



UNIVERSIDAD DEL PAPALOAPAN.

Campus Loma Bonita

INGENIERÍA EN COMPUTACIÓN

MODULARIDAD DE REDES REGULATORIAS GENÉTICAS
EN BACTERIAS

TESIS

QUE PARA OBTENER EL TÍTULO DE:
INGENIERO EN COMPUTACIÓN

PRESENTA:
VIRGINIA ORTIZ GÓMEZ

DIRECTOR DE TESIS:
DRA. BEATRIZ CARELY LUNA OLIVERA

LOMA BONITA, OAXACA,

SEPTIEMBRE 2012

Agradecimientos

A mis padres Eduardo y Domitila por su apoyo incondicional, por su esfuerzo constante porque son mi motor e inspiración para seguir luchando día a día. A ellos que han puesto sus esperanzas en mí. Gracias por motivarme siempre a ser una mejor persona, una mejor profesionalista.

A mis 4 más grandes amores: Ana Mariví, Luis Manuel, Ángel Némesis y José Alberto, que junto que con mis innumerables amigos me han enseñado día a día que las familias no sólo se constituyen de sangre, sino de amor, soy afortunada por haberlos conocido y por saber que están ahí a cada paso que doy, aunque no estén a mi lado.

Al Dr. Eduardo Sánchez que fue quien me planteo la idea de realizar este proyecto, para él mi respeto, cariño y admiración. A la Dra. Beatriz Carely, mi directora de tesis, gracias por el apoyo, la paciencia y por compartir sus conocimientos conmigo.

Agradezco a PROMEP por el apoyo otorgado para la realización del presente proyecto, a los sinodales de este trabajo, el maestro Eduardo Ortiz Hernández y la maestra Carolina Antonio Estrada, por sus valiosos comentarios, y a los revisores de última hora.

Resumen

Los avances de la biología nos permiten conocer los genomas completos de diversas especies, sin embargo, las interacciones entre los genes que conforman dichos genomas permanecen sin esclarecerse, estas interacciones suelen ser representadas mediante un grafo, usando vértices como representación de los genes, y aristas como representación de las interacciones. Además de contar con la información de la forma de la red (su topología), se cuenta también con información dinámica, donde cada gen que interacciona con otro, provoca la producción o inhibición de una cierta proteína. Esta parte de las redes regulatorias genéticas se estudia mediante modelos matemáticos que permiten la comprensión, reproducción y predicción de diversos fenómenos dinámicos. Hasta ahora sólo han sido modeladas redes regulatorias de pocos componentes, ya que el análisis de grandes redes requiere de una estrategia diferente. Como un primer paso hacia la solución de este problema, este trabajo explora metodologías formales para descomponer la red en módulos funcionales, es decir, subgrafos con una función bien determinada en la red regulatoria, que puedan ser estudiados individualmente para después ser incluidos en la red completa. Seleccionamos una de las metodologías, la analizamos y aplicamos a las tres redes más conocidas en bacteria: *E. Coli*, *B. Subtilis*, y *P. Aeruginosa*. Los módulos obtenidos se contrastan con la información conocida.

Palabras clave: *Modularidad, grafos, redes regulatorias*

Abstract

Advances in biology allow us to know the complete genome of species, however, it is important to clarify the role of every gene and its interaction with other genes. These interactions are usually represented by a graph, where vertices represent genes, and arrows represent their interactions. This representation allows us to analyze the static network (the topology of network). However, we have also a dynamical representation, where every gene is associated with a variable that represents the quantity of protein expressed by the gene. Then, we can study regulatory networks using mathematical models, which allow us the comprehension, reproduction and prediction of dynamical phenomena. Until now only networks with a few components have been modeled, because the analysis of big networks requires other tools. As a first step to the modelling of big networks we explore formal methodologies to decompose networks in functional modules, that is to say. Subgraphs with a well characterized function, which can be analyzed separately and then included in the whole network. We select a methodology and we apply this methodology to the three better known bacteria networks: *E. Coli*, *B. Subtilis*, and *P. Aeruginosa*. Results show the contrast between modules obtained by a theoretical methodology and real functional modules.

Keywords: *Modularity, graphs, regulatory networks*

Índice general

| | |
|--|------------|
| Agradecimientos | II |
| Resumen | III |
| Abstract | IV |
| 1. Introducción | 1 |
| 1.1. Antecedentes | 1 |
| 1.2. Objetivos | 3 |
| 1.2.1. Objetivo General. | 3 |
| 1.2.2. Objetivos Particulares. | 3 |
| 1.3. Hipótesis | 4 |
| 1.4. Justificación | 4 |
| 2. Preliminares | 5 |
| 2.1. Conceptos de teoría de grafos | 5 |
| 2.2. Modularidad | 7 |
| 2.3. Bacterias | 10 |
| 2.3.1. <i>Escherichia Coli</i> | 11 |
| 2.3.2. <i>Bacillus Subtilis</i> | 12 |
| 2.3.3. <i>Pseudomonas Aeruginosa</i> | 12 |

| | |
|--|-----------|
| <i>ÍNDICE GENERAL</i> | VI |
| 3. Métodos y materiales | 15 |
| 3.1. El algoritmo seleccionado | 15 |
| 3.2. El algoritmo implementado | 17 |
| 3.3. Las redes regulatorias | 20 |
| 3.4. Metodología | 21 |
| 4. Resultados | 23 |
| 5. Conclusiones | 38 |
| Bibliografía | 39 |

Capítulo 1

Introducción

1.1. Antecedentes

El tema de este trabajo se encuentra ubicado en un área emergente de la biología conocida como biología de sistemas, esta área tiene aplicaciones en biomedicina y biología sintética. La biología de sistemas es un campo de estudios interdisciplinario basado en la biología, que estudia las interacciones complejas entre sistemas biológicos [11]. Considera la integración de los conocimientos de diversas áreas como computación, matemáticas, física, biología, química, etc., al contrario del paradigma reduccionista. Los conjuntos de datos complejos que se analizan e integran provienen de la genómica, epigenómica, transcriptómica, interferómica, proteómica, metabolómica, glicómica, lipidómica, interactómica, biómica, flujómica, etc. Por su parte, el cómputo científico y la informática son usados en biología de sistemas para: desarrollo de técnicas de extracción de información, desarrollo de bases de datos en línea, y construcción de herramientas de software [3].

La biología de sistemas tiene sus orígenes en: el modelado cuantitativo de cinética enzimática, el modelado matemático de crecimiento de población, las simulaciones desarrolladas para estudiar neurofisiología, la teoría de control y la cibernética. Su prin-

La principal diferencia con la bioinformática, es que se considera el uso de la teoría de sistemas dinámicos aplicado a la biología molecular, el modelado permite la integración del conocimiento biológico y la comprensión del funcionamiento del sistema, así como la predicción de ciertos fenómenos. La biología de sistemas se estableció como disciplina alrededor del año 2000, cuando la biología molecular permitió la secuenciación de genomas completos, los cuales contienen la información que determina las características físicas y en parte psicológicas e intelectuales de un ser vivo. A partir de lo anterior se ha vuelto relevante entender el papel de cada uno de los genes que conforman el genoma, así como sus interacciones, ya que esto puede ayudar a la comprensión de diversas enfermedades y procesos biotecnológicos.

Las relaciones entre los genes son estudiadas en redes regulatorias, cada gen que se expresa da lugar a una proteína, que a su vez promueve o inhibe la expresión de otro gen. Estas complejas relaciones pueden representarse gráficamente como una red de interacciones, y modelarse como un sistema dinámico de varios subsistemas acoplados.

Diversas estrategias de modelado han sido utilizadas para comprender el comportamiento dinámico de las redes regulatorias [4]. Los modelos describen el nivel de expresión de cada gen a lo largo del tiempo, es decir, la cantidad de proteína que producen. Los modelos más usuales son: ecuaciones diferenciales acopladas, que consideran tiempo continuo y niveles de expresión continuos y redes lógicas o booleanas [23], que consideran tiempo discreto y niveles de expresión discretos. Ejemplos de interacciones modeladas son: redes regulatorias responsables de patrones de segmentación, ritmos circadianos, diferenciación muscular, desarrollo embrionario, especiación, supresión de tumores, etc.

En este trabajo exploramos el concepto de modularidad biológica, que implica que la red regulatoria puede ser dividida en partes biológicamente funcionales. Estudiar este concepto nos permitirá el estudio de redes de gran tamaño, ya que hasta ahora sólo se realizan modelos de redes pequeñas, debido a que el modelado de redes con un gran

número de genes requiere de herramientas computacionales de las que no se dispone. Si pensamos en el modelo más sencillo, el modelo booleano, en el que solo se consideran dos estados del gen, 0 y 1, podemos observar que la información crece exponencialmente, así pues, tratar con un gen implica el uso de 2 estados, tratar con dos genes el uso de 4 estados, y tratar con 20 genes implica el uso de 1048576 estados. El análisis de la dinámica no solo consiste en conocer el número de estados sino las transiciones entre estos, por lo que tal cantidad de información es de difícil manejo.

En este trabajo analizamos diversos métodos para separar la red en subredes, algunos métodos plantean el uso de la información topológica de la red, es decir las conexiones de la red, ya sea buscando subgrupos que cumplen ciertas características o cascadas regulatorias, otros métodos plantean el uso de información dinámica, es decir los valores de expresión de los genes. En este trabajo seleccionamos un método topológico, el método fué programado y probado en tres redes regulatorias de bacterias: *E. Coli*, *B. Subtilis* y *P. Aeruginosa*.

1.2. Objetivos

1.2.1. Objetivo General.

Identificar módulos funcionales en redes regulatorias de bacteria.

1.2.2. Objetivos Particulares.

- Analizar algunos métodos algorítmicos para encontrar módulos funcionales en redes regulatorias.
- Aplicar un método basado en la topología a tres redes regulatorias de bacterias: *E. Coli*, *B. Subtilis*, *P. Aeruginosa*.

- Comparar los módulos obtenidos algorítmicamente con los resultados biológicos conocidos.

1.3. Hipótesis

Las características topológicas de las redes regulatorias genéticas pueden ser de importancia para el desarrollo de una metodología viable para obtener módulos funcionales en redes regulatorias de bacterias.

1.4. Justificación

Con los avances en varias áreas de la biología la información disponible ha crecido considerablemente, redes regulatorias con un gran número de genes son ahora conocidas, sin embargo, solo partes de estas redes, o redes de menor tamaño han sido consideradas para su estudio. Esto debido a la falta de herramientas computacionales y matemáticas que permitan tratar este problema de una manera sencilla y sin requerir de grandes recursos. Debido a esta situación se ha considerado que una de las maneras adecuadas para estudiar redes regulatorias de tamaño considerable es la modularidad.

