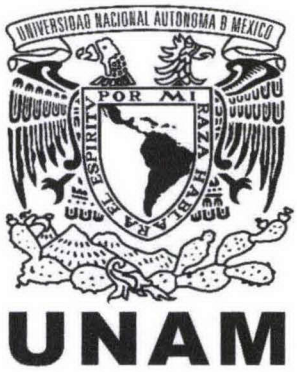


UNAM

CCG

Centro de Ciencias Genómicas

Universidad Nacional Autónoma de México

ACTIVIDAD DE INVESTIGACIÓN PARA OBTENER EL TÍTULO EN:

Licenciatura en Ciencias Genómicas

"Evaluando la Universalidad y la Robustez del Enfoque de Descomposición Natural en un Amplio Repertorio de Bacterias"

Miguel Angel Ibarra Arellano

Tutor:

Dr. Julio Augusto Freyre González

Cuernavaca, Morelos

Agosto de 2015

Agradecimientos

This research was supported by grant IA200614 from PAPIIT-DGAPA-UNAM to Julio A. Freyre Gonzalez. I acknowledge an undergraduate fellowship from the same grant.

Este trabajo fue financiado por el donativo IA200614 de la PAPIIT-DGAPA-UNAM otorgado a Julio A. Freyre Gonzalez. Agradezco también a dicho proyecto la beca otorgada.

Honor a quien honor merece.

Siempre es complicado escribir, y resulta mucho más cuando se trata de agradecer, porque el espacio es muy breve y resta mucho que decir. Espero que en estas breves líneas pueda expresar el amor y profundo agradecimiento que tengo por ustedes.

Esta tesis se la dedico a mis padres que con empeño, sacrificio y cariño me han sacado adelante. A ustedes les debo todo.

A mí madre, la persona más especial en mi vida. . Gracias por enseñarme lo básico, por enseñarme a vivir y a valerme por mí mismo. Gracias por fomentar en mí la curiosidad, y despertar en mí el interés por el mundo. Gracias por todas las horas de espera, por todos los sacrificios y por todos esos momentos grandiosos que tuvimos y tendremos. Gracias por enseñarme que los sacrificios son necesarios y que la familia es primero. Gracias por enseñarme el valor del trabajo!. Gracias por enseñarme a respetar tanto a mí como a los demás, a ser generoso y a dar sin esperar recibir. Gracias por tu dedicación, tiempo, cariño, por tu paciencia y comprensión. Todo lo que soy es gracias a ti, gracias por formar de ese niño a un hombre (espero de bien). GRACIAS POR TODO MAMÁ. TE AMO!

A mi padre, que aun en la distancia nos ama y se preocupa por mí y mis hermanos. Gracias padre por enseñarme a ser duro, por enseñarme a convivir y defenderme, por enseñarme a no confiar en nadie. Y Gracias por enseñarme a amar a la familia. Te amo padre.

A mi abuelita, Abuelita, gracias por cuidarme, protegerme, alimentarme y estar ahí cuando más te necesité. Gracias por tus consejos e historias, gracias por tu compañía y amor. Y muchas gracias por tu arroz! :3.

A mis hermanos, esos entes de caos y terror en mi vida. A ti Gustavo, muchas gracias por ser mi compañero. Mi compañero de travesuras y cosas serias... sí sí sí muy serias. Gracias por todas las risas y momentos maravillosos que pasamos juntos, nos falta muchos más... no te salvarás de mí. Te amo hermanito. A mi Yamis, uno cree conocer la diversión hasta que tu hermana te hace cosquillas ^_^. Muchas gracias Yamile, gracias por estar ahí, por escucharme, por ser mi confidente y por confiar en mí. Te amo hermanita!

Youmy!!!! Condenada!!!! Muchas gracias por ser mi amiga. Gracias por estar en mis mejores y peores momentos. Gracias por preocuparte por mí, y por ayudarme como nadie lo ha hecho. Gracias por estar conmigo en mi momento más oscuro y darme luz para salir de él. Gracias amiga.

Tomy, Larousse! A ustedes les debo las gracias por estar ahí más de una vez. Por acompañarme en diversas y muy variadas aventuras jajaja por brincar, hacernos bolita y caer. Gracias por todas esas horas de Gears y tortas de doña Mary®. Gracias por su amistad!

Compañero, el camino ha sido largo... 2 horas desde aeropuerto no son cosa sencilla y tú hiciste de esos viajes la mejor experiencia posible. Gracias por ser

una figura de inspiración y respeto para mí. Incontables horas hemos pasado juntos y no cambiaría ni un solo segundo por nada más en el universo. Gracias por tu ayuda y confianza. Gracias por tu amistad.

Rul, A ti te agradezco tus consejos, tu siempre acertado punto de vista. Tu actitud me ha motivado y mucho has cambiado en mí. Gracias!

Erika, Es difícil mencionar cuan agradecido estoy contigo. Tú me viste en mis peores momentos y nunca te rendiste conmigo, siempre conté en ti con una amiga incondicional, una sonrisa sincera y muchas risas y diversión, te pido una disculpa por no haber sido yo así para ti.

Ricardito, eeeeee, eeeeee, eeeeeee, eeeeee! Hermano! Muchas gracias por vivir conmigo este viaje, con todos sus tropiezos, valles tenebrosos, caos y destrucción. Gracias por escucharme, aconsejarme y nunca darme la razón. Gracias por todas esas noches de desvelo, de pláticas profundas y de banalidades mundanas. Todos esos miércoles de espositos y esos jueves de ERREO intenso, gracias por abrirme las puertas de tu casa y más importante aún de tu amistad. Nunca estaré sólo en el camino...nunca estaremos solos en el camino, nunca más. Gracias hermano!

Perezosita hermosa, es complicado plantear un escenario donde no estés tú, tú que has hecho de mi vida la más feliz del universo, tú que me has apoyado y has confiado en mí incondicionalmente. Tú me has enseñado cualidades de nobleza, dedicación y convicción que nunca antes había visto, así como darme el valor de hacer cosas de las que me encontraba temeroso y que quizá no hubiera hecho de otra forma. Gracias por dejarme amarte y amarme de la forma en la que lo haces.

Julio, mi tutor, mi mentor, mi maestro, mi amigo. En estos años si he aprendido de alguien ha sido de ti y estoy eternamente agradecido de todo tu esfuerzo, tiempo y dedicación. Por nunca rendirte y siempre encontrar soluciones, por siempre exigirme porque sabías que podía dar más, por tus consejos y sabiduría. Gracias por compartirme tu visión de la vida y formar así parte de la mía.

Nación de Chi! Gracias por aceptarme y dejarme convivir con ustedes, por los momentos divertidos y por su ayuda en los momentos difíciles.

A toda la doceava generación!!!!!

Agradezco a todos los profesores que he tenido a lo largo de mi vida, que han formado parte elemental de lo que soy ahora. No sería nada sin sus enseñanzas y a ustedes va también este logro.

Y gracias MONESVOL sin tu ayuda y guía esto no hubiera sido posible.

Índice

1. INTRODUCCIÓN

El enfoque reduccionista por años ha buscado entender los sistemas biológicos enfocando su estudio a partes cada vez más pequeñas. Este enfoque ha identificado exitosamente en los diferentes sistemas biológicos la mayoría de los componentes que los forman y muchas de las interacciones entre ellos, sin embargo, el todo no es la suma de las partes. El enfoque reduccionista no ofrece conceptos convincentes o métodos que permitan entender cómo es que emergen las propiedades de un sistema. Con el surgimiento de las tecnologías ómicas y la gran cantidad de datos generados por estas, resulta necesaria tener una visión más global de las propiedades de los sistemas biológicos, la pluralidad de causas y efectos en las redes biológicas son mejor entendidos observando múltiples componentes simultáneamente y por medio de una rigurosa integración de datos utilizando modelos matemáticos [Sauer et al. 2007].

La biología de sistemas, es el modelado matemático y computacional de sistemas biológicos complejos. Se encarga del estudio de las interacciones entre los componentes de un sistema, y de cómo estas dan origen a las funciones y comportamientos del mismo. Uno de sus principales objetivos es el descubrimiento, descripción y modelado de las propiedades emergentes de un sistema. Dentro de los recursos usados por la biología de sistemas podemos destacar la teoría de grafos. La teoría de grafos ha provisto un buen marco teórico para analizar grandes conjuntos de interacciones moleculares (como interacciones regulatorias), buscando leyes universales que los gobiernen y organicen [Julio A. Freyre-González, et al. 2008].

El Enfoque de Descomposición Natural (Natural Decomposition Aproach, NDA) es un método para inferir la organización intrínseca de redes regulatorias. Este método clasifica los elementos de una red a partir de sus características de conectividad y agrupamiento, obteniendo las siguientes categorías: reguladores globales, elementos modulares, maquinaría basal y elementos intermodulares. El método, como su nombre lo indica, consiste en la descomposición progresiva de la red analizada. En primer lugar se define un valor de corte (*kappa*), definido como el punto de equilibrio de la distribución de coeficiente de agrupamiento promedio C (k), esto es, el punto en el que la variación de la C (k) es igual a la variación en la conectividad con signo contrario. Una vez calculado este punto de corte se procede a categorizar a todos los nodos con conectividad mayor a *kappa* como reguladores globales. Estos reguladores globales se retiran junto con todas sus conexiones revelando así: islas aisladas compuestas de nodos interconectados los cuales serán clasificados cómo elementos modulares, y nodos desconectados que son clasificados cómo maquinaría basal. Los elementos intermodulares se obtienen removiendo todos los genes estructurales, genes que no codifican reguladores. Esto separa el megamódulo en islas aisladas (pre-submódulos). Posteriormente los genes previamente removidos son reintegrados siguiendo la siguiente regla: si el gen se encuentra regulado exclusivamente por reguladores de un mismo módulo entonces pertenece a este módulo, por el contrario si es regulado por reguladores de módulos distintos recibe la clasificación de gen intermodular [Figura 1] [Julio A. Freyre-González, et al. 2012].

La ventaja del NDA sobre otros métodos de inferencia topológica, es que se sobrepone a los problemas usuales de éstos: Ignorar genes que no codifican TFs, obtener módulos diferentes al modificar los parámetros del análisis o ubicar incorrectamente a genes pleiotrópicos conocidos dentro de módulos. Por este motivo, el NDA resulta un método ideal para el estudio de redes de regulación.

A lo largo de los años ha aumentado considerablemente la cantidad de compendios de interacciones regulatorias y su curación masiva se ha hecho presente. Ejemplos como RegulonDB, CoryneRegNet, DBTBS, RegTransBase y diversos estudios más en literatura muestran este avance. Sin embargo, resulta imperante una visión de sistemas para el estudio de estas interacciones, de forma que podamos comprender mejor los principios de organización y evolución que los gobiernan.

El Enfoque de Descomposición Natural ha sido utilizado en anteriores ocasiones para la identificación y anotación de los sistemas que componen la red regulatoria de *Escherichia coli* y de la bacteria Gram positiva *Bacillus subtilis*, proveyendo de esta manera un marco para estudiar los principios que gobiernan la organización de las redes de regulación. Con estos antecedentes y debido a la importancia del estudio global de las redes regulatorias bacterianas en el campo de la biología evolutiva y de sistemas, se decidió estudiar la robustez de las predicciones del Enfoque de Descomposición Natural. Es necesario evaluar la sensibilidad de las predicciones en función del grado de completez de las redes desde diversos ángulos.

2. Universalidad

2.1. Obtención de datos

En este estudio se buscó reconstruir la mayor cantidad de redes de regulación abarcando la mayor cantidad de géneros bacterianos disponibles, siempre priorizando aquellas redes mejor curadas y más completas. Para lograr este objetivo se recurrieron a diversas fuentes, como bases de datos especializadas y generales, así como a la literatura, obteniendo un total de 485 redes de regulación.

2.1.1. *Escherichia coli*

RegulonDB es conocida por ser la base de datos de referencia de interacciones regulatorias de la bacteria Gram negativa *Escherichia coli (E. coli)*. La red regulatoria de *E. coli* fue obtenida directamente de las tablas proporcionadas por RegulonDB, en un archivo de texto delimitado por tabuladores que contienen signo de la regulación y evidencia [Salgado H, 2012].

2.1.2. *Bacillus subtilis*

Bacillus subtilis (*B. subtilis*) es la representante del genero *Bacillus* y principal representante de las bacterias Gram positivas. Su red de regulación fue reconstruida usando datos de la base de datos DBTBS. La base de datos fue proporcionada como un archivo XML por el equipo de DBTBS entre los datos contenidos están el signo de regulación y la evidencia [Sierro N. et al 2008].

2.1.3. *Corynebacterium glutamicum*

Corynebacterium glutamicum (*C. glutamicum*) es una bacteria Gram Positiva de gran importancia económica debido a su uso en la producción de glutamato a escala industrial. Está red regulatoria fue obtenida de la base de datos CoryneRegNet [Pauling J et al, 2012], el archivo en formato texto fue provisto por el Dr. Andreas Tauch.

2.1.4. *Pseudomonas aeruginosa*

Pseudomonas aeruginosa (*P. aeruginosa*), es una bacteria Gram negativa e importante modelo bacteriano debido a sus características patogénicas y metabólicas que le permiten colonizar un amplio rango de organismos, incluidos plantas y animales. Su red de regulación fue obtenida de la literatura (208964-EGV) [Edgardo Galán-Vásquez et al, 2011] y complementada con la red obtenida de RegTransBase (208964-RTB) [Alexei E. Kazakov et al,2006].

2.1.5. *Mycobacterium tuberculosis*

Mycobacterium tuberculosis (M. tuberculosis) es una bacteria Gram positiva, ácido-alcohol resistente e importante patógena causante de la tuberculosis. La virulencia de este organismo radica en su habilidad para cambiar entre los estados replicativo y durmiente, evitando a conveniencia la respuesta inmune del organismo infectado. Se estima que una tercera parte de la población del mundo se encuentra infectado por esta bacteria. Para esta bacteria se obtuvieron cuatro redes de regulación diferentes 83332-2008 [Gábor Balázsi et al, 2008], 83332-2011 [Joaquín Sanz et al, 2011], 83332-2012 [Kyle H. Rohde et al, 2012] y 83332-2015 [Kyle J. Minch et al, 2015] las cuales se analizaron para su posterior integración en una red más completa (Tabla 1).

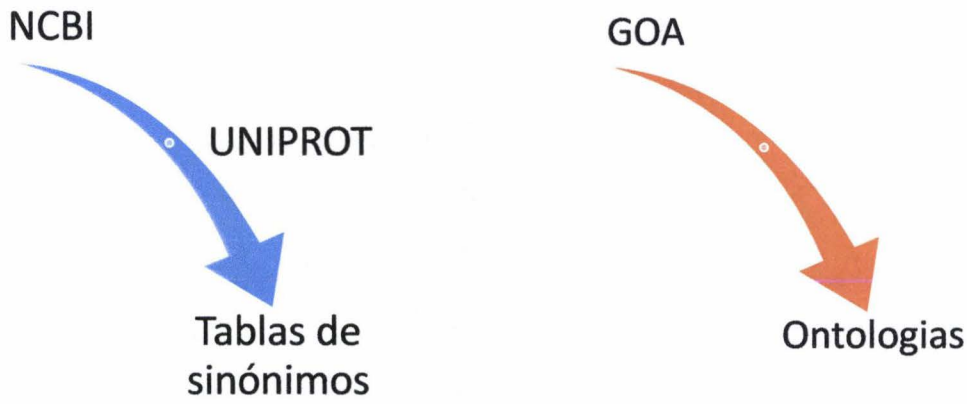
Tabla 1: Redes regulatorias de *M. tuberculosis*

	Condición	Genes	Cobertura	Interacciones	Fuentes
Gábor Balázsi et al, 2008	Adaptación a fase estacionaria e hipoxia.	783	20%	937	Microarrego Literatura Expansión por operón Ortología con *E.coli*
Joaquín Sanz et al, 2011	-	1624	40%	3212	Literatura
Kyle H. Rohde et al, 2012	Infección de 14 días	1133	28%	1801	Ortología con *C.glutamicum* Expansión por operon
Kyle J. Minch et al, 2015	Medio de cultivo Lowenstein-Jensen.	2547	63%	6581	ChIP-ChIP ChIP-Seq Microarreglo

2.1.6. RegTransBase

Adicionalmente a las redes obtenidas de las bases de datos especializadas y de las obtenidas por literatura se hizo uso de la base de datos RegTransBase. RegTransBase es una base de datos manualmente curada de interacciones regulatorias en procariontes [Alexei E. Kazakov et al, 2006]. Debido a que el archivo SQL provisto en la web se encontraba obsoleto y a la nula respuesta de parte de los autores se hizo uso de un robot (WinHTTrack) para descargar todos los archivos HTML que contenían las interacciones regulatorias para todos los organismos. La información fue extraída de los archivos HTML por medio de un *script,* que convertía estos a archivos a formato texto por columnas separadas por tabuladores con los campos regulador, producto del regulador, gen regulado, producto del gen regulado e ID usado en RegTransBase. Dando un total de 477 redes regulatorias (Tabla1).

