyrT*, *alaT* and *tyrV* share the same characteristics except that they have three or two Fis binding sites. In the case of the genes *alaU*, *ileU* and *thrV*, a site for the nucleoid-structuring protein (HNS) has been characterized. It has been reported that the Fis site located near the promoter (between -71 and -78) is essential for promoter activation [35]. We observed another group of Fis-regulated genes that share their regulatory region with accessory TFs and additional Fis sites. The group of genes including *truB*, *b3170*, *nusA*, *infB* and *rpsO*, are co-transcribed by the complex regulon – ArgR(-), CRP(-) and Fis(+) --. According to our data, this group

appeared coordinately induced. We assume that this induction is caused by Fis activation together with no repressing effect of CRP (inactive in the presence of glucose) or ArgR.

The *nuo* genes, encoding the proton-translocating NADH:quinone oxidoreductase, appeared coordinately expressed, and all of the *nuo* genes are organized in a 13 genes operon (one of the longest transcription units in the genome). It has been reported that regulation of the expression of the *nuo* operon is subject to ArcA, that mediates anaerobic repression and NarL that mediates anaerobic activation in the presence of nitrate [36]. FNR and IHF act as weak repressors under anaerobic conditions [36], and Fis has been reported to stimulate expression of the operon in early exponential phase and to a lesser extent in the late exponential and stationary growth phases [37]. No significant difference in dissolved oxygen tension is expected when comparing cultures in LB or LB+G. Therefore, it can be speculated that transcriptional downregulation of the *nuo* operon is caused by medium composition or cell growth rate by an unknown mechanism. We detected an increase in the activity of *marA*, a gene that codes for the MarA TF, which is known to regulate its own expression [36]. Previous reports demonstrated that Fis stimulates expression of *marA* when MarA acts as an activator [38].

CRP has been described as the master regulator largely responsible for the expression pattern when *E. coli* is grown in glucose as the carbon source. However, very little is known about the influence of Fis on the gene expression pattern under the same conditions. We found a previous report showing that Fis is the factor mostly responsible for catabolite repression at the *nrf* promoter [39]. Experiments from other groups revealed that Fis assists both Mlc repression and CRP-cAMP activation of *ptsG* through the formation of Fis-CRP-Mlc or Fis-CRP nucleoprotein complexes at the *ptsG* promoter depending on the glucose availability in the growth medium [39,40]. Considering the large fraction of genes regulated by Fis identified in our study, it is clear that this TF has an important role in the cellular response to glucose.

Cluster analysis of transcriptome data for selected genes of wild type and crp strains

It has been proposed that most of the genes affected by the presence of glucose are directly or indirectly modulated by CRP. Glucose has an inactivating effect on CRP activity mainly by virtue of depressing cAMP levels. An analysis that compares transcriptome patterns between a *crp* mutant and the isogenic wild type strain grown in the presence of glucose could give clues about what genes are differentially expressed under these conditions. The results obtained from such analyses should identify which

genes have a CRP-dependent response to glucose. To help in identifying the role of CRP in the response to glucose, in this analysis we included transcriptome data from a *crp* mutant strain grown in LB conditions. A subset of 83 genes that displayed a significant response to both WTg/WT and *crp*/WT conditions was used in this analysis. For this purpose, using the results of microarrays conducted under these two previous conditions, we used a hierarchical clustering algorithm to evaluate the behavior of the genes shared under both conditions [41]. Figure 2, presents the clustering results, including labels with gene names and the corresponding regulating TFs. The cluster results showed that nearly all genes present the same response under both conditions. This indicates that the observed transcriptional response is dependent on CRP; however, it is not possible to determine from these results if the effect is direct or indirect. From this group of 83 genes, 25 displayed higher transcript level in the presence of glucose, with 13 being regulated solely by Fis, and 9 by this and other TFs. Among the latter group is the gene *fis*, regulated by CRP, Fis and IHF. It is noteworthy that under both conditions, the genes up-regulated by Fis, including *fis*, are significantly induced, suggesting that CRP plays an important role in the regulation of this gene. This result indicates that CRP is acting as a repressor of *fis* transcription. It has been reported that CRP together with Fis represses *fis* transcription during the exponential growth phase[42]

The TU including genes *aceE* and *aceF* is positively regulated by CRP, dually regulated by FNR and negatively regulated by PdhR. Considering that upregulation was also observed in the *crp* mutant, it can be inferred that CRP is not participating in this response. No changes in dissolved oxygen tension are expected when comparing cultures in LB or LB+G; therefore regulation by FNR can be ruled out. On the other hand, in LB+G, glucose catabolism should cause an increase in pyruvate concentration when compared to growth in LB medium. If this is the case, pyruvate can bind to and inactivate the repressor PdhR, thus causing the observed induction.

Another remarkable observation resulted from examination of the genes that appeared repressed, but a binding site for CRP or for other TFs regulated by CRP has not been identified (considering the information available in Regulon DB or EcoCyc). This was the case for the *pckA*, *lon*, *gatA*, *gatZ*, *gatY*, *gcvH*, *gcvT*, *osmE*, *dppA*, *pspE*, *ilvC*, *rpoD*, *lysU*, and *tdh* genes. Some of them, as mentioned before, are carrier proteins related to the import of alternative carbon and nitrogen sources (*gatA*, *gatZ*, *gatY* and *dppA*). The genes *aceA* and *pckA* deserve special attention because their regulator, the fructose repressor (FruR), is known to be partially inactivated in the presence of glucose. Fructose-1-phosphate and fructose-1-6-diphosphate, (direct

products of glycolysis), bind to FruR and inactivate its DNA-binding capacity [41,43]. As FruR positively regulates the expression of these two genes, the inactivation of the regulator causes the gene to be down regulated, a result that can be observed in our data. In addition, we found the gene *fruB* to be upregulated by the presence of glucose. This gene is repressed by FruR. In this case, we again find evidence of FruR inactivation by glycolytic intermediates[44]. These are significant results, as they allowed us to infer that a higher internal level of the glycolytic intermediate fructose-1-6-bisphosphate is present in the cells growing in the LB+G medium, when compared to the LB grown cells.

The genes *osmE* and *ompF* displayed a significant change in their levels of expression being induced in the *crp* mutant and repressed in the presence of glucose. It has not been reported that CRP directly regulates these well characterized genes. Instead, CRP directly controls the expression of the *ompR* gene, whose product controls the expression of *ompF*. Our result is consistent with a report showing an increment in the expression level of *ompF* under glucose limitation [45]. The effect is caused by the absence of cAMP that increases the levels of phosphorylated OmpR, which repress expression of *ompF*.

We have presented some of the relevant observations that can be extracted from table 1 and the cluster analysis comparing the wild type and the *crp* mutant. This analysis has shown that, as has been pointed out before, catabolic repression is mainly controlled by CRP, but that a small set of genes respond as a consequence of the intervention of TFs that have no described relationship with CRP. On the other hand, the prevalent role of Fis in the activation of genes under the LB+G conditions becomes evident in this analysis. It is known that *fis* gene transcription levels respond to growth rate, as can be expected since cells in LB+G medium grow 5% faster than cells in LB. Interestingly, it was also found that in the *crp* mutant, a strain that grows 5% slower than the wild type strain in the same LB medium, *fis* transcript levels are increased 3 fold (Table 1). Thus, these results show that CRP is playing an important role in *fis* regulation, resulting in its derepression when glucose is present.

Topological analysis of the regulatory network involved in the glucose response

The experimental results revealed that transcription factors CRP and Fis, are major regulators causing an extended response to glucose. However, it is clear that other TFs are also involved in controlling the genes found to respond to glucose. To help in identifying the relative roles of these TFs, an analysis of the properties of the regulatory network and its subnetworks (modules) is required. Resendis *et al* [28], demonstrate that the analysis of the regulatory net-

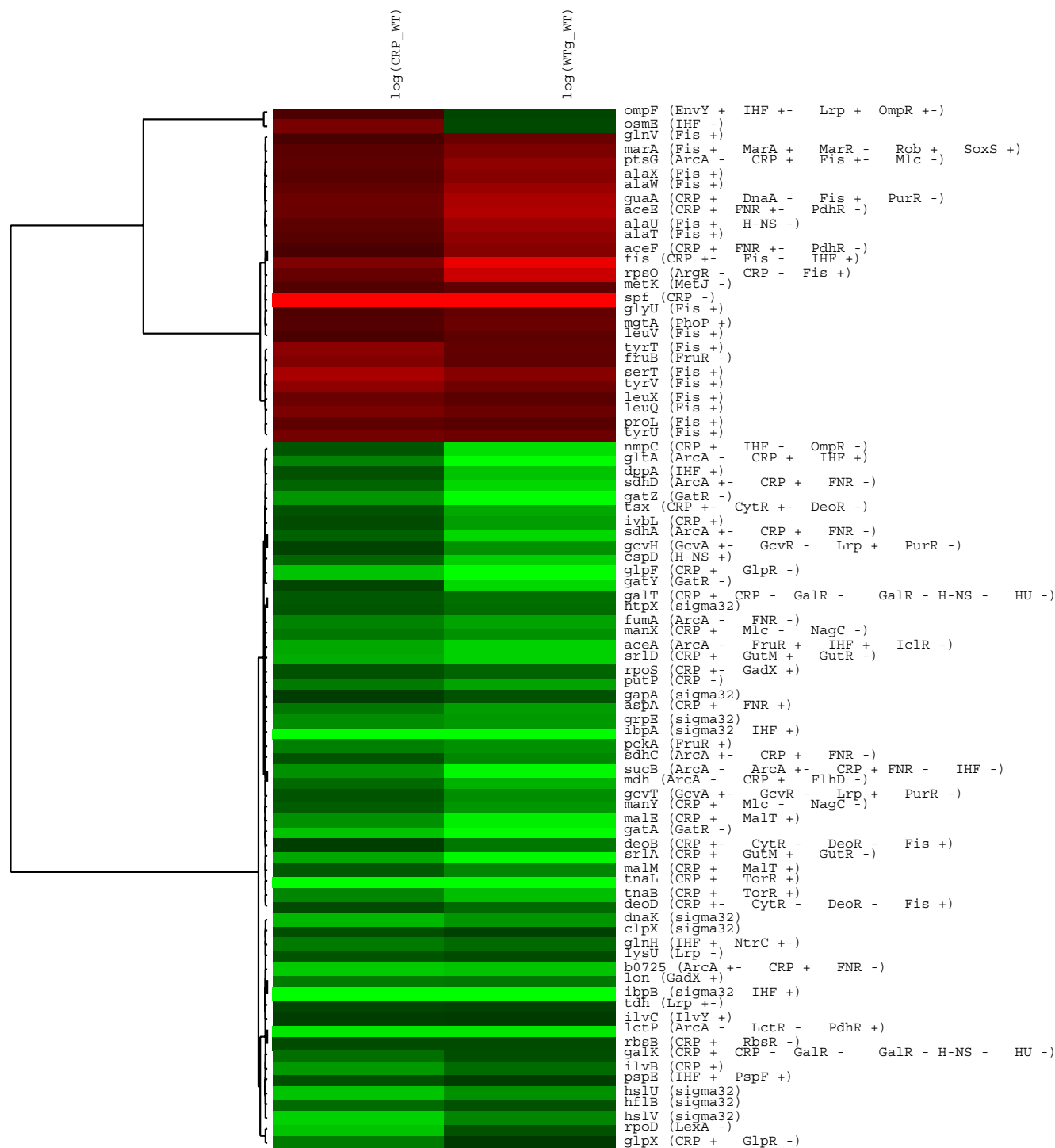


Figure 2
Cluster analysis of genes responding to both the presence of glucose and the loss of CRP. Red color indicates the induction state of the genes, and green color represents the repressed state. The name of each gene and the transcriptional factors involved in its regulation are indicated on the right side.

work in terms of its topology will evidence the relationship between modules and physiological functional classes [28]. Starting from the identified RN, we then performed a topological analysis to identify modules involved in the observed transcriptional responses.

Our study revealed sets of genes grouped into one independent unit (figure 12) and eight connected topological modules (figure 3). Module 1 includes genes regulated primarily by the sigma factor RpoH (sigma32), related to the heat shock response and chaperone proteins. The levels of the HSPs are tightly coupled to the metabolic and environmental status of the cell by regulation at the transcriptional level. Homogeneous patterns of gene expression were observed for 10 affected genes that were repressed in the presence of glucose (figure 4). Even though no direct effect of CRP has been reported for these genes, it has been reported that the active form of CRP directly stimulates the expression of sigma32 [46]. This result is consistent with the observed repression of this group of genes that can be explained by inactive CRP in the presence of glucose or in a *crp* background.

Module 2 is controlled by Integration Host Factor (IHF). It includes 4 genes that are homogeneously expressed, and are mainly related with amino acids metabolism and laterally transferred elements (figure 5).

Module 3 is composed of five genes mainly regulated by the oxidative stress protein OxyR with some subnetworks also regulated by methionine repressor MetJ or IHF. We can observe in figure 6 that the genes coregulated by OxyR are homogeneously repressed. We also observed that all of the proteins coregulating this set except MetJ, function as activators. The gene *metK*, which appeared to be solely regulated by MetJ, was induced under this condition, suggesting that all of the regulators should be inactive in the presence of glucose.

Module 4 (figure 7), is mainly composed of the PurR regulon that controls expression of the *gcvTHP* operon that is involved in glycine metabolism. These genes were down regulated in the presence of glucose, a phenomenon that has been studied by Wonderling and Stauffer [47]. The authors demonstrate that *crp* inactivation caused a reduction in *gcvT* expression in the presence of glucose. The other three genes (*guaA*, *glnB*, *spe*), appeared induced under this condition. In Table 1, the genes *glnB* and *speB* are exclusively regulated by PurR, acting as a repressor. If no other TF or alternative regulatory processes are affecting these genes, it would be possible to predict that the state of PurR is in *off* in the presence of glucose, therefore, it is not repressing transcription.

Module 5 is largely regulated by leucine-responsive regulatory protein Lrp (figure 8). Genes *kbl* and *tdh* belong to the same operon, according to the data extracted from RegulonDB. Lrp represses the expression of the operon. These two genes appeared down-regulated in the presence of glucose suggesting that Lrp is repressing their expression. The *ompF* and *fimA* genes exhibit a very complex regulation. EnvY and Lrp that act as activators, and IHF and OmpR that function as dual regulators of the *ompF* gene, are repressed under glucose conditions. A search of the literature revealed that our results are consistent with previous data that report that the expression of *ompF* is increased more than 40-fold higher under glucose limitation conditions [45]. In the same work, the authors reported expression of *ompF* in the absence of cAMP. The induced OmpR resulted in the production of more OmpR-P, which represses the expression of *ompF* gene.

Module 6 is composed of 12 genes related to respiratory or energy generation functions (figure 9). It was amazing to find that all the genes that constitute this module are homogeneously expressed considering that, as for the CRP module, the genes are regulated by several factors controlling expression of these genes. Module 7 has 41 genes, regulated by Fis. Within subdivisions observed in the tree for these sub-branches, homogeneous gene expression patterns are observed that correspond to genes coregulated by the same set of regulatory proteins (figure 10).

Module 8 includes genes involved in carbon source assimilation. As mentioned before, the 50 genes of the CRP module do not present homogeneous gene expression patterns. However, by searching lower branches of the tree (figure 3 or table 1), we found that except for 2 sub-branches, the rest correspond to groups of genes coregulated by more than one TF. Following this criterion, we located 19 sub-branches (figure 11) in which 6 subgroups show non-homogeneous gene expression patterns. It is interesting to note that two of them are exclusively regulated by only one TF, CRP in the first case, and the fructose regulator FruR in the second. The differences in gene expression are given because the clustering algorithm does not consider the fact that the proteins can exert opposite effects on regulated genes, positive or negative. Interestingly, genes positively regulated appear closer to each other in the cluster than for negatively regulated genes. The other groups cluster genes coregulated by CRP and particularly accessory TFs.

