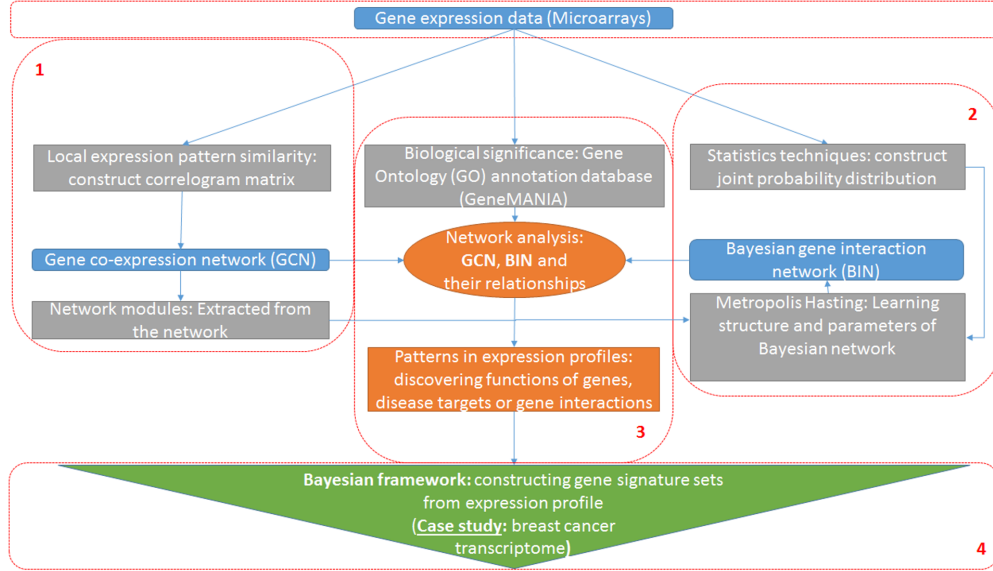


Figura 1: Marco metodológico.



1. Desarrollo y validación del marco de trabajo de redes Bayesianas

1.1. Metodología

Se siguieron cuatro pasos metodológicos para el desarrollo de este marco de trabajo. A continuación, se muestra el diagrama de flujo en la Figura 1, y seguidamente se detalla cada paso.

1.1.1. Construcción y visualización de redes de co-expresión de genes

Primero se caracterizó y adoptó un método general para construir y visualizar redes de co-expresión génica, a partir de los datos de expresión de genes generados por micro-arreglos, y después, se aplicó este método a datos de expresión de E.coli (para mayor detalle de este proceso ver el capítulo 2, sección materiales y métodos).

1.1.2. Construcción de la arquitectura óptima de una red Bayesiana y sus parámetros, a partir de los datos de expresión de genes

Paso a seguir se caracterizó un método general de aprendizaje de redes Bayesianas y sus parámetros a partir de datos de expresión de genes, basado en el enfoque de simulación de Markov Chain Monte Carlo (MCMC), en el cual definiendo una cadena de Markov sobre el espacio de posibles estructuras de la

red Bayesiana, se realiza una caminata aleatoria en dicha cadena de Markov y se genera un conjunto de estructuras viables probabilísticamente (que mejor se ajustaron a los datos según una función de verosimilitud ó una distribución posterior) para la red (para más detalle ver el capítulo 3.).

1.1.3. Implementación de las herramientas de análisis de redes complejas y bayesianas de expresión de genes y sus relaciones

En este paso metodológico se analizan los módulos (o clusters) obtenidos a partir de la red de co-expresión (CGN) obtenida en el primer paso y después del paso anterior, de manera pos-óptima se filtran las redes Bayesianas óptimas empleando algunas estrategias para evitar sobre-entrenamiento (overfitting, es decir, grafos con muchas aristas o estructuras muy complejas) de la red, empleando análisis de redes complejas (ver en el capítulo 3, la sección de sobre-entrenamiento y la sección de red Bayesiana minimal).

Con la construcción de las redes de co-expresión (CGN) y las redes bayesianas de interacción de genes (BIN), se procede con la selección de genes de interés biológico usando las redes mencionadas para realizar análisis de redes complejas sobre los módulos o agrupamientos (clusters) de genes en dichas redes, donde cada módulo es caracterizado topológicamente midiendo propiedades que indican la relevancia biológica de los genes al interior de ellos, y con información de bases de datos de anotaciones de ontologías de los genes (GO) , la tarea es identificar genes relevantes dentro de procesos biológicos de interés mediante la caracterización de los módulos a nivel funcional y de su forma (topología).

1.1.4. Caso de estudio: Análisis de datos de expresión de genes de micro-arreglos de DNA de transcriptomas

En el último paso, se somete el marco de trabajo construido a experimentación y análisis de datos de expresión de genes de micro-arreglos a través de un caso de estudio. Los pasos a seguir se describen a continuación.

Selección del problema Se seleccionó para la experimentación del marco de trabajo bayesiano la exploración de transcriptomas de *Escherichia coli*, ya que es un organismo que ha sido objeto de estudio y se encuentra documentado. Además, por ser un organismo bacteriano cuenta con un metabolismo más acelerado que otros organismos como humanos, chimpances o ratas, esto permite que en los experimentos de micro-arreglos para medir la expresión de sus genes se haya capturado de manera más evidente la interacción entre transcriptomas en comparación de otros organismo con un metabolismo más lento. Esta última característica nos inclinó más hacia este estudio.