Figura 1. Obtención de tablas de sinónimos y ontologías para anotación funcional.



2.2. Procesamiento de datos y selección de redes

2.2.1. RegTransBase

Los archivos de textos resultantes del procesamiento de los archivos HTML fueron cortados de manera que únicamente contuvieran los campos regulador y gen regulado, homogenizando así el formato de las redes [Figura 2]. Los nombres de los archivos pasaron a ser los TaxonID de cada organismo. Adicionalmente estas redes fueron seleccionadas para proceder con los siguientes análisis.

Figura 2. Esquema del procesamiento general de las redes obtenidas de RegTransBase.

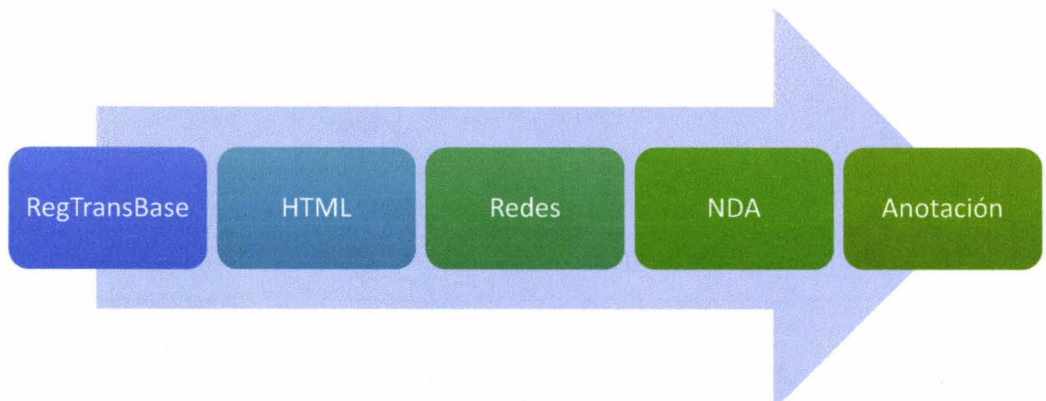


Tabla 1: Información de una fracción de las redes regulatorias obtenidas.

Organismo	Genes	TFs	Interacciones Regulatorias
Escherichia coli B171	279	52	390
Escherichia coli B7A	287	52	406
Escherichia coli CFT073	256	50	347
Escherichia coli O1:K1 / APEC	247	46	337
Escherichia coli O6:K15:H31 (strain 536 / UPEC)	259	51	348
Escherichia coli O9:H4 (strain HS)	297	59	427
Escherichia coli str. K-12 substr. DH10B	304	55	435
Escherichia coli str. K-12 substr. MG1655	3228	198	7759
Mycobacterium tuberculosis (H37Rv) (2008)	738	45	937
Mycobacterium tuberculosis (H37Rv) (2011)	1624	83	3213
Mycobacterium tuberculosis (H37Rv) (2012)	1133	85	1801
Mycobacterium tuberculosis (H37Rv) (2015)	816	110	971
Pseudomonas aeruginosa (PAO1)	905	125	1362
Pseudomonas aeruginosa (strain PA7)	130	29	125
Staphylococcus aureus (strain JH1)	375	47	546
Staphylococcus aureus (strain USA300)	368	46	560
Streptococcus pneumoniae (strain Hungary19A-6)	204	27	253
Streptococcus pneumoniae SP9-BS68	202	25	236
Streptococcus pyogenes serotype M1	165	19	196
Streptomyces coelicolor (A3(2) / 145)	311	64	296

2.1.1. Sinónimos y GOs.

Adicionalmente a las redes de regulación, se obtuvieron tablas de sinónimos provenientes de UniProt que fueron completadas con tablas de NCBI (ver métodos). Estas tablas de sinónimos fueron usadas para evitar redundancias debidas a posibles interacciones duplicadas en las redes. Además de estas tablas de sinónimos fueron necesarias tablas con las ontologías de genes (*Gene Ontologies, GOs*), esto con el fin de llevar a cabo una anotación automatizada empleando un vocabulario controlado de los sistemas identificados en las redes. Las ontologías fueron obtenidas de la base de datos GOA [Figura 1].

2.2.2. *Escherichia coli y Bacillus subtilis*

Dada la cantidad y diversidad de las fuentes de datos fue necesario estandarizar a un formato homogéneo todas las redes obtenidas. El formato utilizado fue un archivo de texto separado por columnas (regulador/tabulador/regulado) como lista de aristas. También se agregó el signo de la regulación y la evidencia de la interacción, si la información estaba disponible.

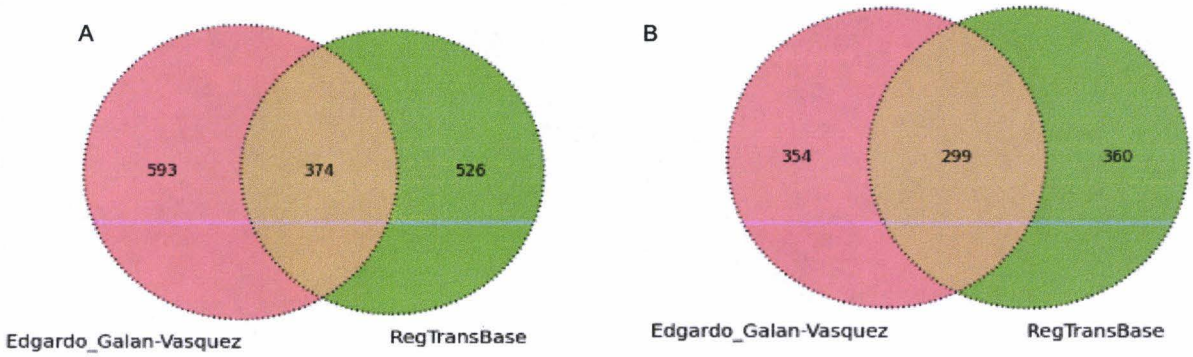
Para *E. coli* y para *B. subtilis,* se obtuvieron dos redes a partir de los datos obtenidos: Una red con todas las interacciones regulatorias, tanto de evidencia fuerte como de evidencia débil, (511145_*All y* 224308_All) y una con sólo las interacciones regulatorias con evidencia fuerte (511145_*Strong* y 224308_*Srong*). Adicionalmente, para *E. coli* se obtuvo una nueva red (511145_*RNA*) que contenía, además de todas las evidencias experimentales, una nueva capa de regulación, la de ARNs pequeños (smallRNA, sRNA).

2.2.3.*Pseudomonas aeruginosa.*

Con el fin de obtener redes regulatorias cepa específica, la red regulatoria 208964-EGV fue desprovista de todas las interacciones no pertenecientes a la cepa PAO1 (33 interacciones eliminadas) dejando una red con 967 interacciones entre 653 genes. Esta red 208964-EGV corregida, junto con la red 208964-RTB (900 interacciones entre 659 nodos), fueron analizadas con el programa SetAnalizer (ver métodos) con la finalidad de analizar su similitud [Figura 3]. Usando el índice de Jaccard como medida de similitud, el análisis entre estas dos redes arrojó índices de 0.25 y 0.3 para interacciones y nodos

respectivamente. Tomando estos resultados se puede observar tanto complementariedad como concordancia entre ambas redes. Por este motivo se decidió tomar la unión de estas redes para los posteriores análisis dentro de este estudio. Siendo esta nueva red regulatoria la más completa para *P. aeruginosa* hasta la fecha.

Figura 3. (A) Interacciones entre ambas redes de *P. aeruginosa* PAO1. (B) Nodos entre ambas redes de *P. aeruginosa* PAO1.



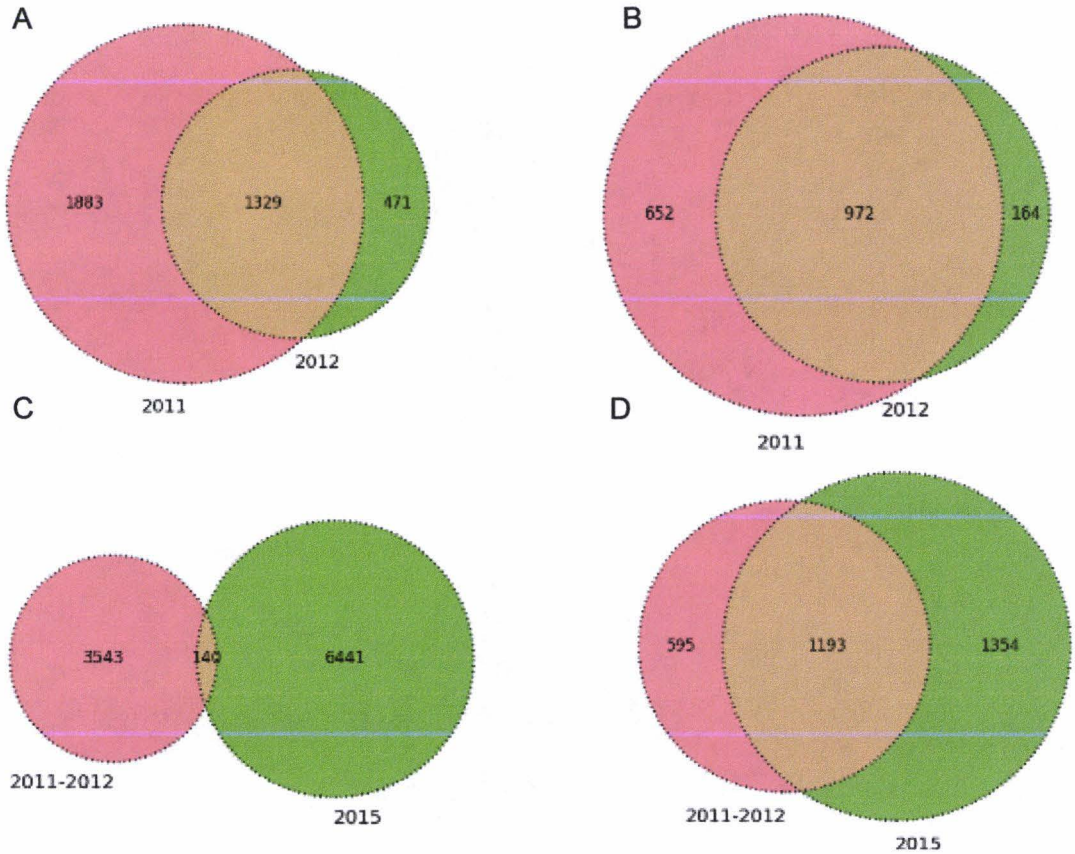
2.2.4. *Mycobacterium tuberculosis*

Debido a su patogenicidad, esta bacteria tiene un gran interés clínico y médico, y tiene sentido que sus redes de regulación se encuentren principalmente enfocadas en aspectos relacionados a infección y patogenicidad (83332-2008 y 83332-2012). Sin embargo, siendo nuestro objetivo reconstruir las redes más completas realizamos una serie de análisis para obtener más información de estas redes. En el caso de las redes 83332-2011 y 83332-2012, ambas fueron una extensión de la red 83332-2008. De esta manera fuimos capaces de descartar esta red debido a su inclusión completa en sus contrapartes más recientes. A ambas redes 83332-2011 y 83332-2012 las analizamos con la herramienta SetAnalizer (ver métodos). Los resultados del análisis arrojaron un total de 1329 interacciones y 972 nodos compartidos entre ambas redes, del total de 3683 interacciones y 1788 nodos, con índices de Jaccard de 0.36 y

0.54 respectivamente [Figura 4]. Con estos resultados y tomando en cuenta el núcleo de ambas redes (83332-2008) decidimos tomar la unión de las mismas teniendo así una red más completa (83332-2011-2012).

La red 83332-2015 difiere del resto de redes para este organismo, ya que fue obtenida *de novo* y en su construcción no se usó inferencia por ortología. Esta red contiene aproximadamente el 80% de los 206 TFs predichos para este organismo, por lo tanto, la consideramos como un buen candidato para complementar la red 83332-20211-2012. Los resultados arrojados por el SetAnalizer para 83332-2011-2012 y 83332-2015 fueron 140 interacciones y 1193 nodos compartidos de un total de 10124 interacciones y 3142 nodos, resultado en índices de Jaccard de 0.01 y 0.38 respectivamente [Figura 4]. Cabe destacar que la cantidad de interacciones compartidas no excede el 1% del total de interacciones entre ambas redes, sugiriendo que son redes casi completamente diferentes, a pesar de compartir alrededor del 40% de los nodos. Esto puede deberse a las diferentes condiciones en las fue expuesto *M. tuberculosis* durante el momento de obtención de las redes de regulación anteriormente mencionadas: una red pudo haber sido obtenida en condiciones de estrés e infección y la otra en condiciones estables [Tabla 1]. Estos resultados podrían indicar que estas redes representan programas regulatorios alternativos dentro de *M. tuberculosis*. Tomando en consideración lo anterior, optamos por tomar ambas redes por separado y generar una tercera red que fuera la unión de estas dos redes (83332-2011-2012-2015). Está nueva red consta de 10124 interacciones y 3142 nodos (75% del genoma de *M. tuberculosis*), siendo la red más completa del análisis.

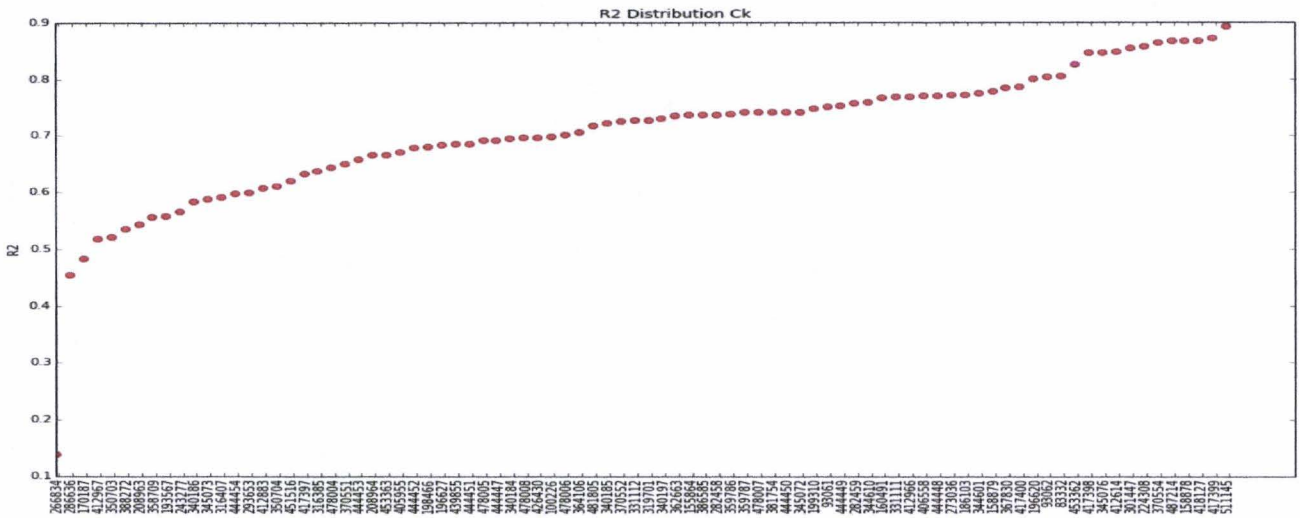
Figura 4. (A) Interacciones entre 83332-2011 y 83332-2012, (B) Nodos entre 83332-2011 y 83332-2012. (C) Interacciones entre 83332-2011-2012 y 83332-2015, el conjunto total de interacciones es la nueva red 83332-2011-2012-2015, (D) Nodos entre 83332-2011-2012 y 83332-2015.