We found four cases (submodules 8.2, 8.6, 8.8 and 8.19), in which one gene presents the opposite expression pattern compared to the other members of the group. In all cases, the gene with opposite expression pattern lacks one of the TF binding sites present in the other genes. An example is the subbranch containing the genes *aceE*, *aceF*

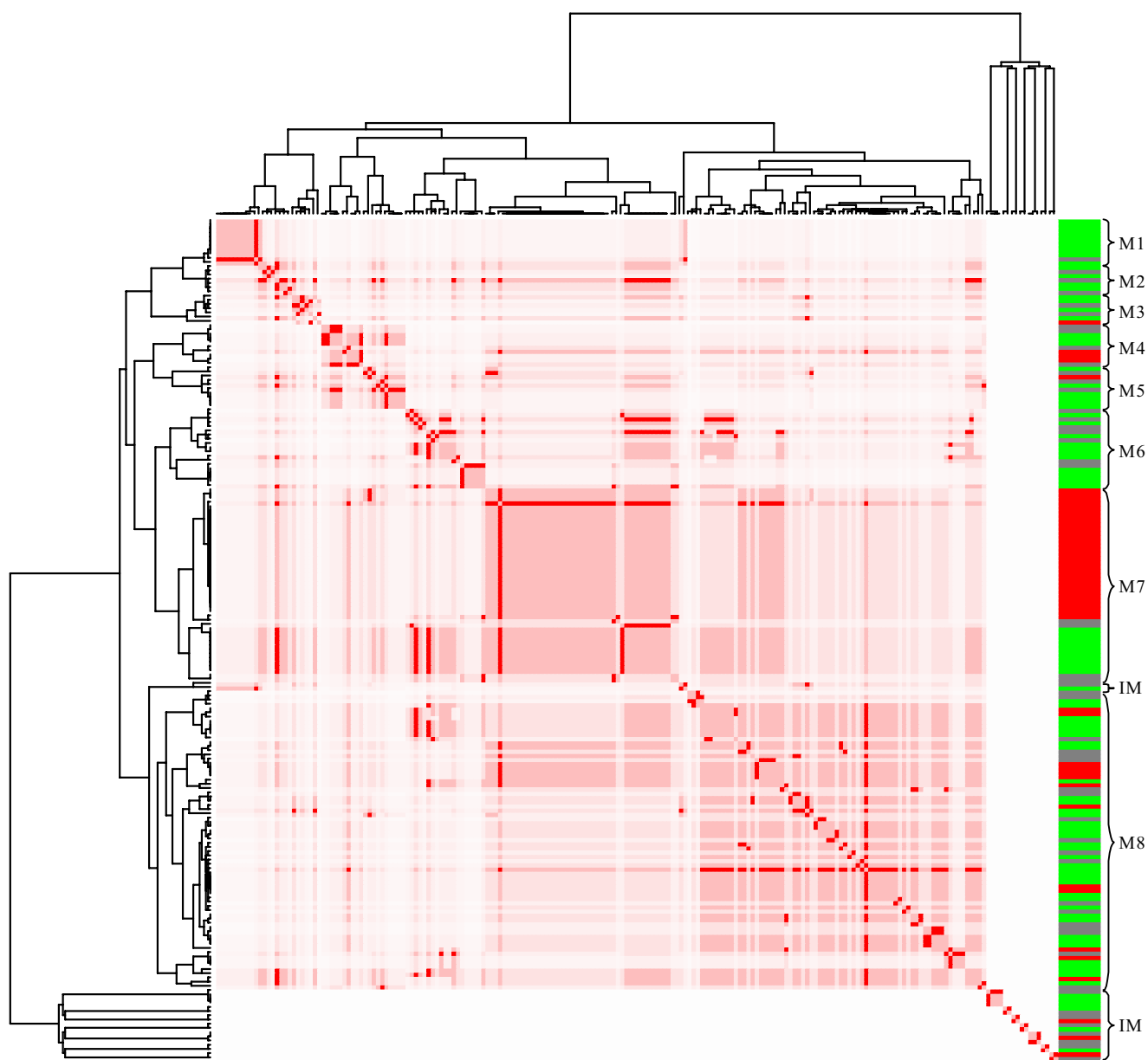


Figure 3
Representation of the regulatory network as topological modules. The figure illustrates the presence of several isolated modules and eight interconnected modules. The relative transcript levels of genes integrating the RN are shown in the right edge. Increased and decreased transcript levels are indicated by red and green, respectively.

and *aspA*, in which the first two genes are coregulated positively by CRP, positively or negatively by FNR, and negatively by PdhR. If we consider only the information found in RegulonDB and EcoCyc, the increased levels of expression of *aceE* and *aceF* should be a consequence of the inactivation of CRP and PdhR. Considering the low levels of cAMP, and the increase of pyruvate as a product of glycolysis [48], we can assume that FNR might activate or not repress the *aceE* and *aceF* genes. The *aspA* gene,

which is positively regulated by CRP and FNR, appears down regulated in the presence of glucose. This result is consistent with the finding that *aspA* is under catabolite repression control [49,50].

The preceding analysis provides a view of the roles and interactions of specific TFs in response to glucose. An important question related to this subject is: how many different pathways/mechanisms exist in *E. coli* to detect

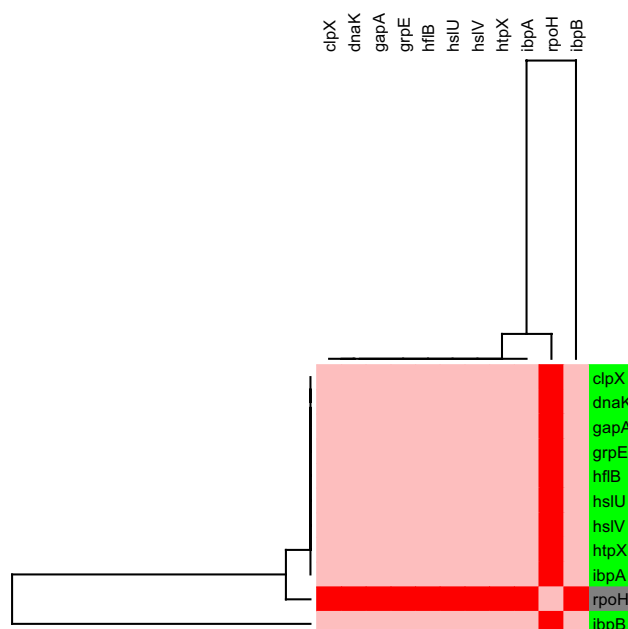


Figure 4
Individual modular component of the regulatory network under the control of Sigma 32 (M1). The figure represents a detailed view of the individual topological module extracted from figure 3. The relative transcript levels and names of genes integrating the subnetwork are shown on the right. Decreased transcript levels are indicated in green. Grey indicates transcription factors that did not respond to glucose in the transcriptome experiment.

the presence of glucose and relay this information to the RN? Figure 13 was generated mainly from previously published works and it is supported by some of our current results. This figure shows a summary of the signals generated by the consumption of glucose and their effects on specific TFs. Specific glucose detection is dependent on the phosphorylation state of the glucose-specific PTS protein IIB^{Glc}. This protein is involved in the phosphorylation of glucose that is transported by the IIC^{Glc} integral membrane protein domain. When glucose is present in the medium, IIB^{Glc} is mainly in a non-phosphorylated state. Under this condition, IIB^{Glc} binds the Mlc repressor protein, thus relieving its repression of the *ptsHI* and *ptsG* genes, among others[16]. Other signals generated by the presence of glucose, such as a relatively low level of cAMP, increased levels of certain metabolites, and an increased growth rate are caused directly or indirectly. A clear effect of this phenomenon can be seen in figure 13 with fructose-1-6-biphosphate and pyruvate that induce the expression of genes under FruR and PdhR control. Sugars other than glucose can also cause some of these effects, but these will vary depending on their quality as carbon

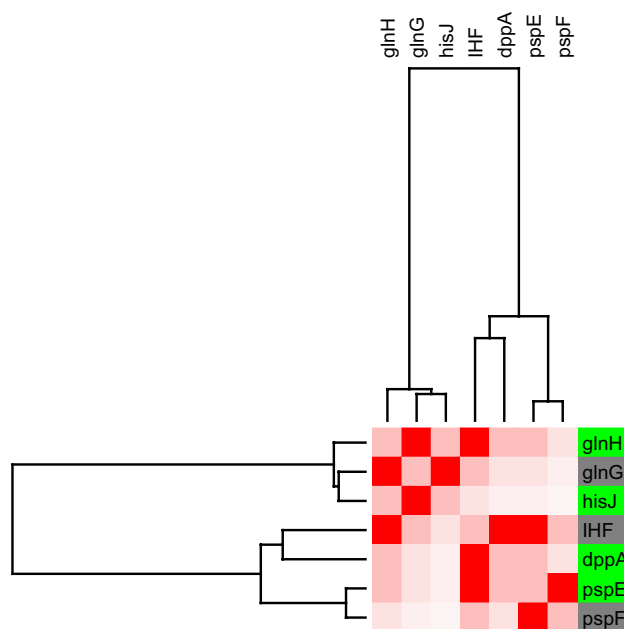


Figure 5
Modular component of the regulatory network controlled by IHF. (M2). Module 2, shows a zoom in to a subnetwork, that clustered a set of genes mainly regulated by IHF. As in figure 4, the relative transcript levels and names of genes integrating the subnetwork are shown on the right. Decreased transcript levels are indicated in green. Grey indicates transcription factors that did not respond to glucose in the transcriptome experiment.

and energy sources. All these signals are detected by specific TFs that in turn regulate other TFs or structural genes. As shown in figure 13, some TFs can simultaneously receive and thus integrate inputs from different pathways, such as is the case with Fis, which displays growth-rate regulation and is also regulated by Crp. It should be emphasized, though, that we are still far from a complete understanding of how the glucose signal is propagated through this network, and how other environmental signals are integrated to modulate the overall response. The combined analyses of transcriptome data and the RN involved in the observed responses, as has been performed here, should contribute to the identification of signaling pathways and their integration by the RN.

Conclusion

The analysis of transcriptome data collected under conditions of glucose deficiency and sufficiency in a complex medium enabled us to identify functions involved in the adaptation of *E. coli* to these two different growth conditions. The known repressive effects of glucose on gluconeogenesis and on alternative carbon source import and metabolism were clearly demonstrated. Furthermore,

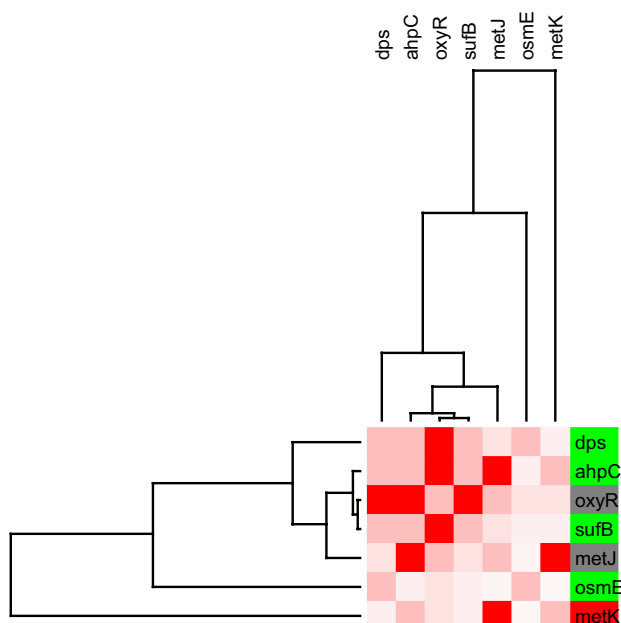


Figure 6
The OxyR modular component of the regulatory network. (M3). The figure illustrates the genes in the module that were up regulated (in red) and down regulated (in green). Grey indicates transcription factors that did not respond to glucose in the transcriptome experiment.

when glucose was present in the medium, an increase in overall protein synthesis capacity was observed. Also, responsive to the presence of glucose were genes encoding different cellular functions including cell division, replication, transcription, and the biosynthesis of cofactors, nucleic acids, amino acids and lipids. This analysis also revealed that functions related to proteolysis and protein folding are apparently more important when *E. coli* is growing in LB medium as compared with LB+G medium.

The topological analysis of the RN involved in the regulation of a subset of glucose-responsive genes, revealed eight modules including 37 TFs. Most of the RN topological modules include genes encoding functions with similar physiological roles, and together they represent a significant part of the glucose stimulon. The modules we identified partially correspond to the regulatory subnetworks originating at sensor TFs (origons) that have been identified in the complete *E. coli* RN[29]. The difference can be explained considering that we have limited our analyses to specific growth conditions and a subset of the RN. It can be assumed that this is still a partial representation of the RN involved in this response, since the functions of a significant number of TFs in *E. coli* are still unknown [30,51]. In spite of this shortcoming, our results and those previously reported by other groups indicate

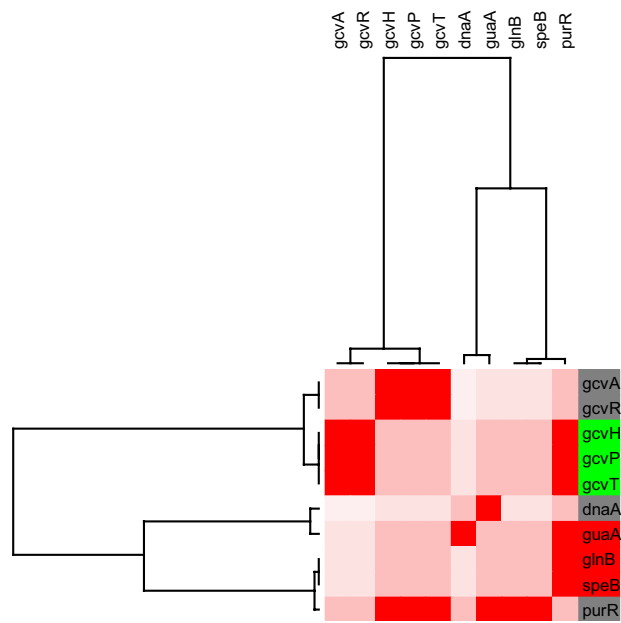


Figure 7
The PurR modular component. (M4). The figure shows the relative transcript levels and names of genes integrating the subnetwork at the right. Increased and decreased transcript levels are indicated by red and green, respectively. Grey indicates transcription factors that did not respond to glucose in the transcriptome experiment.

that CRP and Fis play a dominant role in the transcriptional responses detected in this study. This analysis places CRP and Fis as central TFs in the subset of the *E. coli* RN that senses and responds to glucose and other sugars. These two regulatory proteins integrate different types of signals that reflect the nutritional composition of the medium and the physiological state of the cell, causing a corresponding genome-wide transcriptional response.

Current limits in sensitivity and specificity for transcriptome analysis methodologies, together with our incomplete knowledge of the properties and interactions of TFs, still do not allow a thorough understanding of the cellular response to specific stimuli. However, integrative analysis of transcriptome and RN data as performed here, should provide a framework for the future generation of models representing the cell's capacity to respond to a changing environment.

Methods

Source of experimental data

Transcriptome data was obtained from previously reported experiments performed with *E. coli* strain BW25113 and an isogenic *crp* mutant (LJ3017)[10]. Briefly, strains were grown at 37°C with agitation in Luria-

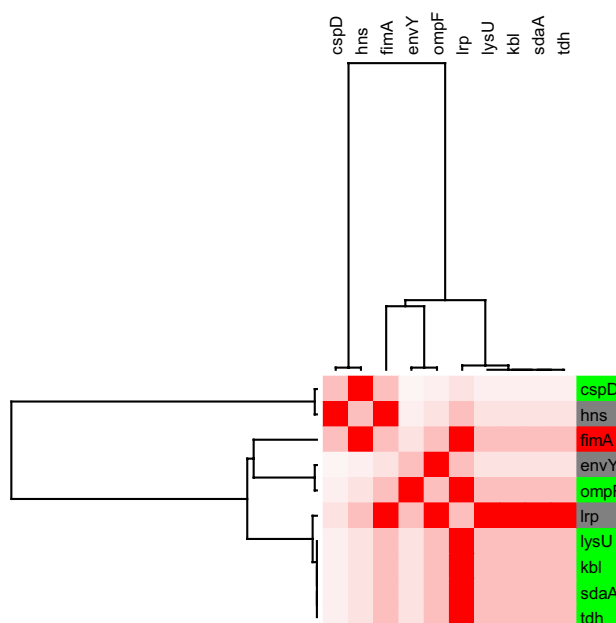


Figure 8
Individual modular component of the regulatory network controlled by Lrp. (M5). The relative transcript levels and names of genes integrating the subnetwork mainly regulated by Lrp, are shown on the right. Six genes in the figure decreased their relative transcript levels (in green), and only one gene was induced in this module (in red). Grey indicates transcription factors that did not respond to glucose in the transcriptome experiment.

Bertani (LB) broth containing 50 mM potassium phosphate, pH 7.4, and 0.2 mM Lcysteine with or without 0.4% glucose. Cells were grown in triplicate in 25 ml of medium in shake flasks starting at an OD₆₀₀ of 0.05 and harvested in the exponential growth phase when cultures reached an OD₆₀₀ of 0.5. When grown in LB medium, generation times for strains BW25113 and LJ3017 corresponded to 37 and 43 min., respectively. In LB+G medium, generation times for strains BW25113 and LJ3017 corresponded to 35 and 41 min., respectively[10,21]. Total RNA was extracted from each sample, processed and hybridized to the Affymetrix *E. coli* array which includes 4327 genes and intergenic regions[11].