Capítulo 2

Preliminares

2.1. Conceptos de teoría de grafos

La teoría de grafos tuvo sus orígenes en el problema de los Puentes de Königsberg planteado en 1736 por el matemático Euler y enunciado de la siguiente forma: *Dado el mapa de Königsberg, con el río Pregolya dividiendo el plano en cuatro regiones distintas, que están unidas a través de los siete puentes, ¿es posible dar un paseo comenzando desde cualquiera de estas regiones, pasando por todos los puentes, recorriendo sólo una vez cada uno y regresando al mismo punto de partida?* [29].

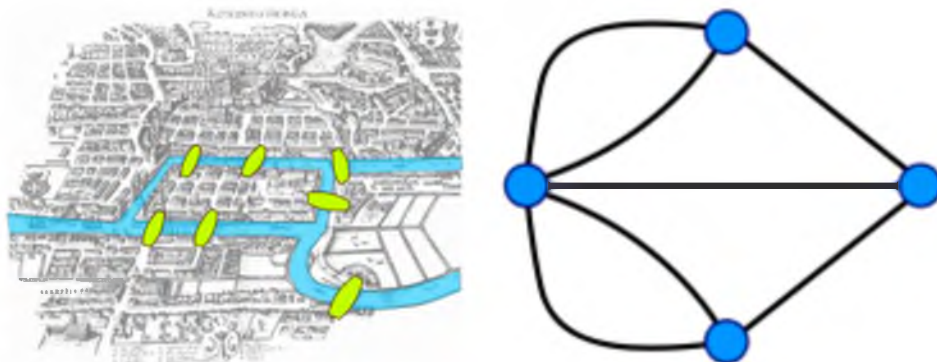


Figura 2.1: Problema de los puentes de Königsberg.

Euler recurrió a una abstracción del mapa que indicaba el problema, considerando las islas como vértices, también llamados nodos (puntos o círculos en el plano), y los puentes como aristas (líneas entre los vértices). Usando esta abstracción Euler resolvió el problema y a partir de esta representación se desarrolló toda una teoría que sigue vigente y que permite el análisis de los más diversos problemas.

De esta manera, un grafo queda definido por sus vértices, sus aristas y las interacciones entre estos objetos. Se define un grafo como la tripleta $G = \{V, A, \psi_G\}$ donde V es un conjunto de vértices no vacío, $A \subset V \times V$ es el conjunto de aristas, y ψ_G es la función de incidencia que relaciona una arista con un sólo par de vértices $(u, v) = (v, u)$. Cuando el grafo es dirigido se considera $D = \{V, A, \psi_D\}$ donde V es el conjunto de vértices, $A \subset V \times V$ es el conjunto de aristas dirigidas, y ψ_D es la función de incidencia que relaciona una arista con un sólo par de vértices ordenados (u, v) , donde la arista sale de u y llega a v .

Matriz de adyacencia: Se considera como matriz de adyacencia a la matriz $A = [A_{ij}]$ que indica que hay una interacción del vértice i al vértice j .

Grado de entrada (k_u^{in}): Es el número de aristas que llegan a un vértice u , es decir, la cardinalidad del conjunto $\{v | (v, u) \in A\}$.

Grado de salida (k_u^{out}): Es el número de aristas que salen de un vértice u , es decir, la cardinalidad del conjunto $\{v | (u, v) \in A\}$.

Grado de un vértice (k_u): Es el número de aristas que se conectan a un vértice considerando que las aristas no tienen dirección, es decir, no importa si salen del vértice o llegan a él.

Trayectoria dirigida: Una trayectoria dirigida en D es una secuencia finita no nula de vértices y aristas $v_0, a_1, v_1, \dots, a_k, v_k$ tal que la arista a_i sale de v_{i-1} y llega a v_i .

Donde además todos los vértices y aristas son distintos.

Subgrafos: Decimos que $S = \{V', A', \psi_S\}$ es un subgrafo de $D = \{V, A, \psi_D\}$ si $V' \subset V$, $A' \subset A$, y $\psi_S \subset \psi_D$.

Coeficiente de clusterización: El coeficiente de clusterización C_i para un vértice v_i está dado por la razón entre el número de aristas que existen en los nodos adyacentes, y el número de aristas que habría si todos sus vecinos se conectaran entre si.

$$C_i = \frac{e}{k_i(k_i - 1)},$$

e indica el número de aristas que hay entre los vecinos del vértice i .

Distancia entre dos vértices: Es la longitud mínima de todos los caminos posibles entre dos vértices.

2.2. Modularidad

A continuación mencionaremos cómo se clasifican los diversos métodos que permiten subdividir una red:

- **Métodos basados en redes:** Son aquellos que identifican subgrafos conectados a las redes biológicas en forma de módulos. Este enfoque identifica módulos a partir de la estructura topológica de una red biológica. Algunas veces se adiciona información como: expresión, interacciones genéticas o fenotipos. Se identifican submódulos o grafos altamente conectados [14].
- **Métodos basados en expresión:** Estos métodos usan datos de expresión génica para inferir módulos de genes que muestran patrones similares de expresión. Se

basa en la suposición de que los genes coexpresados son coordinados regularmente y muestran módulos de genes con funciones similares. Tradicionalmente se incluyen técnicas de aglomeración y partición. Estos métodos son extensamente usados para identificar módulos de genes coexpresados a través de condiciones experimentales [16].

- **Métodos basados en trayectorias:** El objetivo de este enfoque es identificar trayectorias, es decir, grupos de genes, en lugar de genes individuales involucrados en procesos biológicos. El análisis a nivel de genes individuales generalmente utiliza pruebas estadísticas para detectar genes con cambios de expresión significativos, en contraste, el enfoque basado en trayectorias detecta cambios sutiles y coordinados de un grupo de genes en una trayectoria, o relacionados con una función. Los métodos para identificar los genes en una trayectoria incluyen un análisis de enriquecimiento[22] y autómatas [10].

Entre los distintos tipos de métodos elegimos uno que pertenece a los métodos basados en redes, por lo que profundizaremos en la descripción de este tipo de algoritmos. Dentro de los métodos basados en redes encontramos: agrupación jerárquica y gráfica.

- **Métodos de agrupación jerárquica.**

El enfoque de agrupación jerárquica clasifica primero pares de vértices con un mismo peso, después agrupa los vértices basado en los pesos. Los pesos pueden ser definidos de distintas maneras, por ejemplo, dados dos vértices u y v , si hay muchos caminos posibles para ir de u a v , se considera que u y v tienen un alto peso y tienen más probabilidad de ser agrupados juntos. Alternativamente, algunos autores usan medidas invariantes de la red como el coeficiente de clusterización, la distancia, el número de interacciones comunes, etc. para pesar los vértices y usan estos pesos para encontrar las regiones más densas en grandes redes de proteínas.

Aunque este enfoque es poco práctico para redes a escala de genoma, se ha aplicado con éxito a subredes como la red de interacciones entre proteínas y las redes regulatorias.

Hay dos enfoques para la agrupación jerárquica: podemos ir *de abajo hacia arriba*, es decir, agrupando a pequeños grupos con pesos similares en grupos más grandes, o *de arriba hacia abajo*, dividiendo grupos grandes en pequeños. Estos métodos son llamados aglomerativos y divisivos respectivamente [20].

Después de la transformación jerárquica de la topología de la red se usan definiciones cualitativas o cuantitativas de comunidades para obtener finalmente los módulos. Por ejemplo una típica definición cuantitativa requiere que cada vértice en una comunidad esté más conectado a vértices dentro de la comunidad y menos con los de afuera [25].

■ Método de agrupación gráfica

En el ordenamiento gráfico podemos encontrar los siguientes algoritmos:

Spirin y Mirny [26] proponen dos algoritmos de este tipo. El primero de ellos se basa en ordenamiento supramagnético y el segundo es el algoritmo Monte Carlo que maximiza la densidad del agrupamiento obtenido. Ambos algoritmos requieren el tamaño del conjunto buscado como entrada. El algoritmo supramagnético muestra específicamente como realizar de manera correcta la detección de estructuras densas, débilmente conectadas a otras zonas de la red. Además de detectar los complejos de proteínas, Spirin y Mirny también muestran que su método es capaz de detectar módulos funcionales conectados débilmente.

Przhijl [17] en 2004 utilizó el algoritmo de subgrafos altamente conectados para determinar módulos en los datos de interacción proteína-proteína. Un subgrafo altamente conectado se define como un subgrafo con n nodos de tal manera que más de $n/2$ vértices deben ser removidos para desconectarlo.

El algoritmo de búsqueda restringida por zona de agrupamiento (RNSC) propuesto por King [17] divide el conjunto de nodos de la red en grupos basados en una función de costos que se utiliza para evaluar la partición. El algoritmo comienza con una asignación de grupos aleatoriamente y procede por reasignación de nodos, a fin de maximizar la puntuación de la partición. Además, una lista tabú se mantiene para evitar los ciclos de regreso a particiones analizadas con anterioridad. Por último, los grupos se filtran en función de su tamaño, densidad y homogeneidad funcional.

El algoritmo de agrupación de Markov [5] ha sido recientemente propuesto para la detección de módulos en redes de interacción proteína-proteína. El algoritmo simula el flujo en el grafo mediante la construcción de su matriz de adyacencia y el cálculo de sus sucesivas potencias para aumentar el contraste entre las regiones con alto flujo y las regiones con un bajo flujo. Este proceso converge hacia una partición del grafo en las regiones de alto flujo correspondientes a complejos de proteínas, separadas de las regiones de ausencia de flujo.

El algoritmo implementado en esta tesis se encuentra clasificado dentro de los algoritmos de agrupación gráfica, utiliza una función cuantitativa para evaluar que tan bueno es el agrupamiento, parte de un solo conjunto de nodos, y lo va subdividiendo de acuerdo al mejor valor de la función.

2.3. Bacterias

Las bacterias incluidas en este trabajo son las que poseen las tres redes regulatorias mejor caracterizadas. Distinguimos en la red regulatoria principalmente tres tipos de componentes: los factores σ , que son proteínas que permiten iniciar la transcripción de un determinado gen, y que responden a condiciones ambientales; los reguladores, que actúan sobre los genes blanco u otros reguladores inhibiéndolos o activándolos; y los

genes blanco que no actúan sobre ningún otro gen.

A continuación incluimos una tabla con la información taxonómica de las bacterias [30][1][31], así como una descripción que nos permite conocer su importancia.

| | <i>E.coli</i> | <i>B. subtilis</i> | <i>P aeruginosa</i> |
|---------|---------------------|--------------------|---------------------|
| Reino | Bacteria | Bacteria | Bacteria |
| Filo | Proteo bacteria | Firmicutaee | Proteobacteria |
| Clase | Gammaproteobacteria | Bacilli | Gammaproteobacteria |
| Orden | Enterobacteriales | Bacilliales | Pseudomonadales |
| Familia | Enterobacteriaceae | Bacillaceae | Pseudomonadaceae |
| Género | Escherichia | Bacillus | Pseudomonas |
| Especie | E. coli | Subtilis | P. aeruginosa |

2.3.1. *Escherichia Coli*

E. coli es la bacteria más estudiada en biotecnología. Normalmente, esta bacteria vive en el intestino del hombre y de los animales sin causar ningún tipo de problema, sin embargo, debido al intercambio genético, algunas cepas son capaces de causar infecciones y provocar diarrea. Las cepas más peligrosas de esta bacteria, por ejemplo la denominada O157:H7, provocan, además de diarreas graves, insuficiencia renal y, en el peor de los casos, la muerte, especialmente en niños y en adultos con sistemas inmunológicos debilitados [15].