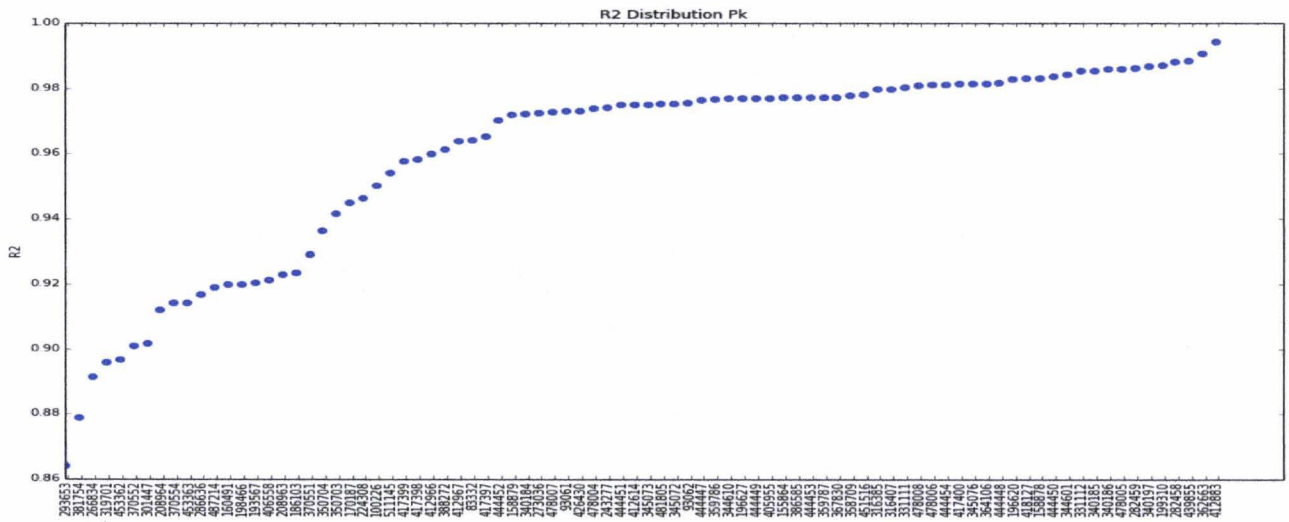


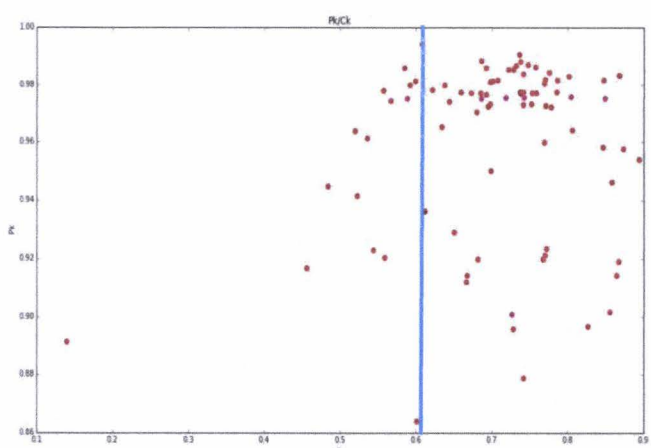
2.2.5. Selección de redes.

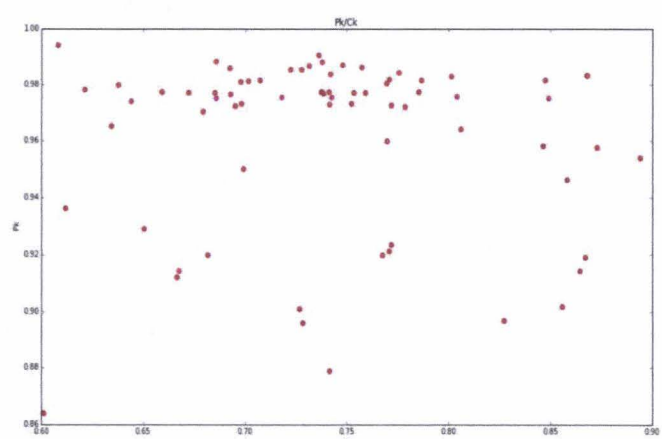
El Enfoque de Descomposición Natural al ser un dependiente de la conectividad y el coeficiente de agrupamiento de las redes, es incapaz de analizar redes inconexas o árboles. Se descartaron 363 redes por este motivo, dejando un total de 114 redes, además de las redes de *E. coli*, *B. subtilis*, *C. glutamicum*, *M. tuberculosis* y *P. aeruginosa*. Se recuperaron todas las redes que tuvieran una bondad de ajuste de la C (k) mayor a 0.6 y una bondad de ajuste de la P (K) mayor a 0.8, ambas bondades de ajuste son dadas por la R^2 [Albert-László Balasi et al. 2004].

Figura 5. Superior, distribución de bondad de ajuste (dado por la R^2) de la C (k) a una ley de potencia. Medio, distribución de bondad de ajuste (dado por la R^2) de la P (k) a una ley de potencia.. Inferior izquierda, gráfica de P(k) contra C (k) antes del corte; margen para el corte en C (k) (línea azul). Inferior derecha, gráfica de P (k) vs C (k) después de realizar el corte.









Estas 48 redes englobaban un total de 9 especies bacterianas diferentes [Tabla 3], pertenecientes a 4 clases bacterianas [Tabla 4] tanto Gram positivas (31 redes) como Gram Negativas (18 redes) englobando así la mayor cantidad posible de especies bacterianas [Figura 6].

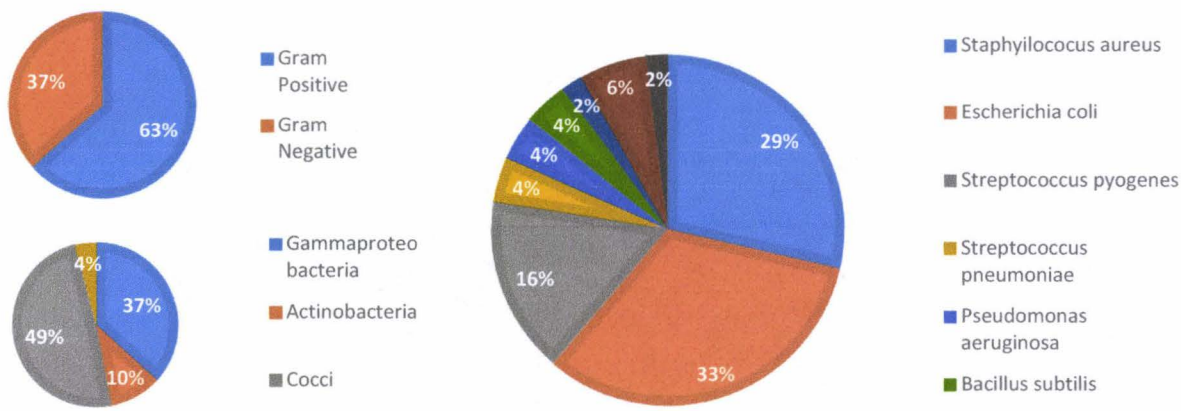
Tabla 3: Especies bacterianas únicas y cantidad de redes por especie.

Organismo	Cantidad de Redes
Staphylococcus aureus	14
Escherichia coli	15
Streptococcus pyogenes	8
Streptococcus pneumoniae	2
Pseudomonas aeruginosa	2
Bacillus subtilis	2
Corynebacterium glutamicum	1
Mycobacterium tuberculosis	3
Streptomyces coelicolor	1

Tabla 4: Clases bacterianas abarcadas en el estudio y cantidad de redes por clase

Clase Bacteriana	Número de redes
Gammaproteobacteria	17
Actinobacteria	5
Cocci	24
Bacilli	2

Figura 6: Superior izquierda, relación bacterias gram positivas (azul celeste) y gram negativas (naranja). Inferior izquierda, clases bacterianas gammaproteobacteria (azul celeste), actinobacteria (naranja), cocci (gris), bacilli (amarillo). Derecha, Redes por organismo, *S.aureus* (azul celeste), *E. coli* (naranja), *S. pyogenes* (gris), S. *pneomoniae* (amarillo), *P. aeruginosa* (azul), *B. subtilis* (verde), *C. glutamicum* (azul marino), *M. tuberculosis* (café), *S. coelicolor* (Gris oscuro).

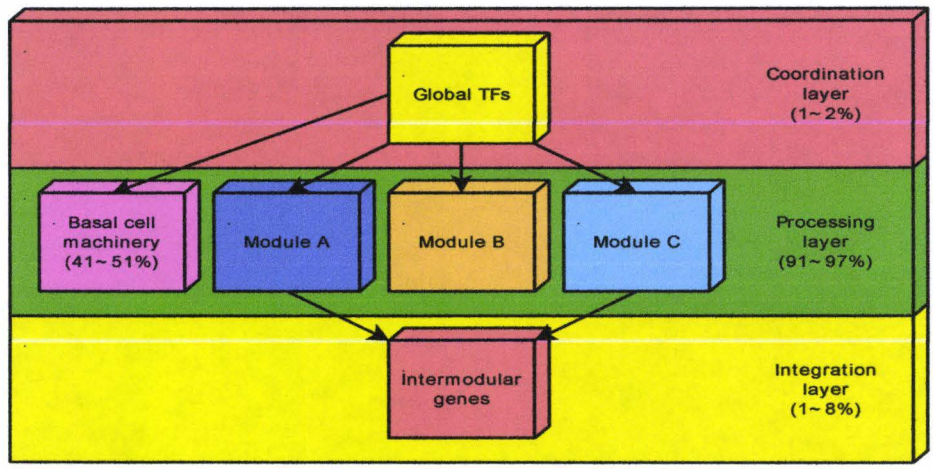


2.3. Descomposición y anotación de redes

2.3.1. Aplicación del Enfoque de Descomposición Natural

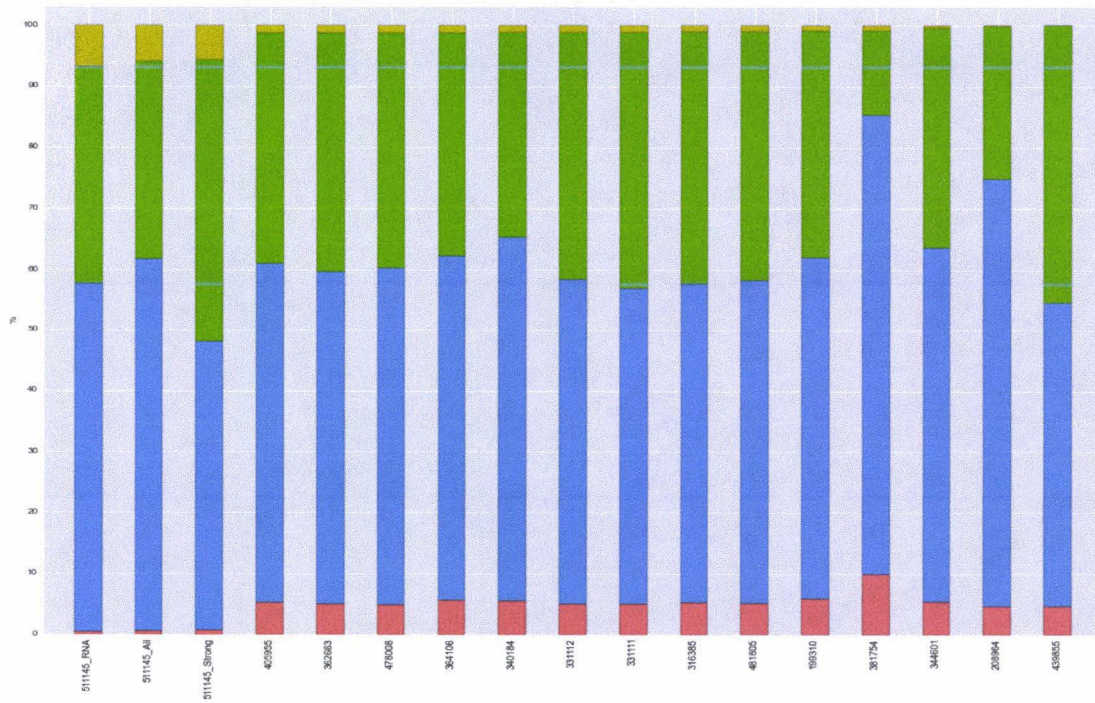
Una vez obtenidas las redes procedimos a aplicar una versión automatizada del Enfoque de Descomposición Natural desarrollado en nuestro grupo de investigcación a cada una de ellas. En estudios anteriores se había mencionado que del 1% al 2% de los nodos en una red pertenecían a la capa de coordinación (reguladores globales), que del 91% al 97% pertenecían a la capa de procesamiento (maquinaria basal y genes modulares) y el restante 1% a 8% a la capa de integración (genes intermodulares) [Figura 7][Julio A. Freyre-González, et al. 2012]. En el estudio actual se obtuvieron resultados promedio de 4.8% de reguladores globales, 27.3% para modulares, 66% para maquinaria basal (93.3% para la capa de procesamiento) y 1.74% para la capa de integración, mostrando así concordancia con los datos previamente encontrados. De las 49 redes presentadas sólo 4 redes no muestran las 3 capas de regulación: 370551 (Streptococcus pyogenes MGAS9429), 439855 (Escherichia coli (strain SMS-3-5 / SECEC)), 487214 (Streptococcus pneumoniae (strain Hungary19A-6)) y 186103 (Streptococcus pyogenes serotype M18 (strain MGAS8232)), las cuales carecen de genes intermodulares. Todas estas redes están compuestas por menos de 250 genes, lo cual muestra una carencia significativa de datos que podría explicar este comportamiento, ahondaremos en este punto más adelante [Figura 8].

Figura 7: capas regulatorias obtenidas por el NDA y sus componentes.



Las redes fueron analizadas en dos conjuntos, el conjunto de las bacterias Gram positivas y el de las Gram negativas, con el fin de identificar diferencias notorias entre los porcentajes de sus componentes y por lo tanto en la organización de sus redes [Figura 8,9].

Figura 8: Categorías en baterías Gram negativas, Intermodulares (Amarillo), maquinaria basal (Azul), modulares (verde).Reguladores globales (Rojo).



Se encontró que los porcentajes promedio para reguladores globales, maquinarias basales, modulares e intermodulares fue de: 5%, 72%, 21% y 2% para las bacterias Gram positivas y de 5%, 57%, 36% y 2% para las Gram Negativas [Figura 10]. Estos resultados muestran una clara diferencia en las proporciones de nodos modulares y maquinaria basal entre bacterias Gram positivas y Gram Negativas [Figura 11].

Figura 9: Categorías en baterías Gram positivas, Intermodulares (Amarillo), maquinaria basal (Azul), modulares (verde).Reguladores globales (Rojo).