Data analysis

Array scanning, data collection and normalization were performed following the procedure described by Caldwell et al. 2001[52]. Three data sets were obtained for each of three experimental conditions: wild type grown in LB medium (WT), wild type grown in LB medium + glucose (WTg) and a *crp* mutant grown in LB medium (CRP). The data sets for each strain and condition were compared

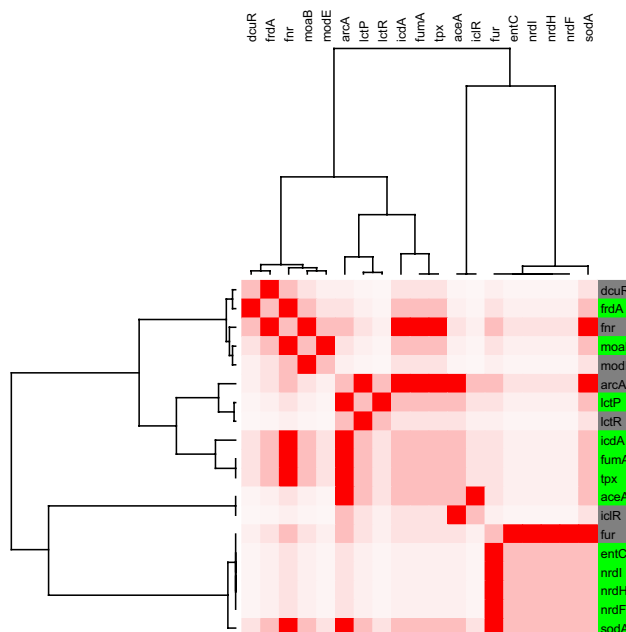


Figure 9
Respiration module. (M6). In this case, the module is controlled by a set of proteins known to be related to different respiration forms in *E. coli*. The expression levels and color remains as previously described.

pair-wise to determine the Pearson correlation coefficient. For each triplicate data set, the two sets with the highest Pearson correlation coefficient were retained for further analysis.

For each pair of data sets of all experimental conditions, the reliability of the data for each gene was calculated according to the Affymetrix statistical algorithms reference guide (Affymetrix, Inc., 2004). A "Present" absolute call is assigned to a gene when the signal/noise ratio is higher than an internally calculated threshold. When signal value data for each gene displayed a "Present" absolute call in both duplicate experiments, both values were considered to be reliable. The two signal values for that gene were averaged, and the resulting data were used in subsequent analyses. Using this approach, the number of genes considered for further analysis corresponded to 1908, 1910 and 3083 for WT, WTg and CRP conditions, respectively. Using the signal averages for each condition, we then calculated the WTg/WT and CRP/WT log ratios.

Identification of differentially transcribed genes

Differentially transcribed genes were selected using an outlier iteration method [12-14]. The method consists in calculating the average and the standard deviation (SD) of the log ratio for all sets of genes under the four conditions. In order to identify significant levels of gene expression,

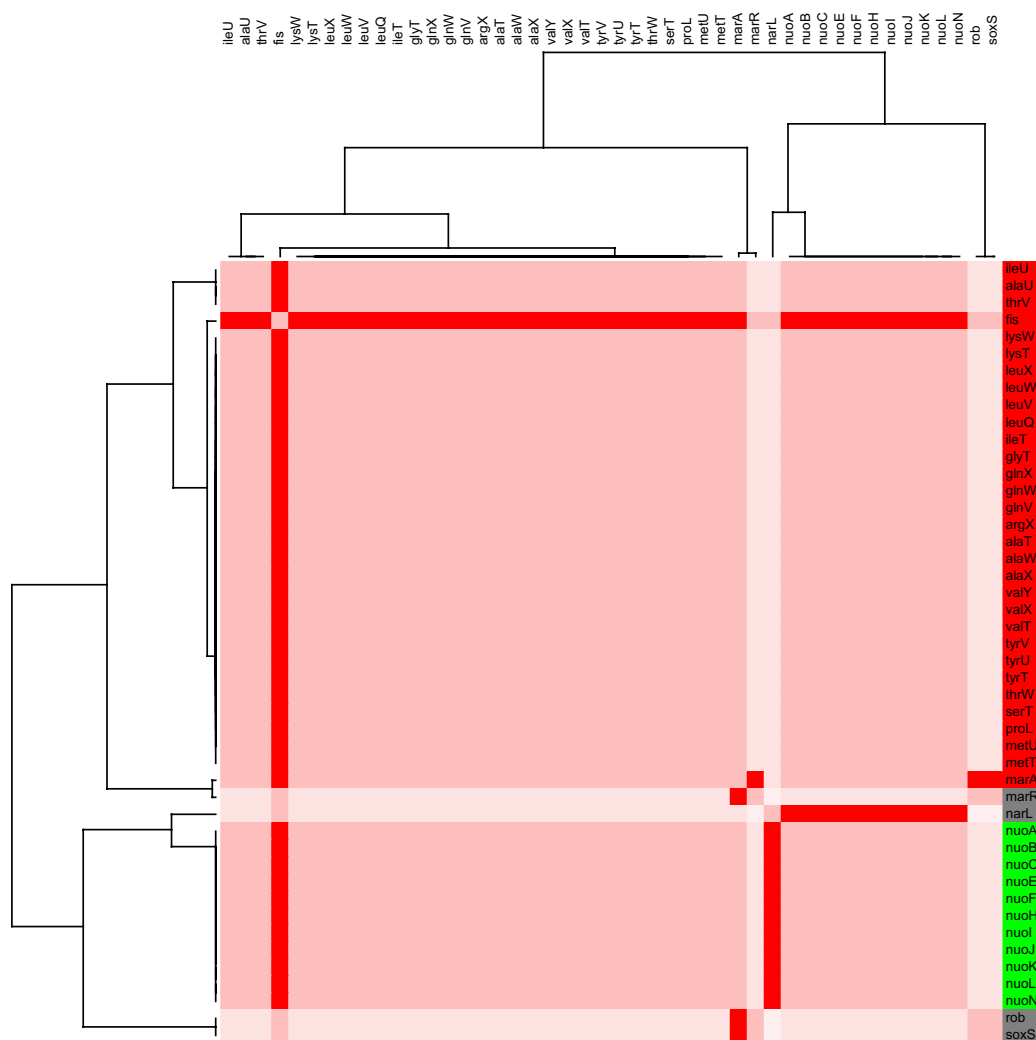


Figure 10
Individual modular component of the regulatory network controlled by Fis. (M7). A modular component mainly regulated by the transcriptional factor Fis. Increased and decreased transcript levels are indicated by red and green, respectively. Grey indicates transcription factors that did not respond to glucose in the transcriptome experiment.

we assumed that the threshold value of significance is two SD. Thus, any gene with a log ratio higher than two SD from the mean is considered an outlier. Outliers were removed from the population and gathered in a differentially expressed subset. For the rest of the genes, we calculated again the averages of their log ratios and their SD values. Selection of the outliers was determined as in the previous case. The process was repeated until no outliers were detected in each situation. Using this method, the number of genes selected corresponded to 380 for WTg/WT and 333 for CRP/WT. For CRP/WT, 196 genes were down regulated and 137 up regulated. Table S1 shows the genes identified in this study, where values for WTg/WT and CRP/WT log ratios are provided. In addition, when

known, the regulatory phrase for each gene is indicated and also, when a gene is part of an operon, the genes belonging to it are indicated. Information about gene functions and operon organization was obtained from RegulonDB [53] and EcoCyc [54]. It should be pointed out that the terms "induced" and "repressed" are used in this work to indicate increased or decreased transcript levels, respectively. These terms do not imply a particular mechanism of gene regulation.

Clustering of microarray data

We applied a hierarchical centroid linkage clustering algorithm [41,55] with correlation uncentered as similarity measure, to the WTg/WT and CRP/WT log ratios. The clus-

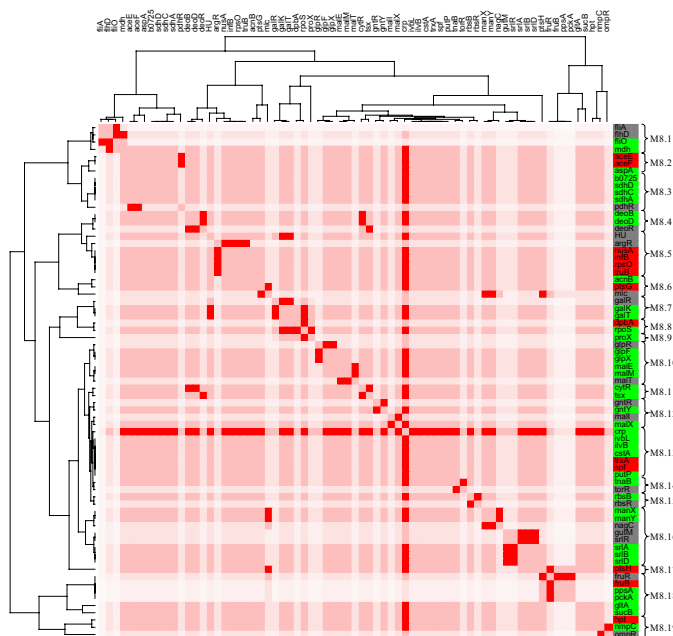


Figure 11
Carbon sources assimilation module. (M8). The figure illustrates a module in which most of the genes are interconted by the CRP transcription factor. It is also divided in 19 submodules, that are related with other Tf's that coregulate with CRP. The expression levels and color remains as previously described.

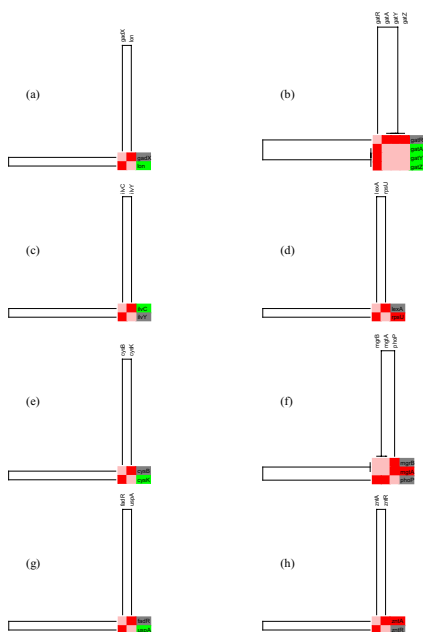


Figure 12
Eight mini-modules disconnected from the giant component. (IM). We illustrated a series of subnetworks that do not show any connection with the giant component in figure 3. The expression levels and color remains as previously described.

tering results were visualized using the Treeview program[56].

Extraction of condition-specific subnetworks

For each microarray condition (WTg/WT or CRP/WT), we reconstructed a condition specific subnetwork as follows. From the transcriptional regulatory network (RN) of *E. coli*, we extracted the genes identified for each microarray condition, the TFs regulating their expression, and the transcriptional interactions between TFs and regulated genes. In these subnetworks, nodes represent genes, and edges represent the transcriptional interactions. Known regulatory sites and transcriptional unit organization were obtained from RegulonDB [30] and EcoCyc[57].

Identification of condition-specific modules

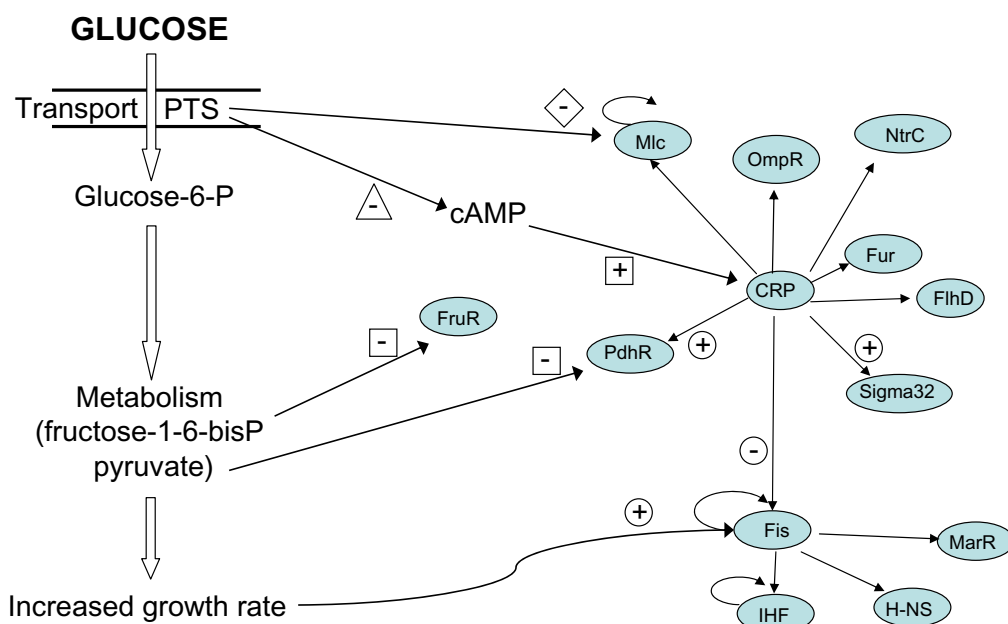
We identified the WTg/WT condition-specific modules applying to the condition specific subnetwork the methodology described in Resendis-Antonio et al[28]. That is to say, we clustered the genes based on their shortest distance within the network. Afterwards, we annotated each gene with its corresponding microarray expression level.

Abbreviations

LB, LB+G WTg, WT, TF, RN, PTS.

Authors' contributions

RMG contributed with the computer analysis and interpretation of the microarray data in terms of the regulatory

**Figure 13**

Glucose sensing pathways and the TFs involved. Plus and minus signs indicate positive and negative effects: (○) transcriptional control, (□) metabolites interaction, (◇) protein-protein interaction, and (△) adenylate cyclase control.

network, elaboration of programs for data management and discussion for the selection and processing methods. JF contributed with computer analysis of microarray data and with the construction of topological modules. OR processed microarray data, evaluating different approaches for the identification of differentially transcribed genes. JCV participated in data analysis and the discussion of every section of the manuscript. MS participated in the experimental design, supplying the microarray data and collaborating in the discussion of every section of the manuscript. GG participated in the experimental design and contributed with the analysis and interpretation of microarray data for every section of the manuscript.

Acknowledgements

We thank Heladia Salgado Nancy Mena and Verónica Jiménez for technical assistance. We also thank the Computational Unit and the 'Macroproyecto de Tecnologías de la Información y la Computación de la Universidad Nacional Autónoma de México' for the use of their computer facilities. This work was supported by grants IN205005-2 and IN203705-3 from PAPIIT-UNAM, and CONACyT. J.A. Freyre-González is supported by Ph.D. fellowship number 176341 from CONACyT-México.

References

- Preston RD: *The physical biology of plant cell walls* London, Chapman and Hall; 1974.
- Saier MH Jr., Ramseier TM, Reizer J: **Regulation of carbon utilization.** In *Escherichia coli and Salmonella. Cellular and Molecular Biology*. Edited by: Neidhardt FC. Washington, D.C., American Society for Microbiology; 1996:1325-1343.
- Postma PW, Lengeler JW, Jacobson GR: **Phosphoenolpyruvate: Carbohydrate phosphotransferase systems.** In *Escherichia coli and Salmonella. Cellular and Molecular Biology*. Edited by: Neidhardt FC. Washington, D.C., American Society for Microbiology; 1996:1149-1174.
- Saier MH Jr.: **Vectorial metabolism and the evolution of transport systems.** *J Bacteriol* 2000, **182**:5029-5035.
- Tchieu JH, Norris V, Edwards JS, Saier MH Jr.: **The complete phosphotransferase system in Escherichia coli.** *J Mol Microbiol Biotechnol* 2001, **3**:329-346.
- Hua Q, Yang C, Oshima T, Mori H, Shimizu K: **Analysis of gene expression in Escherichia coli in response to changes of growth-limiting nutrient in chemostat cultures.** *Appl Environ Microbiol* 2004, **70**:2354-2366.
- Liu M, Durfee T, Cabrera JE, Zhao K, Jin DJ, Blattner FR: **Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli.** *J Biol Chem* 2005, **280**:15921-15927.
- Tao H, Bausch C, Richmond C, Blattner FR, Conway T: **Functional genomics: expression analysis of Escherichia coli growing on minimal and rich media.** *J Bacteriol* 1999, **181**:6425-6440.
- Oh MK, Rohlin L, Kao KC, Liao JC: **Global expression profiling of acetate-grown Escherichia coli.** *J Biol Chem* 2002, **277**:13175-13183.
- Gosset G, Zhang Z, Nayyar S, Cuevas WA, Saier MH Jr.: **Transcriptome analysis of Crp-dependent catabolite control of gene expression in Escherichia coli.** *J Bacteriol* 2004, **186**:3516-3524.
- Selinger DW, Cheung KJ, Mei R, Johansson EM, Richmond CS, Blattner FR, Lockhart DJ, Church GM: **RNA expression analysis using a 30 base pair resolution Escherichia coli genome array.** *Nat Biotechnol* 2000, **18**:1262-1268.
- Britton RA, Eichenberger P, Gonzalez-Pastor JE, Fawcett P, Monson R, Losick R, Grossman AD: **Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of Bacillus subtilis.** *J Bacteriol* 2002, **184**:4881-4890.
- Loos A, Glanemann C, Willis LB, O'Brien XM, Lessard PA, Gerstmeier R, Guillouet S, Sinskey AJ: **Development and validation of corynebacterium DNA microarrays.** *Appl Environ Microbiol* 2001, **67**:2310-2318.