La primera secuencia completa del genoma de *E. coli* fue publicada en 1997 (CEPA K-12 MG1655). Ésta consiste de 4639221 pares de bases, que contienen 4288 genes que codifican proteínas. El número total de factores de transcripción predichos es alrededor de 300 y 7 factores sigma. Algunas bases de datos que contienen información molecular acerca de esta bacteria son las siguientes: RegulonDB, y Ecocyc.



Figura 2.2: Micrografía electrónica (coloreada) de bacterias *Escherichia coli*, que forman parte habitual de la flora intestinal. STEVE GSCHMEISSNER (CORDON PRESS) [27].

2.3.2. *Bacillus Subtilis*

B. subtilis es una bacteria Gram-positiva, aerobio facultativo comúnmente encontrada en el suelo. Esta bacteria es bien conocida por su habilidad para diferenciarse de células en crecimiento en esporas inactivas que le permiten sobrevivir por largos periodos de tiempo bajo condiciones ambientales extremas. Su genoma está constituido por 4214810 pares de bases dentro de los cuales encontramos 4100 genes que codifican proteínas. La cepa distintiva de la especie es la 168. Su número estimado de factores de transcripción es alrededor de 275, y el número de factores sigma es 14.

2.3.3. *Pseudomonas Aeruginosa*

P. aeruginosa es una bacteria Gram-negativa metabólicamente versátil, que expresa un alto rango de factores de virulencia. Esto también permite a *P. aeruginosa* crecer en el suelo y en hábitats marinos, al igual que en tejidos de plantas y animales, lo que convierte a esta bacteria en un patógeno oportunista.

P. aeruginosa es capaz de permanecer por tiempos prolongados en líquidos y su-

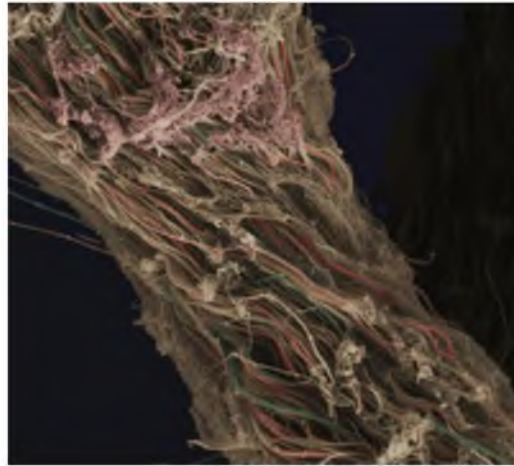


Figura 2.3: Biopelícula de bacterias: esta micrografía confocal muestra a *B. subtilis*, una bacteria con forma de bastón que se encuentra comúnmente en el suelo. Distintos linajes de bacterias que expresan diferentes proteínas fluorescentes fueron inicialmente mezclados aleatoriamente sobre una placa de Petri [28].

perfiles como antisépticos, alimentación parenteral, equipo de inhaloterapia, fluidos de diálisis y grifos de agua, entre otros. En contraste, es excepcional encontrarla como parte de la microflora normal de los individuos sanos, por lo que ataca principalmente a individuos enfermos provocando neumonía en hospitales y otros sitios donde se presente debilidad inmunológica [7]. Es también una fuente de bacterimia en víctimas quemadas, infecciones del tracto urinario y neumonía adquirida en hospitales; y causa de mortalidad en pacientes con fibrosis quística. La secuencia del genoma de *P. aeruginosa* cepa PAO1 fue reportado en el año 2000 con 5570 genes predichos, entre los que hay al menos 14 factores sigma. La base de datos PseudoCAP está dedicada a la información genética de esta bacteria [19].



Figura 2.4: Micrografía electrónica de barrido de *P. aeruginosa* [24].

Las bases de datos más importantes que contienen información acerca de cada uno de los genes y sus interacciones son NCBI, DDBJ y EBI. Cabe mencionar que los datos usados en este trabajo no fueron adquiridos gen por gen, sino a partir del trabajo de muchos otros investigadores, que al realizar experimentos propusieron y confirmaron cada interacción posible de las que ahora conocemos, o que se ocuparon de la recolección de los datos reportados y construyeron las redes regulatorias más actualizadas de las tres bacterias que consideramos: *E.coli* [9], *B. subtilis* [21], *P. aeruginosa* [7].

Capítulo 3

Métodos y materiales

A continuación mostraremos los fundamentos del algoritmo seleccionado, el programa desarrollado en Matlab para su implementación y la metodología. Añadimos 3 Anexos donde se muestran los datos ocupados en el programa. En el Anexo I se muestran las interacciones de las tres redes ocupadas en este trabajo, en la columna 1 se muestra el regulador, en la columna 2 se muestra el gen regulado y en la columna 3 su tipo de interacción. En el Anexo II se muestran los módulos a los que cada gen pertenece según la información biológica conocida [8], en la columna 1 se muestra el nombre del gen, en la columna 2 el módulo funcional al que pertenece y en la columna 3 el módulo asignado por el algoritmo. En los Anexos III, IV y V se muestran las matrices de adyacencia para cada red regulatoria.

3.1. El algoritmo seleccionado

Una vez analizados los diversos métodos para obtener sub-estructuras topológicas, consideramos el algoritmo de detección de comunidades propuesto por E. A. Leicht y M. E. J. Newman [12]. El método fue seleccionado debido a que se encontraron las siguientes ventajas:

- Es un método diseñado para obtener comunidades en redes dirigidas, mientras que muchos métodos trabajan en redes en las que no se considera la dirección de las interacciones y esto causa una pérdida de información.
- No requiere información a priori de cuántos módulos se buscan o de cuántos elementos pertenecen a cada módulo, por lo que puede aplicarse a cualquier red.
- Usa optimización exacta en lugar de optimización aproximada, que utiliza por ejemplo algoritmos greedy, simulated annealing u otros similares, que pueden estacionarse en un máximo local.
- Es computacionalmente eficiente.

El método consiste en maximizar una función de beneficio Q llamada la modularidad. Q =fracción de aristas dentro de la comunidad - fracción esperada de aristas.

Se espera que dentro de la comunidad haya un gran número de conexiones, o al menos un número mayor a las conexiones esperadas, en base al número total de aristas. En este caso consideramos $A = [A_{ij}]$, donde $A_{ij} = 1$ si existe la interacción de j a i , y Q está dada por

$$Q = \frac{1}{m} \sum_{ij} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right] \delta_{c_i, c_j},$$

donde k_i^{in} representa el grado de entrada del vértice i , k_j^{out} representa el grado de salida del vértice j , m representa el número total de aristas pertenecientes a la red y $\delta_{c_i, c_j} = 1$ si los vértices i y j pertenecen a la misma comunidad, en caso contrario $\delta_{c_i, c_j} = 0$.

Como podemos observar A_{ij} está indicando las conexiones reales en la red, y la cantidad $\frac{k_i^{in} k_j^{out}}{m}$ se refiere a las conexiones esperadas.

Dada la función a maximizar podríamos buscar soluciones aleatorias, sin embargo, se puede considerar también un algoritmo determinista, que puede ser computacionalmente

eficiente.

Si consideramos sólo la división iterativa en dos comunidades hay un algoritmo que maximiza Q . Éste consiste en calcular el vector propio que corresponde al valor propio positivo más grande de la matriz simétrica $\mathbf{B} + \mathbf{B}^T$, y asignar comunidades basado en los signos de los elementos de este vector propio. $\mathbf{B} = [B_{ij}]$, donde

$$B_{ij} = A_{ij} - \frac{k_i^{in} k_j^{out}}{m}.$$

3.2. El algoritmo implementado

El algoritmo descrito anteriormente fue implementado en Matlab, este algoritmo pide como entrada las matrices de adyacencia de cada bacteria, realiza una primera división y calcula el valor inicial de Q , a partir de estas subredes verifica para cada una si haciendo una subdivisión el valor de Q aumenta o disminuye. Si el valor de Q aumenta considera la subdivisión, y así sucesivamente, hasta que más divisiones de las subredes ya no incrementan el valor de Q .

El código que implementa el programa se muestra a continuación:

```
--Algoritmo de deteccion de comunidades
function S = comunidad(A)

tiempo_inicial= clock;
k_in = sum(A,2);
k_out = sum(A,1);
m = sum(sum(A));

-- Calculando valores y vectores propios
[V, LAMBDA] = eig (A);
```

```

[max_Eig_Val, max_Eig_Val_pos] = max(diag(LAMBDA));
Eig_Vec_max=V(:,max_Eig_Val_pos);
S=ones(size(Eig_Vec_max));
S(Eig_Vec_max >= 0)=2;

Q=calidad(A, S, k_in, k_out, m)

Delta_Q = 1
Q_total=[]

-- Condicionante para evaluar el incremento de la calidad

while (Delta_Q > 0)
    minimo = min(S);
    maximo = max(S)
    [cuantos, donde]=hist(S, [1:maximo]);
    for i=1:max(S)
        A_2 = A(S==donde(i), S==donde(i));
        [lin_A2, col_A2] = size(A_2);
        if lin_A2 >= 1 % >1
            i
            S_2 = separacion(A_2, min(S(S==donde(i))), maximo+i)
            if (min(S_2) ~= max(S_2)) % entonces hubo separacion
                S(find(S==donde(i))) = S_2 % crear nueva S
                Q_nueva=calidad(A, S, k_in, k_out, m);
            end
        end
    end
end

```

```

        end
    end

    Delta_Q = Q_nueva - Q;
    Q = Q_nueva;
    Q_total=[Q_total Q]
end

tiempo_final=clock;
tiempo_total=tiempo_final-tiempo_inicial

-- Se divide la matriz en nuevas comunidades
function S=separacion(A, actual, i)
    A;
    [V, LAMBDA] = eig (A);
    [max_Eig_Val, max_Eig_Val_pos] = max(diag(LAMBDA));
    Eig_Vec_max=V(:, max_Eig_Val_pos);
    S = actual * ones(size(Eig_Vec_max));
    S(Eig_Vec_max>0) = i;

-- Se calcula la calidad de la nueva matriz

function Q=calidad(A, S, k_in, k_out, m)
    [lin_A, col_A]=size(A);
    Q = 0;
    for i=1:lin_A
        for j=1:col_A
            Q = Q + (A(i,j)-k_in(i)*k_out(j)/m)*delta_kronecker(S(i),S(j));

```

```

        end
    end
    Q=Q/m;

-- Funcion delta de kronecker

function d = delta_kronecker(i,j)
    if (i == j)
        d = 1;
    else
        d = 0;
    end
end

```

3.3. Las redes regulatorias

Las redes regulatorias, así como las funciones, fueron tomadas de [8]. Las características topológicas de las redes regulatorias conocidas de las tres bacterias se incluyen a continuación:

| | <i>E. Coli</i> | <i>B. Subtilis</i> | <i>P. Aeruginosa</i> |
|-----------------------------|----------------|--------------------|----------------------|
| Número de vértices | 151 | 93 | 82 |
| Número de aristas | 420 | 227 | 122 |
| Número de módulos conocidos | 12 | 13 | 10 |

Para la red de *E. Coli* se reconocen los módulos de: genes sin categoria asignada, uso alternativo de fuentes de carbono, módulo de respiración, responsables de productos

tóxicos, regulación de aminoácidos, quimiotaxis y regulación de motilidad, regulación de biofilms, respuesta a choque frío, respuesta externa, nitrógeno molecular, replicación de ADN y diversas actividades.

