1

ADOPCIÓN DE UN MÉTODO PARA CONSTRUIR Y VISUALIZAR REDES DE CO-EXPRESION GÉNICA EN ECOLI

1.1 INTRODUCCION

En este capítulo, revisaremos la definición de redes pero esta vez en un contexto biológico, así como las redes de co-expresión génica. Como ya mencionamos en la sección **??** una red o un grafo es un conjunto de puntos llamados *nodos* y enlazados a través de lineas llamadas *aristas*, las cuales representan algún tipo de relación. En el ámbito biológico es posible representar relaciones entre biomoléculas a través de estas redes, de manera que cada nodo represente un gen o una proteína y las aristas indican sus interacciones. Las redes han sido utilizadas en biología para representar rutas metabólicas, co-expresión génica, regulación génica e interacción de proteínas, entre otras aplicaciones, como mencionan Barabasi y Vibal en [**?**, **?**].

Que es una red de co-expresion?

Las redes de co-expresión de Genes (GCN), se definen como un grafo no dirigido en el que cada nodo corresponde a un gen, más exactamente a su perfil de expresión, y las aristas que los conectan representan relaciones de co-expresión, entendiéndose esta relación como la expresión de dos o más genes de manera simultánea. Así, dos genes se conectan si sus perfiles de expresión están asociados entre las perturbaciones estudiadas.

Estas redes proveen información de relaciones de asociación, similitud de expresión y vecindad entre genes que eventualmente permiten inferir interacciones entre las proteínas que codifican. Generalmente estas redes se construyen a partir de datos de expresión provenientes de micro-arreglos de ADN y pueden ser de dos tipos según la conexión entre los nodos: (1) sin pesos (o no-ponderadas), en las que sus aristas denotan si hay una asociación entre un par de nodos, y (2) con pesos (o ponderadas), en las que se cuantifica el grado de asociación entre los nodos por medio de un atributo llamado peso, el cual corresponde comúnmente a un valor en el rango [0, 1].

Una aplicación de las redes de co-expresión génica es la identificación de grupos de nodos latamente co-expresados llamados *modulos*. Estos modulos o *clusters* indican funciones génicas comunes .

Que es WGCNA?

Horvart y Zhang presentan en [**?**], un método para la construcción de redes con pesos denominado Análisis de Redes Ponderadas de Co-expresión de Genes (WGCNA del inglés Weighted Gene Coexpression Network Analysis) en donde se usa el coeficiente de correlación de Pearson como medida de co-expresión entre los genes (nodos).

Como se construyen y visualizan?

Básicamente este método consiste en la creación de una matriz que codifica el grado o fuerza de conexión entre pares de nodos. De manera general el método WGCNA comprende los siguientes pasos (para mayor detalle ver la sección 1.2):

1. Preparación de los datos de expresión (ej: normalización).
2. Definición de una medida de similitud entre nodos (ej: coeficiente de correlación de Pearson).
3. Definición de una función de adyacencia entre nodos (ej: la función potencia).
4. Definición de los parámetros de adyacencia (ej: el exponente de la función potencia).
5. Definición de una medida de disimilitud entre los nodos (ej: sobreposición topológica TOM del inglés Topological Overlap Measure).
6. Identificación de los módulos (ej: el algorítmo de agrupamiento jerarquíco llamado «poda dinámica de árbol», del íngles dynamic tree cut algorithm [**?**]).

Los autores del método han desarrollado WGCNA (R / Bioconductor), un paquete de lenguaje R que cuenta con diferentes funciones para la construcción y análisis de las GCN [**?**].La visualización de estas redes se puede lograr con aplicaciones como Cytoscape (http://www.cytoscape.org/) presentada por Shannon en [**?**], Gephi (http://gephi.github.io/), o gViz (http://urbm-cluster.urbm.fundp.ac.be/webapps/gviz/).

Estado del arte

Xue et al. en [**?**] usan las CGN para identificar y seleccionar genes candidatos claves en el proceso de desarrollo embrionario en humanos y ratones. Por otra parte, Davidson et al. en [**?**] identifican genes candidatos para estudios sobre desarrollo reproductivo y mejoramiento del potencial de producción en maíz. También, Hollender et al. en [**?**] a través del análisis de las CGN, identifican genes y rutas metabólicas involucrados en la floración y desarrollo de frutos en fresa silvestre. Además, Liu et al. en [**?**] empleando la base de datos de [**?**] y WGCNA,

construyen exitosamente una red de co-expresión génica de E. coli e identifican diversos clusters y los comparan con el trabajo previo de Treviño et al. en [**?**], encontrando genes de interes biológico en sus predicciones. Por último, una plataforma de análisis y visualización de redes para analizar diversos datos de las Omicas (ciencias genómicas) que ha surgido en los últimos es presentada por Jang et al. en [**?**].