14. Zheng D, Constantinidou C, Hobman JL, Minchin SD: **Identification of the CRP regulon using in vitro and in vivo transcriptional profiling.** *Nucleic Acids Res* 2004, **32**:5874-5893.
http://www.ibt.unam.mx/biocomputo/gutierrez_rios.htm
15. http://www.ibt.unam.mx/biocomputo/gutierrez_rios.htm 2000 [http://www.ibt.unam.mx/biocomputo/gutierrez_rios.htm].
16. Plumbidge J: **Expression of ptsG, the gene for the major glucose PTS transporter in Escherichia coli, is repressed by Mlc and induced by growth on glucose.** *Mol Microbiol* 1998, **29**:1053-1063.
17. Cronan JE, LaPorte D: **Tricarboxylic acid cycle and glyoxylate bypass.** In *Escherichia coli and Salmonella: cellular and molecular biology*. 2nd edition. Edited by: Neidhardt FC, Curtiss III R, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff VWS, Riley M, Schaechter M and Umberger HE. Washington, D.C., ASM Press; 1996:206-216.
18. Goff SA, Goldberg AL: **Production of abnormal proteins in E. coli stimulates transcription of lon and other heat shock genes.** *Cell* 1985, **41**:587-595.
19. Parsell DA, Sauer RT: **Induction of a heat shock-like response by unfolded protein in Escherichia coli: dependence on protein level not protein degradation.** *Genes Dev* 1989, **3**:1226-1232.
20. Gourse RL, Gaal T, Bartlett MS, Appleman JA, Ross W: **rRNA transcription and growth rate-dependent regulation of ribosome synthesis in Escherichia coli.** *Annu Rev Microbiol* 1996, **50**:645-677.
21. Zhang Z, Gosset G, Barabote R, Gonzalez CS, Cuevas WA, Saier MH Jr.: **Functional interactions between the carbon and iron utilization regulators, Crp and Fur, in Escherichia coli.** *J Bacteriol* 2005, **187**:980-990.
22. Yamanaka K, Inouye M: **Growth-phase-dependent expression of cspD, encoding a member of the CspA family in Escherichia coli.** *J Bacteriol* 1997, **179**:5126-5130.
23. Yamanaka K, Zheng W, Crooke E, Wang YH, Inouye M: **CspD, a novel DNA replication inhibitor induced during the stationary phase in Escherichia coli.** *Mol Microbiol* 2001, **39**:1572-1584.
24. Szumanski MB, Boyle SM: **Influence of cyclic AMP, agmatine, and a novel protein encoded by a flanking gene on speB (agmatine ureohydrolase) in Escherichia coli.** *J Bacteriol* 1992, **174**:758-764.
25. Silverman M, Simon MI: **Bacterial flagella.** *Annu Rev Microbiol* 1977, **31**:397-419.
26. Liu X, Matsumura P: **Differential regulation of multiple overlapping promoters in flagellar class II operons in Escherichia coli.** *Mol Microbiol* 1996, **21**:613-620.
27. Blais A, Dynlacht BD: **Constructing transcriptional regulatory networks.** *Genes Dev* 2005, **19**:1499-1511.
28. Resendis-Antonio O, Freyre-Gonzalez JA, Menchaca-Mendez R, Gutierrez-Rios RM, Martinez-Antonio A, Avila-Sanchez C, Collado-Vides J: **Modular analysis of the transcriptional regulatory network of E. coli.** *Trends Genet* 2005, **21**:16-20.
29. Balazsi G, Barabasi AL, Oltvai ZN: **Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli.** *Proc Natl Acad Sci U S A* 2005, **102**:7841-7846.
30. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A, Collado-Vides J: **RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions.** *Nucleic Acids Res* 2006, **34**:D394-D397.
31. Gutierrez-Rios RM, Rosenbluth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR, Collado-Vides J: **Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles.** *Genome Res* 2003, **13**:2435-2443.
32. Martinez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria.** *Curr Opin Microbiol* 2003, **6**:482-489.
33. Barnard A, Wolfe A, Busby S: **Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes.** *Curr Opin Microbiol* 2004, **7**:102-108.
34. Richet E, Vidal-Ingigliardi D, Raibaud O: **A new mechanism for coactivation of transcription initiation: repositioning of an activator triggered by the binding of a second activator.** *Cell* 1991, **66**:1185-1195.
35. Travers A, Muskhelishvili G: **DNA microloops and microdomains: a general mechanism for transcription activation by torsional transmission.** *J Mol Biol* 1998, **279**:1027-1043.
36. Bongaerts J, Zoske S, Weidner U, Uden G: **Transcriptional regulation of the proton translocating NADH dehydrogenase genes (nuoA-N) of Escherichia coli by electron acceptors, electron donors and gene regulators.** *Mol Microbiol* 1995, **16**:521-534.
37. Wackwitz B, Bongaerts J, Goodman SD, Uden G: **Growth phase-dependent regulation of nuoA-N expression in Escherichia coli K-12 by the Fis protein: upstream binding sites and bioenergetic significance.** *Mol Gen Genet* 1999, **262**:876-883.
38. Martin RG, Rosner JL: **Fis, an accessory factor for transcriptional activation of the mar (multiple antibiotic resistance) promoter of Escherichia coli in the presence of the activator MarA, SoxS, or Rob.** *Journal of Bacteriology* 1997, **179**:7410-7419.
39. Browning DF, Grainger DC, Beatty CM, Wolfe AJ, Cole JA, Busby SJ: **Integration of three signals at the Escherichia coli nrf promoter: a role for Fis protein in catabolite repression.** *Mol Microbiol* 2005, **57**:496-510.
40. Shin D, Cho N, Heu S, Ryu S: **Selective regulation of ptsG expression by Fis. Formation of either activating or repressing nucleoprotein complex in response to glucose.** *J Biol Chem* 2003, **278**:14776-14781.
41. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**:14863-14868.
42. Nasser W, Schneider R, Travers A, Muskhelishvili G: **CRP modulates fis transcription by alternate formation of activating and repressing nucleoprotein complexes.** *J Biol Chem* 2001, **276**:17878-17886.
43. Bledig SA, Ramseier TM, Saier MH Jr.: **FruR mediates catabolite activation of pyruvate kinase (pykF) gene expression in Escherichia coli.** *J Bacteriol* 1996, **178**:280-283.
44. Ramseier TM, Negre D, Cortay JC, Scarabel M, Cozzone AJ, Saier MH Jr.: **In vitro binding of the pleiotropic transcriptional regulatory protein, FruR, to the fru, pps, ace, pts and icd operons of Escherichia coli and Salmonella typhimurium.** *J Mol Biol* 1993, **234**:28-44.
45. Liu X, Ferenci T: **An analysis of multifactorial influences on the transcriptional control of ompF and ompC porin expression under nutrient limitation.** *Microbiology* 2001, **147**:2981-2989.
46. Jenkins DE, Auger EA, Matin A: **Role of RpoH, a heat shock regulator protein, in Escherichia coli carbon starvation protein synthesis and survival.** *J Bacteriol* 1991, **173**:1992-1996.
47. Wonderling LD, Stauffer GV: **The cyclic AMP receptor protein is dependent on GcvA for regulation of the gcv operon.** *J Bacteriol* 1999, **181**:1912-1919.
48. Quail MA, Haydon DJ, Guest JR: **The pdhR-aceEF-lpd operon of Escherichia coli expresses the pyruvate dehydrogenase complex.** *Mol Microbiol* 1994, **12**:95-104.
49. Bell PJ, Andrews SC, Sivak MN, Guest JR: **Nucleotide sequence of the FNR-regulated fumarase gene (fumB) of Escherichia coli K-12.** *J Bacteriol* 1989, **171**:3494-3503.
50. Woods SA, Miles JS, Roberts RE, Guest JR: **Structural and functional relationships between fumarase and aspartase. Nucleotide sequences of the fumarase (fumC) and aspartase (aspA) genes of Escherichia coli K12.** *Biochem J* 1986, **237**:547-557.
51. Perez-Rueda E, Collado-Vides J, Segovia L: **Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea.** *Comput Biol Chem* 2004, **28**:341-350.
52. Caldwell R, Sapolsky R, Wweyler VV, Maile RR, Causey SC, Ferrari E: **Correlation between Bacillus subtilis scoC phenotype and gene expression determined using microarrays for transcriptome analysis.** *J Bacteriol* 2001, **183**:7329-7340.
53. <http://regulondb.ccg.unam.mx/> 2007 [<http://regulondb.ccg.unam.mx/>].
54. <http://www.ecocyc.org/> 2007 [<http://www.ecocyc.org/>].
55. De Hoon MJ, Imoto S, Nolan J, Miyano S: **Open source clustering software.** *Bioinformatics* 2004, **20**:1453-1454.
56. Saldanha AJ: **Java Treeview--extensible visualization of microarray data.** *Bioinformatics* 2004, **20**:3246-3248.
57. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive**

database resource for *Escherichia coli*. *Nucleic Acids Res* 2005, **33**:D334-D337.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp



Apéndice C

Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach

Referencia:

FREYRE-GONZÁLEZ, J.A., ALONSO-PAVÓN, J.A., TREVIÑO-QUINTANILLA, L.G.,
Y COLLADO-VIDES, J. Functional architecture of *Escherichia coli*: new insights
provided by a natural decomposition approach. *Genome Biol* **9**(10):R154 (2008)

Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach

Julio A Freyre-González, José A Alonso-Pavón, Luis G Treviño-Quintanilla and Julio Collado-Vides

Address: Programa de Genómica Computacional, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México. Av. Universidad s/n, Col. Chamilpa 62210, Cuernavaca, Morelos, México.

Correspondence: Julio A Freyre-González. Email: jfreyre@ccg.unam.mx. Julio Collado-Vides. Email: collado@ccg.unam.mx

Published: 27 October 2008

Received: 28 September 2008

Genome Biology 2008, **9**:R154 (doi:10.1186/gb-2008-9-10-r154)

Accepted: 27 October 2008

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2008/9/10/R154>

© 2008 Freyre-González et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Previous studies have used different methods in an effort to extract the modular organization of transcriptional regulatory networks. However, these approaches are not natural, as they try to cluster strongly connected genes into a module or locate known pleiotropic transcription factors in lower hierarchical layers. Here, we unravel the transcriptional regulatory network of *Escherichia coli* by separating it into its key elements, thus revealing its natural organization. We also present a mathematical criterion, based on the topological features of the transcriptional regulatory network, to classify the network elements into one of two possible classes: hierarchical or modular genes.

Results: We found that modular genes are clustered into physiologically correlated groups validated by a statistical analysis of the enrichment of the functional classes. Hierarchical genes encode transcription factors responsible for coordinating module responses based on general interest signals. Hierarchical elements correlate highly with the previously studied global regulators, suggesting that this could be the first mathematical method to identify global regulators. We identified a new element in transcriptional regulatory networks never described before: intermodular genes. These are structural genes that integrate, at the promoter level, signals coming from different modules, and therefore from different physiological responses. Using the concept of pleiotropy, we have reconstructed the hierarchy of the network and discuss the role of feedforward motifs in shaping the hierarchical backbone of the transcriptional regulatory network.

Conclusions: This study sheds new light on the design principles underpinning the organization of transcriptional regulatory networks, showing a novel nonpyramidal architecture composed of independent modules globally governed by hierarchical transcription factors, whose responses are integrated by intermodular genes.

Background

Our understanding of transcriptional control has progressed a long way since Jacob and Monod unraveled the mechanisms that control protein synthesis [1]. These mechanisms allow bacteria to be robust and able to respond to a changing environment. In fact, these regulatory interactions give rise to complex networks [2], which obey organizational principles defining their dynamic behavior [3]. The understanding of these principles is currently a challenge. It has been suggested that decision-making networks require specific topologies [4]. Indeed, there are strong arguments supporting the notion of a modular organization in the cell [5]. A module is defined as a group of cooperating elements with one specific cellular function [2,5]. In genetic networks, these modules must comprise genes that respond in a coordinated way under the influence of specific stimuli [5-7].

Topological analyses have suggested the existence of hierarchical modularity in the transcriptional regulatory network (TRN) of *Escherichia coli* K-12 [7-10]. Previous works have proposed methodologies from which this organization could be inferred [9-11]. These works suggested the existence of a pyramidal top-down hierarchy. Unfortunately, these approaches have proven inadequate for networks involving feedback loops (FBLs) or feedforward motifs (FFs) [10,11], two topological structures relevant to the organization and dynamics of TRNs [2,12-16]. In addition, module identification approaches frequently have been based on clustering methods, in which each gene must belong to a certain module [6,7,17]. Although analyses using these methods have reported good results, they have revealed two inconveniences: they rely on certain parameters or measurement criteria that, when modified, can generate different modules; and a network with scale-free properties foresees the existence of a small group of strongly connected nodes (hubs), but to what modules do these hubs belong? Maybe they do not belong to a particular module, but do they serve as coordinators of module responses?

Alternatively, we developed a novel algorithm to enumerate all the FBLs comprising two or more nodes existing in the TRN, thus providing the first systems-level enumeration and analysis of the global presence and participation of FBLs in the functional organization of a TRN. Our results show, contrary to what has been previously reported [9,10], the presence of positive and negative FBLs bridging different organizational levels of the TRN of *E. coli*. This new evidence highlights the necessity to develop a new strategy for inferring the hierarchical modular organization of TRNs.

To address these concerns, in this work we propose an alternative approach founded on inherent topological features of hierarchical modular networks. This approach recognizes hubs and classifies them as independent elements that do not possess a membership to any module, and reveals, in a natural way, the modules comprising the TRN by removing the

hubs. This methodology enabled us to reveal the natural organization of the TRN of *E. coli*, where hierarchical transcription factors (hierarchical TFs) govern independent modules whose responses are integrated at the promoter level by intermodular genes.

Results

The TRN of *E. coli* K-12 is the best characterized of all prokaryote organisms. In this work, the TRN was reconstructed using mainly data obtained from RegulonDB [18], complemented with new sigma factor interactions gathered from a literature review on transcriptional regulation mediated by sigma factors (see Materials and methods). In our graphical representation, each node represents a gene and each edge a regulatory interaction. The TRN used in this work was represented as a directed graph comprising 1,692 nodes (approximately 40% of the total genes in the genome) with 4,301 arcs (directed regulatory interactions) between them. Neglecting autoregulation and the directions of interactions between genes, the average shortest path of the network was 2.68, supporting the notion that the network has small-world properties [2]. The connectivity distribution of the TRN tends to follow a power law, $P(k) \sim k^{-2.06}$, which implies that it has scale-free properties (Figure S1a in Additional data file 1). In addition, the distribution of the clustering coefficient shows a power law behavior, with $C(k) \sim k^{-0.998}$ (Figure S1b in Additional data file 1). In the latter, the exponent value is virtually equal to -1, strongly suggesting that the network possesses a hierarchical modular architecture [2,19].

The TRN has FBLs that involve mainly global and local TFs

The pioneering theoretical work of René Thomas [15,16,20,21] and experimental work [14,22] have shown the topological and dynamic relevance of feedback circuits (FBLs). In regulatory networks, FBLs are associated with biological phenomena, such as homeostasis, phenotypic variability, and differentiation [14,16,20,22]. Previous studies have established the importance of FBLs for both the modularity of regulatory networks [21] and their dynamics [14-16,20,22]. Ma *et al.* [9,10] suggested that FBLs that exist in the TRN of *E. coli* are not relevant for the topological organization of the TRN. Using an *E. coli* TRN reconstruction that included sigma factor interactions, they claimed to have identified only seven two-node FBLs (that is, FBLs with the structure $A \rightarrow B \rightarrow A$) and no FBLs comprising more than two nodes [10]. However, given that their approach requires, *a priori*, an acyclic network [23], genes involved in an FBL are placed in the same hierarchical layer, under the argument that they are in the same operon [10].