Selección de los datos Se realizó el análisis de un conjunto de datos de expresión de genes de acuerdo a la problemática de aplicación escogida (ver capítulo 2, sección de materiales y métodos).

Aplicación del marco de trabajo bayesiano Se aplicó al conjunto de datos seleccionado al marco de trabajo de redes Bayesianas.

Verificación de los resultados Por último se realizó un análisis y discusión de los resultados obtenidos.

1.2. Arquitectura general de la solución

Siguiendo los pasos metodológicos de la sección anterior se estructuró una arquitectura general para la solución de este marco de trabajo de redes Bayesianas. Esta sección consta de dos partes: La primera introduce los conceptos básicos de arquitectura como configuración y propiedades arquitectónicas, el estado del arte donde se mencionan algunos trabajos relacionados y se describe el diagrama general de arquitectura del marco. En la segunda parte, se detalla la especificación de la implementación computacional del marco e incluye algunas vistas arquitectónicas del mismo.

1.2.1. Introducción

Que es una arquitectura de software? Hablaremos de la arquitectura general en términos de una abstracción de los elementos de un sistema de software durante alguna fase de su operación, como afirma Fielding en [4].

Por que se necesita una arquitectura? y de que se compone? Este marco está compuesto de elementos de diferente naturaleza algunos estadística, computacionales, biológica e informática. El objetivo de esta sección consiste en especificar una arquitectura general del marco que capture sus elementos, las relaciones entre ellos y las propiedades externamente visibles, de tal manera que esta pueda ser entendida por distintos interesados como estadísticos, biólogos e informáticos.

Que es la configuración de un sistema? y que son las propiedades arquitectónicas? El primer concepto arquitectónico es la configuración, entendiéndose como la estructura de las relaciones entre los diferentes elementos del sistema. El segundo concepto es el de propiedad arquitectónica también conocido como atributo de calidad, el cual hace referencia a cualquier característica externamente visible y que se desprende de una configuración particular [11].

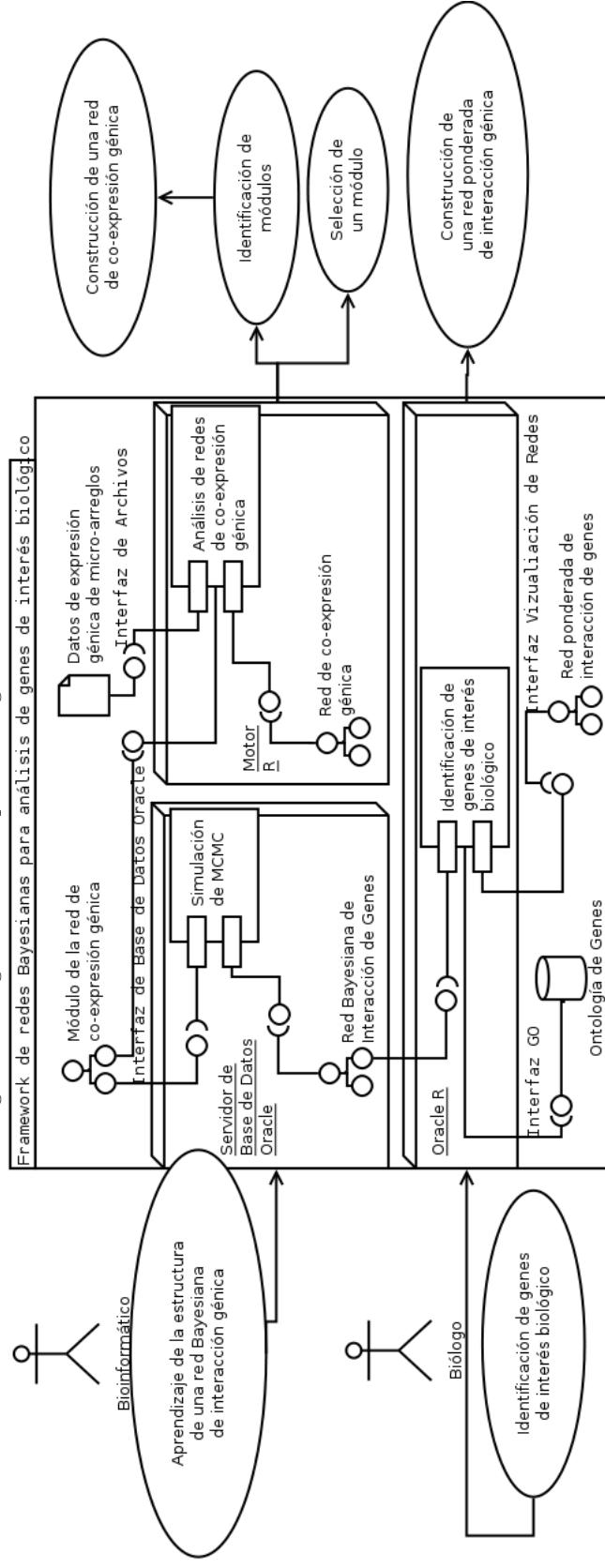
Trabajos relacionados Existen buenos trabajos relacionados con redes Bayesianas y análisis de datos de expresión, por ejemplo, [10] presenta un algoritmo basado en restricciones para el aprendizaje de la red Bayesiana, [5] se enfoca más en la utilización de los algoritmos de simulación de *Markov Chain Monte Carlo (MCMC)* para aprender la estructura de una red Bayesiana, [3, 7] hacen énfasis

en el modelo estadístico empleado para evaluar el *score* de las estructuras en el *espacio de búsqueda*, mientras que [9, 8] son una buena referencia en la literatura con las bases teóricas de las redes Bayesianas y los fundamentos de su aprendizaje con aplicaciones en expresión génica. Desafortunadamente estos autores no mencionaron, ni documentaron la arquitectura computacional de la implementación que usaron para obtener sus resultados.