1.2 MATERIALES Y MÉTODOS

A continuación, se detallan las fuentes de datos empleadas en la experimentación con E.coli y la caracterización del método WGCNA para la construcción y análisis de redes de co-expresión génica.

1.2.1 *Preparación de los datos*

Para construir la red de co-expresión génica de Escherichia Coli (E.coli) fue utilizada la base de datos de expresión de genes denominada M^{3D} (del inglés Many Microbe Microarrays Database) [**?**]. Los conjuntos de datos continenen 524 arreglos medidos bajo 264 condiciones experimentales como confirma Allen en [**?**]. Estos datos fueron medidos usando micro-arreglos de ADN (GeneChip) de la compañía estadounidense Affymetrix diseñados para el organismo E.coli (con clasificación MG1655) midiendo la expresión de 4292 genes, por medio de sondas genéticas (o en inglés gene probes). Por ultimo, se promediados los arreglos medidos bajo las mismas condiciones experimentales, estos experimentos reportan el efecto en la expresión génica de 380 perturbaciones diferentes, de las cuales 152 fueron repetidos al menos 3 veces. Los experimentos inclutenperturbaciones ambientales como niveles de PH, fase decrecimiento o propagación, presencia de antibíoticos, temperatura, media de crecimiento o propagación y concentración de oxigeno, así como tambien, perturbaciones genéticas [**?**]. Otras fuentes publicas de datos de expresión génica obtenidos a partir de micro-arreglos son: GEO [**?**], ArrayExpress [**?**], Colombos [**?**].

1.2.2 *Construcción de las redes de co-expresión génica*

Para efectos de la construcción de la red en meción es posible calcular un coeficiente de correlación (como Perarson, Kendall o Spearman) entre cada pár de vectores de expresión (ver la sección **??**) de genes para obtener una matriz de co-relación y seleccionar un método para determinar un umbral. Algunos métodos disponibles para esto podrían ser:

1. Elegir un umbral de corte (del inglés cutoff threshold) de co-relación en donde si la entrada de la matrix tiene valores que

excedan dicho umbrál, entonces los genes correspondientes a dicha entrada se considerarán co-expresados.

2. La z-transformación de Fisher (del inglés Fisher's Z-transformation, la cual basado en el número de muestras calcula un puntaje llamado *z-score* para cada co-relación) o una variación de este como la presentada por Weirauch en [**?**].

En este trabajo se adoptó el análisis de redes ponderadas de co-expresión génica (WGCNA del inglés Weighted Gene Co-expression Network Analysis) propuesto por Zhang y Horvath en [**?**] para pasar de un enfoque estadístico a uno sistémico quer permite incluir conocimiento externo en contribución al logro de resultados de interes biológico. Para esto se utilizó el paquete de R/Bioconductor llamado WGCNA (para mayor detalle ver la documentación de Langfelder y Horvath en [**?**]) presentado en la sección anterior. En WGCNA de definen los parametros que se detallan a continuación:

- Se define la matriz de similitud entre los perfiles de expresión génica como el valor absoluto de la función de correlación de Pearson: $S_{ij} = [s_{ij}]$
- Se define la matriz de adyacencia como la transformación de la matriz de similitud ele vado a una potencia β , así:$A_{ij} = [a_{ij}]$, donde $a_{ij} = |corr(x_i, x_j)|^\beta = |s_{ij}|^\beta$.
- Cada entrada de la matriz de adyacencia se corresponde al peso o fuerza de conexión entre los genes correspondientes.
- El exponente o potencia β también llamado umbral suave (del inglés soft-threshold) se selecciona siguiendo el criterio de topología libre de escala presentado por Zhang y Horvath en [**?**] y se evalua el ajuste de la red a la propiedad de independencia de escala (para más detalle ver el trabajo de Watts en [**?**]) y los valores de conectividad media para diferentes valores de β.
- Se define la matriz de sobreposición topológica [TOM del inglés Topological Overlap Matrix] como una medida de similitud basada en la inter-conectividad relativa entre dos nodos (del inglés relative inter-connectedness), así: (TOM) $\Omega = [\omega_{ij}]$, donde