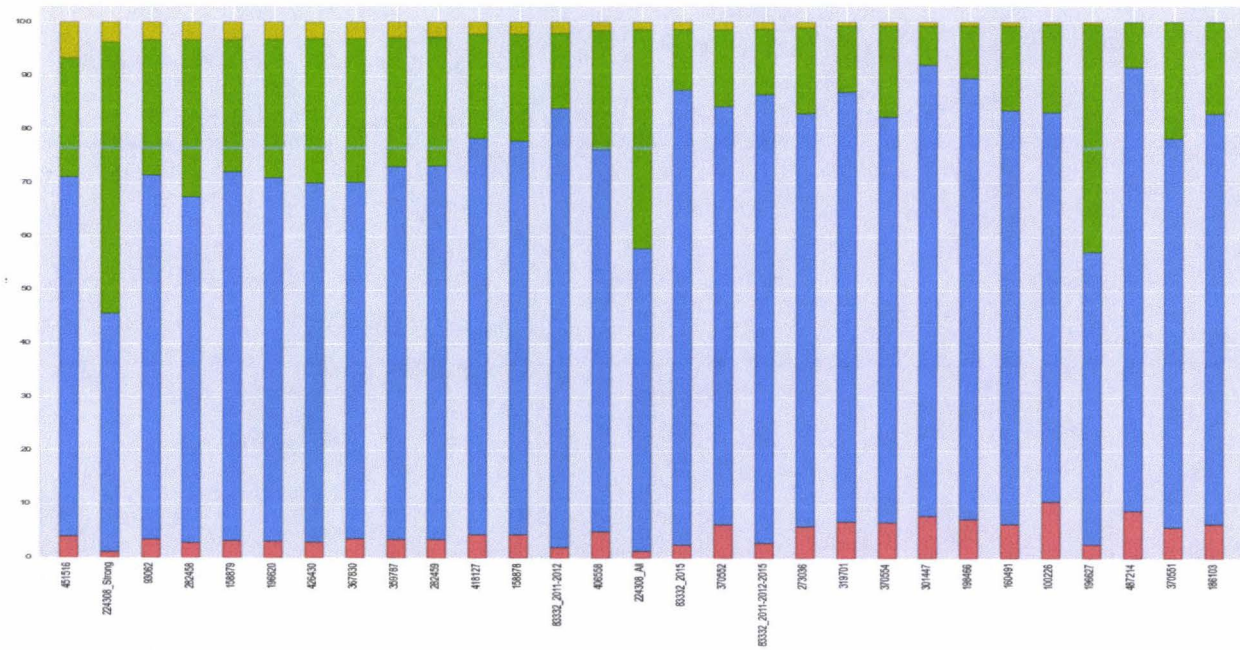
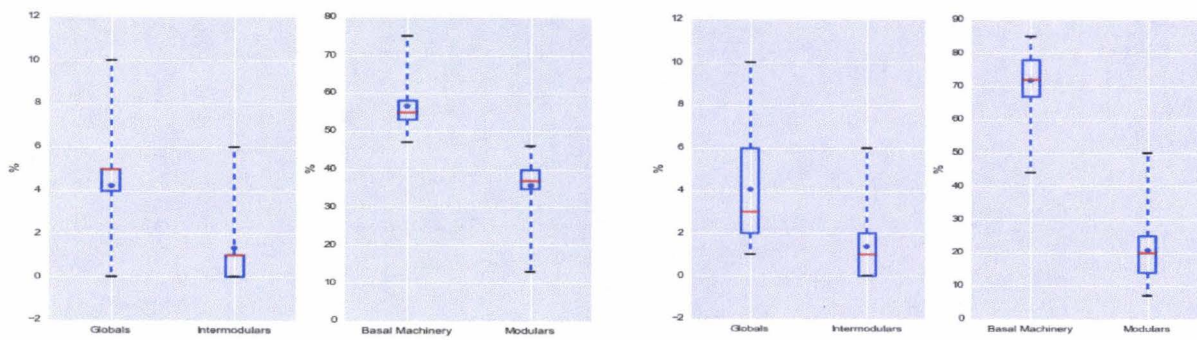


Figura 10: Diagramas de caja y brazos para cada categoría del Enfoque de Descomposición Natural mostrando medías (punto azul) y mediana (línea roja) para organismos Gram negativos (Izquierda) y organismos Gram positivos (Derecha).



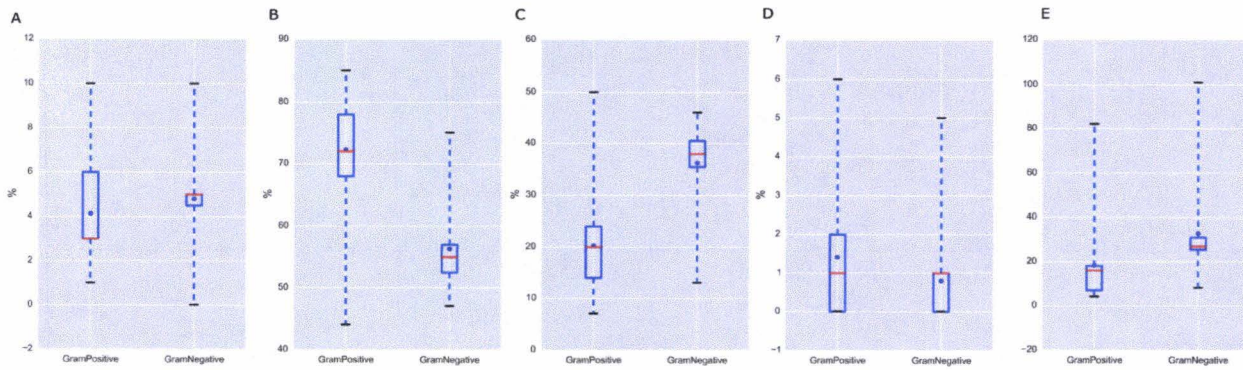
Para validar que las diferencias observadas entre nodos modulares e intermodulares para bacterias Gram positivas y Gram negativas tienen una significancia estadística realizamos pruebas Kolmogrov-Smirnov para determinar la bondad de ajuste de dos distribuciones de probabilidad y la U de Mann-Whitney para comprobar la heterogeneidad entre ambas muestras.

Los resultados arrojados por estas pruebas (Tabla 5) indican a un nivel de significancia de 0.0005 para la prueba Kolmogrov-Smirnov y 0.01 para Mann-Whitney que los genes modulares pertenecen a dos distribuciones diferentes en bacterias Gram negativas y Gram Positivas. Esto sugiriere la existencia de una diferencia en las organizaciones entre ambos tipos bacterianos, adicional a las diferencias en su regulación genética [Lozada-Chavez et al., 2008; Price et al., 2007; Sonenshein et al.2002].

Tabla 5: Heterogeneidad entre distribuciones y medias de las categorías obtenidas para los conjuntos de bacterias Gram positivas y Gram negativas.

	Globales	Maquinaria Basal	Modulares	Intermodulares	Módulos
Kolmogrov-Smirnov	0.0839	0.0053	0.0004	0.0119	1.1419e-06
Mann-Whitney U	0.2214	0.1047	0.0093	0.2050	0.0001

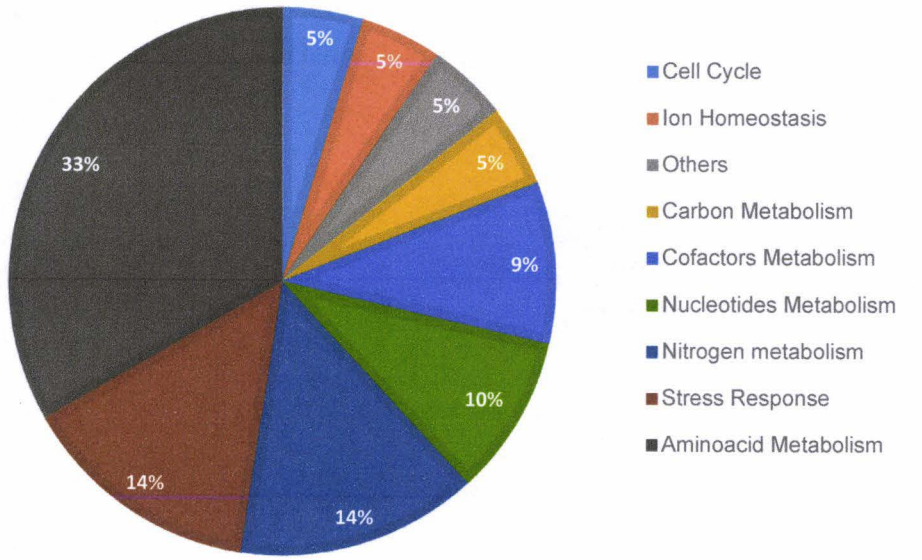
Figura 11: Diagrama de caja y brazos comparativo entre bacterias Gram positivas y Gram negativas reguladores globales (A), maquinaria basal (B), modulares (C), intermodulares (D) y cantidad de modulos (E).

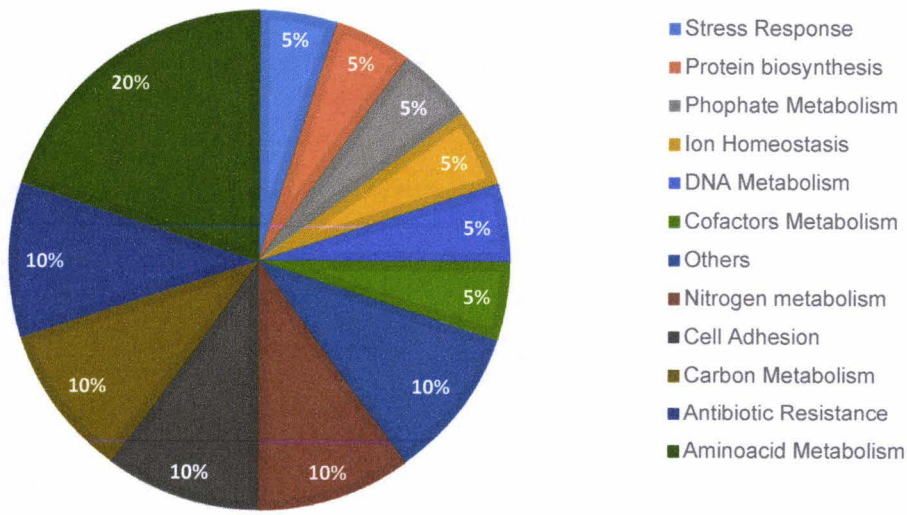


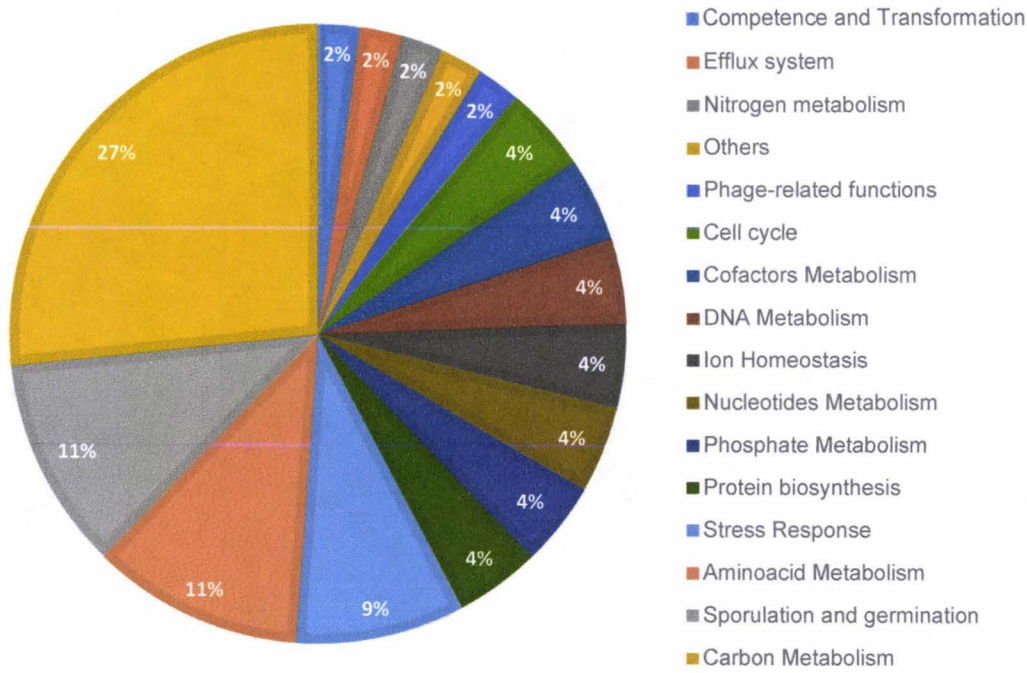
2.3.2. Anotación funcional de los módulos.

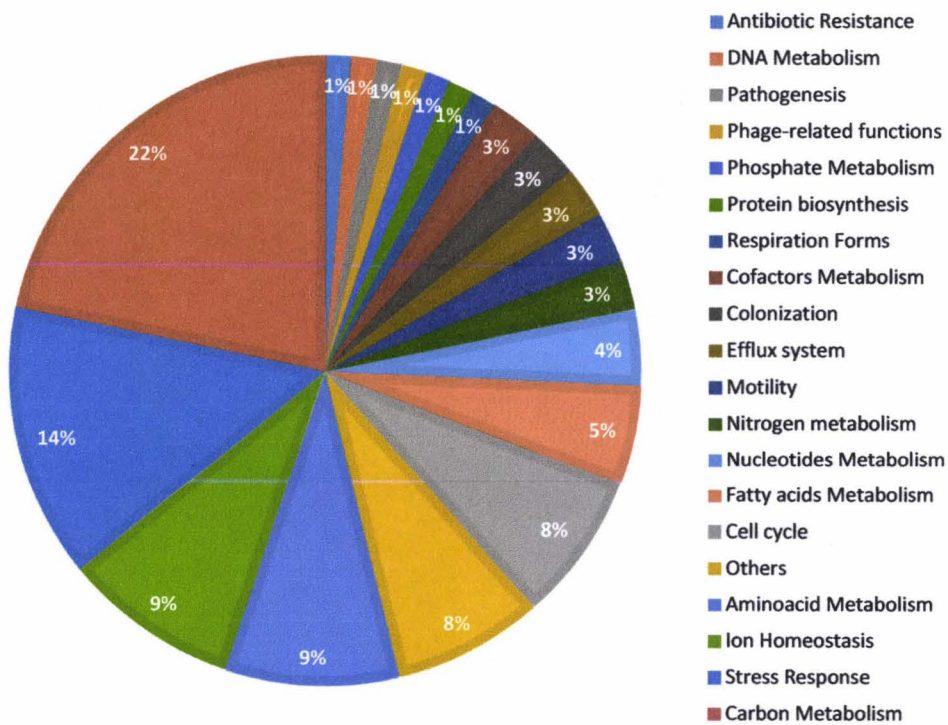
Cómo parte de este estudio realizamos una anotación computacional para todos los módulos obtenidos por este método. Primero le asignamos a cada gen su *Gene ontology*. Posteriormente calculamos los *p-values*, como medida de aleatoriedad en la distribución de las clases funcionales de los módulos identificados. Este *p-value* fue calculado basándose en una distribución hipergeométrica. Este método de anotación fue automatizado e implementado por nuestro grupo de investigación. Se anotaron 529 módulos a un nivel de significancia de 0.05 de un total de 1471 (35%) módulos para todos los organismos. Posteriormente, se procedió a realizar una categorización manual de los módulos anotados para los organismos: *M. tuberculosis* (18 Módulos), *E. coli* (78 Módulos), *P. aeruginosa* (23 Módulos), *B. subtilis* (45 Módulos) *y C. glutamicum* (21 Módulos). En estas categorizaciones es más fácil apreciar el poder predictivo del método al recuperar funciones generales de los módulos [Figura 12]. Las funciones más representadas en *E.coli* y *B. subtilis* fueron módulos relacionados con respuesta a estrés (14%) y metabolismo de carbono (22%) para *E. coli;* y módulos de metabolismo de carbono (27%), y esporulación (11%) para *B. subtilis*, concordando con resultados previamente publicados [Julio A. Freyre-González et al. 2012]. Para el caso de *C. glutamicum, P. aeruginosa* y *M. tuberculosis* sus funciones mayoritariamente representadas fueron: metabolismo de aminoácidos (33%) y respuesta a estrés (14%); metabolismo de aminoácidos (20%) y resistencia a antibióticos (10%); y respuesta a estrés (28%) y patogénesis (17%) respectivamente. Estos resultados concuerdan con los ciclos de vida y nichos preferidos de estos organismos.