Por el contrario, en [1] se desarrolló *Mergeomics* una arquitectura computacional de estilo arquitectónico *pipeline* que integra datos multidimensionales de asociación de enfermedades con genómica funcional y redes moleculares para obtener rutas metabólicas, redes de genes y reguladores centrales críticos para el desarrollo de enfermedades. Por otra parte, en [13] se desarrolló un nuevo método computacional Bayesiano con una búsqueda *MCMC* e implementado en el *Bayesian High-order Interaction Toolkit (BHIT)* para detectar interacciones epistáticas entre *Single Nucleotide Polymorphisms (SNPs)*. También en [12], se presentó el *GlobalMIT*, un *toolbox* para aprender la estructura óptima global de una red Bayesiana dinámica a partir de datos de expresión génica. Este método usa una métrica de *scoring* basado en teoría de la información llamada *Mutual Information Test (MIT)*, y con su utilización la tarea de aprendizaje es lograda en tiempo polinomial. Estos trabajos tienen en común que sus autores mencionaron y documentaron características arquitectónicas de las herramientas computacionales utilizadas en sus investigaciones y que están relacionados con este trabajo en temas como la simulación *MCMC* y las redes Bayesianas para detectar interacciones génicas a partir de datos de expresión o transcritomas. En la arquitectura propuesta, se integran demás elementos importantes como el análisis de redes complejas a través de la construcción de redes de co-expresión y detección de clusters que habilita un aprendizaje de redes Bayesianas un poco más enfocado en grupos de genes que comparten alguna función génica ya que se encuentran co-expresados. Adicionalmente, se incluyen también filtros pos-óptimos como estrategia para mitigar el sobre-entrenamiento de las redes Bayesianas.

En la figura 2 se presenta el diagrama de la arquitectura general y en la siguiente sección se describirán brevemente los elementos, componentes, conectores, datos y vistas arquitectónicas que se estructuraron en esta arquitectura de la solución.

Figura 2: Diagrama de la arquitectura general



1.2.2. Estructuración de la solución

A continuación veremos la descripción del diagrama de la arquitectura general de la figura 2, su configuración y propiedades arquitectónicas.

Elementos Los elementos arquitectónicos incluyen datos, procesos y conexiones. El marco incluye tres elementos arquitectónicos fundamentales: las redes de co-expresión génica, las redes Bayesianas y redes ponderadas de interacción génica.

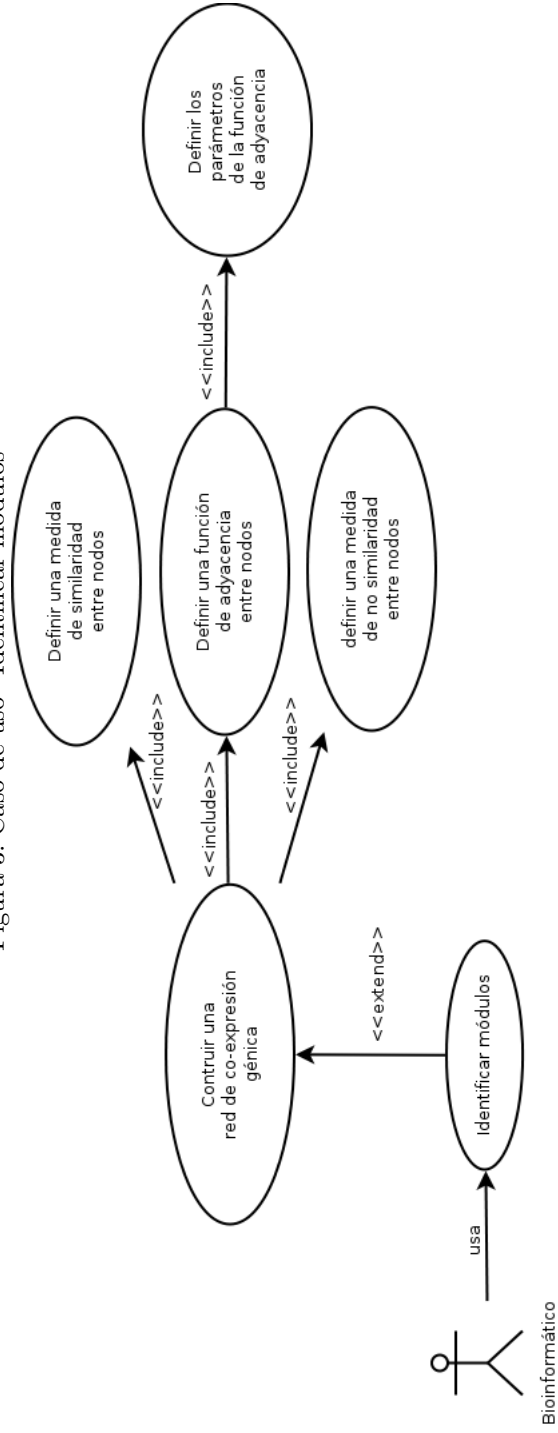
Componentes Los componentes arquitectónicos proporcionan una transformación de los datos por medio de sus interfaces, como es el caso de los componentes de simulación MCMC y el de análisis de redes de co-expresión génica del marco.