Para la red de *B. Subtilis* se reconocen los módulos de: genes sin categoría asignada, desarrollo de competencias, replicación de ADN, módulo de respiración, nitrógeno molecular, fuente de nitrógeno, uso alternativo de fuentes de carbono, respuesta externa, regulación de aminoácidos, esporulación, calor y respuesta a choque frío, respuesta a productos tóxicos y diversas actividades.

Para la red de *P. Aeruginosa* se reconocen los módulos de: genes sin categoría asignada, movimiento celular, procesos metabólicos de nitrógeno, procesos de biosíntesis, hierro regulador metabólico, biofilms, resistencia a antibióticos, proceso celular de aminoácidos, factores de virulencia y QS.

Como ya mencionamos, las redes de interacciones y los módulos funcionales conocidos se muestran en los Anexos II y III.

3.4. Metodología

Una vez obtenidas las interacciones que conforman las redes regulatorias [8], generamos a partir de ellas las matrices de adyacencia de dichas redes. Esta información fue introducida en el programa que implementa el algoritmo de detección de comunidades. Como resultado se obtuvo para cada matriz una lista, que indica a que subgrupo pertenece cada vértice en la red. Con esta información se realizó una comparación entre los módulos funcionales conocidos y los módulos obtenidos por el algoritmo.

Adicionalmente, se utilizó un algoritmo aleatorio, los algoritmos aleatorios son aquellos donde algunas elecciones son realizadas al azar. En este caso, se realizarón divisiones en módulos de manera aleatoria, considerando dos valores fijos de los números posibles de módulos: el número de módulos funcionales en la red real, y el número de

módulos obtenido por el algoritmo de detección de comunidades. Para cada división obtenida se calculó el valor de Q , y se seleccionó la partición en módulos que reportó el mejor valor dentro de 15000 corridas para *E. coli* y *P. aeruginosa* y 25000 para *B. subtilis*.

El número de corridas fue calculado a partir de la técnica del intervalo de confianza, que nos permite evaluar el número óptimo de simulaciones de acuerdo a la formula: $N = \frac{\sigma^2}{0,0001^2}$, donde $\sigma = 3,55 \times 10^{-3}$ y $N = 1266,7$ para *E. coli*, $\sigma = 0,014188$ y $N = 20130$ para *B. subtilis* y $\sigma = 0,010040$ y $N = 10080$ para *P. aeruginosa*. El valor de σ fue estimado previamente en base a los valores máximos de Q , obtenidos de correr el programa 100 veces en bloques de 100 experimentos.

La prueba indica que el parámetro Q se encuentra con probabilidad 95 % en los intervalos $0,284 \pm 0,0001$ para *E. coli*, $0,3166 \pm 0,0001$ para *B. subtilis* y $0,2868 \pm 0,0001$ para *P. aeruginosa*.

Capítulo 4

Resultados

Para ejemplificar la partición de una red en módulos usando el algoritmo seleccionado, mostramos en las Figuras 4.1 a 4.10 la partición de la red regulatoria de *E.coli*.

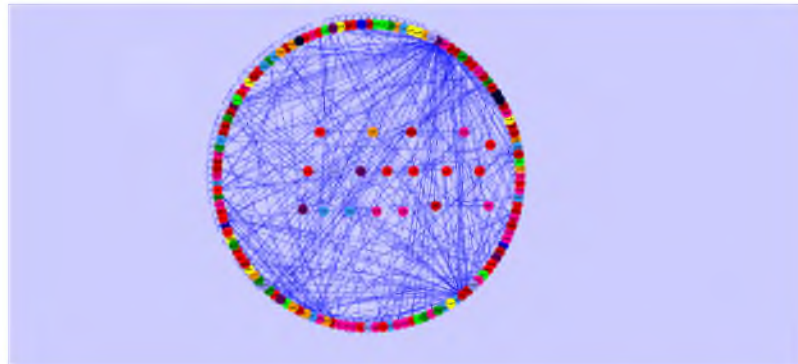


Figura 4.1: Todos los módulos conforman un sólo grupo.

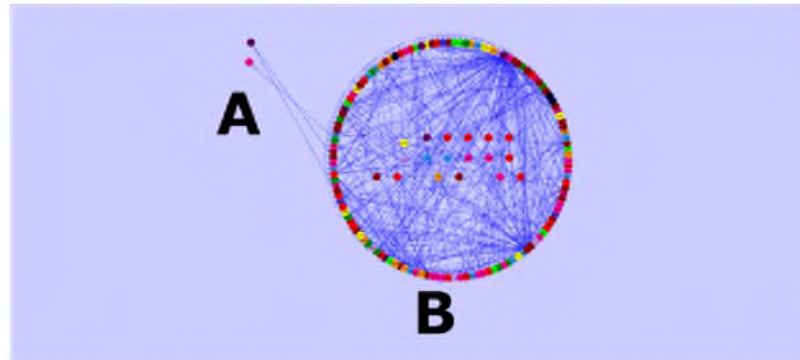


Figura 4.2: Partición en dos módulos A y B.

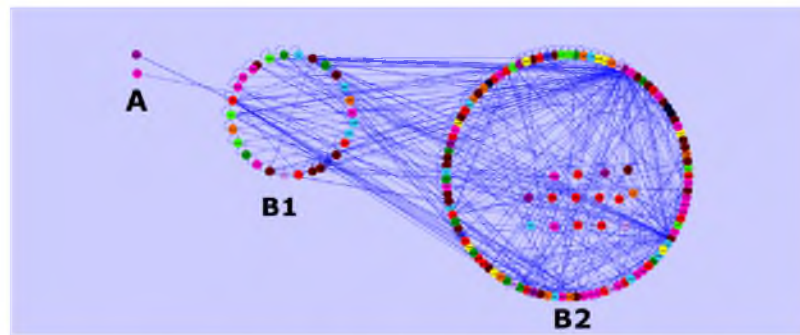


Figura 4.3: Subdivisión del módulo B en dos módulos: B1 y B2.

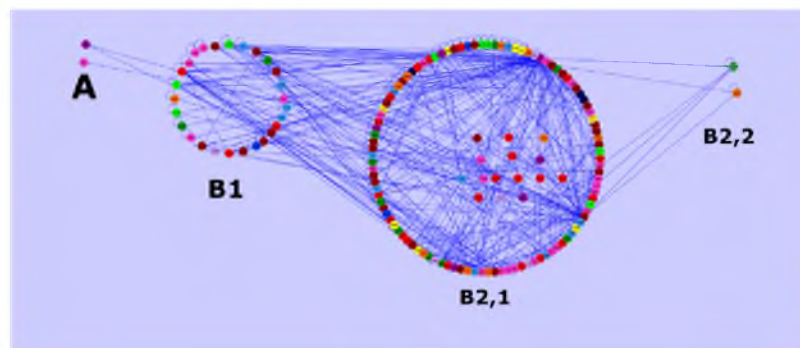


Figura 4.4: Subdivisión del módulo B2 en B2,1 y B2,2

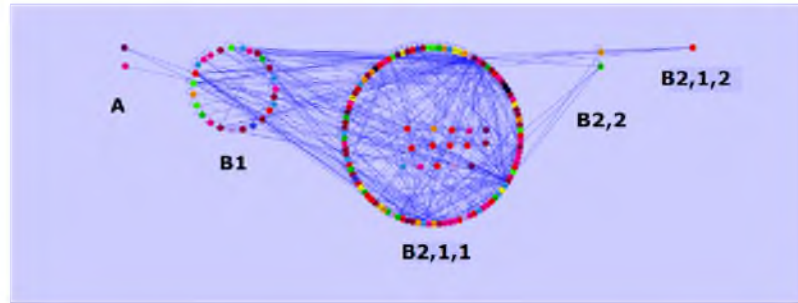


Figura 4.5: Subdivisión del módulo B2,1 para obtener los módulos B2,1,1 y B2,1,2

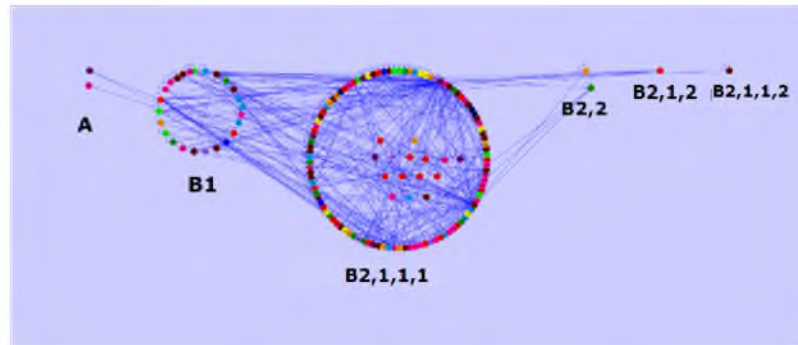


Figura 4.6: Subdivisión del módulo B2,1,1 para obtener los módulos B2,1,1,1 y B2,1,1,2

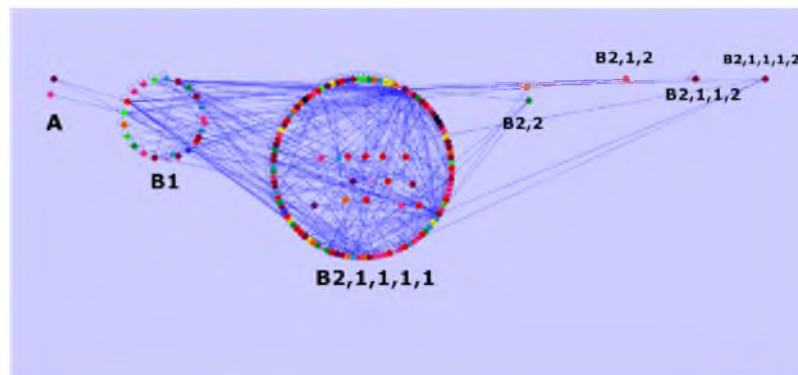


Figura 4.7: Subdivisión del módulo B2,1,1,1 para obtener los módulos B2,1,1,1,1 y B2,1,1,1,2