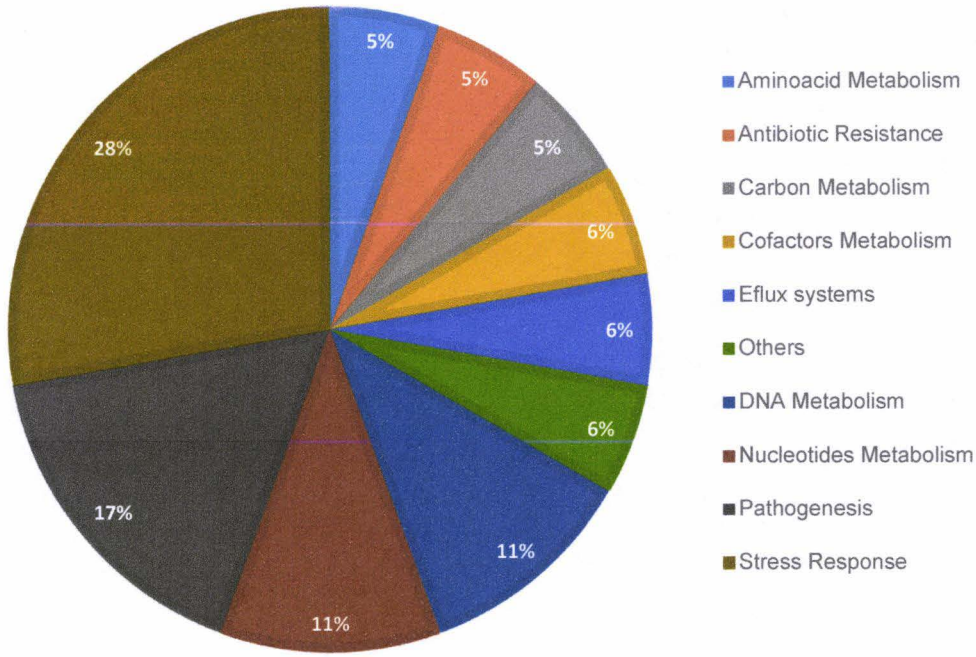
Figura 12: categorización de los módulos anotados con porcentajes con respecto al total de módulos anotados. En orden, *C. glutamicum, P. aeruginosa, B. subtilis, E. coli* y *M. tuberculosis.*






2%
2%
2%
2%
2%
4%
4%
4%
4%
4%
4%
4%
4%
9%
11%
11%
27%
Competence and Transformation
Efflux system
Nitrogen metabolism
Others
Phage-related functions
Cell cycle
Cofactors Metabolism
DNA Metabolism
Ion Homeostasis
Nucleotides Metabolism
Phosphate Metabolism
Protein biosynthesis
Stress Response
Aminoacid Metabolism
Sporulation and germination
Carbon Metabolism


1%
1%
1%
1%
1%
1%
1%
3%
3%
3%
3%
3%
4%
5%
8%
8%
9%
9%
14%
22%
Antibiotic Resistance
DNA Metabolism
Pathogenesis
Phage-related functions
Phosphate Metabolism
Protein biosynthesis
Respiration Forms
Cofactors Metabolism
Colonization
Efflux system
Motility
Nitrogen metabolism
Nucleotides Metabolism
Fatty acids Metabolism
Cell cycle
Others
Aminoacid Metabolism
Ion Homeostasis
Stress Response
Carbon Metabolism


5%
5%
5%
6%
6%
6%
11%
11%
17%
28%
Aminoacid Metabolism
Antibiotic Resistance
Carbon Metabolism
Cofactors Metabolism
Eflux systems
Others
DNA Metabolism
Nucleotides Metabolism
Pathogenesis
Stress Response

3. Robustez

Con el fin de conocer los alcances, limitaciones y parámetros óptimos del Enfoque de Descomposición Natural, así como el comportamiento de sus predicciones en redes en constante actualización, resultó necesario llevar el método a sus límites. Para lograr este objetivo series de perturbaciones biológicas y teóricas fueron realizadas.

3.1. Perturbaciones biológicas.

La biología es cambiante y el estudio de las entidades biológicas tiene que ser lo suficientemente flexible para tolerar estos cambios. Por eso mismo decidimos evaluar el Enfoque de Descomposición Natural considerando la actualización de las redes, el uso de distintas evidencias alternas en las mismas y la integración de capas alternas de regulación.

3.1.1. Actualización de las redes

Al encontrarse incompletas las redes de regulación, estas se someten constantemente a actualizaciones de sus componentes, ya sea añadiendo o eliminando nodos o interacciones. Mientras más se estudia una red más completa es. En este estudio se actualizó la red de *E. coli* con respecto al estudio realizado en 2012 por nuestro grupo [Julio A. Freyre-González, et al. 2012]. La red paso de contener el 37% (1692) de cobertura genómica y 4390 interacciones en la versión 2012, a tener un 42% (1889) de los nodos y 4006 interacciones para la red de interacciones con evidencia experimental fuerte en este estudio. Las proporciones de las categorías obtenidas por el NDA para la red estudiada en [Julio A. Freyre-González, et al. 2012, Julio A. Freyre-

González, et al. 2008] fueron de 0.89% para reguladores globales, 45.74% para genes modulares, 50.63% para genes solo regulados por globales y 2.5 % para genes intermodulares. Por su parte, la red actualizada usada en este estudio mostró proporciones de 0.74% para reguladores globales, 46.2% para genes modulares, 47% para genes sólo regulados por globales y un 5.6% para genes intermodulares [Figura 13]. Adicionalmente a estas proporciones notamos un aumento en la cantidad de módulos obtenidos, de 99 módulos en la red estudiada en [Julio A. Freyre-González, et al. 2012, Julio A. Freyre-González, et al. 2008] a 101 módulos en la red de este estudio.

3.1.2. *Uso de evidencias fuertes y débiles en las redes regulatorias de Escherichia coli y Bacillus subtilis.*

Hay diversas maneras de inferir una red regulatoria, y cada interacción dentro de la red tiene un nivel de veracidad dependiente al método con el que fue observada experimentalmente. Aquellas interacciones inferidas por patrones de expresión o inferencias del curador, son clasificadas como evidencias débiles, mientras que las obtenidas con ensayos de Inmunoprecipitación de cromatina (ChIP) con validación estadística, ensayos de transcripción *in vitro* o mutación *in situ,* son clasificadas como evidencias fuertes, debido a que son ensayos que muestran directamente la interacción entre el regulador y el ADN. Aunque para las redes de *E. coli* y *B. subtilis* la información sobre las evidencias de cada interacción está disponible, lamentablemente, no todas las redes de regulación cuentan con esta información.

En el caso de *E. coli* la red que sólo cuenta con evidencias fuertes consta de 1889 nodos y 2006 interacciones (42% del genoma), mientras que la red

formada con evidencias fuertes y evidencias débiles (red completa) consta de 3228 nodos y 7911 interacciones (71.7% del genoma). Para *B. subtilis* la red de evidencias fuertes se encuentra formada por 1414 nodos y 2590 interacciones (32% del genoma), y la red completa, por su parte, consta de 1696 nodos y 3132 interacciones (38% del genoma). Con el objetivo de evaluar la robustez de las predicciones del NDA en función de la confiabilidad de las interacciones, se procedió a descomponer las redes mencionadas anteriormente.

En el caso de *E. coli* la red de interacciones fuertes (511145_Strong) muestra una cantidad de 101 módulos de los cuales 84 (83%) presentan un enriquecimiento funcional. La red completa, por otro lado, tiene 114 módulos de los cuales 100 (88%) tienen enriquecimiento funcional. Para *B. subtilis* en su versión con interacciones fuertes (224308_Strong) se recuperó una cantidad de 82 módulos, de los cuales 46 (56%) se encuentran enriquecidos funcionalmente. En su versión completa (224308_All), de 87 módulos totales 48 (55%) se encuentran anotados.

Un ejemplo del cambio generado por la integración de las evidencias débiles a la red es el gen *oppA*. En la red con evidencias fuertes, sste gen es categorizado cómo un gen de maquinaria basal (solo regulado por genes globales), regulado por los genes fur y lrp. En el caso de la red completa, este mismo gen es categorizado como un gen modular y 3 interacciones son añadidas; dos dadas por los reguladores globales *fliA* y *arcA*, y una más por el gen modular *modE*. El módulo al cual pertenece *modE* (2.16) se encuentra estadísticamente enriquecido con funciones relacionadas al transporte de proteínas y respiración anaerobia.

3.1.3. Integración de RNAs regulatorios a la red regulatoria de *Escherichia coli.*

Muchos estudios se enfocan en redes de regulación que toman en cuenta solamente interacciones moduladas por TFs, sin embargo el rol de los RNAs regulatorios ya ha sido remarcado en otros estudios [Lauren S. Waters and Gisela Storz, 2009]. Algunas redes, como la de *B. subtilis* ya incluyen interacciones mediadas por RNAs regulatorios; sin embargo este no es el caso para *E.coli.* Por este motivo se decidió tomar todas aquellas relaciones regulatorias que incluyeran RNAs pequeños (sRNA) y añadirlas a la red preexistente. Esta información fue obtenida directamente de RegulonDB. Un total de 227 interacciones regulatorias con evidencia experimental débil fueron añadidas a la red completa de *E coli.* De esta manera una nueva red fue generada (511145_All+RNA) la cual consta de 3279 nodos y 8134 interacciones (73% del genoma).

En los resultados podemos apreciar que no hay un cambio radical en las proporciones de los componentes obtenidos por el NDA, sin embargo añadir esta capa alterna de regulación resulta enriquecedor para el poder predictivo del método. Una muestra del rol enriquecedor que nos otorga la integración de esta capa alterna de regulación es el gen *oppA*. En esta última red, *oppA* es categorizado como un gen intermodular, a comparación de su clasificación en la red que incluye todas las interacciones donde se clasifica como gen modular. Esto es debido a que en esta nueva red se agrega una interacción con el gen *gcvB* que es perteneciente al módulo 1.24, mientras que el gen *modE* es perteneciente al módulo 1.23 [Figura 14]. Este efecto se puede apreciar

también cuando se toman todas las evidencias regulatorias tanto fuertes cómo débiles contra cuando sólo se toman las evidencias fuertes. En ese caso el gen oppA es clasificado cómo un gen de maquinaria basal en la red fuerte y un gen modular en la red completa. Este fenómeno tiene lugar en 348 genes cuando se incluyen todas las interacciones en la red y en 244 genes cuando se agrega la capa de regulación mediada por RNAs regulatorios.

Figura 13. Observando los cambios en las proporciones de las categorías obtenidas para las diferentes redes.

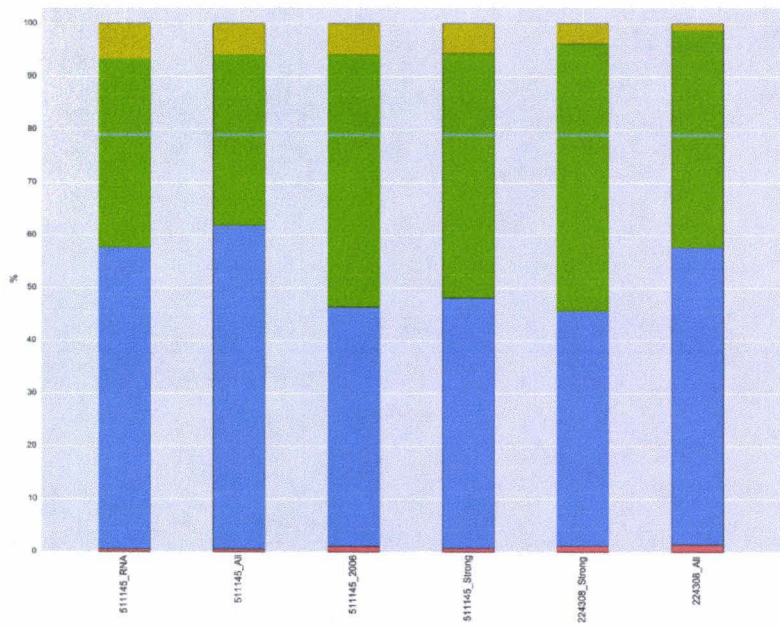
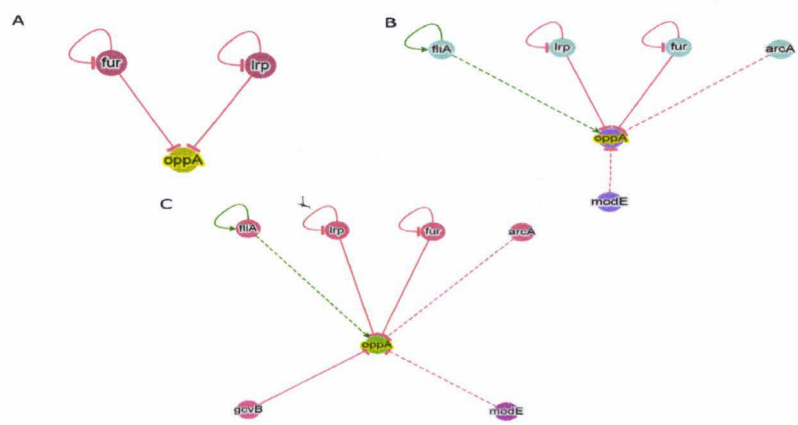


Figura 14. Efecto de la adición de nuevos elementos regulatorios en un gen. Red de interacciones fuertes de *E. coli* (A), red de interacciones fuertes más interacciones débiles *E. coli* (B) y red con interacciones fuertes, débiles y sRNAs *E. coli* (C).



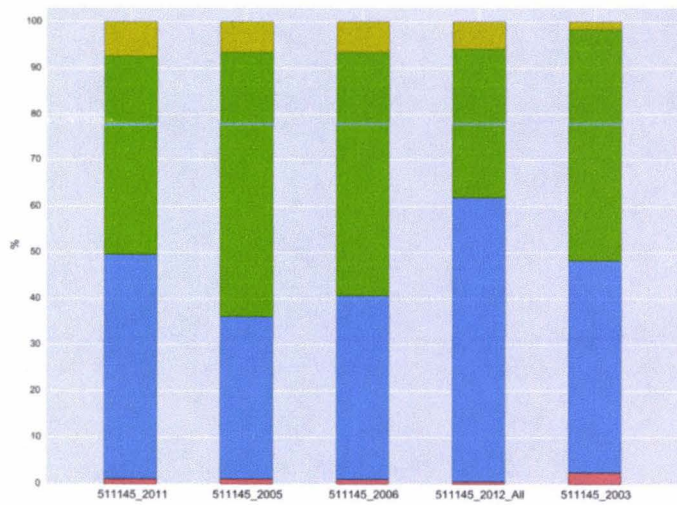
3.1.4. Sensibilidad de las predicciones para *Escherichia coli* a través del tiempo.

En este estudio ya se hemos realizado pruebas de cómo son afectadas las predicciones debido a la actualización de las redes, sin embargo no se ha observado el comportamiento de una misma red a lo largo del tiempo. Por este motivo se decidió aplicar el enfoque de descomposición natural a diversas redes pertenecientes a la bacteria *E coli.* Estas redes regulatorias obtenidas de RegulonDB en diversos momentos, fueron analizadas con el NDA [Tabla 6] y sus resultados graficados [Figura 15].

Tabla 6. Se observa un aumento en la cantidad de genes intermodulares, así como aumento en la cantidad de los módulos, también se observa un incremento de los genes modulares indicando que aumentan tanto la cantidad de módulos cómo el tamaño de estos.

Red	Globales	Modulares	Intermodulares	Maquinaria Basal	Módulos	Completes
2003	19	397	13	363	68	792 (17%)
2005	13	685	79	420	81	1197 (26%)
2006	16	864	108	650	98	1638 (36%)
2011	23	910	158	1027	113	2118 (47%)
2012	18	1040	190	1985	114	3233 (72%)

Figura 15. Cambio de proporciones en redes regulatorias de *Escherichia coli* a lo largo del tiempo.



Se observa un aumento en la cantidad de genes intermodulares, así como aumento en la cantidad de los módulos, también se observa un incremento de los genes modulares indicando que aumentan tanto la cantidad de módulos cómo el tamaño de estos. En el caso de la red de 2011 comparada con la red de 2012 se puede observar un aumento de casi 100 genes modulares y un solo módulo indicando que estos genes añadidos pertenecen a su mayoría a módulos previamente identificados. Esto sugiere la red regulatoria de *E.coli* este llegando probablemente a su completez y que su organización topológica no sufrirá muchos cambios.