To get a global image of FBLs, an original algorithm was developed and implemented (see Materials and methods). This algorithm allowed us to enumerate all FBLs, comprising two or more nodes, existing in the TRN (Table 1). A total of 20

Table 1**FBLs identified in the TRN of *Escherichia coli***

Type of FBL	Number of genes	Genes	Interactions	Are genes in the same operon?
+	2	<i>arcA fnr</i>	- -	No
-	2	<i>arcA fnr</i>	- +	No
-	2	<i>gadX hns</i>	+ -	No
+	2	<i>gadX rpoS</i>	+ +	No
-	2	<i>gutM srlR</i>	+ -	Yes
-	2	<i>lexA rpoD</i>	- +	No
-	2	<i>marA marR</i>	+ -	Yes
-	2	<i>marA rob</i>	- +	No
+	2	<i>rpoD rpoH</i>	+ +	No
+	3	<i>crp rpoH rpoD</i>	+ + +	No
-	3	<i>crp rpoH rpoD</i>	- + +	No
-	3	<i>cytR rpoH rpoD</i>	- + +	No
+	3	<i>gadE gadX rpoS</i>	+ + +	No
+	3	<i>marA rob marR</i>	- + -	No
+	3	<i>rpoD rpoN rpoH</i>	+ + +	No
-	4	<i>cpxR rpoE rpoH rpoD</i>	- + + +	No
-	4	<i>crp cytR rpoH rpoD</i>	+ - + +	No
-	5	<i>IHF fis hns gadX rpoS</i>	+ + - + +	No
-	5	<i>argP dnaA rpoH rpoD phoB</i>	+ - + + +	No
-	5	<i>cpxR rpoE rpoN rpoH rpoD</i>	- + + + +	No

Eighty percent of the total FBLs involve, at least, one global TF. The longest FBL comprises five TFs. Only two FBLs have genes encoded in the same operon, contrary to what was previously reported by Ma et al. [10], thus suggesting that these FBLs work as uncoupled systems. In addition, seven positive FBLs were identified, which potentially could give rise to multistability.

FBLs were found: 9 (45%) with two nodes and 11 (55%) with more than two nodes. It was found that FBLs in the TRN tend mainly to connect global TFs with local TFs (at this point we used the definitions of global and local TFs given by Martinez-Antonio and Collado-Vides [24]). It was also found that only 2 FBLs (10%) are located in the same operon, 4 (20%) involve only local TFs, 10 (50%) involve both global and local TFs, and 6 (30%) involve only global TFs. We observed a couple of dual FBLs, the first comprising *arcA* and *fnr* and the second comprising *crp*, *rpoH*, and *rpoD*. These dual FBLs comprise dual regulatory interactions, thus giving rise to two overlapping FBLs, one positive and the other negative. However, each of these overlapping FBLs was enumerated as a different FBL, given that the dynamic behaviors of positive and negative FBLs are quite different.

Nodes of hierarchical modular networks can be classified into one of two possible classes: hierarchical or modular nodes

The characteristic signature of hierarchical modularity in a network is the clustering coefficient distribution, which must follow a power law, $C(k) \sim k^{-1}$ [2,19]. This coefficient measures how much the nearest neighbors of a TF affect each other, thus providing a measure of the modularity for the TF. In the extreme limits of the clustering coefficient distribution, nodes follow two apparently contradictory behaviors [2] (Figure 1a).

At low connectivity, nodes show high clustering coefficients. On the contrary, at high connectivity, nodes show low clustering coefficients. Previous work with the *E. coli* metabolic network [17] suggested that the first behavior is due to network modularity but the latter is due to the presence of hubs. In addition, a previous analysis of the TRN of *Saccharomyces cerevisiae* found that direct connections between hubs tend to be suppressed while connections between hubs and poorly connected nodes are favored [25], suggesting that modules tend to be organized around hubs. This evidence suggested two possible roles for nodes: nodes that shape modules (they have low connectivity and a high clustering coefficient, which will be called modular nodes); and nodes that bridge modules (they have high connectivity and a low clustering coefficient, which will be called hierarchical nodes), establishing in this way a hierarchy that dynamically governs module responses.

It can be observed in $C(k)$ distributions following a power law that initially slight increments in the connectivity value (k) will make the clustering coefficient decrease quickly. However, eventually a point is reached where the situation is inverted. Then, a larger increment in connectivity is needed to make the clustering coefficient decrease. From this behavior the existence of an equilibrium point in the $C(k)$ distribution is inferred, where the variation of the clustering

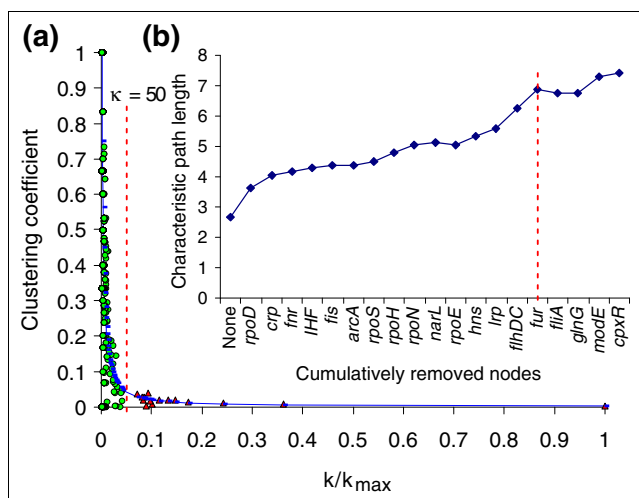


Figure 1
 Identification of hierarchical and modular nodes. **(a)** Distribution of the clustering coefficient, $C(k)$, and calculated κ value. The blue line represents the $C(k)$ power law. The dashed red line indicates the κ value obtained for this $C(k)$ distribution. Red triangles represent hierarchical nodes, while green circles indicate modular nodes. **(b)** The characteristic path length after cumulative removal of all hierarchical nodes and some modular ones. The red dashed line indicates the sudden change in the original increasing tendency when the last hierarchical TFs (*flhDC* and *fur*) were removed. This suggests that the removal of hierarchical nodes broke the connections bridging modules, thus disintegrating the TRN.

coefficient is equal to the variation of connectivity but with the opposite sign:

$$dC(k)/dk = -1$$

Solving this equation gives the connectivity value (κ) where such an equilibrium is reached (see Material and methods). Herein, κ is proposed as a cutoff value that disaggregates the set of nodes into two classes (Figure 1a). Hierarchical nodes are those with connectivity greater than κ . On the other hand, modular nodes are those with connectivity less than κ .

The κ value can be calculated with the formula (see Materials and methods):

$$\kappa = \alpha + \sqrt{\alpha\gamma} \cdot k_{max}$$

This formula relates the equilibrium point (κ) of the $C(k)$ distribution with its exponent ($-\alpha$) and its proportionality constant (γ). It has been shown that in 'ideal' hierarchical modular networks the exponent $-\alpha$ is equal to -1 [2,19]. Thus, substituting this value into the previous formula gives:

$$\kappa = \sqrt{\gamma} \cdot k_{max}$$

Therefore, in 'ideal' networks the equilibrium point depends exclusively on the proportionality constant of $C(k)$. To the best of our knowledge, this is the first time that a relevant top-

ological interpretation has been given to the proportionality constant.

Hierarchical nodes correlate highly with known global TFs

After computing the κ value for the TRN, the following 15 TFs were identified as hierarchical nodes (nodes with connectivity greater than 50; Figure 1): *RpoD* (σ^{70}), *CRP*, *FNR*, *IHF*, *Fis*, *ArcA*, *RpoS* (σ^{38}), *RpoH* (σ^{32}), *RpoN* (σ^{54}), *NarL*, *RpoE* (σ^{24}), *H-NS*, *Lrp*, *FlhDC*, and *Fur*. All these TFs, except *FlhDC* and *Fur*, have been reported several times as global TFs [13,24,26,27]. In addition, Madan Babu and Teichmann [27] have previously reported *Fur* as a global TF. *FlhDC* and *Fur* regulate genes with several physiological functions, which makes them potential candidates to be global TFs [28]. *Fur* regulates amino acid biosynthesis genes [29], Fe^+ transport [30-32], flagellum biosynthesis [29], the Krebs cycle [33], and $Fe-S$ cluster assembly [34]. On the other hand, *FlhDC* mainly regulates membrane genes. Nevertheless, these genes take part in several physiological functions, such as motility [35], glutamate [36] and galactose [37] transport, anaerobiosis [37], and 3-P-glycerate degradation [37]. When connectivity was less than κ , genes encoding local TFs (herein called modular TFs) and structural genes were found. *FliA* (σ^{28}) and *FecI* (σ^{19}) sigma factors are in the group of modular nodes. This is understandable, because both respond to very specific cell conditions (flagellum biosynthesis and citrate-dependent Fe^+ transport, respectively), and they affect the transcription of few genes (43 and 6 genes, respectively). These results suggest that the κ value may be a good predictor for global TFs.

Hierarchical nodes act as bridges keeping modules connected

The characteristic path length is defined as the average of the shortest paths between all pairs of nodes in a network. It is a measure of the global connectivity of the network [38]. Using an *in silico* strategy, the effect on the characteristic path length when attacking hierarchical nodes was analyzed. In order to do this, all hierarchical nodes and some modular ones were removed one by one in decreasing order of connectivity (Figure 1b). The removal of hierarchical nodes increased, following a linear tendency, the characteristic path length from 2.7 to 6.9. However, when the last two hierarchical nodes (*flhDC* and *fur*) were removed, a sudden change was observed in the tendency, followed by a stabilization when some modular nodes were removed, therefore supporting the idea that removal of hierarchical nodes disintegrates the TRN by breaking the bridges that keep modules together.

Identification of modules in the TRN

The removal of hierarchical nodes revealed 62 subnetworks or modules (see Materials and methods; Additional data file 2) and left 691 isolated genes. An analysis of the biological function of the isolated genes showed that many of them are elements of the basal machinery of the cell (tRNAs and its charging enzymes, DNA and RNA polymerases, ribosomal

proteins and RNAs, enzymes of the tricarboxylic acid cycle and respiratory chain, DNA methylation enzymes, and so on). The regulation of these genes, whose products must be constantly present in the cell, is mediated only by hierarchical TFs. One of the identified modules (module 5) comprises 606 genes (35% of the analyzed TRN). This megamodule suggested the existence of other elements, in addition to hierarchical nodes, that connect modules. We know that a TRN that has been reconstructed while neglecting structural genes does not show the existence of a megamodule (JAF-G, unpublished data). Therefore, an intermodular gene was defined as a structural gene whose expression is modulated by TFs belonging to two or more submodules. To identify these intermodular genes, the megamodule was isolated and structural genes removed. This revealed the submodule cores (islands of modular TFs) shaping the megamodule (see Materials and methods). The megamodule comprises 39 submodules connected by the regulation of 136 intermodular genes, which are organized into approximately 55 transcriptional units (Additional data file 3).

To determine the biological relevance of the theoretically identified modules, two independent analyses were performed. On the one hand, one of us (LGT-Q) used biological knowledge to perform a manual annotation of identified modules. On the other hand, two of us (JAF-G and JAA-P) made a blind-automated annotation based on functional class, according to the MultiFun system [39], that showed a statistically significant enrichment (p -value < 0.05 ; see Materials and methods). Both analyses showed similar conclusions. The blind-automated method found that 97% of modules show enrichment in terms of functional classes. However, it was observed that the manual analysis added subtle details that were not evident in the automated analysis due to incompleteness in the MultiFun system (Additional data file 2). At the module level, it was found that *E. coli* mainly has systems for carbon source catabolism, cellular stress response, and ion homeostasis. In addition, it was found that the 39 submodules comprising the megamodule could be grouped according to their biological functions into seven regions interconnected by intermodular genes (Figure 2). The most interconnected regions involve nitrogen and sulfur assimilation, carbon source catabolism, cellular stress response, respiration forms, and oxidative stress.

Inference of the hierarchy governing the TRN

For more than 20 years it has been recognized that regulatory networks comprise complex circuits with different control levels. This makes them able to control different subroutines of the genetic program simultaneously [28,40]. Recently, global topological analyses have suggested the existence of hierarchical modularity in TRNs [2,7,8]. Previous works proposed methodologies to infer this hierarchical modular organization [9-11]. Unfortunately, the previous methodological approaches have been shown to be inadequate to deal with FFs and FBLs [10,11], two relevant topological struc-

tures. On the other hand, biological conclusions obtained with these approaches were counterintuitive, as they placed, in the highest hierarchical layers, TFs that respond to very specific conditions of the cell and which, therefore, lack pleiotropic effects.

Gottesman [28] defined a global TF as one that: regulates many genes; entails regulated genes that participate in more than one metabolic pathway; and coordinates the expression of a group of genes when responding to a common need (for detailed definitions of global and local TFs please refer to the work of Martinez-Antonio and Collado-Vides [24]). Based on Gottesman's ideas, it could be asked if a modular organization requires a hierarchy to coordinate module responses. To address this concern, based on the definition proposed by Gottesman and using the concept of pleiotropy, a methodology to infer the hierarchy governing the TRN was developed. For this methodology, nodes belonging to the same module were shrunk into a single node, and a bottom-up approach was used (see Materials and methods). This approach places each hierarchical TF in a specific layer, depending on two factors: theoretical pleiotropy (the number of regulated modules and hierarchical TFs); and the presence of direct regulation over hierarchical TFs placed in the immediate lower hierarchical layer. This second factor was taken into account because a hierarchical TF may indirectly propagate its control to other modules, by changing the expression pattern of a second hierarchical TF that directly controls them. Given that a hierarchical layer does not depend on the number of genes regulated by a hierarchical TF, but on the number of modules, it is worth mentioning that this approach is not based on connectivity. Therefore, given that each module is in charge of a different physiological response, it can be argued that this approach is founded on pleiotropy.