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\text{mín}\{k_i, k_j\} + 1 - a_{ij}}$$

  donde $l_{ij} = \sum_u a_{iu} a_{uj}$ y $k_i = \sum_u a_{iu}$. Es decir, que l_{ij} es el número de nodos para los cuales tanto i como j comparten arista. Osea que, si $\omega_{ij} = 1$ significa que todos los vecinos de i son también vecinos de j o de otra manera si $\omega_{ij} = 0$ significa que i y j estan desconectados.
- La medida de no-similaridad se define como: $d^{\omega}_{i,j} = 1 - \omega_{ij}$

Para efectos practicos se utilizó el paquete WGCNA (Versión 2.3.2) para construir la red e identificr los módulos. Los parámetros mas releveantes fueron: TOMType (se parametrizó con valor 'unsigned'), soft-

Power (con valor 6), minModuleSize (con valor 4), powerVector (con valores 1,2,3,4,5,6,7,8,9,10,12,14,16,18,20) en sus funciones *pickSoftThreshold* y *blockwiseModules* (para mayor detalle revisar la documentación en [**?**]).

1.2.3 *Análisis de redes complejas*

Identificación de clusters y anotaciones

Ya que el método WGCNA se basa en el criterio de topología libre de escala de la red, tanto para la construcción de la red como la identificación de clusters, se consideraron solo las redes que cumplen las siguientes condiciones:

- Se seleccionó un valor para $\beta = 6$ en el cual el $R^2 = 0.942$ de la regresión fue lo más cercano a 1 (lo recomiendable es $R^2 >= 0.8$) y la pendiente lo más cercana a -1(-1.55, para el β escogido)
- Además, se encontró equilibrio entre el ajuste al modelo lineal y el número promedio de conexiones en la red (conectividad media) de 19.5, ya que valores de R^2 muy cercanos a 1, generan redes con muy pocas conexiones.

Por otra parte, por medio de AmiGO [**?**], se encontraron los términos de la base de datos biológica de ontologías del gen (o GO del inglés Gene Ontology, para mayor detalle ver la sección **??**) más representativos asociados a los módulos y genes enncontrados y seleccionados según las condiciones establecidas en esta sección.

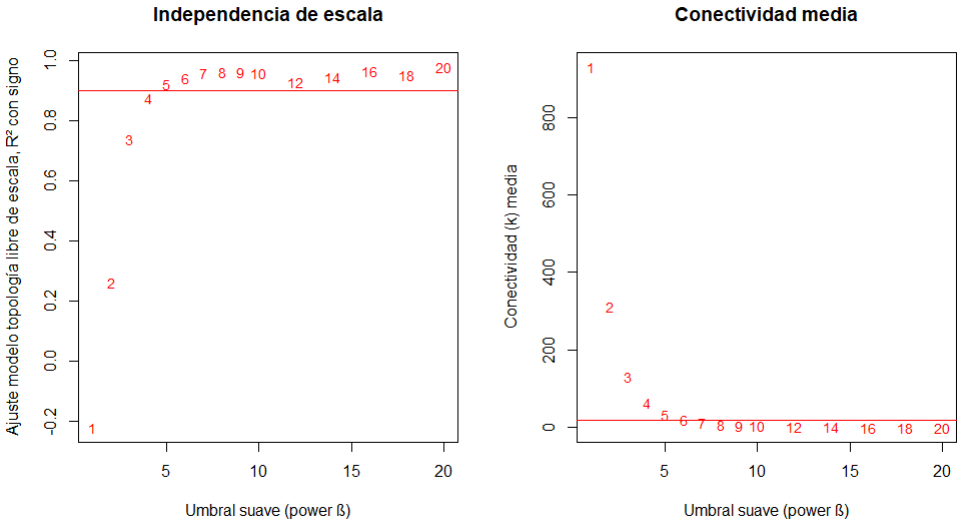
Como validar los resultados obtenidos con WGCNA en E.Coli?