Conectores Los conectores arquitectónicos son los encargados de la coordinación, comunicación y cooperación entre componentes. Los principales conectores del marco son las interfaces de base de datos del componente de simulación MCMC, la interfaz de archivos de datos de expresión de micro-arreglos, la interfaz de visualización de redes génicas y la interfaz de base de datos de ontología del gen.

Datos Los datos arquitectónicos son los elementos transferidos entre dos componentes por medio de un conector. En el marco se trabajó con dos elementos de este tipo: los datos de expresión de genes de micro-arreglos y los datos de muestras generados por simulación computacional.

Vistas Las vistas son perspectivas o puntos de vista arquitecturales que capturan diferentes preocupaciones de los diferentes interesados en el sistema, dichas preocupaciones pueden ser la perspectiva lógica del sistema, el modelo de base de datos, la vista de despliegue o de casos de uso. En este marco se seleccionó la vista de casos de uso para caracterizar de forma general la capacidad de esta solución. Primero vemos en la figura 3 las capacidades para identificar clusters en redes de co-expresión génica por medio del análisis de redes complejas de la adopción del método WGCNA (ver el capítulo 2, para mayor detalle). Por otra parte, la figura 4 muestra como se extiende el caso de uso para identificar genes de interés biológico por medio del aprendizaje de redes Bayesianas y algunos filtro pos-óptimos para mitigar el sobre-entrenamiento (para mayor detalle ver el capítulo 3)

Figura 3: Caso de uso "Identificar módulos"



1.3. Casos de estudio: utilización del marco de trabajo de redes Bayesianas en *Escherichia coli*.

1.3.1. Flujo de trabajo (WF del ingles Workflow o pipeline) de las experimentaciones

Para cada gen del cluster donde se aplicó el marco de trabajo de redes Bayesianas, se cargaron las muestras a de datos de expresión génica, previamente discretizadas. y después, se ejecutó el proceso de aprendizaje de redes Bayesianas, para luego tomar los resultados y realizar por último el respectivo análisis bioinformático de los mismos, ver la figura 5. Para lograr este flujo de trabajo, cada vector de expresión génica fue sometido a un proceso preliminar de discretización por medio del método de cuantiles en donde se definieron tres niveles de expresión (alto, medio, bajo) o cuantiles ver el trabajo de Dey en [2].

Una vez realizadas las simulaciones MCMC los resultados fueron analizados por medio de técnicas bioinformáticas como el análisis de redes complejas, donde se revisaron las redes obtenidas, el logaritmo de su función distribución posterior y otras propiedades de la red como el valor máximo de grado de entrada (*máx. in-grade*), el valor mínimo de dependencia entre nodos hijos y sus padres (*near-independence*) y el número de aristas de la red.

Finalmente, filtramos las redes encontradas con un procedimiento de transformación a una red Bayesiana minimal (ver el capítulo 3) mitigando el sobreentrenamiento en este tipo de aprendizaje y por último reducimos la red Bayesiana minimal y apoyandonos en conocimiento biológico como el de la base de datos de interacción génica bacteriana en ABASY para *Escherichia coli* [6] y poder así dar una interpretación coherente de los resultados desde el punto de vista biológico.

1.3.2. Descripción de los casos seleccionados para análisis

Como se dijo en el marco metodológico, previo al aprendizaje de la red Bayesiana se construye una red de co-expresión génica, se identifican los clusters y se seleccionan los de interés según las funciones génicas anotadas en GO. Para el análisis del caso de estudio se seleccionaron aleatoriamente algunos clusters de los 90 identificados en el ejemplo de la red de co-expresión génica de *E.coli* construida en capítulo 2, en el cuadro 1 se muestra uno de los clusters seleccionados y como puede verse en las funciones génicas presentadas en el cuadro 2 el cluster seleccionado esta relacionado con el transporte de maltose.

1.3.3. Resultados de los casos

Experimentos con redes de 9 nodos o genes. Uno de los cluster seleccionados en la sección anterior para el análisis del caso de estudio fue el relacionado con el transporte de maltose. En la figura 6 puede verse el resultado de la red Bayesiana minimal para este cluster. Así mismo, las figuras 7 y 8 muestran el

Figura 4: Caso de uso "Identificar genes de interés biológico"