Five global chains of command were found, showing the regulatory interactions between hierarchical TFs (Figure 3). Each of the chains of command is in charge of global functions in the cell. In addition, in the highest hierarchical layers, the presence of six hierarchical TFs was observed, three of them (RpoD, CRP, and FNR) governing more than one of these global chains of command. The expression of IHF, in spite of the fact that it only governs one global chain of command, can be affected by a different chain from a lower hierarchy (RpoS) [41]. Each of these TFs sends signals of general interest to a large number of genes in the cell. RpoD (σ^{70}) is the housekeeping sigma factor, and it can indicate to the cellular machinery the growth phase of the cell or the lack of any stress [42]. CRP-cAMP alerts the cell to low levels of energy uptake, allowing a metabolic response [43]. IHF (besides Fis and H-NS) senses DNA supercoiling, thus indirectly sensing many environmental conditions (growth phase, energy level, osmolarity, temperature, pH, and so on) that affect this DNA property [44]. This supports the idea that DNA supercoiling itself might act as a principal coordinator of global gene expression [45,46]. Finally, FNR senses extracellular oxygen

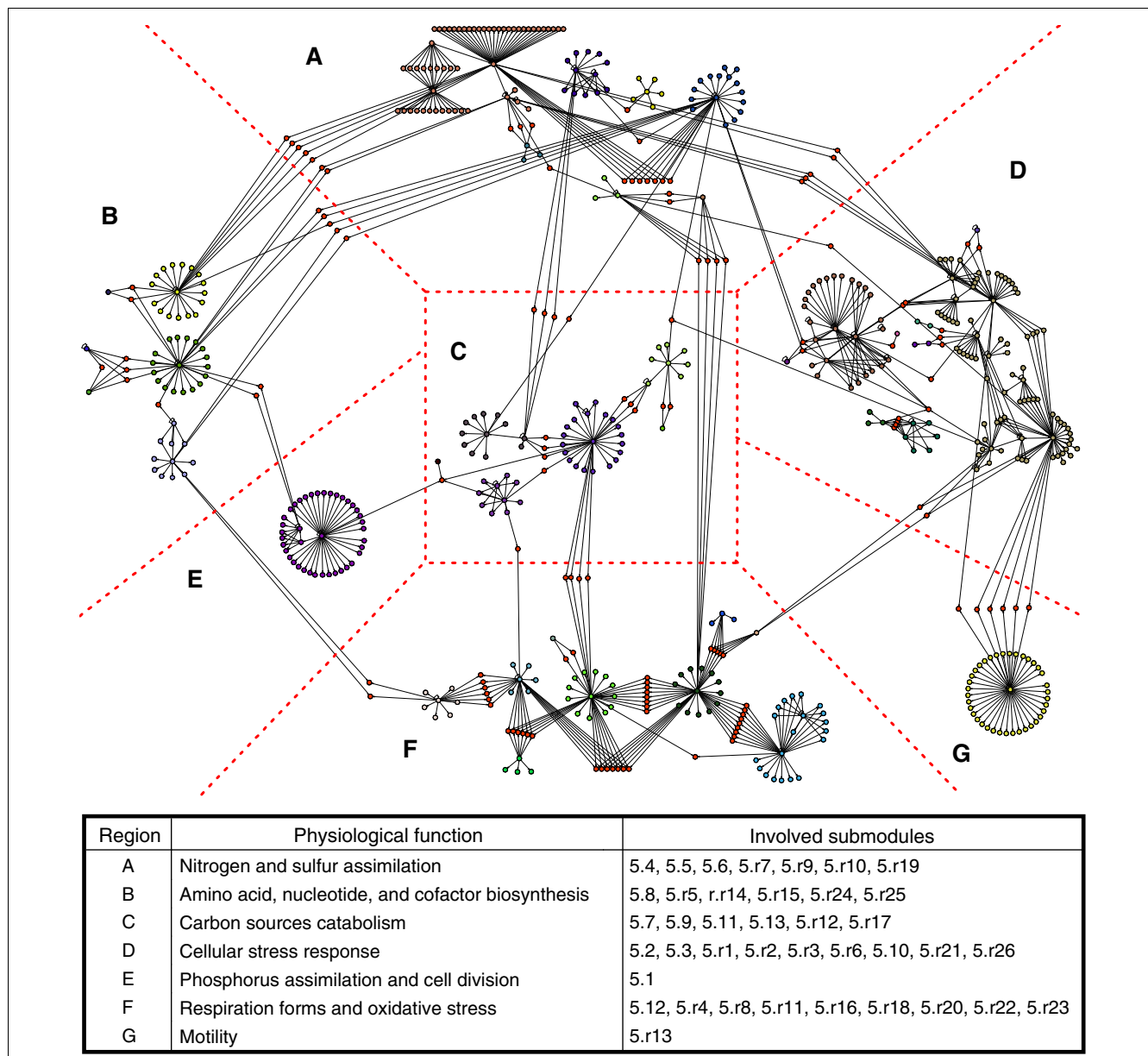


Figure 2
 Empirical grouping, into seven regions, of submodules comprising the megamodule. Each color represents a submodule, while intermodular genes are shown in orange. Intermodular genes are placed inside the region that best associates with its most important physiological function. For example, the intermodular gene *amtB*, positively regulated by NtrC (region A) and GadX (region D), encodes an ammonium transporter under acidic growing conditions. Therefore, this gene was placed in the nitrogen and sulfur assimilation region (region A).

levels, permitting, through coregulation with ArcA and NarL, a proper respiratory response [47,48]. RpoN, with σ^{54} -dependent activators, controls gene expression to coordinate nitrogen assimilation [49]. RpoE (σ^{24}) reacts to stress signals outside the cytoplasmic membrane by transcriptional activation of genes encoding products involved in membrane protection or repair [50].

FFs mainly bridge modules shaping the TRN hierarchical backbone

A remarkable feature of complex networks is the existence of topological motifs [12,13]. It has been previously suggested that they constitute the building blocks of complex networks [8,12]. Nevertheless, recent studies have provided evidence that overabundance of motifs does not have a functional or evolutionary counterpart [51-54]. Indeed, some studies have suggested that motifs could be by-products of biological network organization and evolution [52,53,55]. In particular,

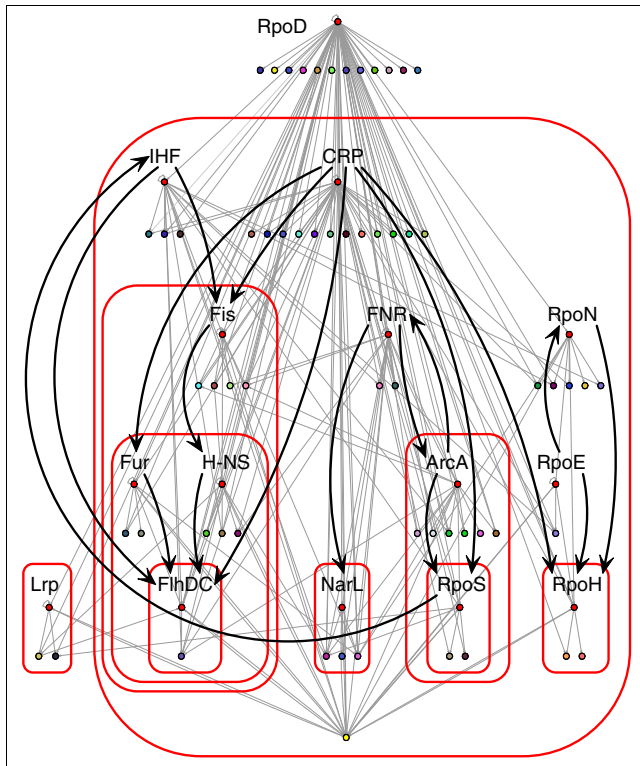


Figure 3
Hierarchical modular organization map of subroutines comprising the genetic program in *E. coli*. Each color represents a module, while hierarchical TFs are shown in red. Black arrows indicate the regulatory interactions between hierarchical TFs. For the sake of clarity, RpoD interactions are not shown, and the megamodule is shown as a single yellow node at the bottom. However, according to our data, RpoD affects the transcription of all hierarchical TFs, except RpoE, while RpoD, RpoH, and LexA (a modular TF) could affect RpoD expression. Red rounded-corner rectangles bound hierarchical layers. The presence of five global chains of command is noted: host/free-life sensor and type I fimbriae (Lrp); replication, recombination, pili, and extracytoplasmic elements (Fis, Fur, H-NS, FlhDC); respiration forms (NarL); starvation stress (ArcA, RpoS); and heat shock (RpoH). Lrp appears disconnected from other hierarchical TFs because, to date, it is only known that RpoD, Lrp, and GadE (a modular TF) modulate its expression.

work by Ingram *et al.* [54] has shown that the bi-fan motif can exhibit a wide range of dynamic behaviors. Given that, we concentrated our analysis on three-node motifs.

We identified the entire repertoire of three-node network motifs present in the *E. coli* TRN by using the mfinder program [12]. Thus, we identified two three-node network motifs: the FF; and an alternative version of an FF merging an FBL between the regulatory nodes. It suggests that the FF is the fundamental three-node motif in the *E. coli* TRN. In order to analyze FF participation in the hierarchy inferred by our methodology, the effect of the removal of hierarchical nodes on the total number of FFs in the TRN was analyzed (Figure 4a). The fraction of remaining FFs after cumulative removal of hierarchical nodes, in decreasing connectivity order, was computed. It was found that the sole removal of *rpoD* (σ^{70})

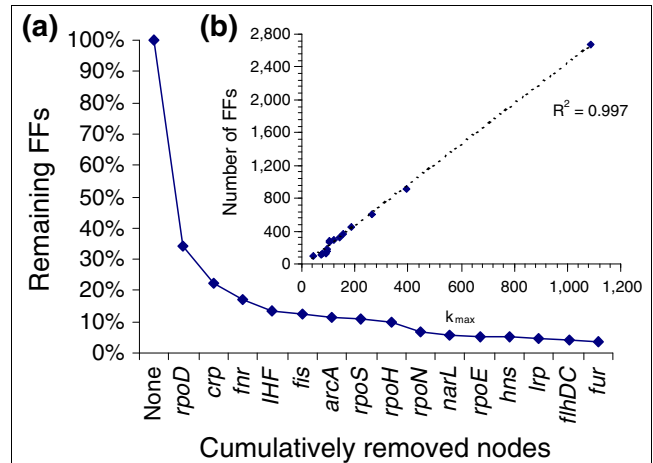


Figure 4
FFs bridge modules and shape the backbone of the hierarchy governing the TRN. **(a)** Remaining TFs after cumulative removal of hierarchical nodes. The removal of all hierarchical nodes decreased to 3.5% the total FFs. **(b)** Correlation between FF number and maximum connectivity for each attacked network. The FF number is proportional to the number of links of the most-connected hierarchical node, thus suggesting that FFs are the backbone of the hierarchy in the TRN.

and *crp*, the two most-connected hierarchical nodes in the TRN, decreased to 22% the total FFs. However, the removal of all hierarchical nodes decreased the total FFs to 3.5%, in agreement with previous work suggesting that FFs tend to cluster around hubs [56]. Our results showed that 96.5% of the total FFs are in the TRN bridge modules, while the remaining 3.5% are within modules. This evidence suggests that the FF role is to bridge modules, shaping a hierarchical structure governed by hierarchical TFs.

The correlation between FF number and maximum connectivity (number of links of the most-connected node, k_{max}) for each attacked network was analyzed (Figure 4b). It was found that the FF number linearly correlated with the maximum connectivity. As hierarchical nodes were removed, the FF number decreased proportionally with the maximum connectivity of the corresponding attacked network. All this shows that hierarchical TFs are intrinsically related to FFs, suggesting that, in addition to bridging modules, FFs are the backbone of the hierarchical organization of the TRN.

Discussion

Contrary to what has been previously reported [9,10], we found FBLs involving different hierarchical layers, which implies that the expression of some hierarchical TFs also may depend on modular TFs, thus allowing the reconfiguration of the regulatory machinery in response to the fine environmental sensing performed, through allosterism, by modular TFs. On the other hand, a network with FBLs poses a paradox when inferring its hierarchy. Given the circular nature of interactions, what nodes should be placed in a higher hierar-

chical layer? This paradox was solved using the κ value to identify hierarchical and modular elements and then using the theoretical pleiotropy to infer the hierarchy governing the TRN.

Global TFs have been proposed using diverse relative measures [9,10,13,24,27,28]; unfortunately, currently there is not a consensus on the best criteria to identify them. Gottesman's seminal paper [28] was the first to define the properties for which a TF should be considered a global TF. Martínez-António and Collado-Vides [24] conducted a review and analyzed several properties, searching for diagnostic criteria to identify global TFs. Nevertheless, while these authors did shed light on relevant properties that could contribute to identification of global TFs, they did not reach any explicit diagnostic criteria. The κ value showed high predictive power, as all known global TFs were identified, and even more, the existence of two new global TFs is proposed: FlhDC and Fur. Recently, an analysis of the TRN of *Bacillus subtilis* supported the predictive ability of this method (JAF-G, unpublished data), offering the possible first mathematical criterion to identify global TFs in a cell. This criterion allowed us to show that, in spite of its apparent complexity, the TRN of *E. coli* possesses a singular elegance in the organization of its genetic program. Only 15 hierarchical TFs (0.89% of the total nodes) coordinate the response of the 100 identified modules (50.23% of the total nodes). All the modules identified by Resendis-António *et al.* [7] were recovered by our methodology. However, given that in this study the TRN includes structural genes, we could identify 87 new modules. Therefore, our approach allows fine-grain identification of modules, for example, modules responsible for catabolism of specific carbon sources. There are 691 genes (40.84% of the total nodes) that mainly encode cellular basal elements. The existence of one megamodule led us to define intermodular genes and to identify 136 of them (8.04% of the total nodes). It was found that submodules with similar functions tend to agglomerate into seven regions, thus shaping the megamodule. Therefore, at a TRN level, data processing follows independent casual chains for each module, which are globally governed by hierarchical TFs. Thus, hierarchical TFs coordinate the cellular system responses as a whole by letting modules get ready to react in response to external stimuli of common interest, while modules retain their independence, responding to stimuli of local interest. On the other hand, intermodular genes integrate, at the promoter level, the incoming signals from different modules. These promoters act as molecular multiplexers, integrating different physiological signals in order to make complex decisions. Examples of this are the *aceBAK* and *carAB* operons. The *aceBAK* operon encodes glyoxylate shunt enzymes. The expression of this operon is modulated by FruR [57] (module 5.11, gluconeogenesis) and IclR [58] (module 5.13, aerobic fatty acid oxidation pathway). This operon could integrate the responses of these two modules in order to keep the balance between energy production from fatty acid oxidation and gluconeogenesis activation for biosynthesis of building blocks.

On the other hand, the *carAB* operon encodes a carbamoyl phosphate synthetase. The expression of this operon is controlled by PurR [59] (module 5.r25, purine and pyrimidine biosynthesis), ArgR [60] (module 5.r5, L-ornithine and L-arginine biosynthesis), and PepA [59] (5.r24, carbamoyl phosphate biosynthesis and aminopeptidase A/I regulation). This is an example where different modules could work as coordinators of a shared resource. The promoter of this operon could integrate the responses of the modules to coordinate the expression of an enzyme whose product, carbamoyl phosphate, is a common intermediary for the *de novo* biosynthesis of pyrimidines and arginine. This evidence shows a novel nonpyramidal architecture in which independent modules are globally governed by hierarchical transcription factors while module responses are integrated at the promoter level by intermodular genes.

The clustering coefficient is a strong indicator of modularity in a network. It also quantifies the presence of triangular substructures. The TRN shows a high average clustering coefficient, implying a high amount of triangular substructures. Indeed, the probability of a node being a common vertex of n triangles decreases as the number of involved triangles increases, following the power law $T(n) \sim n^{-1.95}$ (Figure S1c in Additional data file 1). In other words, if a node is arbitrarily chosen, the probability of it being the vertex of a few triangles is high. This also implies that many triangles have as a common vertex a small group of nodes. On the other hand, in a directed graph there are only two basic triangular substructures: FFs and three-node FBLs. By merging two-node FBLs with these two triangular substructures, it is possible to create variations of them. It was found that the number of two-node and three-node FBLs (eight and five FBLs, respectively) was much lower than the total number of FFs (2,674 FFs). These results imply that triangular substructures are mainly FFs or variations of them. Besides, FFs mainly comprise, at least, one hierarchical node [56] (Figure 4). This is in agreement with the observation that many triangles possess as a common vertex a small group of nodes. Here it was shown that hierarchical nodes and their interactions shape the backbone of the TRN hierarchy. Therefore, FFs are strongly involved in the hierarchical modular organization of the TRN of *E. coli*, where they act as bridges connecting genes with diverse physiological functions. Resendis-António *et al.* [7] showed that FFs are mainly located within modules. Nevertheless, given that in this study it was determined that hubs do not belong to modules, it was found that FFs shape the hierarchy of the TRN bridging modules in a hierarchical fashion. This supports the findings of Mazurie *et al.* [52], showing that FFs are a consequence of the network organization and they are not involved in specific physiological functions.

Conclusions

The study of the topological organization of biological networks is still an interesting research topic. Methodologies for

node classification and natural decomposition, such as the one proposed herein, allow identification of key components of a biological network. This approach also enables the analysis of complex networks by using a zoomable map approach, helping us understand how their components are organized in a meaningful way. In addition, component classification could shed light on how different networks (transcriptional, metabolic, protein-protein, and so on) interface with each other, thus providing an integral understanding of cellular processes. The herein-proposed approach has promising applications for unraveling the functional architecture of the TRNs of other organisms, allowing us to gain a better understanding of their key elements and their interrelationships. In addition, it provides a large set of experimentally testable hypotheses, from novel FBLs to intermodular genes, which could be a useful guide for experimentalists in the systems biology field. Finally, network decomposition into modules with well-defined inputs and outputs, and the suggestion that they process information in independent casual chains governed by hierarchical TFs, would eventually help in the isolation, and subsequent modeling, of different cellular processes.

Materials and methods

Data extraction and TRN reconstruction

To reconstruct the TRN, structural genes, sigma factor-encoding genes, and regulatory protein-encoding genes were included (the full data set is available as Additional data file 4). Two flat files with data (NetWorkSet.txt and SigmaNetWorkSet.txt) were downloaded from RegulonDB version 5.0 [18,61]. From the NetWorkSet.txt file, 3,001 interactions between regulatory proteins and regulated genes were obtained. From the SigmaNetWorkSet.txt file, 1,488 interactions between sigma factors and their transcribed genes were obtained. Next, this information was complemented with 81 new interactions found in a literature review of transcribed promoters by the seven known sigma factors of *E. coli* (these interactions account for 5.4% of the total sigma factor interactions in the reconstructed TRN and currently are integrated and available in RegulonDB version 6.1). The criteria used to gather the additional sigma factor interactions from the literature were the same as those used by the RegulonDB team of curators. In our graphic model, sigma factors were included as activator TFs because their presence is a necessary condition for transcription to occur. Indeed, some works [62-64] have shown that there are TFs that are able to interact with free polymerase before binding to a promoter, in a way reminiscent of the mechanism used by sigma factors. To avoid duplicated interactions, heteromeric TFs (for example, IHF encoded by *ihfA* and *ihfB* genes, HU encoded by *hupA* and *hupB*, FlhDC encoded by *flhC* and *flhD*, and GatR encoded by *gatR_1* and *gatR_2*) were represented as only one node, given that there is no evidence indicating that any of the subunits have regulatory activity *per se*.