Una vez obtenidos los clusters con WGCNA se hizo una selección aleatoria de 10 clusters de diferentes tamaños entre 4 y 13. Despúes, cada cluster seleccionado fue comparado con su par en el atlas bacteriano ABASY (del inglés Across-bacteria system y con un 88 % de módulos anotados del genoma del organismo Escherichia Coli str. K-12 substr. MG1655 - 2017, RDB16, Strong. Para mayor detalle ver [**?**]). Para la comparación se empleó el *índice de Jaccard* tambien conocido como la intercepción sobre la unión (del inglés Intersection over Union) definido como:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

, donde $0 \leq J(A,B) \leq 1$. Solo los clusters con índice de Jaccard mayor o igual a 0.8 (80 %) fueron considerados.

Figura 1.1: Ajuste al modelo de topologia libre de escala y conectividad media.



1.3 RESULTADOS

1.3.1 *Construcción de la red*

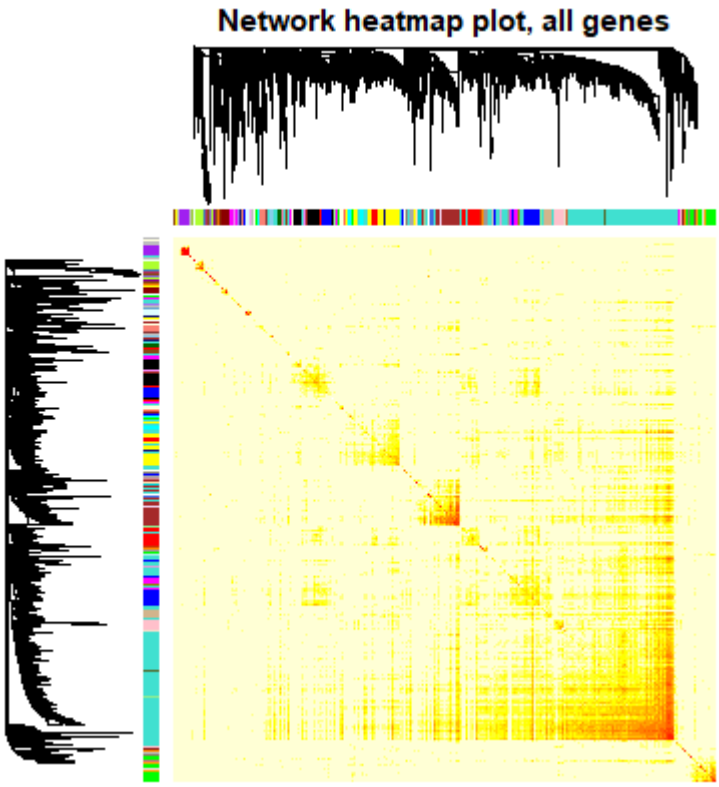
Un paso importante en la construcción de la red con el método WGCNA es la elección de una potencia β ó umbral ligero o liviano. La función *pickSoftThreshold* (mencionada en la sección anterior) devuelve un conjunto de indices de redes que podrían ser seleccionados y de los cuales la potencia $\beta = 6$ fue seleccionada de acuerdo a las condiciones fijadas en la sección anterior. Los valores de ajuste para el parámetro del vector de potencias $1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20$ para la función *pickSoftThreshold* (recomendado por el tutorial de Langfelder y Horvath [**?**]) se muestra en el cuadro 1.1, y como puede observarse, la potencía 6 ajusta al modelo a $R^2 = 0.942$ cumpliendo la condición de cercania a 1. Así mismo, esta otencia ajusta el modelo a una pendiente de -1.55 cumpliendo la condición de cercania a -1. Además, el valor de conectividad media para la selección da 19.5 para la cual la red resultante no es una red con pocas conexiones.

Por otra parte, la figura 1.1 muestra para las diferentes potencias el valor de ajuste R^2. Igualmente, en la parte derecha de la figura en mención podemos ver la conectividad media para cada potencia del vector. En ambos gráficos se trazó una línea verticar para encerrar los valores de ajuste que convienen de acuerdo a las condiciones mencionadas en la sección anterior.

Cuadro 1.1: Valores de ajuste al modelo de topologìa libre de escala y conectividad media de la red construida.