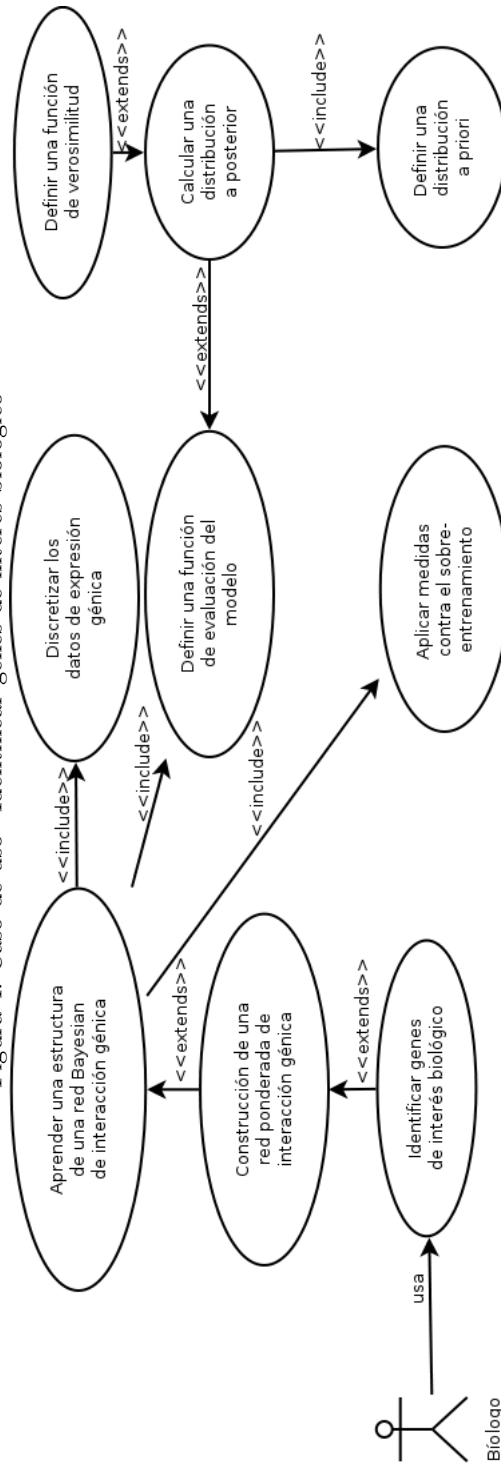
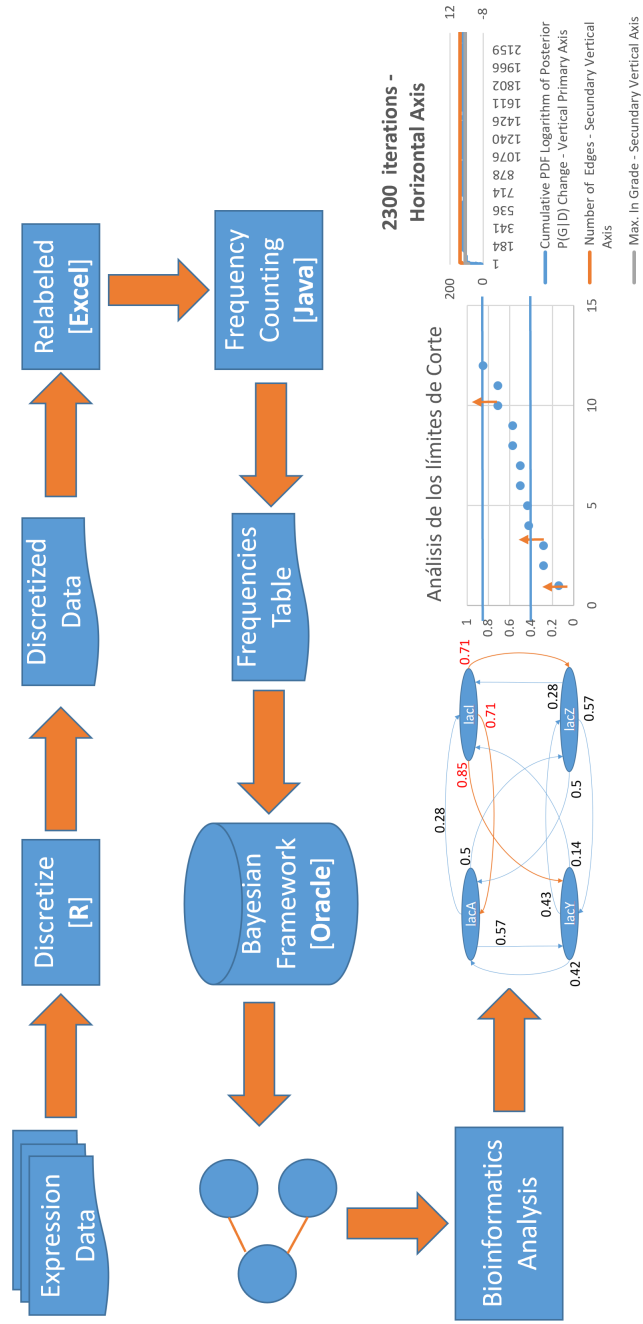


Figura 5: Flujo de trabajo para la experimentación del marco con E.coli



Cuadro 1: Clusters seleccionados de la red de co-expresión génica de E.coli construida en el capítulo 2, para el caso de estudio.

<i>Módulo</i>	<i>Término GO (ID, Descrip.)</i>	<i>G. Anotados</i>	<i>G. Obs.</i>
34 - Darkmagenta	GO:0008643 carbohydrate transport	9	13
	GO:0015768 maltose transport		
	GO:0042956 maltodextrin transport		

Cuadro 2: Genes seleccionados pra el análisis del caso de estudio.