Software

For the analysis and graphic display of the TRN, Cytoscape [65] was used. To identify FFs, the mfinder program [12] was used. To calculate κ values, computational annotations, and other numeric and informatics tasks, Microsoft Excel and Microsoft Access were used.

Algorithm for FBL enumeration

First, The TRN was represented, neglecting autoregulation, as a matrix of signs (**S**). Thus, each $S_{i,j}$ element could take a value in the set $\{+, -, \mathbf{D}, \mathbf{0}\}$, where '+' means that *i* activates *j* transcription, '-' means that *i* represses *j* transcription, **D** means that *i* has a dual effect (both activator and repressor) over *j*, and **0** means that there is no interaction between *i* and *j*. Second, All nodes with incoming connectivity or outgoing connectivity equal to zero were removed. Third, the transitive closure matrix of the TRN (**M**) was computed using a modified version of the Floyd-Warshall algorithm [23]. Each $M_{i,j}$ element could take a value in the set $\{0, 1\}$, where **0** means that there is no path between *i* and *j* and **1** means that, at least, there is one path between *i* and *j*. Fourth, for each $M_{i,i}$ element equal to **1**, a depth-first search beginning at node *i* was done, marking each visited node. The depth-first search stopping criterion relies on two conditions: first, when node *i* is visited again, that is, an FBL ($i \rightarrow \dots \rightarrow i$) is identified; second, when a previously visited node, different from *i*, is visited again. Fifth, isomorphic subgraphs were discarded from identified FBLs.

κ value calculation

For each node in the TRN, connectivity (as a fraction of maximum connectivity, k_{\max}) and the clustering coefficient were calculated. Next, the $C(k)$ distribution was obtained using least-squares fitting. Given $C(k) = \gamma k^{-\alpha}$, the equation:

$$dC(k)/dk = -1$$

has as its solution the formula:

$$\kappa = \alpha + 1 \sqrt{\alpha \gamma} \cdot k_{\max}.$$

Module identification

The algorithm to identify modules used a natural decomposition approach. First, the κ value was calculated for the TRN of *E. coli*, yielding the value of 50. Then, all hierarchical nodes (nodes with $k > \kappa$) were removed from the network. Therefore, the TRN breaks up into isolated islands, each comprising interconnected nodes. Finally, each island was considered a module.

Identification of submodules and intermodular genes comprising the megamodule

The megamodule was isolated and all structural genes were removed, breaking it up into isolated islands. Next, each island was identified as a submodule. Finally, all the removed structural genes and their interactions were added to the net-

work according to the following rule: if a structural gene *G* is regulated only by TFs belonging to submodule *M*, then gene *G* was added to submodule *M*. On the contrary, if gene *G* is regulated by TFs belonging to two or more submodules, then gene *G* was classified as an intermodular gene.

Manual annotation of identified modules

Manual annotation of physiological functions of identified modules was done using the biological information available in RegulonDB [18,61] and EcoCyc [66,67].

Computational annotation of identified modules

Each gene was annotated with its corresponding functional class according to Monica Riley's MultiFun system, available via the GeneProtEC database [39,68]. Next, *p*-values, as a measure of randomness in functional class distributions through identified modules, were computed based on the following hypergeometric distribution: let $N = 1,692$ be the total number of genes in the TRN and A the number of these genes with a particular F annotation; the *p*-value is defined as the probability of observing, at least, x genes with an F annotation in a module with n genes. This *p*-value is determined with the following formula:

$$p\text{-value} = \sum_{i=x}^n \frac{\binom{A}{i} \binom{N-A}{n-i}}{\binom{N}{n}}$$

Thus, for each module, the *p*-value of each functional assignment present in the module was computed. The functional assignment of the module was the one that showed the lowest *p*-value, if and only if it was less than 0.05.

Inference of the hierarchy

To infer the hierarchy, a shrunken network was used, where each node represents a module or a hierarchical element. Hierarchical layers were created following a bottom-up approach and considering the number of regulated elements (theoretical pleiotropy) by hierarchical nodes, neglecting autoregulation, as follows. First, all nodes belonging to the same module were shrunk into a single node. Second, for each hierarchical element, the theoretical pleiotropy was computed. Third, the hierarchical element with lower theoretical pleiotropy and its regulated modules were placed in the lower hierarchical layer. Fourth, each hierarchical element and its regulated modules were added one by one in order of increasing theoretical pleiotropy. Fifth, if the added hierarchical element regulated, at least, one hierarchical element in the immediate lower layer, a new hierarchical layer was created; otherwise, the hierarchical element was added to the same hierarchical layer.

Abbreviations

FBL, feedback loop; FF, feedforward topological motif; TF, transcription factor; TRN, transcriptional regulatory network.

Authors' contributions

JAF-G and JC-V designed the research; JAF-G conceived the approach and designed algorithms; JAA-P and LGT-Q contributed to the algorithm to infer hierarchy; JC-V proposed the computational annotation of modules; JAF-G, JAA-P, and LGT-Q performed research; JAF-G, JAA-P, and LGT-Q contributed analytic tools; JAF-G, JAA-P, and LGT-Q analyzed data; JAF-G, JAA-P, LGT-Q, and JC-V wrote the paper.

Additional data files

The following additional data are available. Additional data file 1 contains the topological properties of the transcriptional regulatory network of *E. coli*. Additional data file 2 is a table listing all the modules identified in this study and their manual and computational annotations. Additional data file 3 contains a listing of all the intermodular genes found in this study, their biological descriptions and roles as integrative elements. Additional data file 4 is a flat file with the full data set for the *E. coli* transcriptional regulatory network reconstructed for our analyses as described in the Materials and methods section.

Acknowledgements

We thank Veronika E Rohen for critical reading of the statistical methodology used for the computational annotation of modules. We thank Mario Sandoval for help in codifying the algorithm for FBL enumeration. We also thank Patricia Romero for technical support. JAF-G was supported by PhD fellowship 176341 from CONACyT-México and was a recipient of a graduate complementary fellowship from DGEP-UNAM. This work was partially supported by grants 47609-A from CONACyT, IN214905 from PAPIIT-UNAM, and NIH ROI GM071962-04 to JC-V.

References

- Jacob F, Monod J: **Genetic regulatory mechanisms in the synthesis of proteins.** *J Mol Biol* 1961, **3**:318-356.
- Barabási AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nat Rev Genet* 2004, **5**:101-113.
- Varianto EA, McCoy JH, Lipson H: **Networks, dynamics, and modularity.** *Phys Rev Lett* 2004, **92**:188701.
- Oosawa C, Savageau MA: **Effects of alternative connectivity on behavior of randomly constructed Boolean networks.** *Physica D* 2002, **170**:143-161.
- Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology.** *Nature* 1999, **402**:C47-C52.
- Gutiérrez-Ríos RM, Freyre-González JA, Resendis O, Collado-Vides J, Saier M, Gosset G: **Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*.** *BMC Microbiol* 2007, **7**:53.
- Resendis-Antonio O, Freyre-González JA, Menchaca-Méndez R, Gutiérrez-Ríos RM, Martínez-Antonio A, Avila-Sánchez C, Collado-Vides J: **Modular analysis of the transcriptional regulatory network of *E. coli*.** *Trends Genet* 2005, **21**:16-20.
- Dobrin R, Beg QK, Barabási AL, Oltvai ZN: **Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network.** *BMC Bioinformatics* 2004, **5**:10.

9. Ma HW, Buer J, Zeng AP: **Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach.** *BMC Bioinformatics* 2004, **5**:199.
10. Ma HW, Kumar B, Ditges U, Gunzer F, Buer J, Zeng AP: **An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs.** *Nucleic Acids Res* 2004, **32**:6643-6649.
11. Yu H, Gerstein M: **Genomic analysis of the hierarchical structure of regulatory networks.** *Proc Natl Acad Sci USA* 2006, **103**:14724-14731.
12. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network motifs: simple building blocks of complex networks.** *Science* 2002, **298**:824-827.
13. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of *Escherichia coli*.** *Nat Genet* 2002, **31**:64-68.
14. Smits WK, Kuipers OP, Veening JW: **Phenotypic variation in bacteria: the role of feedback regulation.** *Nat Rev Microbiol* 2006, **4**:259-271.
15. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J: **From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*.** *Bioessays* 1998, **20**:433-440.
16. Thomas R, Kaufman M: **Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior.** *Chaos* 2001, **11**:170-179.
17. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**:1551-1555.
18. Salgado H, Gama-Castro S, Peralta-Gil M, Díaz-Peredo E, Sánchez-Solano F, Santos-Zavaleta A, Martínez-Flores I, Jiménez-Jacinto V, Bonavides-Martínez C, Segura-Salazar J, Martínez-Antonio A, Collado-Vides J: **RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions.** *Nucleic Acids Res* 2006, **34**(Database issue):D394-D397.
19. Ravasz E, Barabási AL: **Hierarchical organization in complex networks.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **67**:026112.
20. Thomas R: **Laws for the dynamics of regulatory networks.** *Int J Dev Biol* 1998, **42**:479-485.
21. Thieffry D, Romero D: **The modularity of biological regulatory networks.** *Biosystems* 1999, **50**:49-59.
22. Kaern M, Elston TC, Blake WJ, Collins JJ: **Stochasticity in gene expression: from theories to phenotypes.** *Nat Rev Genet* 2005, **6**:451-464.
23. Lipschutz S: *Schaum's Outline of Data Structures* First edition. New York: McGraw-Hill; 1986.
24. Martínez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria.** *Curr Opin Microbiol* 2003, **6**:482-489.
25. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296**:910-913.
26. Browning DF, Busby SJ: **The regulation of bacterial transcription initiation.** *Nat Rev Microbiol* 2004, **2**:57-65.
27. Madan Babu M, Teichmann SA: **Evolution of transcription factors and the gene regulatory network in *Escherichia coli*.** *Nucleic Acids Res* 2003, **31**:1234-1244.
28. Gottesman S: **Bacterial regulation: global regulatory networks.** *Annu Rev Genet* 1984, **18**:415-441.
29. Stojiljkovic I, Bäumlér AJ, Hantke K: **Fur regulon in gram-negative bacteria. Identification and characterization of new iron-regulated *Escherichia coli* genes by a fur titration assay.** *J Mol Biol* 1994, **236**:531-545.
30. Angerer A, Braun V: **Iron regulates transcription of the *Escherichia coli* ferric citrate transport genes directly and through the transcription initiation proteins.** *Arch Microbiol* 1998, **169**:483-490.
31. Escolar L, Pérez-Martín J, de Lorenzo V: **Coordinated repression in vitro of the divergent *fepA-fes* promoters of *Escherichia coli* by the iron uptake regulation (Fur) protein.** *J Bacteriol* 1998, **180**:2579-2582.
32. Lavrrar JL, Christoffersen CA, McIntosh MA: **Fur-DNA interactions at the bidirectional *fepDGC-entS* promoter region in *Escherichia coli*.** *J Mol Biol* 2002, **322**:983-995.
33. Zhang Z, Gosset G, Barabote R, Gonzalez CS, Cuevas WA, Saier MH Jr: **Functional interactions between the carbon and iron utilization regulators, Crp and Fur, in *Escherichia coli*.** *J Bacteriol* 2005, **187**:980-990.
34. Outten FW, Djaman O, Storz G: **A *suf* operon requirement for Fe-S cluster assembly during iron starvation in *Escherichia coli*.** *Mol Microbiol* 2004, **52**:861-872.
35. Liu X, Matsumura P: **The FlhD/FlhC complex, a transcriptional activator of the *Escherichia coli* flagellar class II operons.** *J Bacteriol* 1994, **176**:7345-7351.
36. Stafford GP, Ogi T, Hughes C: **Binding and transcriptional activation of non-flagellar genes by the *Escherichia coli* flagellar master regulator FlhD₂C₂.** *Microbiology* 2005, **151**:1779-1788.
37. Prüss BM, Liu X, Hendrickson W, Matsumura P: **FlhD/FlhC-regulated promoters analyzed by gene array and *lacZ* gene fusions.** *FEMS Microbiol Lett* 2001, **197**:91-97.
38. Albert R, Jeong H, Barabási AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406**:378-382.
39. Serres MH, Goswami S, Riley M: **GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins.** *Nucleic Acids Res* 2004, **32**(Database issue):D300-D302.
40. Neidhardt FC, Savageau M: **Regulation beyond the operon.** In *Escherichia coli and Salmonella: Cellular and Molecular Biology* Second edition. Edited by: Neidhardt FC. Washington DC: American Society for Microbiology; 1996:1310-1324.
41. Aviv M, Giladi H, Schreiber G, Oppenheim AB, Glaser G: **Expression of the genes coding for the *Escherichia coli* integration host factor are controlled by growth phase, *rpoS*, ppGpp and by autoregulation.** *Mol Microbiol* 1994, **14**:1021-1031.
42. Jishage M, Iwata A, Ueda S, Ishihama A: **Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions.** *J Bacteriol* 1996, **178**:5447-5451.
43. Park YH, Lee BR, Seok YJ, Peterkofsky A: **In vitro reconstitution of catabolite repression in *Escherichia coli*.** *J Biol Chem* 2006, **281**:6448-6454.
44. Goosen N, van de Putte P: **The regulation of transcription initiation by integration host factor.** *Mol Microbiol* 1995, **16**:1-7.
45. Blot N, Mavathur R, Geertz M, Travers A, Muskhelishvili G: **Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome.** *EMBO Rep* 2006, **7**:710-715.
46. Travers A, Muskhelishvili G: **DNA supercoiling - a global transcriptional regulator for enterobacterial growth?** *Nat Rev Microbiol* 2005, **3**:157-169.
47. Partridge JD, Sanguinetti G, Dibden DP, Roberts RE, Poole RK, Green J: **Transition of *Escherichia coli* from aerobic to micro-aerobic conditions involves fast and slow reacting regulatory components.** *J Biol Chem* 2007, **282**:11230-11237.
48. Ravcheev DA, Gerasimova AV, Mironov AA, Gelfand MS: **Comparative genomic analysis of regulation of anaerobic respiration in ten genomes from three families of gamma-proteobacteria (Enterobacteriaceae, Pasteurellaceae, Vibrionaceae).** *BMC Genomics* 2007, **8**:54.
49. Reitzer L, Schneider BL: **Metabolic context and possible physiological themes of σ^{54} -dependent genes in *Escherichia coli*.** *Microbiol Mol Biol Rev* 2001, **65**:422-444.
50. Hayden JD, Ades SE: **The extracytoplasmic stress factor, σ^E , is required to maintain cell envelope integrity in *Escherichia coli*.** *PLoS ONE* 2008, **3**:e1573.
51. Dwight Kuo P, Banzhaf W, Leier A: **Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence.** *Biosystems* 2006, **85**:177-200.
52. Mazurie A, Bottani S, Vergassola M: **An evolutionary and functional assessment of regulatory network motifs.** *Genome Biol* 2005, **6**:R35.
53. Solé RV, Valverde S: **Are network motifs the spandrels of cellular complexity?** *Trends Ecol Evol* 2006, **21**:419-422.
54. Ingram PJ, Stumpf MP, Stark J: **Network motifs: structure does not determine function.** *BMC Genomics* 2006, **7**:108.
55. Cordero OX, Hogeweg P: **Feed-forward loop circuits as a side effect of genome evolution.** *Mol Biol Evol* 2006, **23**:1931-1936.
56. Vázquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabási AL: **The topological relationship between the large-scale attributes and local interaction patterns of complex networks.** *Proc Natl Acad Sci USA* 2004, **101**:17940-17945.
57. Ramseier TM, Nègre D, Cortay JC, Scarabel M, Cozzone AJ, Saier MH Jr: **In vitro binding of the pleiotropic transcriptional regulatory protein, FruR, to the *fru*, *pps*, *ace*, *pts* and *icd* operons of**