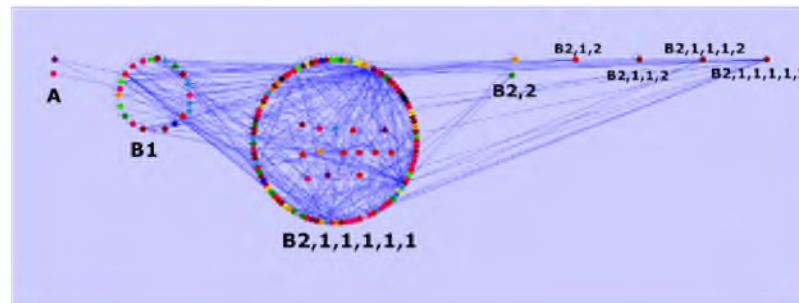


Figura 4.8: Subdivisión del módulo B2,1,1,1,1 para obtener los módulos B2,1,1,1,1 y B2,1,1,1,2

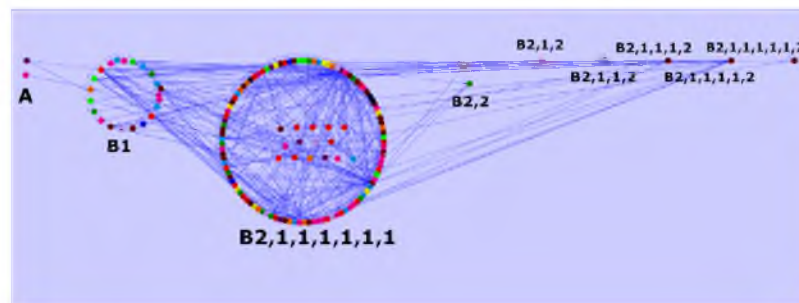


Figura 4.9: Subdivisión del módulo B2,1,1,1,1 para obtener los módulos B2,1,1,1,1,1 y B2,1,1,1,1,2

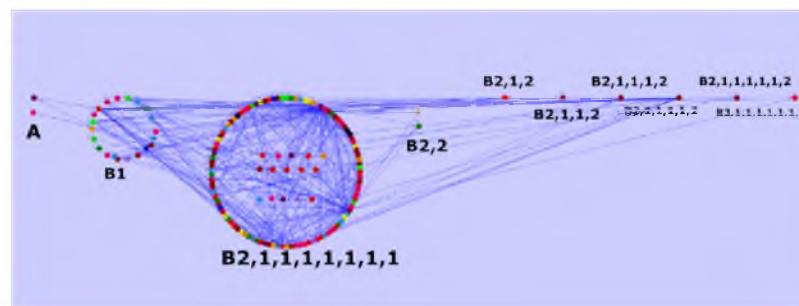


Figura 4.10: Subdivisión del módulo B2,1,1,1,1,1 para obtener el módulo B2,1,1,1,1,1,1 y B2,1,1,1,1,1,1,2

Calculamos Q para la división conocida de la red real en módulos funcionales y 15000 particiones aleatorias para *E. coli* y *P. aeruginosa* y 25000 para *B. subtilis* con la misma cantidad de módulos, reportamos la mejor Q obtenida de manera aleatoria.

Mostramos también la mejor Q obtenida a partir de los resultados del algoritmo, así como los tiempos de corrida del programa y el número de módulos obtenidos. Para este número de módulos también generamos 15000 particiones aleatorias para *E. coli* y *P. aeruginosa* y 25000 para *B. subtilis*, y reportamos la mejor Q obtenida.

| | | <i>E. Coli</i> | <i>B. Subtilis</i> | <i>P. Aeruginosa</i> |
|------------------------------------|--|-----------------------------------|---------------------------------------|-----------------------------------|
| Módulos funcionales en la red real | Número de módulos Q | 12 0.3043 | 13 0.3546 | 10 0.3429 |
| Módulos obtenidos por el algoritmo | Tiempo de corrida Número de módulos Q inicial Q final | 0.1291s 10 0.0189 0.1322 | 0.1017s 2 2.1415e-016 0.1017 | 0.4256s 16 0.3370 0.5224 |
| Mejor red aleatoria I | Número de módulos Q | 12 0.2826 | 13 0.3313 | 10 0.3102 |
| Mejor red aleatoria II | Número de módulos Q | 10 0.2818 | 2 0.2921 | 16 0.2916 |

Se observó que el número de módulos obtenidos es menor al número de módulos funcionales excepto en el caso de *P. aeruginosa*. Se observa también que la mayoría de los vértices están concentrados en pocos módulos que son mucho más grandes que el resto.

En las Figuras 4.11, 4.12 y 4.13 cada columna de la gráfica indica un módulo identificado por el algoritmo, y cada color muestra los genes que corresponden al mismo módulo

funcional. Puede observarse que un módulo del algoritmo contiene a los elementos de varios módulos funcionales y que la mayoría de los elementos de un módulo funcional pertenecen al mismo módulo del algoritmo.

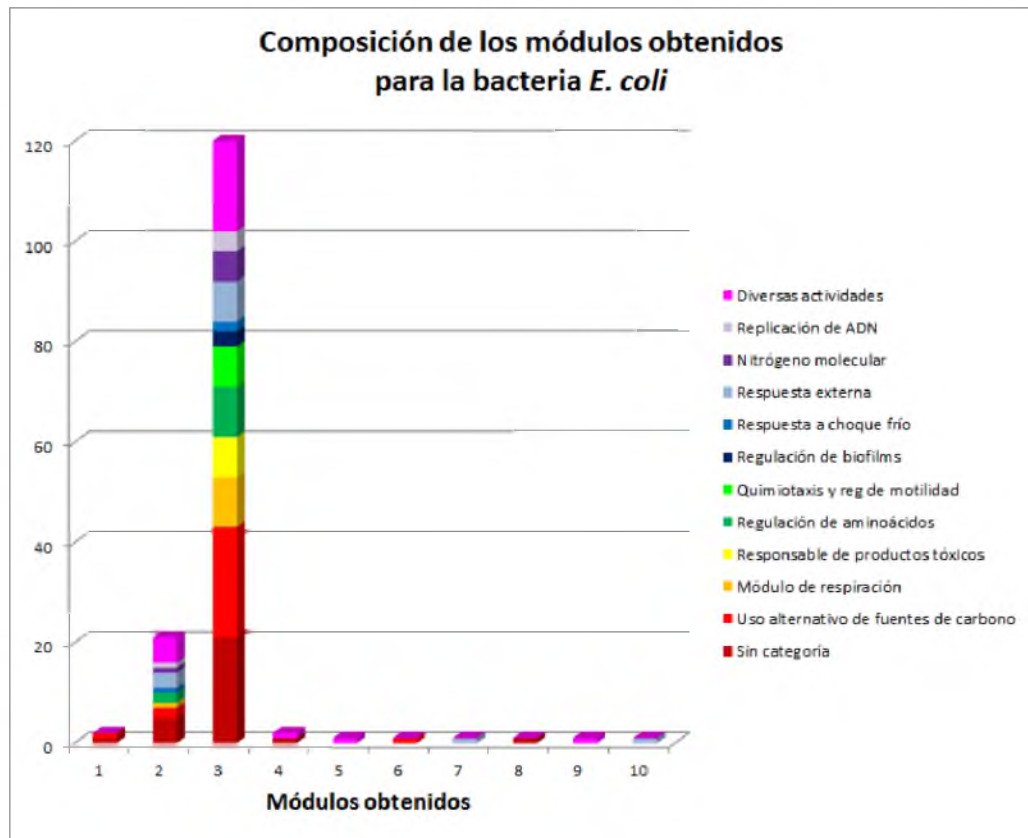


Figura 4.11: El algoritmo divide la red de *E.coli* en 10 módulos.

Por ejemplo, para el caso de *E.coli*: El segundo módulo contiene: 17,2% de los genes sin categoría asignada, 7,69% del módulo funcional de uso alternativo de fuentes de carbono, 9,09% del total del módulo de respiración, 16,7% del módulo de regulación de aminoácidos, 33,3% del módulo de respuesta a choque frío, 23,1% del módulo de respuesta externa, 14,3% del módulo de nitrógeno molecular, 20% del módulo de replicación de ADN y 19,2% del módulo de diversas actividades. El tercer módulo contiene: 72,4% del módulo de genes sin categoría asignada, 84,6% del módulo de uso alternativo

de fuentes de carbono, 90,9% del módulo de módulo de respiración, 100% del módulo de responsables de productos tóxicos, 83,3% del módulo de regulación de aminoácidos, 100% del módulo de quimiotaxis y regulación de motilidad, 100% del módulo de regulación de biofilms, 66,7% del módulo de respuesta a choque frio, 61,5% del módulo de respuesta externa, 85,7% del módulo de nitrogeno molecular, 80% del módulo de replicación de ADN y 69,2% del módulo de diversas actividades.

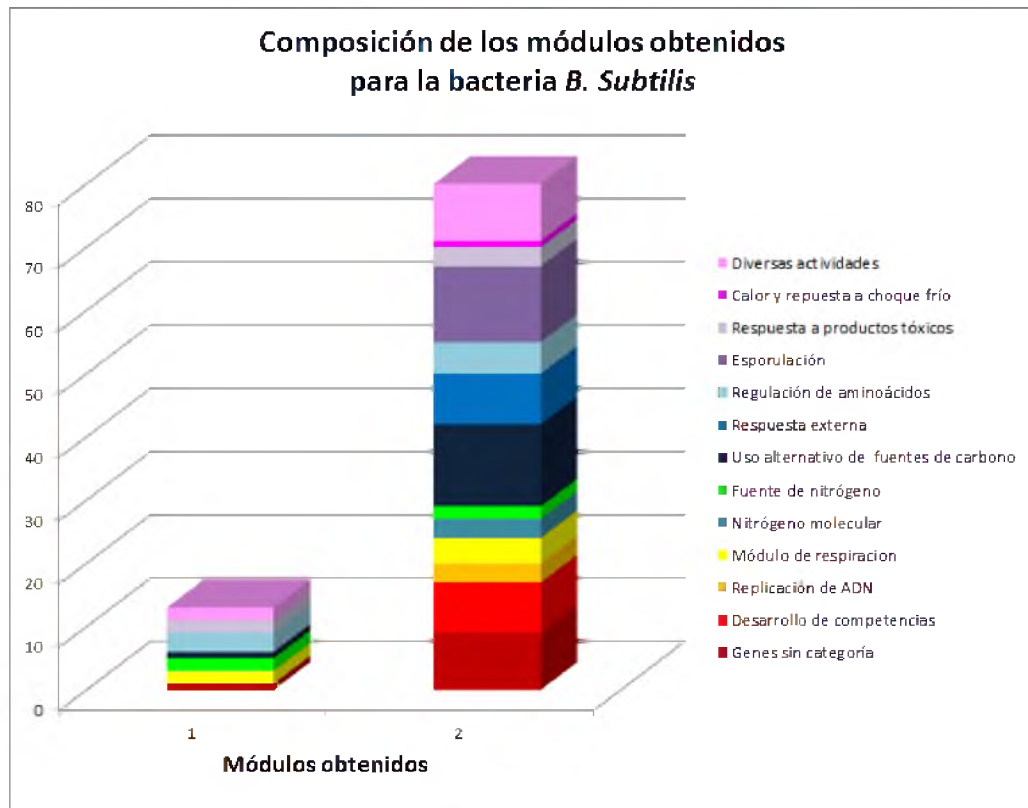


Figura 4.12: Se muestra la composición de los módulos obtenidos por el algoritmo para la red regulatoria de *B. subtilis*.