Potenciaβ	R^2	*Pendiente*	*Conectividad (k) media*	*Máx.*(k)
1	0.220	1.150	929.000	1550.0
2	0.262	-0.638	312.000	776.0
3	0.738	-1.190	131.000	447.0
4	0.875	-1.410	63.100	282.0
5	0.921	-1.500	33.800	188.0
6	0.942	-1.550	19.500	133.0
7	0.958	-1.620	12.000	99.2
8	0.961	-1.660	7.770	77.2
9	0.962	-1.690	5.260	61.7
10	0.958	-1.720	3.690	50.5
12	0.929	-1.740	2.010	35.7
14	0.945	-1.680	1.210	26.6
16	0.965	-1.610	0.796	20.6
18	0.954	-1.540	0.559	16.4
20	0.979	-1.450	0.414	13.3

Figura 1.2: Plot mapa de calor de la matriz de solapamiento topológico (TOM del inglés Topological Overlap Matrix)



1.3.2 *Identificación de clusters y anotaciones GO*

Despues de aplicar el marco de trabajo WGCNA para los datos de E. coli se obtuvieron 90 clusters en total para la red obtenida como se muestra en el cuadro 1.2. De estos 8 clusters fueron validados de acuerdo a las condiciones de la sección anterior. El cuadro 1.4 presenta los modulos identificados y validados en la red construida, con un color titulado en inglés, el tamaño expresado en número de genes y la conectividad media para toda la red.

Adicionalmente, por medio de AmiGO [**?**], se encontraron los términos GO más representativos asociados a los 8 modulos y genes validados en la red. Los cuadros 1.4 y 1.5 presentan los resultados con los términos GO más significados según ABASY [**?**].

Finalmente, las gráficas generadas con el paquete WGCNA (R/Bioconductor) incluyen el dendograma de los 90 clusters identificados y etiquetados con colores (ver figura 1.3) así como el mapa de calor (del inglés Heat Map) correspondiente a la función plot de la matriz TOM (ver figura 1.2)).

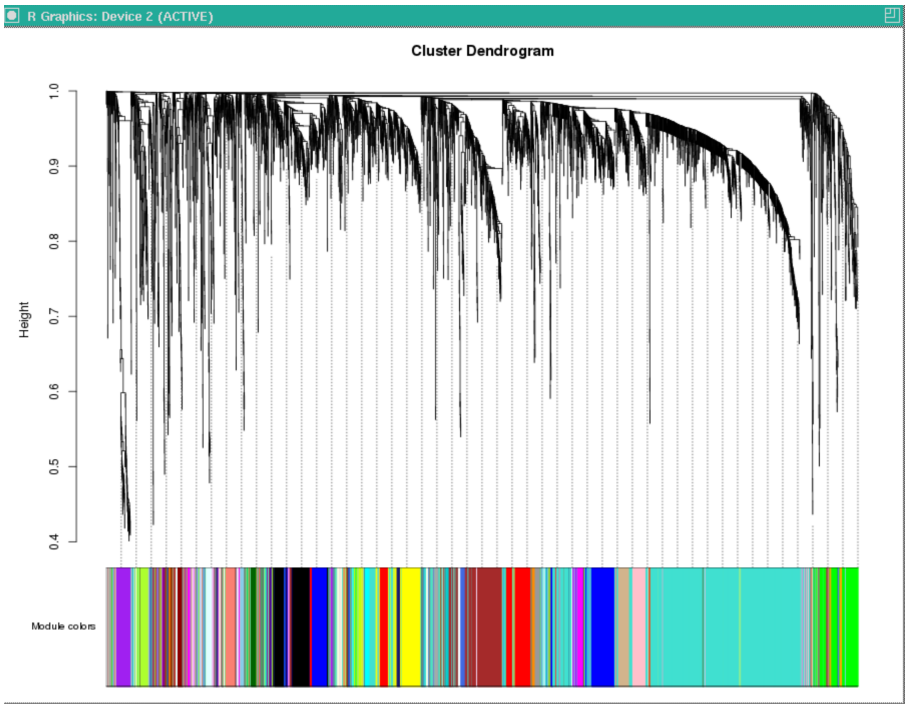
Cuadro 1.2: Resumen de los clusters identificados.

Condición	Dataset	Identificados	Validados
264 condiciones experimentales	M^{3D}	90	8

Cuadro 1.3: Propiedades topológicas de la red y algunos clusters detectados.