- Escherichia coli and Salmonella typhimurium.** *J Mol Biol* 1993, **234**:28-44.
58. Yamamoto K, Ishihama A: **Two different modes of transcription repression of the Escherichia coli acetate operon by IclR.** *Mol Microbiol* 2003, **47**:183-194.
 59. Devroede N, Huysveld N, Charlier D: **Mutational analysis of intervening sequences connecting the binding sites for integration host factor, PepA, PurR, and RNA polymerase in the control region of the Escherichia coli carAB operon, encoding carbamoylphosphate synthase.** *J Bacteriol* 2006, **188**:3236-3245.
 60. Caldara M, Charlier D, Cunin R: **The arginine regulon of Escherichia coli: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation.** *Microbiology* 2006, **152**:3343-3354.
 61. **RegulonDB 6.1** [<http://regulondb.ccg.unam.mx/>]
 62. Griffith KL, Shah IM, Myers TE, O'Neill MC, Wolf RE Jr: **Evidence for "pre-recruitment" as a new mechanism of transcription activation in Escherichia coli: the large excess of SoxS binding sites per cell relative to the number of SoxS molecules per cell.** *Biochem Biophys Res Commun* 2002, **291**:979-986.
 63. Martin RG, Gillette VK, Martin NI, Rosner JL: **Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in Escherichia coli.** *Mol Microbiol* 2002, **43**:355-370.
 64. Griffith KL, Wolf RE Jr: **Genetic evidence for pre-recruitment as the mechanism of transcription activation by SoxS of Escherichia coli: the dominance of DNA binding mutations of SoxS.** *J Mol Biol* 2004, **344**:1-10.
 65. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**:2498-2504.
 66. Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD: **EcoCyc: a comprehensive database resource for Escherichia coli.** *Nucleic Acids Res* 2005, **33(Database issue)**:D334-D337.
 67. **EcoCyc: Encyclopedia of Escherichia coli K-12 Genes and Metabolism** [<http://www.ecocyc.org/>]
 68. **GenProtEC: E. coli Genome and Proteome Database** [<http://genprotec.mbl.edu/>]
 69. **MultiFun** [<http://genprotec.mbl.edu/files/MultiFun.txt>]

Bibliografía

- ALBERT, R., JEONG, H., Y BARABÁSI, A.L. Error and attack tolerance of complex networks. *Nature* **406**(6794):378–382 (2000)
- ALBERT, R. Scale-free networks in cell biology. *J Cell Sci* **118**(Pt 21):4947–4957 (2005)
- ALBERT, R. Y BARABÁSI, A.L. Statistical mechanics of complex networks. *Rev Mod Phys* **74**(1):47–97 (2002)
- ALBERT, R., JEONG, H., Y BARABÁSI, A.L. Internet: Diameter of the World-Wide Web. *Nature* **401**(6749):130–131 (1999)
- ALON, U. Biological networks: the tinkerer as an engineer. *Science* **301**(5641):1866–1867 (2003)
- ALON, U. Network motifs: theory and experimental approaches. *Nat Rev Genet* **8**(6):450–461 (2007)
- BABU, M.M. Y TEICHMANN, S.A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res* **31**(4):1234–1244 (2003)
- BARABÁSI, A.L. *Linked: The New Science of Networks*. 1^a edición. Perseus Books Group (2002)
- BARABÁSI, A.L. Y ALBERT, R. Emergence of scaling in random networks. *Science* **286**(5439):509–512 (1999)
- BARABÁSI, A.L. Y OLTVAI, Z.N. Network biology: understanding the cell’s functional organization. *Nat Rev Genet* **5**(2):101–113 (2004)

- BROWNING, D.F. Y BUSBY, S.J. The regulation of bacterial transcription initiation. *Nat Rev Microbiol* **2**(1):57–65 (2004)
- CHUNG, F. Y LU, L. The average distances in random graphs with given expected degrees. *Proc Natl Acad Sci U S A* **99**(25):15879–15882 (2002)
- COHEN, R. Y HAVLIN, S. Scale-free networks are ultrasmall. *Phys Rev Lett* **90**(5):058701 (2003)
- CORDERO, O.X. Y HOGEWEG, P. Feed-forward loop circuits as a side effect of genome evolution. *Mol Biol Evol* **23**(10):1931–1936 (2006)
- DOBRIN, R., BEG, Q.K., BARABÁSI, A.L., Y OLTVAI, Z.N. Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics* **5**:10 (2004)
- DODDS, P.S., MUHAMAD, R., Y WATTS, D.J. An experimental study of search in global social networks. *Science* **301**(5634):827–829 (2003)
- DOROGOVTSSEV, S.N., GOLTSEV, A.V., Y MENDES, J.F.F. Pseudofractal scale-free web. *Phys Rev E Stat Nonlin Soft Matter Phys* **65**(6 Pt 2):066122 (2002)
- ERDŐS, P. Y RÉNYI, A. On random graphs. *Publ Math* **6**:290–297 (1959)
- ERDŐS, P. Y RÉNYI, A. On the evolution of random graphs. *Bulletin of the Institute of International Statistics* **38**:343–347 (1961)
- EULER, L. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum imperialis Petropolitanae* **8**:128–140 (1741)
- FREYRE-GONZÁLEZ, J.A., ALONSO-PAVÓN, J.A., VÁZQUEZ-HERNANDEZ, D., SANDOVAL-CALDERON, M., MATUS-GARCÍA, M., ORTEGA-DEL VECCHYO, D., Y COLLADO-VIDES, J. Modular and hierarchical organization of the transcriptional regulatory network of *Escherichia coli* K-12. 5th International Workshop on Bioinformatics and Systems Biology, Poster Session, Berlín, Alemania (2005)
- FREYRE-GONZÁLEZ, J.A., ALONSO-PAVÓN, J.A., TREVIÑO-QUINTANILLA, L.G., Y COLLADO-VIDES, J. Un criterio topológico original revela la organización natural de las

- redes de regulación transcripcional de *Escherichia coli* y *Bacillus subtilis*. 1^a Reunión Regional de Ciencias Microbiológicas, Puebla, México (2007)
- FREYRE-GONZÁLEZ, J.A., ALONSO-PAVÓN, J.A., TREVIÑO-QUINTANILLA, L.G., Y COLLADO-VIDES, J. Functional architecture of *Escherichia coli*: new insights provided by a natural decomposition approach. *Genome Biol* **9**(10):R154 (2008)
- GOTTESMAN, S. Bacterial regulation: global regulatory networks. *Annu Rev Genet* **18**:415–441 (1984)
- GRIFFITH, K.L. Y WOLF, R.E. Genetic evidence for pre-recruitment as the mechanism of transcription activation by SoxS of *Escherichia coli*: the dominance of DNA binding mutations of SoxS. *J Mol Biol* **344**(1):1–10 (2004)
- GRIFFITH, K.L., SHAH, I.M., MYERS, T.E., O’NEILL, M.C., Y WOLF, R.E. Evidence for “pre-recruitment” as a new mechanism of transcription activation in *Escherichia coli*: the large excess of SoxS binding sites per cell relative to the number of SoxS molecules per cell. *Biochem Biophys Res Commun* **291**(4):979–986 (2002)
- GUELZIM, N., BOTTANI, S., BOURGINE, P., Y KÉPÈS, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* **31**(1):60–63 (2002)
- GUIMERA, R. Y AMARAL, L.A.N. Functional cartography of complex metabolic networks. *Nature* **433**(7028):895–900 (2005)
- GUTIÉRREZ-RÍOS, R.M., ROSENBLUETH, D.A., LOZA, J.A., HUERTA, A.M., GLASNER, J.D., BLATTNER, F.R., Y COLLADO-VIDES, J. Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res* **13**(11):2435–2443 (2003)
- GUTIERREZ-RÍOS, R.M., FREYRE-GONZÁLEZ, J.A., RESENDIS, O., COLLADO-VIDES, J., SAIER, M., Y GOSSET, G. Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. *BMC Microbiol* **7**:53 (2007)

- HARTWELL, L.H., HOPFIELD, J.J., LEIBLER, S., Y MURRAY, A.W. From molecular to modular cell biology. *Nature* **402**(6761 Suppl):C47–C52 (1999)
- HOPKINS, B. Y WILSON, R.J. The truth about Königsberg. *The College Mathematics Journal* **35**(3):198–207 (2004)
- INGRAM, P.J., STUMPF, M.P.H., Y STARK, J. Network motifs: structure does not determine function. *BMC Genomics* **7**:108 (2006)
- IUCHI, S. Y LIN, E.C. *arcA* (*dye*), a global regulatory gene in *Escherichia coli* mediating repression of enzymes in aerobic pathways. *Proc Natl Acad Sci U S A* **85**(6):1888–1892 (1988)
- JACOB, F., PERRIN, D., SANCHEZ, C., Y MONOD, J. [Operon: a group of genes with the expression coordinated by an operator.]. *C R Hebd Seances Acad Sci* **250**:1727–1729 (1960)
- JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z.N., Y BARABÁSI, A.L. The large-scale organization of metabolic networks. *Nature* **407**(6804):651–654 (2000)
- KAERN, M., ELSTON, T.C., BLAKE, W.J., Y COLLINS, J.J. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* **6**(6):451–464 (2005)
- KESELER, I.M., COLLADO-VIDES, J., GAMA-CASTRO, S., INGRAHAM, J., PALEY, S., PAULSEN, I.T., PERALTA-GIL, M., Y KARP, P.D. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res* **33**(Database issue):D334–D337 (2005)
- KUO, P.D., BANZHAF, W., Y LEIER, A. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems* **85**(3):177–200 (2006)
- LESKOVEC, J. Y HORVITZ, E. Planetary-scale views on an instant-messaging network (2008). [arXiv:0803.0939v1](https://arxiv.org/abs/0803.0939v1) [physics.soc-ph]
- LIPSCHUTZ, S. *Estructura de Datos*. Serie Schaum. Mcgraw-Hill (1986)
- LIU, G. Y CHEN, X. Regulation of the p53 transcriptional activity. *J Cell Biochem* **97**(3):448–458 (2006)

- MA, H.W., BUER, J., Y ZENG, A.P. Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics* **5**:199 (2004a)
- MA, H.W., KUMAR, B., DITGES, U., GUNZER, F., BUER, J., Y ZENG, A.P. An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res* **32**(22):6643–6649 (2004b)
- MAAS, W.K. Studies on the mechanism of repression of arginine biosynthesis in *Escherichia coli*. II. Dominance of repressibility in diploids. *J Mol Biol* **8**:365–370 (1964)
- MARCONI, G. Wireless Telegraphic Communication. En *Physics 1901–1921*, Nobel Lectures, págs. 196–222. Elsevier Publishing Company, Amsterdam (1967)
- MARTIN, R.G., GILLETTE, W.K., MARTIN, N.I., Y ROSNER, J.L. Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in *Escherichia coli*. *Mol Microbiol* **43**(2):355–370 (2002)
- MARTÍNEZ-ANTONIO, A. Y COLLADO-VIDES, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* **6**(5):482–489 (2003)
- MAZURIE, A., BOTTANI, S., Y VERGASSOLA, M. An evolutionary and functional assessment of regulatory network motifs. *Genome Biol* **6**(4):R35 (2005)
- MILGRAM, S. The small world problem. *Psychology Today* **2**(1):60–67 (1967)
- MILLAU, J.F., BASTIEN, N., Y DROUIN, R. P53 transcriptional activities: A general overview and some thoughts. *Mutat Res* (2008)
- MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D., Y ALON, U. Network motifs: simple building blocks of complex networks. *Science* **298**(5594):824–827 (2002)
- NEIDHARDT, F.C. Y SAVAGEAU, M. Regulation beyond the operon. En F.C. Neidhardt (editor), *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2^a edición, págs. 1310–1324. American Society for Microbiology, Washington D.C. (1996)

- NELSON, D.L. Y COX, M.M. *Lehninger Principles of Biochemistry*. 3^a edición. W. H. Freeman (2000)
- NEWMAN, M.E.J. Y GIRVAN, M. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **69**(2 Pt 2):026113 (2004)
- OLTVAI, Z.N. Y BARABÁSI, A.L. Systems biology. Life's complexity pyramid. *Science* **298**(5594):763–764 (2002)
- OOSAWA, C. Y SAVAGEAU, M.A. Effects of alternative connectivity on behavior of randomly constructed Boolean networks. *Physica D* **170**(2):143–161 (2002)
- RAVASZ, E., SOMERA, A.L., MONGRU, D.A., OLTVAI, Z.N., Y BARABÁSI, A.L. Hierarchical organization of modularity in metabolic networks. *Science* **297**(5586):1551–1555 (2002)
- RAVASZ, E. Y BARABÁSI, A.L. Hierarchical organization in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **67**(2 Pt 2):026112 (2003)
- RESENDIS-ANTONIO, O., FREYRE-GONZÁLEZ, J.A., MENCHACA-MÉNDEZ, R., GUTIÉRREZ-RÍOS, R.M., MARTÍNEZ-ANTONIO, A., AVILA-SÁNCHEZ, C., Y COLLADO-VIDES, J. Modular analysis of the transcriptional regulatory network of *E. coli*. *Trends Genet* **21**(1):16–20 (2005)
- RIVES, A.W. Y GALITSKI, T. Modular organization of cellular networks. *Proc Natl Acad Sci U S A* **100**(3):1128–1133 (2003)
- SALES-PARDO, M., GUIMERA, R., MOREIRA, A.A., Y AMARAL, L.A.N. Extracting the hierarchical organization of complex systems. *Proc Natl Acad Sci U S A* **104**(39):15224–15229 (2007)
- SALGADO, H., GAMA-CASTRO, S., PERALTA-GIL, M., DÍAZ-PEREDO, E., SÁNCHEZ-SOLANO, F., SANTOS-ZAVALETA, A., MARTÍNEZ-FLORES, I., JIMÉNEZ-JACINTO, V., BONAVIDES-MARTÍNEZ, C., SEGURA-SALAZAR, J., MARTÍNEZ-ANTONIO, A., Y COLLADO-VIDES, J. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* **34**(Database issue):D394–D397 (2006)

- SERRES, M.H., GOSWAMI, S., Y RILEY, M. GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res* **32**(Database issue):D300–D302 (2004)
- SHEN-ORR, S.S., MILO, R., MANGAN, S., Y ALON, U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31**(1):64–68 (2002)
- SMITS, W.K., KUIPERS, O.P., Y VEENING, J.W. Phenotypic variation in bacteria: the role of feedback regulation. *Nat Rev Microbiol* **4**(4):259–271 (2006)
- SOFFER, S.N. Y VÁZQUEZ, A. Network clustering coefficient without degree-correlation biases. *Phys Rev E Stat Nonlin Soft Matter Phys* **71**(5 Pt 2):057101 (2005)
- SOLÉ, R.V. Y VALVERDE, S. Are network motifs the spandrels of cellular complexity? *Trends Ecol Evol* **21**(8):419–422 (2006)
- THEUNS, J. Y BROECKHOVEN, C.V. Transcriptional regulation of Alzheimer’s disease genes: implications for susceptibility. *Hum Mol Genet* **9**(16):2383–2394 (2000)
- THIEFFRY, D. Y ROMERO, D. The modularity of biological regulatory networks. *Biosystems* **50**(1):49–59 (1999)
- THIEFFRY, D. Y THOMAS, R. Qualitative analysis of gene networks. *Pac Symp Biocomput* págs. 77–88 (1998)
- THIEFFRY, D., HUERTA, A.M., PÉREZ-RUEDA, E., Y COLLADO-VIDES, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* **20**(5):433–440 (1998)
- THOMAS, R. Laws for the dynamics of regulatory networks. *Int J Dev Biol* **42**(3):479–485 (1998)
- THOMAS, R. Y KAUFMAN, M. Multistationarity, the basis of cell differentiation and memory. I. Structural conditions of multistationarity and other nontrivial behavior. *Chaos* **11**(1):170–179 (2001)
- THOMAS, R. Y D’ARI, R. *Biological Feedback*. 1^a edición. CRC Press (1990)

- TRAVERS, J. Y MILGRAM, S. An experimental study of the small world problem. *Sociometry* **32**(4):425–443 (1969)
- VARIANO, E.A., MCCOY, J.H., Y LIPSON, H. Networks, dynamics, and modularity. *Phys Rev Lett* **92**(18):188701 (2004)
- VOIGT, C.A., WOLF, D.M., Y ARKIN, A.P. The *Bacillus subtilis sin* operon: an evolvable network motif. *Genetics* **169**(3):1187–1202 (2005)
- WATTS, D.J. Y STROGATZ, S.H. Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684):440–442 (1998)
- WOLF, D.M. Y ARKIN, A.P. Motifs, modules and games in bacteria. *Curr Opin Microbiol* **6**(2):125–134 (2003)
- YU, H. Y GERSTEIN, M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc Natl Acad Sci U S A* **103**(40):14724–14731 (2006)