Para el caso de *B. subtilis* el primer módulo contiene: 10% del módulo de genes sin categoría asignada, 33% del módulo de respiración, 50% del módulo de fuente de nitrógeno, 7% del módulo de uso alternativo de fuentes de carbono, 37,5% del módulo de regulación de aminoácidos, 33,3% del módulo de calor y respuesta a choque frio, 40%

del módulo de respuesta a productos tóxicos y 18% del módulo de diversas actividades. El segundo módulo contiene: 90% del módulo de genes sin categoría asignada, 100% del módulo de desarrollo de competencias, 100% del módulo de replicación de ADN, 67% del módulo de respiración, 50% del módulo de fuente de nitrógeno, 100% del módulo de nitrógeno molecular, 93% del módulo de uso alternativo de fuentes de carbono, 100% del módulo de respuesta externa, 62,5% del módulo de regulación de aminoácidos, 100% del módulo de esporulación, 33,3% del módulo de calor y respuesta a choque frío, 60% del módulo de respuesta a productos tóxicos y 82% del módulo de diversas actividades.

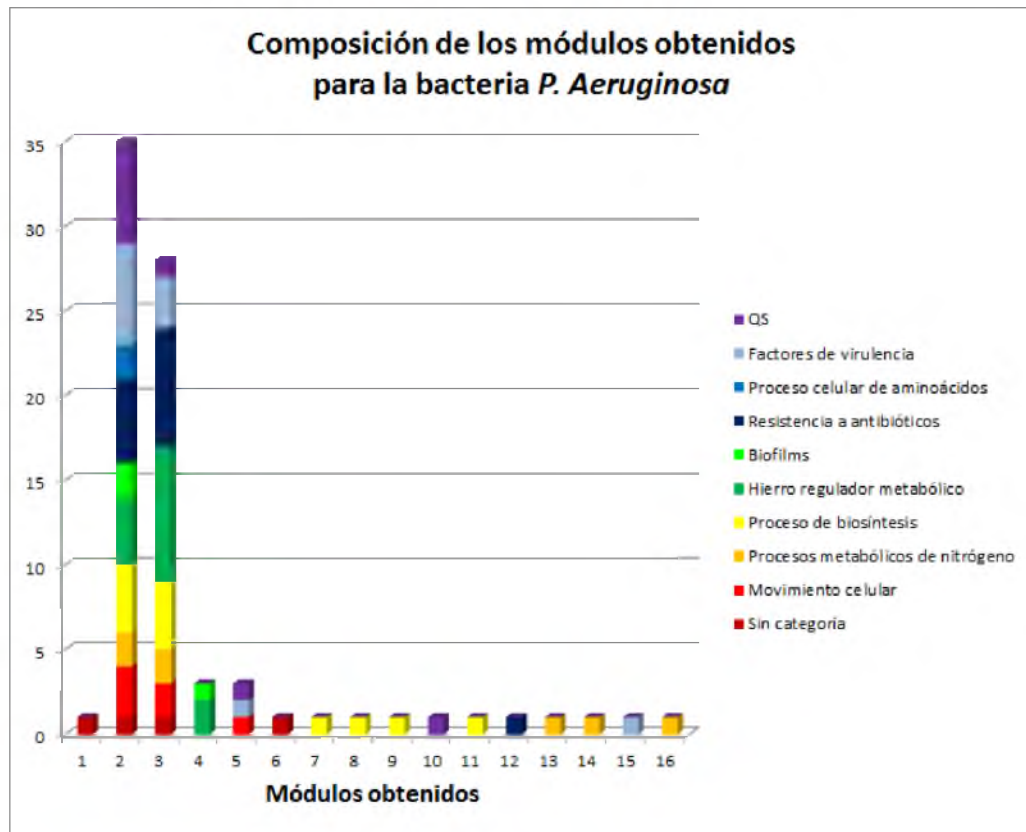
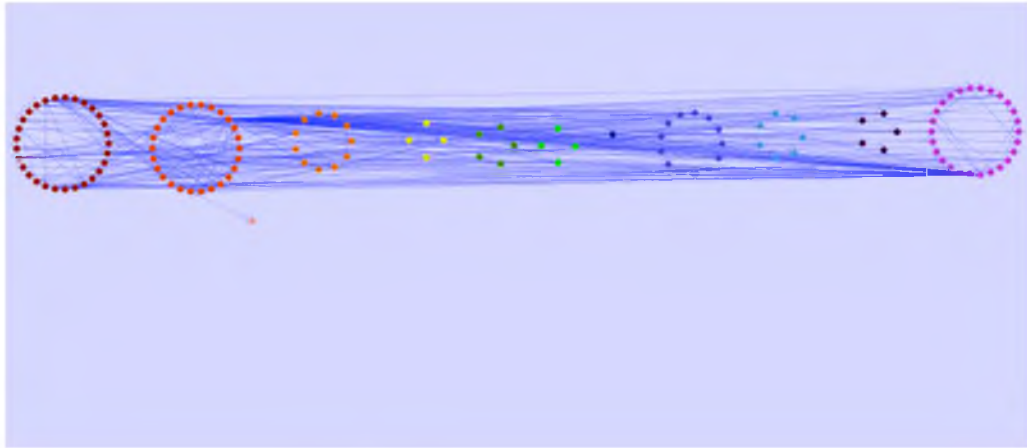
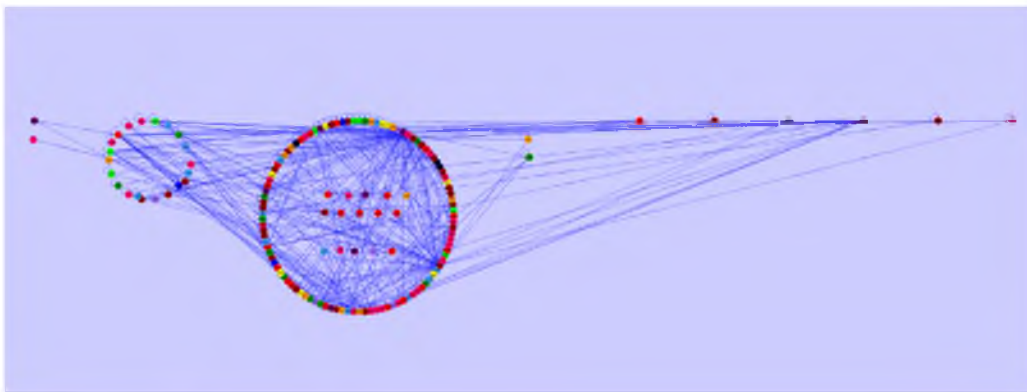


Figura 4.13: Se muestra la composición de los módulos obtenidos por el algoritmo para la red regulatoria de *P. aeruginosa*.

Para el caso de *P. aeruginosa* el algoritmo encontró 16 módulos. El segundo módulo

contiene: 25 % del módulo de genes sin categoría asignada, 50 % del módulo de movimiento celular, 29 % del módulo de procesos metabólicos de nitrógeno, 33 % del módulo de procesos de biosíntesis, 29 % del módulo de hierro regulador metabólico, 67 % del módulo de biofilms, 38 % del módulo de resistencia a antibióticos, 100 % del módulo de proceso celular de aminoácidos, 55 % del módulo de factores de virulencia y 67 % del módulo de QS. El tercer módulo contiene: 25 % del módulo de genes sin categoría asignada, 33,3 % del módulo de movimiento celular, 28,6 % del módulo de procesos metabólicos de nitrógeno, 33,3 % del módulo de procesos de biosíntesis, 57,1 % del módulo de hierro regulador metabólico, 53,8 % del módulo de resistencia a antibióticos, 27,3 % del módulo de factores de virulencia y 11,1 % del módulo de QS.

Para comparar los resultados obtenidos por el algoritmo mostramos para cada bacteria 4 redes, la primera indica la organización real, los genes correspondientes a cada módulo se distinguen por el mismo color, en segundo lugar se ilustra la división realizada por el algoritmo, más adelante se muestran dos divisiones de la red realizadas aleatoriamente, en las últimas tres redes se mantiene el mismo color para los genes pertenecientes a un mismo módulo funcional. Las figuras 4.14, 4.15, 4.16 y 4.17 corresponden a la bacteria *E. coli*. Las figuras 4.18, 4.19, 4.20 y 4.21 corresponden a la bacteria *B. subtilis* y las figuras 4.22, 4.23, 4.24 y 4.25 pertenecen a la bacteria *P. aeruginosa*.

Figura 4.14: Módulos funcionales en la bacteria *E.coli*.Figura 4.15: Módulos obtenidos por el algoritmo para la red regulatoria de *E.coli*

Como podemos observar, este algoritmo particionó en menos módulos de los esperados las redes de *E. coli* y *B. subtilis*. En las tres redes algunos de los módulos obtenidos por el algoritmo agrupan satisfactoriamente varios módulos funcionales en uno solo, sin embargo, necesitaríamos información adicional, relacionada por ejemplo con la dinámica del sistema para realizar una segunda partición de estos grupos. Encontramos entonces, que el algoritmo aplicado a redes biológicas proporciona una solución inicial válida para el problema de modularidad.

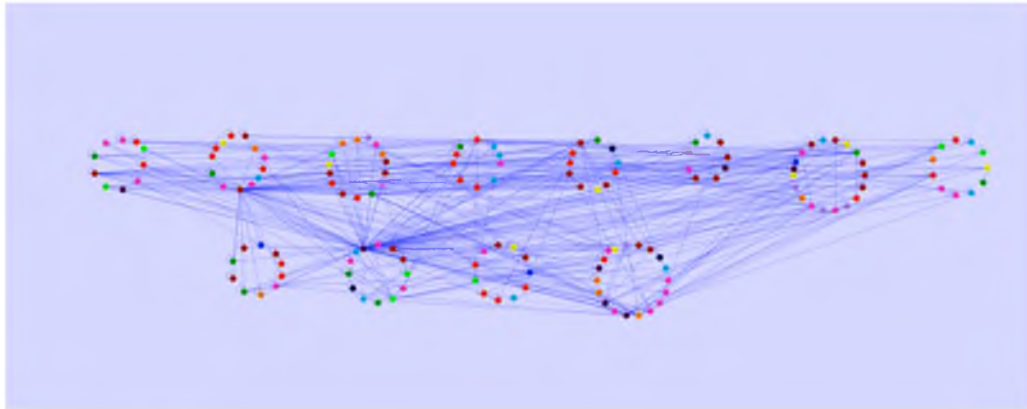


Figura 4.16: Módulos obtenidos aleatoriamente, considerando un número fijo de 12 módulos (*E.coli*).