#	<i>Módulo</i>	<i>Id. Gen</i>	<i>Descripción</i>
34	Darkmagenta	lamB	GO:0042956 - Maltodextrin transport
			GO:0015481 - Maltose transporting porin activity
			GO:0015288 - Porin activity
		malE	maltose ABC transporter - periplasmic binding protein
		malF	maltose ABC transporter - membrane subunit
		malG	maltose ABC transporter - membrane subunit
		malK	maltose ABC transporter - ATP binding subunit
		malM	maltose regulon periplasmic protein
		malP	maltodextrin phosphorylase monomer
		malQ	GO:0000025 - Maltose catabolic process
		malT	MalT transcriptional activator

resultado de aplicar el procedimiento de reducción de la red minimal descrito en el capítulo 3 y al reorganizar esta red Bayesiana reducida puerder verse el resultado una red bipartita, excepto por el nodo malP que tiene aristas provenientes de todos los nodos padres (lamB, malT, malQ y malG) y conecta con dos hijos (malK y malE).

Por último, se tomaron dos conjuntos aleatorios de 9 genes a los que se llamó control 1 y control 2 a los cuales se les aplico el mismo flujo de trabajo que al de transporte de maltose, los resultados son presentados en las figuras 9, 10 y 11, respectivamente. Estos controles se realizaron con el fin de comparar los resultados de un cluster co-expresado vs otros clusters aleatorios y observar lo que sucede con los correspondientes hallazgos.

1.3.4. Análisis bioinformático y discusión de los resultados

La primera que puede notarse en el caso del cluster de maltose y sus controles 1 y 2, es con respecto a la red Bayesiana minimal y es que sus estructuras son idénticas y tienen en común que existe un nodo padre que está conectado con los otros 8 nodos hijos, después viene un nodo hijo como padre de los otros 7 nodos, y a su vez un hijo de este es padre de los otros 6 nodos, y así sucesivamente hasta llegar al nodo 9 quien es el nodo hoja de la estructura jerárquica. Esta situación también se evidenció en otros experimentos llevados a cabo con redes

Figura 6: Red Bayesiana minimal del cluster de transporte de maltose.

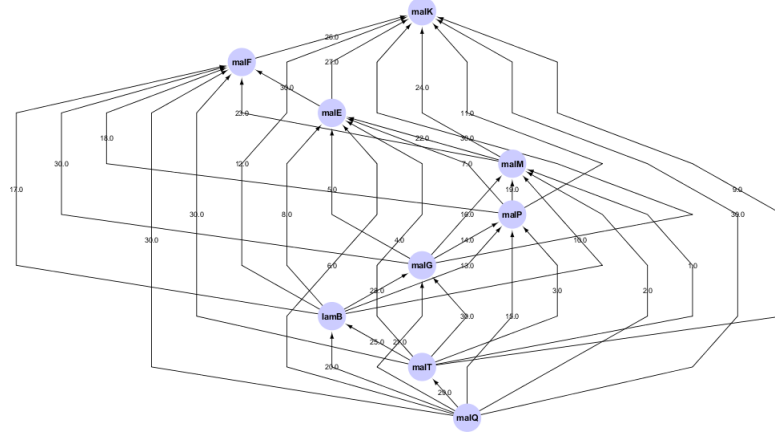


Figura 7: Red Bayesiana minimal reducida del cluster de transporte de maltose.

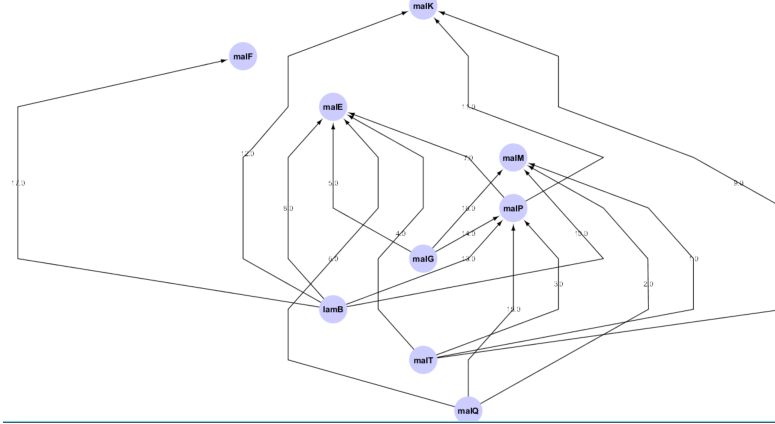


Figura 8: Red Bayesiana bipartida del cluster de transporte de maltose.

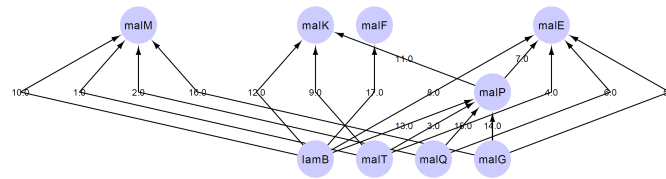


Figura 9: Red Bayesiana minimal del cluster aleatorio de control 1.