No.	*Cluster*	*Tamaño (# Genes)*	*Conectividad (k) media*
	Toda la red	4296	19.5
0	grey	69	0.836
81	salmon2	4	1.37
90	antiquewhite2	4	11.77
73	firebrick4	5	3.67
40	mediumpurple3	12	3.72
34	darkmagenta	13	7.05
35	sienna3	13	18.15
36	yellowgreen	13	8.54

Figura 1.3: Dendograma de los módulos identificados en la red de E.Coli construida con e método WGCNA.



Cuadro 1.4: Clusters identificados en la red de co-expresión génica.

Cluster	*Término GO*	*Genes Anotados*	*Genes Observados*
0 - grey	lactose transport	4	69
81 - salmon2	D-ribose transport D-xylose transport arabinose catabolic process L-arabinose catabolic process to xylulose 5-phosphate D-xylose catabolic process L-arabinose transport carbohydrate phosphorylation	16	4
90 - antiquewhite2	galactose metabolic process	4	4
73 - firebrick4	Protein transport Transporter activity Integral component of membrane ATP binding	5	5
40 - mediumpurple3	cellular protein modification process protein maturation	9	12
34 - darkmagenta	carbohydrate transport maltose transport maltodextrin transport	9	13
35 - sienna3	cysteine metabolic process iron-sulfur cluster assembly	9	13
36 - yellowgreen	Pyrimidine nucleobase catabolic process Uracil catabolic process Nitrogen utilization	223	13

Cuadro 1.5: Genes con función biológica afín dentro de los cluster validados.

#	*Cluster*	*Id. Gen*	*Descripción*	*k*
0	grey	lacA_b0342_15	galactoside O-acetyltransferase monomer	0.647
		lacY_b0343_15	lactose / melibiose:H^+ symporter LacY	0.6
		lacZ_b0344_15	β-galactosidase monomer	0.19
81	salmon2	araC_b0064_15	AraC	0.6
		araF_b1901_15	arabinose ABC transporter - periplasmic binding protein	1.9
		araG_b1900_15	arabinose ABC transporter - ATP binding subunit	1.5
		araH_b4460_28	arabinose ABC transporter - membrane subunit	1.4
90	mediumpurple3	galE_b0759_14	UDP-glucose 4-epimerase monomer Intermodular	14.4
		galK_b0757_14	galactokinase Intermodular	9.26
		galM_b0756_15	galactose-1-epimerase Intermodular	12.18
		galT_b0758_14	galactose-1-phosphate uridylyltransferase Intermodular	11.177
73	firebrick4	oppA_b1243_15	peptide ABC transporter - periplasmic binding protein Basal machinery	3.3
		oppB_b1244_15	murein tripeptide ABC transporter / peptide ABC transporter - putative membrane subunit Basal machinery	3.6
		oppC_b1245_15	murein tripeptide ABC transporter / peptide ABC transporter - putative membrane subunit Basal machinery	3.3
		oppD_b1246_15	murein tripeptide ABC transporter / peptide ABC transporter -	4.2

1.4 DISCUSIÓN

En este trabajo se tiene mucho cuidado con la elección de los parámetros del método WGCNA para la construcción de la red en comparacion con otros trabajos como el de Liu [**?**], donde se seleccionan el tamaño de cluster mínimo muy grade (30), que para el analisis global de la red acertadamente se extraen las cracteristicas y propiedades esenciales, sin embargo, para estudios mas focalizados como las interacciones regulatorios o causales (grafo dirijído) se queda corto, ya que como indica ABASY [**?**] para el organismo E.coli existen módulos como Lactosa (4 genes, LacA, Lacz,LacY) o Maltosa (9 genes), entre otros, que no quedarian filtrados con la parametrizacion de Liu.

Los 8 grupos o clusters de genes que se mostrarón en el cuadro 1.3, fueron seleccionados aleatoriamente entre los 90 clusters identificados por WGCNA para el genoma de E.coli (ver cuadro 1.2 y sección 1.3.1). Adicionalmente, los genes de cada uno de los 8 cluster fueron consultados en GO y filtrados por la afiidad de su función biológica (como se mostró en los cuadros 1.4 y 1.5). A continucación, se detallarán algunos de ellos con una breve reseña biológica:

- Uno de los módulos más estudiados de E.coli es el de la lactosa (del inglés lactose), el cual se entiende y es un punto de referencia para el estudio del sistema regulatorio bacteriano (ver fígura 1.4), el cual es controlado por 4 promotores: P1 (promotor principal) y p2,p3,p4 promotores más debiles. La transcripción de este módulo es regulada positiva (activación) y negativamente (represión). La regulación negativa ocurre cuando el lac represor (lacI) se liga al operador previniendo la transcripción por la polimeraza RNA (encima). Por otra parte, la regulación positiva es mediada por la proteína receptora cAMP, CRP (CAP del inglés catabolite activator protein). En presencia de glucosa la expresión de lactosa no es necesaria y ocurre un mecanismo de control llamado represión (del ingés catabolite repression), es decir, que la glucosa tiene la habilidad de inhibir la expreción de lactosa; como los niveles de cAMP bajan en respuesta a que la glucosa no es tomada, entoces CRP no se activa y la expresión se inhibe. De otra menera, en ausencia de glucosa, los niveles de cAMP se incrementan y estos metabilitos se ligan a CRP, el cual se activa y se liga al DNA cerca de la región promotora facilitando la expresión de lactosa. En resumen, la proteinas de lactosa son inducidas cuando en E.coli hay un incremento de lactosa en ausencia de glucosa y CRP es el principal activador de expresión y estimula la transcripción de P2. De la misma manera CRP solapa completamente P1 y P3 evitando que la polimerasa RNA se lige, como se dijo anterioemente, previniendo la transcripción de estos promoteres.

Figura 1.4: Módulo de transporte de lactosa. Tomado de [?]

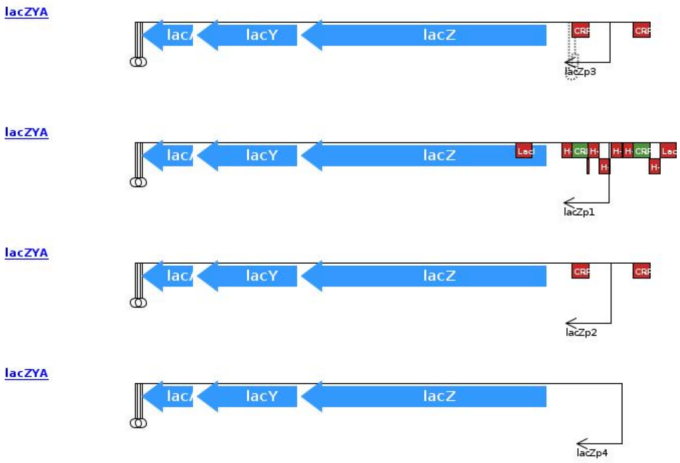
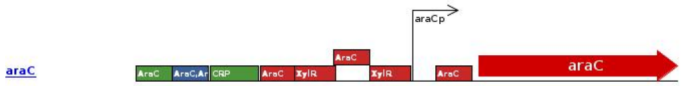


Figura 1.5: Módulo de transporte de arabinose. Tomado de [?]



- De manera similar se validó el funcionamiento de otros clusters identificados como el de Arabinose (ver figura 1.5), Maltose, Hydrogenase, intramodulares y maquinaria basal. La explicación de los diferentes componentes del sistema regulatorio bacteriano del organismo E.coli va más alla del alcance de este trabajo, para mayor detalle se recomienda consultar [?, ?].

1.5 COMENTARIOS CONCLUYENTES

- El objetivo de este capitulo fue revisar y adaptar computacionalmente la construcción y visualización de redes de co-expresión de genes, lo que permitira a otros metodo de predicción de redes de interacción génica, encontrar asociaciones que representen relaciones causales (grafo dirigido) como regulaciones entre los genes estudiados de un cluster particular.
- Además, este estudio permitirá la facil adopción de un método como WGCNA para uso bioinformático sin ser experto en análisis de redes complejas.
- Un indicador de que el análisis del genoma del E.coli con WGCNA fue coherente se observa en los cuadros 1.3 y 1.5, donde los clusters (90-antiquewhite2) correspondientes a genes Intermodulares (galE_b0759_14,galK_b0757_14, galM_b0756_15, galT_b0758_14) presentan un índice de conectividad media más alto (11.7 para el cluster y entre 9.2 y 14.4 para los genes) que el

de otros genes modulares como el de lactose (0.8) o arabinose (1.3).

- Fue posible corroborar la coherencia desde la función biológica para los cluster seleccionados con respecto a la literatura y bases de datos de anotaciones en [**?**, **?**].