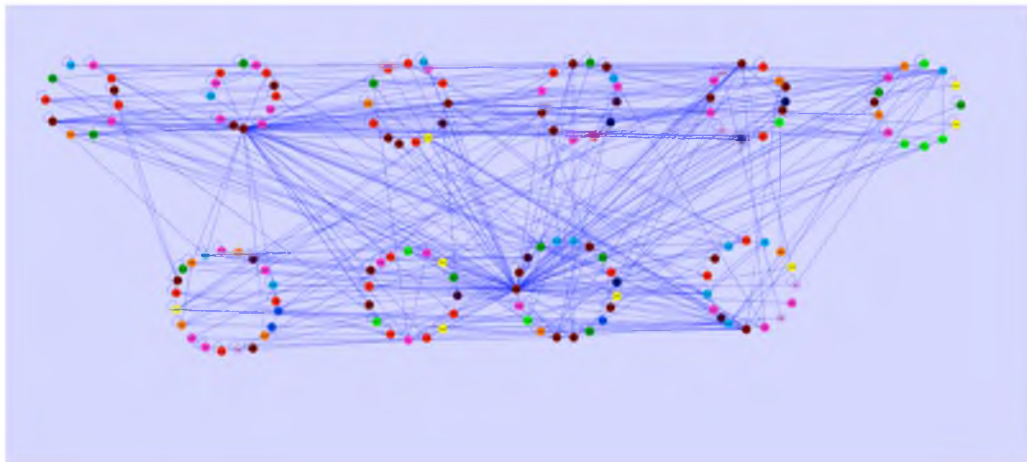


Figura 4.17: Módulos obtenidos aleatoriamente, considerando un número fijo de 10 módulos (*E.coli*).

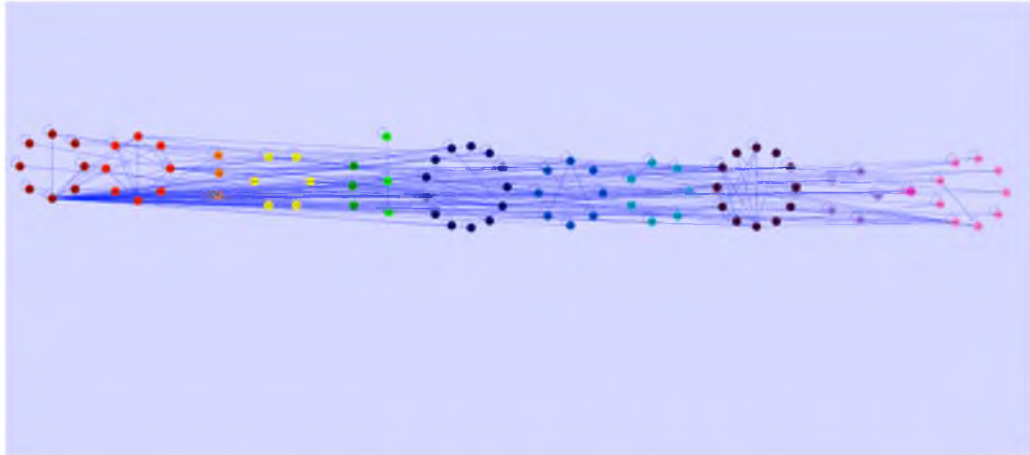


Figura 4.18: Módulos funcionales en la bacteria *B. subtilis*.

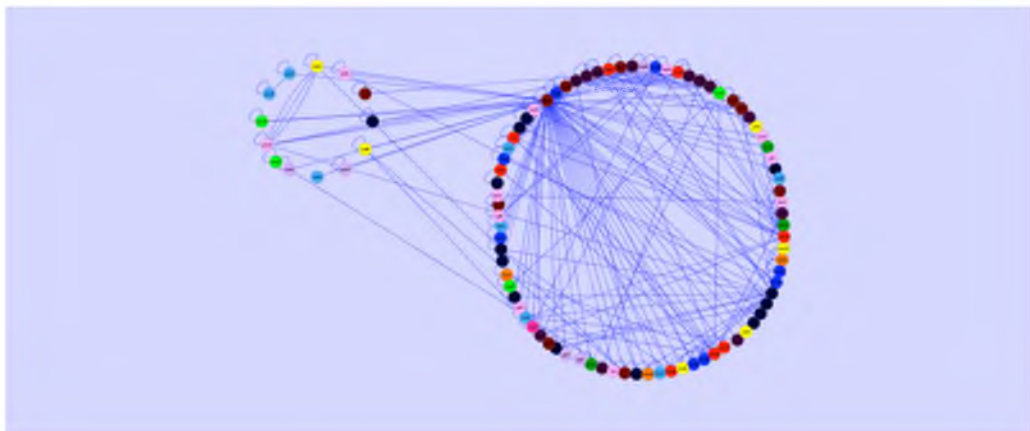


Figura 4.19: Módulos obtenidos por el algoritmo implementado.

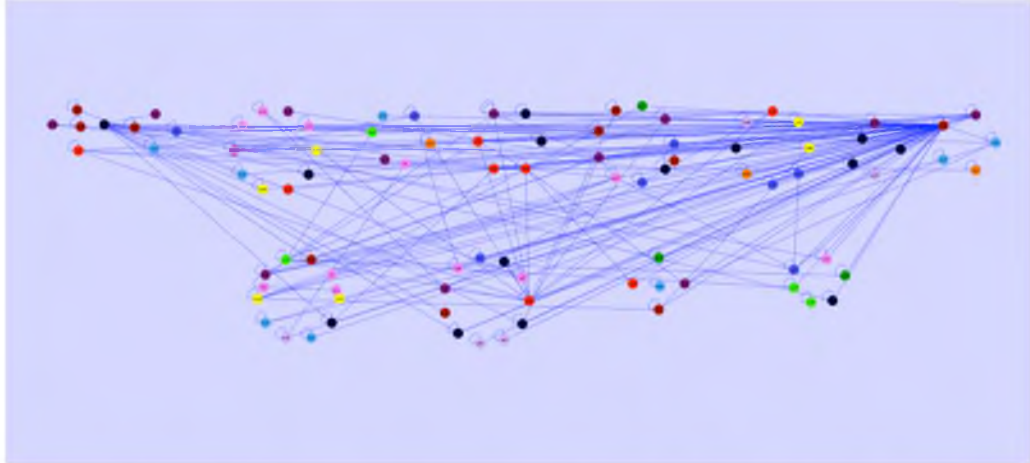


Figura 4.20: Módulos obtenidos aleatoriamente (*B. subtilis*).

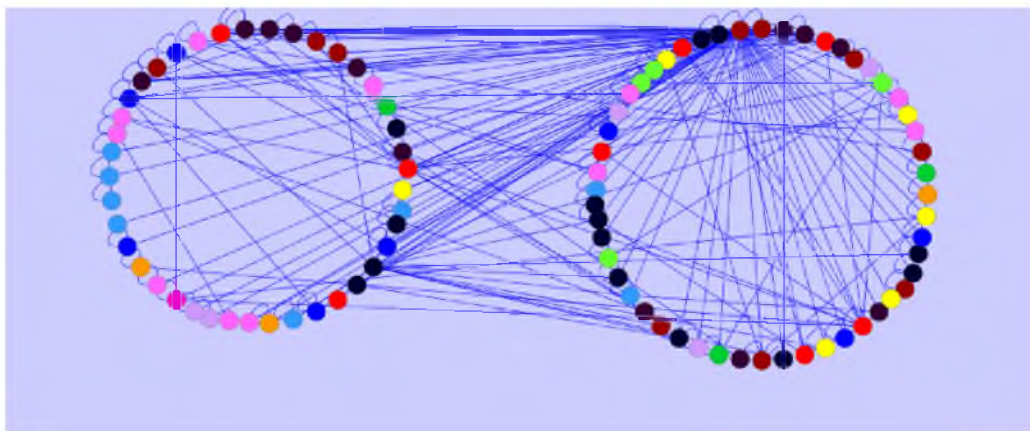


Figura 4.21: Módulos obtenidos aleatoriamente (*B. subtilis*)

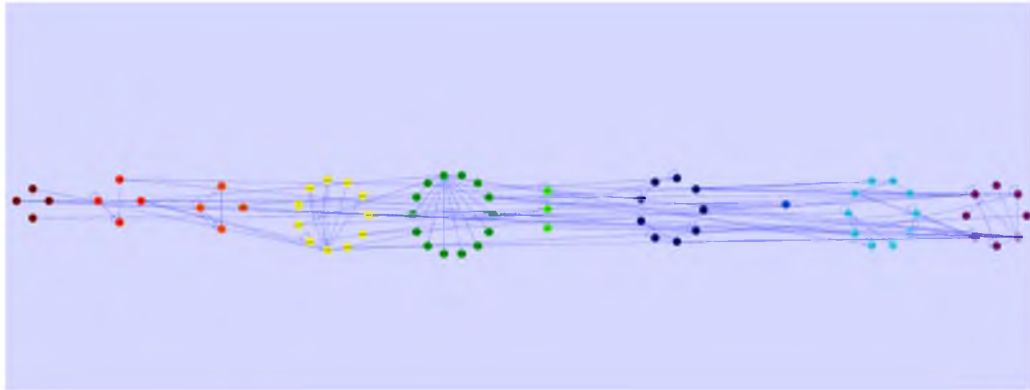


Figura 4.22: Módulos funcionales para la red regulatoria de *P. aeruginosa*.

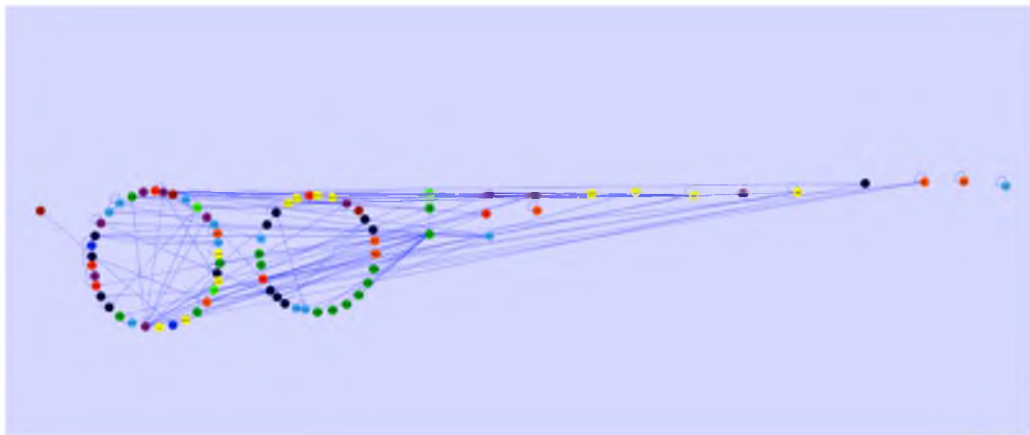


Figura 4.23: Módulos obtenidos por el algoritmo para la red regulatoria conocida de la bacteria *P. aeruginosa*.