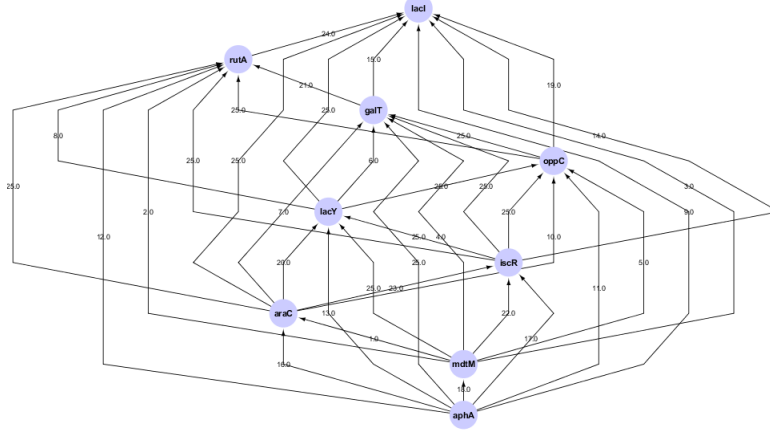


Figura 10: Red Bayesiana reducida del cluster aleatorio de control 1.

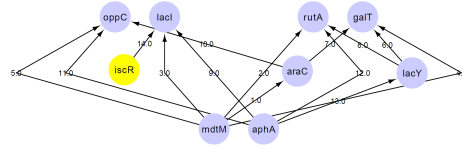
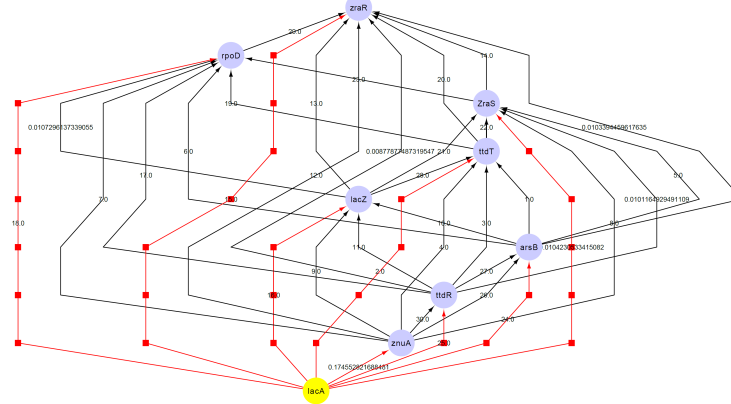


Figura 11: Red Bayesiana minimal del cluster aleatorio de control 2.



de 11,4 y 6 nodos, lo que nos lleva a concluir que las estructuras de redes Bayesianas minimales se pueden predecir dependiendo del número de nodos y que el aprendizaje se remite meramente a establecer el orden en que las variables ó genes en este caso se ajustan a la estructura optima minimal.

Por otra parte desde un contexto más biológico, al analizar la red reducida bipartita del cluster de transporte de maltose podemos ver que en el primer grupo de nodos padres, esta el nodo o gen *lamB* con funciones génicas relacionadas con el transporte de maltodextrin, *malT* que es un gen regulador con función de activación, *malQ* relacionado con el proceso catabólico (conjunto de rutas metabólicas que divide las moleculas en pequeñas unidades que son oxidadas para liberar energía o ser usadas en otras reacciones anabolicas) de maltose y *malG* que esta relacionado con la sub-unidad de membrana de transporte de maltose. Así mismo, es interesante que el nodo que conecta los dos grupos de la red bipartita y corresponde al gen *malP* esta relacionado con una encima para producir maltodextrin, mientras que el segundo grupo conformado por *malK* tiene función de ligamiento de ATP (energía), *malE* relacionado con la proteína de ligamiento del espacio periplasmático (que rodea el citoplasma, parte que rodea a su vez el núcleo de la célula); el nodo *malM* no tiene relación con *malP* y si con una proteína periplasmática del regulon de maltose. En este orden de ideas, esta red Bayesiana minimal reducida del cluster de transporte de maltose, sugiere un orden coherente que podría asociarse a una relación causal entre los genes y las funciones genicas que se acaban de detallar, donde en resumen, los padres tienen funciones reguladoras o de transporte cercano a la membrana, el gen intermedio (*malP*) con enzimas y el grupo de genes hijos con sub-unidades de transporte de maltose o maltodextrin.

Otros análisis similares fueron realizados con otros cluster como el de transporte de lactose y arabinose con 4 nodos (*lacA*,*lacY*,*lac*, *lacI* y *araC*,*araF*,*araG*, *araH*, respectivamente), cluster intermodular (*galE*, *galK*, *galM*, *galT*), maquinaria basal de 6 nodos (*oppA*,*oppB*,*oppC*,*oppD*,*opp*) e hydrogenase de 11 nodos (*hycA*,*hycB*,*hycC*,*hycD*,*hycE*,*hycF*,*hycF*,*hycG*,*hycH*,*hycI*,*hydN*,*fdhF*) llegando siempre a un rden coherente que podría asociarse a una relación causal entre los genes y las funciones genicas analizados en todos los casos.

1.4. Comentarios concluyentes

- Las redes Bayesianas de interacción de genes empleadas para modelar componentes de una red regulatoria bacteriana pueden mostrarnos en sus nodos raíz (por su carácter jerárquico) genes reguladores de interés biológico responsables de la activación o inhibición de un conjunto de genes que en una red de co-expresión génica solo se aprecian co-expresados. Es decir, que una red Bayesiana comparada con la red de co-expresión génica puede explicar (con cierto nivel de incertidumbre, es decir, con una probabilidad) con un análisis causal de la red (por las dependencias modeladas entre sus nodos, soportadas en el concepto de probabilidad condicional) aspectos topológicos de una red de co-expresión como por qué un conjunto de

genes se estan co-expresando.

- Por la naturaleza exploratoria de la caminata aleatoria realizada en la simulación MCMC durante el proceso de aprendizaje de una red Bayesiana, podríamos tener el caso de una red de N nodos, donde un nodo n tiene a los demas $N - 1$ como padres en la topología de esta red. Supongamos que se trata de red con variables binarias, así que tendremos par el nodo n , 2^{N-1} entradas en la tabla de probabilidad condicional asociada al nodo; para efectos del cálculo del criterio de aceptación del algoritmo MH, deberemos evaluar 2^{N-1} funciones de distribución de probabilidad (dirithlet) nos enfrentamos a una complejidad computacional exponencial. esta situación limita a que las redes Bayesianas solo puedan ser utilizadas para analizar un conjunto de genes de un par de decenas a lo más correspondientes a algún componente modular pequeño del sistema regulatorio bacteriano (el cual puede ser del orden de varios miles de genes). Como trabajo futuro se podria implementar en el marco de trabajo técnicas computacionales como la computación paralela a traves de sistemas distribuidos (MPI-Message Processing Interface) para paralelizar la evaluación de las PDF dentro del criterio de evaluación del algoritmo MH, ya que para el caso del modelo multinomial-dirichlet se trata de una sumatoria apta de ser paralelizada.
- Por otra parte, los sistemas regulatorios en general pueden incluir ciclos, dado que un producto del gen como un transcriptoma (RNA) o proteína podría adherirse al citoplasma y posteriormente a una región del ADN que impacte al mismo gen que genero dicho producto (dogma central de la biología). Especialmente en organismos bacterianos este ciclo puede ser mucho mas acelerado que un eukariota. Esta situación limita el modelamiento de componentes de un sistema regulatorio bacteriano al ser este un grafo aciclico. Como trabajo futuro se podría incorporar al marco de trabajo el modelamiento de redes Bayesianas dinamicas para dar tratamiento en este aspecto.
- Como trabajo futuro podría ser interesante explorar la variación del aprendizaje de redes Bayesianas de encontrar un orden de las variables o nodos de estudio partir de los datos asumiendo una estructura óptima predecible dependiendo del número de nodos. Dicha estructura se define de la siguiente manera para un número N de variables o nodos: se tiene un nodo padre que esta conectado con los otros $N - 1$ nodos hijos, después viene un nodo hijo como padre de los otros $N - 2$ nodos, y a su vez un hijo de este , es padre de los otros $N - 3$ nodos, y así sucesivamente hasta llegar al nodo N quien es el nodo hoja de la estructura jerarquica. *Nota:* como parte de este trabajo futuro se podría demostrar que la estructura como se acaba de describir es optima por medio de inducción matemática e igualmente que si se agregará una arista más a dicha estructura ya no sería más un grafo aciclico.

Referencias

- [1] Douglas Arneson, Anindya Bhattacharya, Le Shu, Ville-Petteri Mäkinen, and Xia Yang. Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration. *BMC Genomics*, 17(1):722, Sep 2016.
- [2] Dey, D.K. and Ghosh, S. and Mallick, B.K. *Bayesian Modeling in Bioinformatics*. Chapman & Hall/CRC Biostatistics Series. CRC Press, 2010.
- [3] Byron Ellis and Wing Hung Wong. Learning causal bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103(482):778–789, 2008.
- [4] Roy T. Fielding and Richard N. Taylor. Principled design of the modern web architecture. In *Proceedings of the 22Nd International Conference on Software Engineering*, ICSE '00, pages 407–416, New York, NY, USA, 2000. ACM.
- [5] Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50(1-2):95–125, 2003.
- [6] Miguel A. Ibarra-Arellano, A. I. Campos-G., Luis G. T.-Quintanilla, Andreas Tauch, and Julio A. Freyre-G. Abasy atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database*, 2016:baw089, 2016.
- [7] Faming Liang and Jian Zhang. Learning bayesian networks for discrete data. *Computational Statistics and Data Analysis*, 53(4):865 – 876, 2009.
- [8] Richard E. Neapolitan. *Probabilistic Methods for Bioinformatics: With an Introduction to Bayesian Networks*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2009.
- [9] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.
- [10] E. Prestat, S. R. de Morais, J. A. Vendrell, A. Thollet, C. Gautier, P. A. Cohen, and A. Aussem. Learning the local Bayesian network structure around the ZNF217 oncogene in breast tumours. *Comput. Biol. Med.*, 43(4):334–341, May 2013.
- [11] Nick Rozanski and Eóin Woods. *Software Systems Architecture: Working With Stakeholders Using Viewpoints and Perspectives*. Addison-Wesley Professional, 2005.

- [12] Nguyen Xuan Vinh, Madhu Chetty, Ross Coppel, and Pramod P. Wangikar. Globalmit: learning globally optimal dynamic bayesian network with the mutual information test criterion. *Bioinformatics*, 27(19):2765–2766, 2011.
- [13] Juexin Wang, Trupti Joshi, Babu Valliyodan, Haiying Shi, Yanchun Liang, Henry T. Nguyen, Jing Zhang, and Dong Xu. A bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC Genomics*, 16(1):1011, Nov 2015.