3.2. Perturbaciones teóricas

Una vez probada la flexibilidad del Enfoque de Descomposición Natural a los cambios biológicos en las redes, resta investigar cómo se comporta el método ante diferentes parámetros. Para lograr este objetivo llevamos el método al límite realizando pruebas de reducción de la red por medio de muestreo de nodos y de aristas, crecimiento de las redes al 100% del genoma y pruebas en los parámetros como lo son el uso de conectividad dirigida y coeficiente de agrupamiento dirigido.

3.2.1. Muestreo aleatorio de la red regulatoria de *E. coli* y *B. subtilis.*

Dado que no existe una red de regulación completamente descubierta es importante evaluar cómo cambian las predicciones en función de la incompletez de las redes. Para ello recurrimos a un muestreo de nodos y aristas realizado con las redes regulatorias de *E. coli* y *B. subtilis*. Estas redes fueron reducidas gradualmente tanto en número de nodos cómo de aristas y obtenidos los porcentajes de sus componentes.

3.2.1.1. Nodos

Las redes fueron muestreadas 1000 veces por cada reducción de un 10% la cantidad de nodos [Tablas 7,8]. Podemos observar una dramática reducción en la proporción de los genes intermodulares y modulares conforme se reduce el tamaño de la red, a su vez podemos observar un incremento en la proporción de genes de maquinaria basal [Figura 16], no obstante al tratarse de una eliminación de nodos la cantidad de estos disminuye.

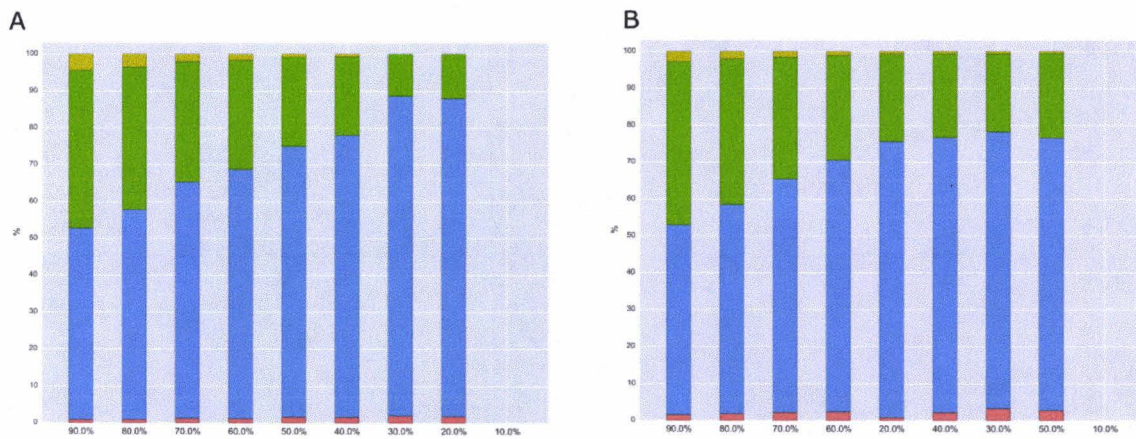
Tabla 7. Promedio de genes por categoría por red muestreada para *E. coli*

Tamaño respecto a la red	Tamaño respecto al genoma	Reguladores Globales	Intermodulares	Modulares	Maquinaria Basal
1889 (100%)	1889 (42%)	14	106	874	895
1700 (90%)	1700 (37%)	14.54	72.83	719.04	893.59
1511 (80%)	1511 (33%)	15.03	47.8	575.07	873.1
1322 (70%)	1332 (29%)	15.18	29.38	439.09	838.35
1133 (60%)	1133 (25%)	14.94	16.65	328.16	773.24
944 (50%)	944 (21%)	13.89	8.14	226.52	695.46
756 (40%)	756 (17%)	13.39	3.65	147.52	591.43
567 (30%)	567 (13%)	6.9	4.9	106.8	448.4
378 (20%)	378 (8%)	4	0	60	314
189 (10%)	189 (4%)	0	0	0	0

Tabla 8. Promedio de nodos por categoría por red muestreada para B. subtilis.

Tamaño respecto a la red	Tamaño respecto al genoma	Reguladores Globales	Intermodulares	Modulares	Maquinaria Basal
1414 (100%)	1414 (32%)	16	53	715	630
1213 (90%)	1213 (27%)	18.95	35.14	563.28	655.63
1131 (80%)	1131 (25%)	19.35	23.99	445.36	642.31
990 (70%)	990 (22%)	20.66	15.13	327.18	627.03
848 (60%)	848 (19%)	19.21	9.27	240.7	578.82
707 (50%)	707 (16%)	18.73	3.73	161.73	522.82
566 (40%)	566 (13%)	11.83	3.44	128.17	422.56
424 (30%)	424 (10%)	13.5	2.5	89.75	318.25
283 (20%)	283 (6%)	2	2	67	212
141 (10%)	141 (3%)	0	0	0	0

Figura 16. Cambio de las proporciones en los componentes en *E. coli*(A) y *B. subtilis* (B) conforme se reduce la cantidad de aristas muestreadas. Se puede observar claramente la pérdida gradual de los nodos intermodulares y modulares, y el aumento gradual de los nodos sólo regulados por globales.



3.2.1.2. Aristas

Las redes fueron muestreadas 1000 veces por cada reducción de un 10% la cantidad de aristas en la red. A cada conjunto de 1000 redes se le aplicó el NDA y los promedios de cada categoría fueron calculados [Tablas 9,10]. De la misma manera que con el muestreo de nodos se puede apreciar una tendencia en la disminución de la proporción de genes modulares e intermodulares, y un aumento tanto en la proporción como con en la cantidad de genes de maquinaria basa. Al tratarse de una eliminación de aristas la cantidad de genes

de maquinaria basal aumenta a diferencia de la eliminación de nodos donde la cantidad de estos disminuye.

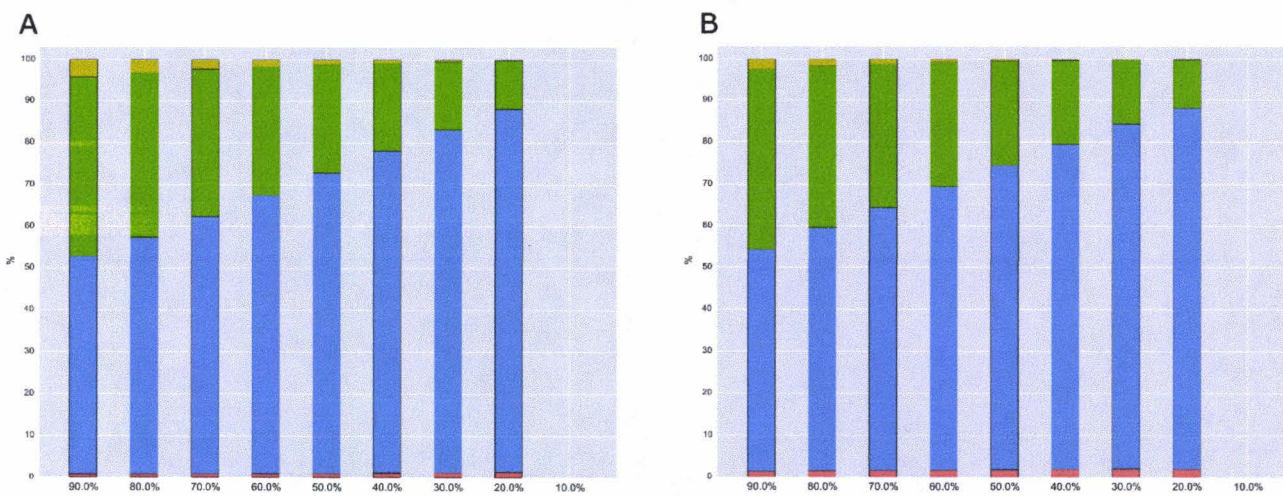
Tabla 9. Promedio de aristas por categoría en muestreo de aristas para *B. subtilis*.

Aristas en la red	Reguladores Globales	Intermodulares	Modulares	Maquinaria Basal
2589 (100%)	14	106	874	895
2330 (90%)	15.24	78.49	811.87	983.41
2071 (80%)	15.98	59.59	745.5	1067.93
1812 (70%)	17.02	42.99	667.57	1161.41
1553 (60%)	18.49	30.45	581.01	1259.06
1294 (50%)	20.22	20.3	491.69	1356.79
1036 (40%)	22.1	12.36	402.28	1452.26
777 (30%)	24.38	5.96	311.98	1546.68
518 (20%)	25.49	2.15	224.6	1636.76
259 (10%)	0	0	0	0

Tabla 10. Promedio de aristas por categoría en muestreo de aristas para *E. coli*.

Aristas en la red	Reguladores Globales	Intermodulares	Modulares	Maquinaria Basal
4006 (100%)	16	53	715	630
3605 (90%)	19.91	34.08	611.23	748.78
3205 (80%)	21.38	23.15	548.45	821.02
2804 (70%)	22.55	15.72	488.15	887.59
2404 (60%)	24.08	9.61	421.64	958.67
2003 (50%)	25.72	5.77	354.44	1029.06
1602 (40%)	27.4	3.33	286.28	1096.99
1202 (30%)	28.63	1.89	220.62	1162.86
801 (20%)	27.1	0.87	167.03	1219
401 (10%)	0	0	0	0

Figura 17. Cambio de las proporciones en los componentes en *E. coli*(A) y *B. subtilis* (B) conforme se reduce la cantidad de aristas muestreadas.



3.2.2. Crecimiento de la red regulatoria de *Escherichia coli y Bacillus subtilis.*

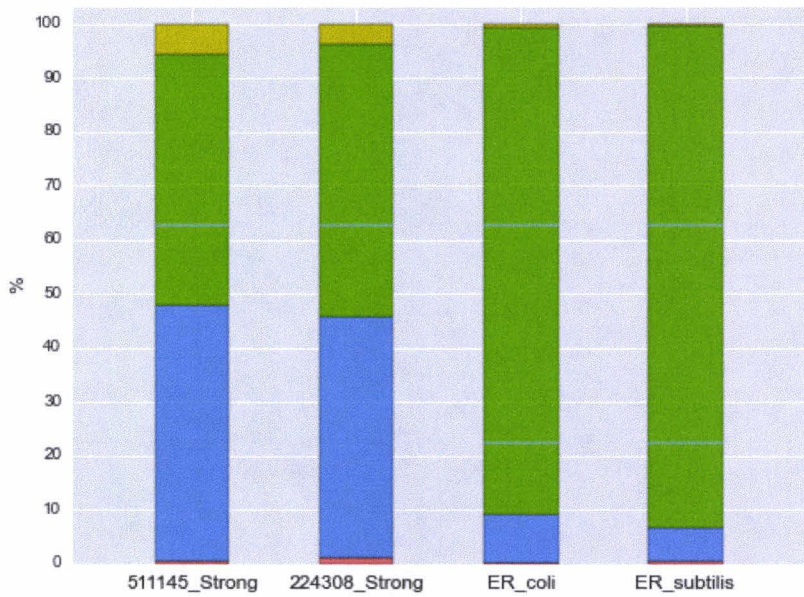
Si bien en el muestreo de nodos y aristas fue observable una tendencia, resulta de igual importancia conocer el comportamiento que seguirán estas redes conforme se completen. Para poder observar el comportamiento del método en redes conteniendo el 100% del genoma recurrimos al crecimiento de estas redes. Para crecer estas redes optamos por dos caminos. El primero de ellos es un crecimiento siguiendo un modelo de Erdos-Rényi, caracterizado por ser un crecimiento aleatorio, el cual empleamos como control negativo. El segundo de ellos es un crecimiento siguiendo un modelo de Barabási-Albert, el cual posee la característica de seguir una ley de potencia. Cabe recalcar que las interacciones y nodos añadidos por estos métodos son sólo teóricos, como una aproximación de cómo podría comportarse la red pero no representan una red verídica. A pesar de esta limitante, y dado que no existe un modelo de crecimiento de redes conforme estas son curadas, el modelo de Barabási-Albert es el más cercano a cómo crece una red de regulación.

3.2.2.1. Erdos-Rényi

El modelo de Erdos-Rényi es un método empleado para la generación de grafos aleatorios. Se basa en la premisa de que un nuevo nodo se enlaza con la misma probabilidad con el resto de la red, es decir que posee una independencia estadística con el resto de los nodos. Usando esté método procedimos a crecer tanto la red de *Escherichia coli* como la de *Bacillus subtilis* hasta el tamaño de su genoma (4497 y 4421 genes respectivamente). Este proceso de crecimiento fue realizado un total de 1000 veces para cada

red. Una vez obtenidas las redes procedimos a realizar el Enfoque de Descomposición Natural a cada una de ellas y calculamos los promedios de cada uno de sus componentes [Figura 18]. Cómo se puede observar existe un aumento dramático en la proporción de genes modulares, debido al aumento en el agrupamiento entre los nodos.

Figura 18. Aumento en la proporción de genes modulares en redes crecidas por el método de Erdos-Rényi.

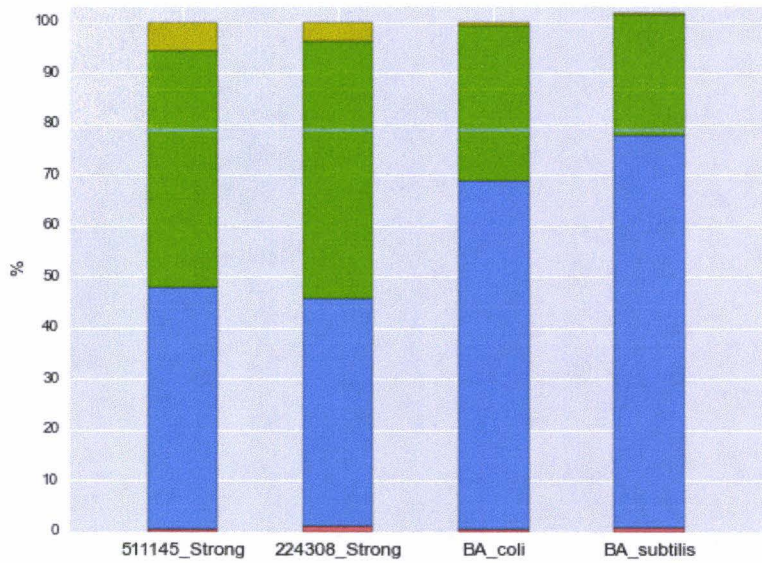


3.2.2.2. Barabási-Albert

El modelo de Barabási-Albert por su parte es un método para la generación de redes aleatorias libres de escala usando un mecanismo conocido como conexión preferencial, en el cual un nodo se conecta con mayor probabilidad a un nodo altamente conectado. De la misma manera que con el crecimiento realizado por Erdos-Rényi se crecieron las redes de ambos organismos al tamaño de su genoma. El crecimiento fue realizado 1000 veces y para cada

red fue aplicado el NDA, los resultados fueron promediados y graficados [Figura 19]. Se puede observar un aumento en los genes sólo regulado por globales.

Figura 19. Aumento en la proporción de genes sólo regulados por globales en redes crecidas siguiendo un método de Barabási-Albert.



3.2.3. Uso del coeficiente de agrupamiento dirigido y conectividad dirigida.

El Enfoque de Descomposición Natural, al ser un método dependiente de la topología, varía en función de la distribución de la conectividad y la distribución de agrupamiento. Ambas distribuciones pueden ser calculadas de diferentes maneras. En el caso de la conectividad esta puede ser dirigida o no dirigida, de la misma manera el agrupamiento puede o no tomar en cuenta la dirección de las interacciones. En una red regulatoria todas las interacciones presentan dirección, es decir, hablamos de una red dirigida, por lo que en teoría deberían de usarse distribuciones de grado y agrupamiento dirigidas. Para probar cual

es el mejor parámetro usamos todas las combinaciones de estos en la red de *E.coli* (sólo interacciones con evidencia experimental fuerte). Se calculó la especificidad y sensibilidad de la combinación formada por estos parametros tomando los reguladores globales por cada combinación y comparándolos con un concenso de reguladores globales obtenidos de la literatura (*fur, arcA, fis, IHF, fnr, lrp, crp, rpoS, rpoE, rpoD, rpoH, rpoN, narL, hns*). Otra medida fue calcular el valor de kappa para cada combinación [Tabla 11].

Tabla 11. Sensibilidad y especificidad de las diferentes combinaciones de parámetros. Se observa una caída del valor de kappa conforme más dirección es incluida en la red, tomando más genes cómo globales que aquellos establecidos en la literatura.

Combinación	Especificidad	Sensibilidad	Globales	Kappa
Conectividad no dirigida Agrupamiento no dirigido	0.999466950959	1	fur, arcA, fis, IHF, fnr, lrp, crp, rpoS, rpoE, rpoD, rpoH, rpoN, narL, hns	50.22
Conectividad dirigida Agrupamiento no dirigido	0.999466950959	1	fur, arcA, fis, IHF, fnr, lrp, crp, rpoS, rpoE, rpoD, rpoH, rpoN, narL, hns	42.67
Conectividad no dirigida Agrupamiento dirigido	0.997867803838	1	fur, arcA, fis, IHF, fnr, lrp, argR, cpxR, crp, rpoS, rpoE, rpoD, rpoH, rpoN, narL, modE, hns	36.52
Conectividad dirigida Agrupamiento dirigido	0.995735607676	1	fur, arcA, fis, pdhR, IHF, fnr, lrp, argR, cpxR, crp, cra, rpoS, rpoE, rpoD, rpoH, rpoN, phoB, narP, narL, modE, hns	30.53

Conclusiones.

Con respecto a los efectos causados por la cantidad de información disponible en los resultados del método, observamos:

1) Cuando las redes biológicas mejor curadas son crecidas al tamaño total del genoma, aumenta la proporción de genes modulares cuando son crecidas usando el modelo de Erdos-Rényi, y la proporción de genes sólo regulados por globales cuando se usa el método de Barabási-Albert. En ambas se reduce la cantidad de genes intermodulares. Esto indica que estos métodos de crecimiento resultan insuficientes para evaluar una red del tamaño del genoma. Sin embargo, aun sin el conocimiento de todas las interacciones y componentes de una red regulatoria podemos describir de manera general su organización modular. Y el aumento de la información regulatoria del organismo permitirá refinar la categorización de los componentes de la red.

2) Por el contrario, cuando las redes regulatorias carecen de información, se observa un aumento en la proporción de genes de maquinaria basal y una disminución en la proporción de genes modulares e intermodulares, llegando estos últimos a desaparecer en algunos casos. Esta pérdida de dicha característica emergente es apreciable cuando las redes son inferiores al 10% del genoma.

Gracias al análisis llevado a cabo en *E. coli* en el tiempo, es evidente que algunos resultados que no concordaron con al modelo de organización regulatorio propuesto, cambien conforme aumente la información sobre las interacciones y componentes de las redes analizadas.

4. Base de datos/BactSystDB

Debido a la gran cantidad de información producida y categorizada en este estudio, surgió en nosotros la iniciativa de desarrollar BactSystDB, una base de datos sistemas en redes regulatorias. Pese a la existencia de bases de datos de interacciones regulatorias cómo RegulonDB, DBTBD, CoryneRegNet y RegTransBase estás no presentan sus datos más allá del nivel de regulón, por lo que no es evidente cómo estas interacciones se organizan en sistemas. BactSystDB es la primera base de datos en su tipo, al identificar los sistemas pertenecientes a las redes regulatorias nos brinda la información necesaria para realizar diversos estudios, también otorga una visión global del organismo.

4.1. Generación de las tablas

Al trabajar con una gran cantidad de datos fue necesario el desarrollo de una línea de obtención y procesamiento de los mismos, logrando así un proceso automatizado para la generación de las tablas usadas en la base de datos [Figura 20]. La obtención de los datos fue realizada de diferentes fuentes como son: UniProt, GOA, EggNOG, NCBI, RegTransBase, CoryneRegNet, DBTBS, RegulonDB y de literatura [Figura 21]. Como se mencionó anteriormente uno de los principales problemas al momento de realizar análisis de redes son los sinónimos. Por esta razón se buscó agrupar la mayor cantidad de sinónimos posibles para cada organismo con el objetivo de generar un compendio de fácil acceso y poder así evitar la redundancia. Estos sinónimos provinieron tanto de NCBI como de UniProt. Y son modelados en la tabla 'genes' con el objetivo de mejorar la búsqueda en la base de datos.

El proceso de generación de datos es dividido en dos partes principales. La primera consiste en la obtención de datos no relacionados con las predicciones como las tablas relacionadas con GOs y COGs. La segunda involucra la aplicación del enfoque de descomposición natural. Para llevar a cabo esta parte se hizo un programa capaz de generar las tablas 'organisms', 'genes', 'computational_annot' y 'modules' con una sola corrida del método por red. Esto con el objetivo de conseguir concordancia en los datos. El programa databasecreator.py es el encargado de este fin [Figura 22].

Figura 20. Diagrama de entidad relación usado en la base de datos.

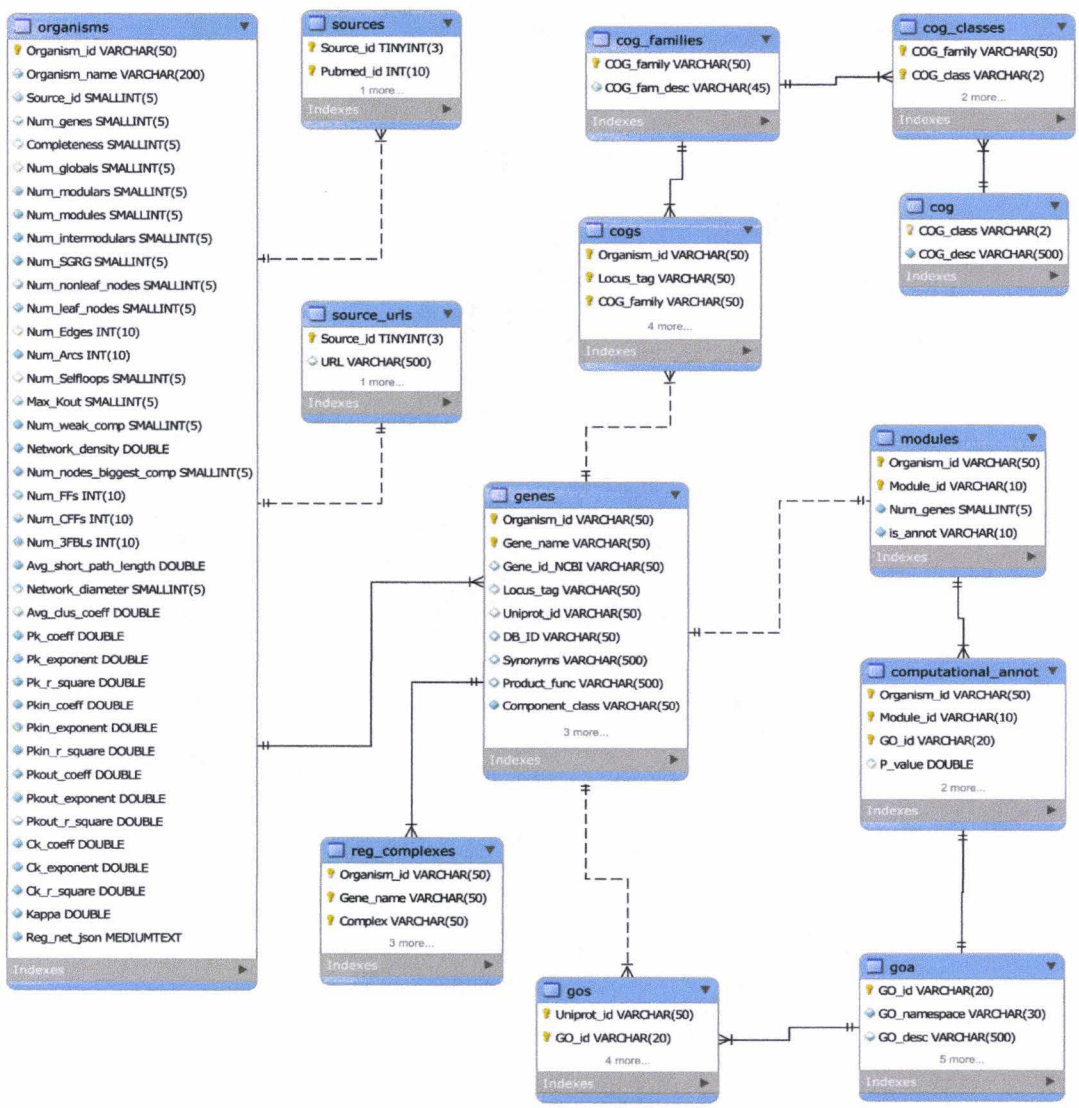


Figura 21. Fuentes de información usadas para el llenado de la base de datos.

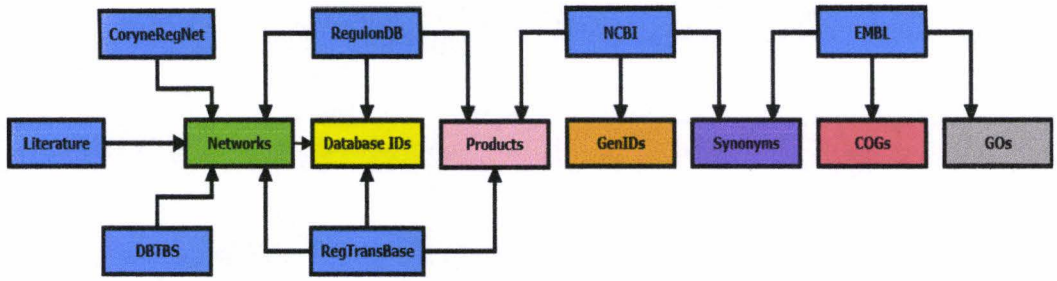
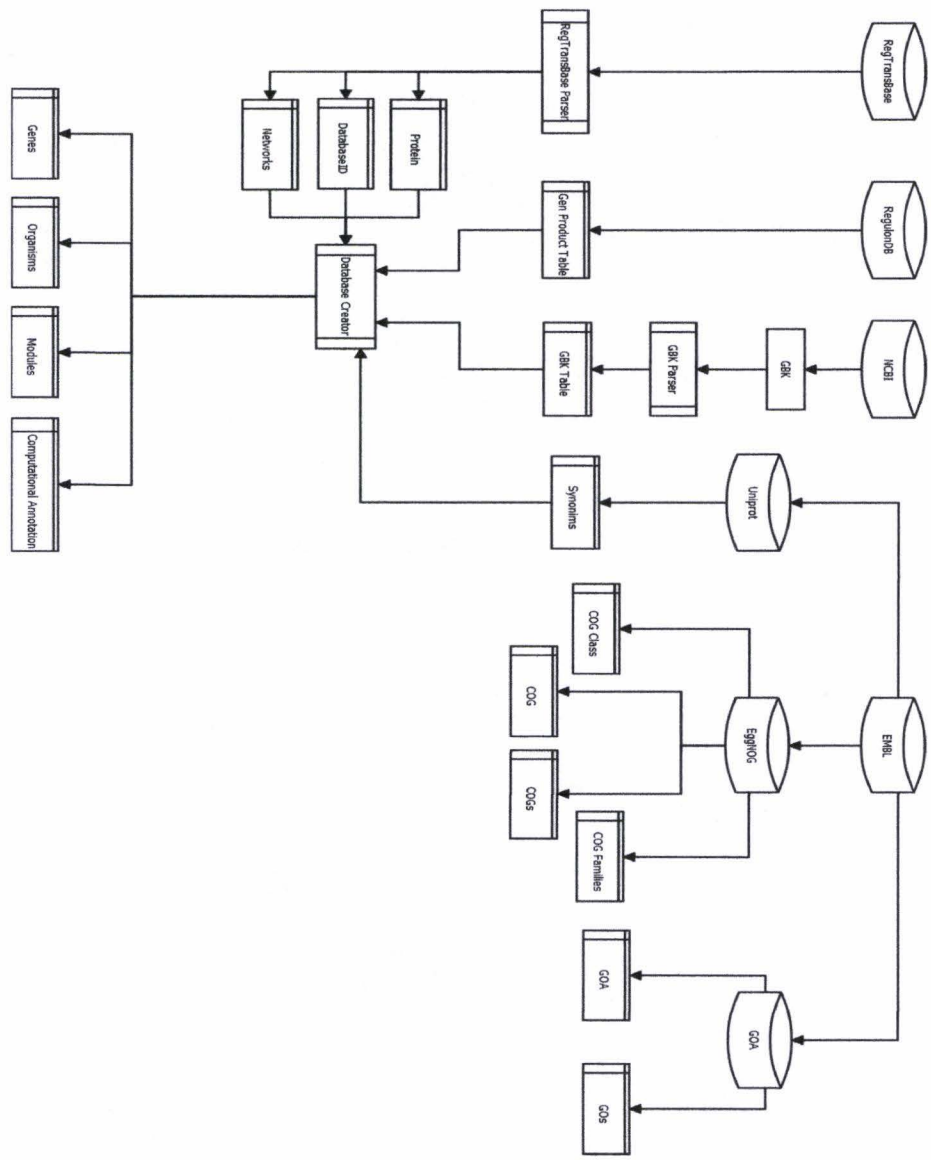


Figura 22. Esquema del procesamiento de datos y generación de las tablas usadas en BactSystDB.



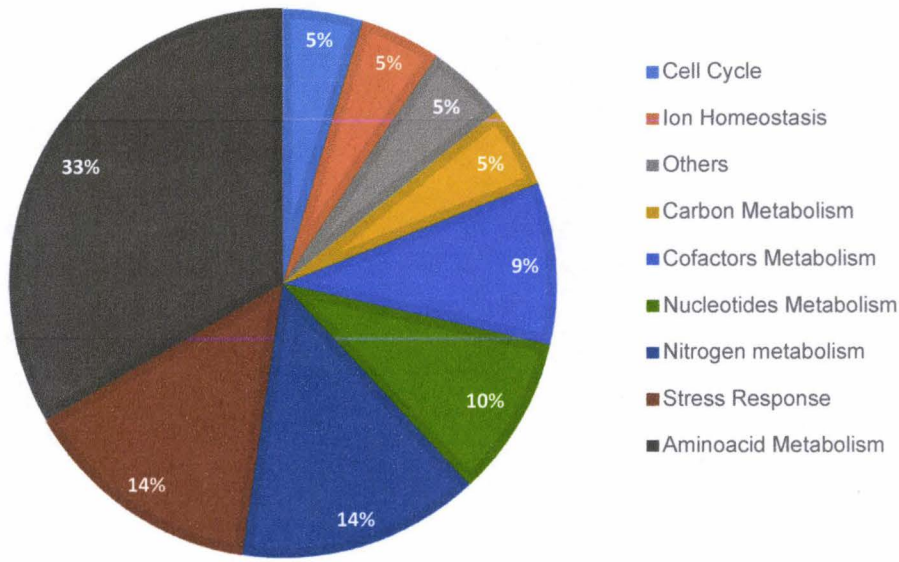
4.2. Casos de estudio

Dado que los repositorios actuales de interacciones regulatorias en bacterias, se limitan a ser descriptivos en interacciones únicas, la base de datos es de gran relevancia, al ser la primera que busca poner a disposición de todos los biólogos una visión integral de las actividades regulatorias coordinadas de las bacterias más estudiadas, y representa una oportunidad para integrar la información emergente de otras. Un ejemplo de esto es la anotación funcional de la bacteria Gram positiva *C. glutamicum,* la cual ha sido estudiada por mucho tiempo pero nunca habían sido identificados y anotados sus sistemas. Dos ejemplos claros de la integración de información emergente son los casos de las bacterias *P. aeruginosa* y *M.* tuberculosis, cuyas redes fueron obtenidas por medio de la unión de dos o más redes regulatorias, proveyendo así una visión más completa del panorama regulatorio de ambos organismos.

4.2.1. *Corynebacterium glutamicum*

Esta bacteria de gran importancia industrial es presentada en esta primera versión de la base de datos. De un total de 60 módulos identificados 21 de ellos (35%) fueron estadísticamente validados con un enriquecimiento funcional. Esta anotación fue llevada a cabo con ayuda de una metodología automatizada usando ontologías génicas. Las funciones más representadas en los 21 módulos anotados en este organismo fueron: Metabolismo de aminoácidos, con 7 módulos anotados, respuesta a estrés, con 3 módulos, y metabolismo de nitrógeno, con 3 módulos. Estas 3 funciones abarcan más de 50% de los módulos anotados en el organismo y representan una buena aproximación de las principales funciones dentro de este organismo [Figura 23].

Figura 23. Módulos enriquecidos funcionalmente para *Corynebacterium glutamicum*.

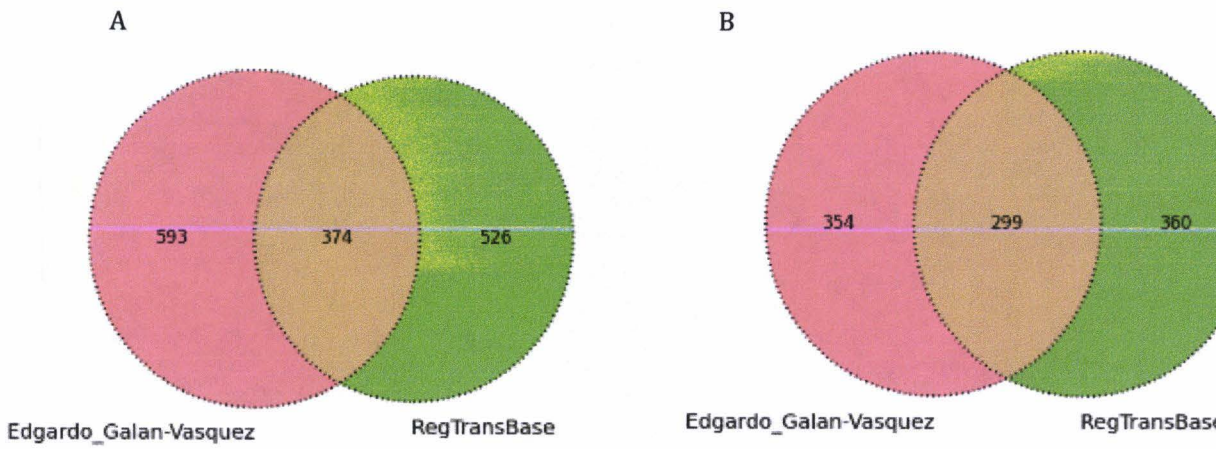


4.2.2. *Pseudomonas aeruginosa*

Una parte importante de la biología de redes es conocer la mayor cantidad de interacciones posibles dentro de un organismo. Es por esto que resulta elemental la generación de datos que den origen a redes regulatorias. Sin embargo, los datos generados por estos estudios ya sea a gran escala o en pequeñas instancias no siempre son reunidos en una misma red. Esta fragmentación impide que podamos ver un todo, o al menos una parte más completa de este. Por este motivo durante la construcción de BactSystDB decidimos unir estos datos dispersos para así tener una red más completa. Una instancia de esta reunión de datos es la representada por la bacteria Gram negativa *Pseudomonas aeruginosa*.

Como se mencionó antes la red presentada en la base de datos y con la que se llevaron a cabo todos los análisis proviene de la unión de dos redes previamente publicadas. La primera de estas redes publicada por Edgardo Galán-Vásquez [Galán-Vásquez et al, 2011] con 967 interacciones regulatorias involucrando a 653 genes. Mientras que la segunda de ellas fue obtenida de la base de datos RegTransBase la cual consta de 900 interacciones y 659 genes [Alexei E. Kazakov et al, 2006]. Ambas redes fueron mostraron una similitud de 0.3 para las interacciones y 0.25 para los nodos (índice de Jaccard), siendo así complementarias y coherentes entre sí [Figura 24]. La nueva red presenta una cantidad de 905 genes (16% del genoma) y un total de 1373 interacciones, siendo esta red la más completa hasta el momento.

Figura 24. Se observa la complementariedad de ambas redes tanto a nivel de nodos como a nivel (A) de aristas (B)



4.2.3. *Mycobacterium tuberculosis*

El caso de esta bacteria requiere de especial énfasis debido a la cantidad de redes encontradas en la literatura. Un total de 4 redes fueron encontradas: 2008, 2011, 2012 y en 2015. La similitud entre estas redes fue analizada y con base en estos resultados fueron agrupadas.

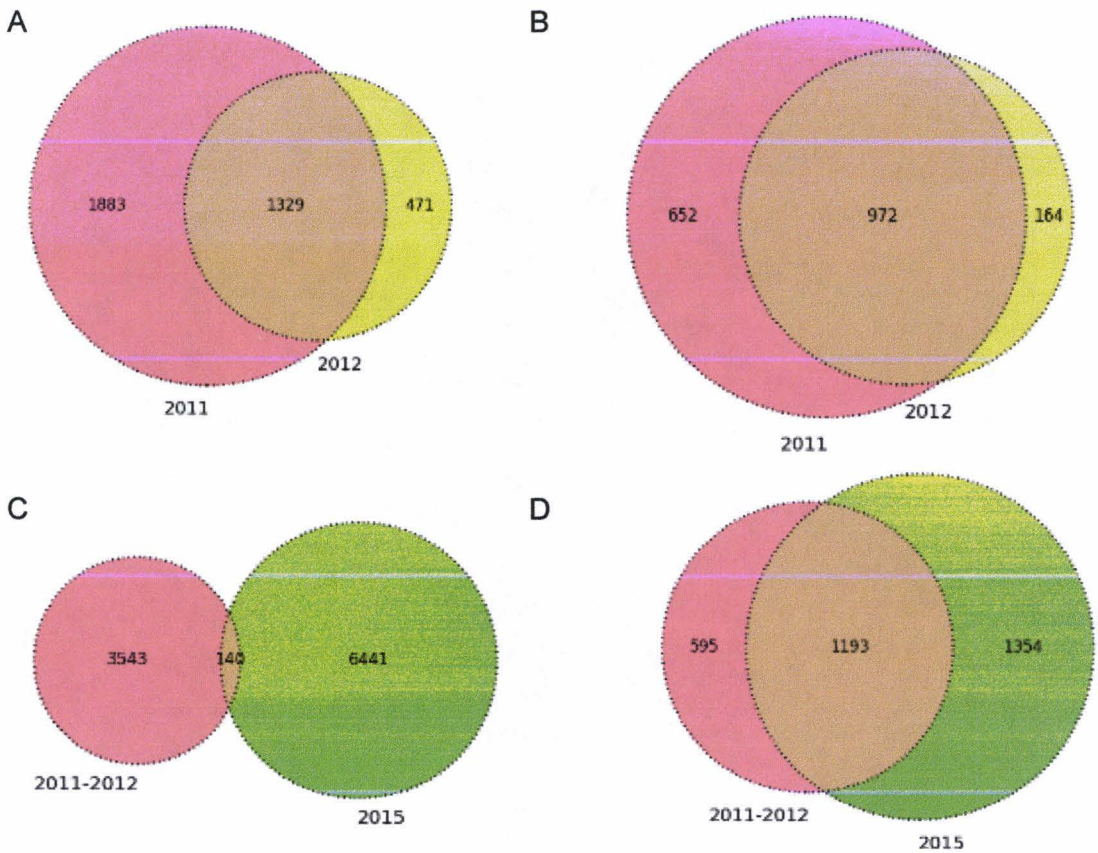
La red de 2008 no fue usada debido a que las redes de 2011 y 2012 la contenían completamente. Cómo se mencionó anteriormente estas dos redes mostraron complementariedad y coherencia entre sus datos exhibiendo índices de Jaccard de 0.36 y 0.54 para interacciones y genes respectivamente. Por este motivo ambas redes fueron fusionadas en una nueva red (83332-2011_2012).

Con respecto a la red obtenida en 2015 y la red 2011_2012 el índice de Jaccard para las interacciones es de 0.01 con sólo 140 interacciones de un total de 10,000. Debido a esta pobre complementariedad entre ambas redes a nivel de interacción, decidimos comprobar qué cantidad de genes eran compartidos entre ambas. Sorprendentemente, 1193 nodos se encontraban presentes en ambas redes de un total de 3142 dado así un índice de Jaccard de 0.38. Al encontrarnos con estos resultados se decidió profundizar más en ambas redes [Figura 25].

En un principio se creyó que esta disparidad de interacciones podría ser debida a las diferentes condiciones en las cuales las redes fueron obtenidas. La red de 2011 fue obtenida como resultado de la expansión mediante curación de la literatura de la red de 2008 por medio de literatura (la red de 2008 fue obtenida

durante la fase estacionaria y en condiciones de hipoxia). La red 2012 fue una expansión de la red de 2008 en condiciones de patogénesis. Y la red de 2015 fue obtenida como resultado de la sobreexpresión durante 18 horas de sus TFs en medio de cultivo Loewenstein-Jensen. Esta diferencia en condiciones podría explicar la baja cantidad de interacciones compartidas. Por este motivo decidimos utilizar 3 redes (2011-2012, 2011-2012-2015 y 2015).

Figura 25 Comparación entre la cantidad de interacciones (A y C) y nodos (B y D) compartidos en las diferentes redes de *Mycobactrium tuberculosis*.



5. Conclusiones

En este estudio demostramos la universalidad y robustez del NDA. Para ello analizamos las predicciones del enfoque de descomposición natural de un conjunto diverso de redes de regulación pertenecientes a diferentes géneros bacterianos. Además, evaluamos los efectos derivados de la cantidad y calidad de información disponible, y los parámetros del NDA, sobre las predicciones obtenidas. La concordancia observada en las proporciones de los componentes predichos brinda confiabilidad en la robustez de las predicciones ante la incompletez de las redes. Mientras que los resultados de la universalidad muestran que la arquitectura funcional y componentes se encuentran conservados a lo largo de diversos géneros bacterianos. Un estudio previo entre *E.coli* y *B. subtilis* mostró que esta conservación es mediada principalmente por convergencia evolutiva [Julio A. Freyre-González, et al. 2012]. Se desconoce cuáles son las fuerzas evolutivas y las restricciones que conducen a converger en esta arquitectura en forma de diamante. Pero el cumulo de resultados presentado en este estudio hace evidente la existencia de principios organizacionales universales en bacterias. Por lo que se hace necesario desarrollar un enfoque de biología de sistemas comparativo para estudiar las adaptaciones y optimizaciones que sistemas análogos han sufrido como consecuencia de la adaptación de los organismos a sus correspondientes nichos.

Podemos concluir que el NDA es capaz de distinguir una organización regulatoria común a lo largo de diversos géneros bacterianos a pesar de la incompletez de los datos. Junto a lo anterior, la obtención de módulos

funcionales en todos los organismos, recalca la importancia biológica, y seguramente dinámica, de una arquitectura funcional jerárquico-modular.

Las predicciones del NDA son importantes en una variedad de vertientes, por ejemplo: 1) La identificación de sistemas con entradas y salidas bien definidas contribuye a la optimización de modelos dinámicos. 2) El estudio comparativo de sistemas análogos permitirá comprender mejor las distintas estrategias regulatorias empleadas en función del modo particular de vida de cada organismo, permitiendo esto optimizar los diseños de la biología sintética. Dada su importancia, decidimos generar una base de datos que almacene las predicciones del NDA. A diferencia de las bases de datos regulatorios actuales, las cuales no contemplan una organización más allá del regulón, nuestra base de datos sienta un precedente al exponer en capas el todo, no sólo como la suma de sus partes sino contemplando también componentes emergentes, como los genes intermodulares. Este estudio ha respondido a su pregunta, pero más importante aún ha abierto las puertas de un sinfín más. Ha clarificado un secreto a voces: Los organismos bacterianos presentan una organización regulatoria similar más allá del regulón. Permitió la generación de una base datos con un enfoque nunca antes realizado, y generó nuevas redes regulatorias más completas y útiles para el estudio de dichos organismos. Si bien la cantidad de datos es aun limitante, las conclusiones y estudios posibles con los datos actuales son abundantes. Mientras más datos poseamos, mejores serán los resultados, por lo tanto, resta un largo camino por recorrer y tanto biólogos teóricos como experimentalistas deberán colaborar para que podamos obtener una vista completa y sencilla de la biología. Actualmente, la biología es una jungla, una jungla de detalles.

6. Materiales y Métodos

6.1. SetAnalizer

Con el objetivo de realizar análisis a gran escala en las diferentes redes obtenidas, era necesario desarrollar una herramienta que nos permitiera hacerlo de la manera más sencilla posible. SetAnalizer es una herramienta desarrollada en Python, con el objetivo de comparar redes (tanto al nivel de interacciones regulatorias como la nivel de nodos) evitando elementos repetidos. Para manejar las redes cómo elementos únicos se hace uso del tipo de dato `set ()`, el cual convierte en conjuntos a los elementos de las redes (interacciones o nodos) evitando así duplicados. Adicionalmente a su manejo de conjuntos se usa un diccionario de sinónimos para homogeneizar los nombres que reciben los nodos en las redes a un mismo tipo (ie. *Locus Tag*, *Gen Name*) y evitar así posibles duplicados debidos a sinónimos, haciéndolo de ésta una herramienta robusta para el análisis de redes.

El método toma cómo entrada dos redes regulatorias y un diccionario (opcional), el análisis resultante arroja la cardinalidad, unión, intersección, diferencia simétrica, ambas diferencias asimétricas, el índice de Jaccard y un diagrama de Venn como representación gráfica (las librerías usadas como dependencias son `matplotlib` y `maplotlib_venn`).

7. Referencias

Sauer, Uwe; Heinemann, Matthias; Zamboni, Nicola (27 April 2007). "Genetics: Getting Closer to the Whole Picture". *Science* **316** (5824): 550–551.

Julio A. Freyre-González. et al (2012). Prokaryotic regulatory systems biology: Common principles governing the functional architectures of Bacillus subtilis and Escherichia coli unveiled by the natural decomposition approach. Journal of Biotechnology, 278-286.

Julio A. Freyre-González., et al (2008). Functional architecture of Escherichia coli: New insights provided by a natural decomposition approach. Genome Biology 9:R154.

Salgado H, et al (2012). RegulonDB (version 8.0): Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Research

Galán-Vásquez et al (2011). The Regulatory Network of Pseudomonas aeruginosa. Microbial Informatics and Experimentation, 1:3.

Alexei E. Kazakov et al (2006). RegTransBase—a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. Nucleic Acids Research, Vol. 35.

Albert-László Barabási et al (2004). Network biology: understanding the cell's functional organization. Nature Reviews, Vol 5, 101-113.

Sierro N. et al (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. Nucleic Acids Res. 2008, 36 (Database issue):D93-D96.

Kyle J. Minch et al (2015). The DNA-binding network of Mycobacterium tuberculosis. Nature Communications 6:5829.

Kyle H. Rohde et al (2012). Linking the Transcriptional Profiles and the Physiological States of Mycobacterium tuberculosis during an Extended Intracellular Infection. PLoS Pathog 8(6): e1002769.

Joaquín Sanz et al (2011). The Transcriptional Regulatory Network of Mycobacterium tuberculosis. PLoS ONE 6(7):e22178.

Gábor Balázsi et al (2008). The temporal response of the Mycobacterium tuberculosis gene regulatory network during growth arrest. Molecular Systems Biology 4:225.

Lauren S. Waters and Gisela Storz (2009). Regulatory RNAs in Bacteria. Cell. 2009 February 20; 136(4): 615–628.

Pauling J et al. (2012). CoryneRegNet 6.0 - Updated database content, new analysis methods and novel features focusing on community demands. Nucleic Acids Res. 2012 Jan;40(1):D610-4.