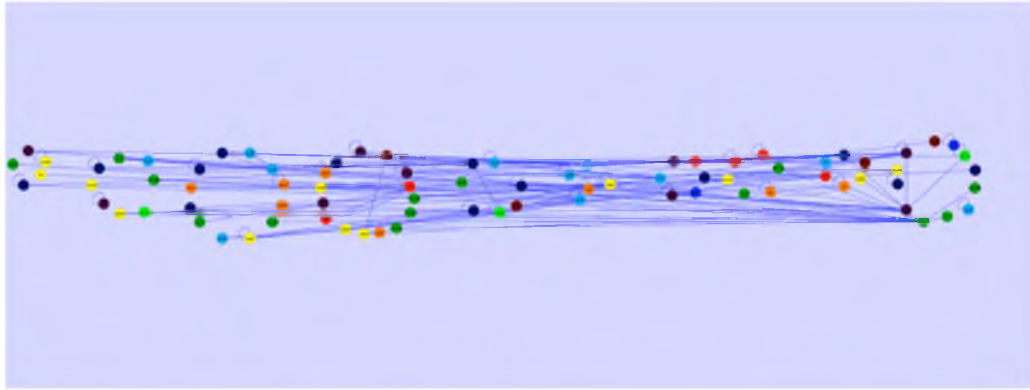


Figura 4.24: Módulos obtenidos aleatoriamente *P. aeruginosa*.

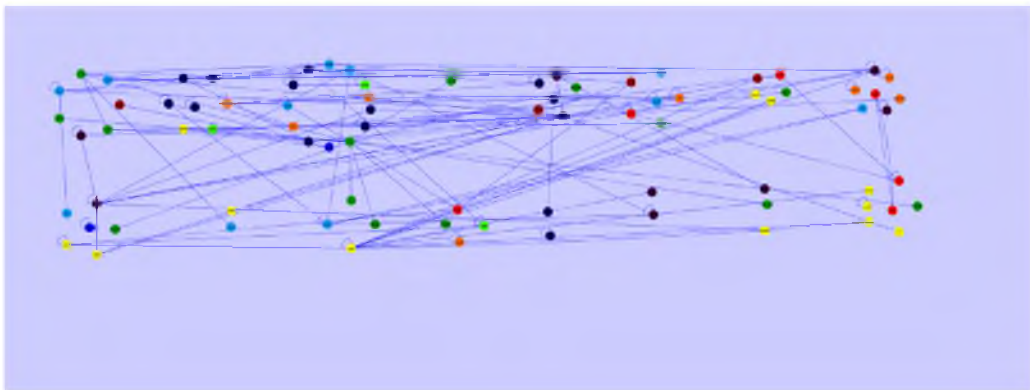


Figura 4.25: Módulos obtenidos aleatoriamente *P. aeruginosa*.

Capítulo 5

Conclusiones

Durante el desarrollo de este trabajo se adquirieron conocimientos nuevos sobre un problema biológico relacionado con el modelado de redes regulatorias dinámicas, de la misma manera se adquirieron nuevos conocimientos sobre la programación en Matlab, así como sobre algoritmos aleatorios y deterministas. Se repasaron además conceptos relacionados con teoría de grafos, álgebra lineal y teoría de algoritmos.

Como un intento para modelar redes grandes se ha propuesto el uso de la modularidad, por lo que la investigación en este sentido es un campo abierto a las propuestas de nuevos algoritmos. En este trabajo se analizaron diversos algoritmos que permiten dividir la red en subgrafos, se estableció su clasificación y sus características y seleccionamos un algoritmo basado sólo en la topología de la red para obtener módulos. Se eligió el algoritmo de detección de comunidades [12] debido a que fue diseñado específicamente para crear comunidades en redes dirigidas, cualidad que también distingue a las redes biológicas, y no requiere información a priori.

Durante la ejecución del programa observamos que es computacionalmente eficiente, y que el valor de Q es sensible a distintos casos, por ejemplo: las autoregulaciones, es decir la inclusión o no de aristas que van de un vértice en si mismo; la matriz transpuesta, que devuelve particiones diferentes; la asignación de los vértices con valor 0 en el

eigenvector seleccionado del máximo eigenvalor.

En cuanto a los resultados obtenidos, se observó que el algoritmo de detección de comunidades podría ser utilizado para realizar una primera división de los vértices en redes regulatorias biológicas, ya que logra concentrar a la mayoría de los vértices relacionados con una función en un sólo módulo, pero no logra distinguirlo de otros módulos funcionales. Por lo que será necesario diseñar un filtro que permita hacer separaciones de estos módulos para tener una división más efectiva.

Debido a que para este trabajo solo se contó con información topológica, este algoritmo resultó de gran utilidad, ya que permite trabajar sin ninguna información a priori, haciendo uso solamente de las matrices de adyacencia sin predeterminedar el tamaño de las comunidades. Por su parte las divisiones aleatorias requieren de un número de módulos a priori, y no muestran ningún tipo de organización en relación a los módulos funcionales.

Muchas pruebas más son posibles con las redes biológicas con las que contamos y el algoritmo de detección de comunidades, sin embargo, el objetivo del trabajo se considera cumplido, ya que durante el desarrollo de esta tesis se avanzó en el conocimiento de la modularidad biológica, los algoritmos existentes, y en particular los alcances y limitaciones de un algoritmo basado en topología, como el algoritmo seleccionado.

Como trabajo a futuro se propone analizar los módulos funcionales conocidos para observar si sus características topológicas coinciden con las características que buscan los algoritmos, así como plantear o mejorar un algoritmo que integre conocimientos dinámicos y topológicos.

Bibliografía

- [1] J. C. Aguavil , S. A. Enriquez **Evaluación del efecto de un probiótico nativo elaborado en base a lactobacillus acidophilus y bacillus subtilis en pollos broiler ross-308 en santo domingo de los tsúchilas.** Escuela politécnica del ejército. Departamento de ciencias de la vida.
- [2] C. Aguirre, J. Martins and R. Vilela Mendes **Dynamics and Coding of a Biologically-Motivated Network.** Bifurcation and Chaos 2:383-394, (2006).
- [3] U. Alon **An introduction to systems biology: Desing principles of biology circuits.** Champman and Hall (2006).
- [4] H. de Jong **Modeling and Simulation of Genetic Regulatory Systems: A Literature Review.** Journal of Computational Biology 9:69-105, (2002).
- [5] A. J. Enrhght, S. Van Dongen and C. A. Ouzounis **An eficiente algorithm for large-scale detection of protein families.** The European Bioinformatics Institute, EMBL Cambridge OutStation.
- [6] J.I Espinosa de los Monteros Fernández. **Caracterización del proceso de crecimiento de Bacillus subtilis bajo condiciones anaerobias.** Instituto de Biotecnología UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO.
- [7] E. Galán-Vásquez, Beatriz Luna and Agustino Martínez-Antonio, **The Regulatory Network of Pseudomonas** Microbial Informatics Experimentation 1, (2011).

- [8] E. Galan Vasquez, B. Luna, and A. Martinez-Antonio. **Description de gene regulatory networks in three bacterial models. E.coli, B. subtilis and P. aeruginosa.** Recent Res. Devel Microbiology 12:19-39, (2012).
- [9] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett and J. Collado-Vides, Nucleic Acids Res, 98105 (2010).
- [10] Z. Jiang, R. Gentleman. **Extensions to gene set enrichment.** Bioinformatics 23:306-313, (2007).
- [11] H. Kitano **Foundations of systems biology** The MIT Press (2001).
- [12] E. A. Leicht and M. E. J. Newman **Community Structure in Directed Networks** (2008).
- [13] W. Longabaugh and H. Bolouri. **Understanding the Dynamic Behavior of Genetic Regulatory Networks by Functional Decomposition.** Curr Genomics 6:333-341 (2006).
- [14] F. Luo, Y. Yang, C. Chen, R. Chang, J. Zhou, R. Scheuermann. **Modular organization of protein interaction networks.** Bioinformatics 23:207-214, (2007).
- [15] F. Martínez-Sandoval, R. Steffen, CD Ericsson, HL DuPont. **Enteroaggregative Escherichia coli as a major etiologic agent in traveler's diarrhea in 3 regions of the world.** Clin. Infect. Dis. 32: 1706-1709, (2001).

- [16] AB Okey **Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data**. *Brief Bioinform* 6:331-343, (2005).
- [17] A á N.Przulj, Department of Computer Science, University of Toronto, Toronto, M5S 3G4, Canada **Protein complex prediction via cost-based clustering** (2004).
- [18] O. Reséndiz-Antonio, A. Freyre-González, R. Menchaca-Méndez, M. Gutiérrez, A. Martínez, C. Sánchez y J. Collado-Vides, **Modular analysis of the transcriptional regulatory network of E. coli**. *Trends Genet* 21:16-20, (2005).
- [19] L. Ruíz Martínez **Pseudomonas aeruginosa: Aportación al conocimiento de su estructura y al de los mecanismos que contribuyan y al de los mecanismos que contribuyen a su resistencia a los antimicrobianos**. Unidad de microbiología . Departamento de patología y terapéutica experimental. Facultad de medicina. Universidad de Barcelona.
- [20] R. Sharan, I. Ulitsky, R. Shamir **Network-based prediction of protein function** School of Computer Science, Tel Aviv University, Tel Aviv, Israel (2007).
- [21] N. Sierro, Y. Makita, M. Hoon and K. Nakai, *Nucleic Acids Res* 36:936, (2008).
- [22] A. Subramanian, P. Tamayo, VK. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander et al.: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 102:15545-15550, (2005).
- [23] R. Thomas: **Boolean formalization of genetic control circuits**. *J Theor Biol* 42:563–85, (1973).
- [24] C. R. Vázquez Quiñones **Actividad antimicrobiana de aceites esenciales de bellotas de ciprés (Cupressus sempervirens var horizontalis) y semilla de**

brócoli (*Brassica oleraceae* var *italica*) sobre *Staphylococcus aureus* y *Pseudomonas aeruginosa* Universidad Autónoma metropolitana-Iztapalapa.

- [25] X. Wang, E.I Dalkic, M. Wu and C. Chan. **Gene module level analysis: identification to networks and dynamics** (2008).
- [26] S. Zhang, X. Ning **Identification of functional modules in a PPI network by clique percolation clustering**
- [27] <http://albherto.wordpress.com/2011/04/27/21692/>
- [28] <http://www.dateriles.com/2012/07/mundo-oculto.html>
- [29] http://es.wikipedia.org/wiki/Problema_de_los_puentes_de_Konisberg
- [30] <http://blog.ciencias-medicas.com/archives/1373>
- [31] <http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax>