**Marco de trabajo de redes Bayesianas para análisis de conjuntos de genes de interés biológico a partir de datos de expresión generados por micro-arreglos**

**PROPUESTA PROYECTO DE GRADO**

**Diego Mauricio García Jiménez**

**Directora**

**Irene Tischer, Ph.D.**

**Profesora ad-honorem, Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle**

**UNIVERSIDAD DEL VALLE**

**FACULTAD DE INGENIERÍA**

**PROGRAMA DE DOCTORADO EN INGENIERÍA**

**SANTIAGO DE CALI**

**2016**

**Marco de trabajo de redes Bayesianas para análisis de conjuntos de genes de interés biológico a partir de datos de expresión generados por micro-arreglos**

**Diego Mauricio García Jiménez**

**Documento presentado como propuesta de tema de investigación del Doctorado en Ingeniería**

**Directora**

**Irene Tischer, Ph.D.**

**Profesora ad-honorem, Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle**

**UNIVERSIDAD DEL VALLE**

**FACULTAD DE INGENIERÍA**

**ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN**

**PROGRAMA DE DOCTORADO EN INGENIERÍA**

**SANTIAGO DE CALI**

**2016**

**Nota de aceptación**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Firma del Presidente del Jurado

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Firma del Jurado

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Firma del Jurado

Santiago de Cali, Septiembre 23 de 2016

**CONTENIDO**

pág.

[1. Presentación del proyecto 8](#_Toc462407829)

[1.1 Planteamiento de problema 8](#_Toc462407830)

[1.2 Propuesta de investigación 9](#_Toc462407831)

[1.3 Pregunta de investigacióne hipótesis 10](#_Toc462407832)

[1.4 Motivación y justificación del trabajo 11](#_Toc462407833)

[2. Objetivos 13](#_Toc462407834)

[2.1 Objetivo General 13](#_Toc462407835)

[2.2 Objetivos Específicos 13](#_Toc462407836)

[3. Materiales, metodología y cronograma 14](#_Toc462407837)

[3.1 Materiales 14](#_Toc462407838)

[3.2 Metodología 14](#_Toc462407839)

[3.2.1 Construcción y visualización de redes de co-expresión de genes 16](#_Toc462407840)

[3.2.2 Construcción de la arquitectura óptima de una red Bayesiana y sus parámetros, a partir de los datos de expresión de genes 16](#_Toc462407841)

[3.2.2.1 Procedimiento de búsqueda de la arquitectura optima de una red Bayesiana 16](#_Toc462407842)

[3.2.2.2 Metodología de utilización del componente de aprendizaje de una arquitectura de red Bayesiana 16](#_Toc462407843)

[3.2.3 Implementación de las herramientas de análisis de redes complejas y bayesianas de expresión de genes y sus relaciones 17](#_Toc462407844)

[3.2.4 Caso de estudio: Análisis de datos de expresión de genes de micro-arreglos de DNA de transcriptomas 18](#_Toc462407845)

[3.3 Cronograma 18](#_Toc462407846)

[4. Resultados esperados 20](#_Toc462407847)

[5. Marco teórico y estado del arte 21](#_Toc462407848)

[5.1 Ontología de genes (Gene Ontology - GO) 21](#_Toc462407854)

[5.1.1 Introducción 21](#_Toc462407855)

[5.1.2 Estado del arte 22](#_Toc462407856)

[5.2 Datos de expresión de genes en micro-arreglos 24](#_Toc462407857)

[5.2.1 Definición y función 24](#_Toc462407858)

[5.2.2 Introducción al procesamiento de imágenes, transformación y normalización 27](#_Toc462407859)

[5.2.2.1 Procesamiento de imágenes y análisis 27](#_Toc462407860)

[5.2.2.2 Transformaciones de la taza de expresión 27](#_Toc462407861)

[5.2.2.3 La normalización de datos 28](#_Toc462407862)

[5.2.3 Análisis de datos de expresión de genes 29](#_Toc462407863)

[5.2.3.1 Representación de datos de expresión de genes 30](#_Toc462407864)

[5.2.4 Estado del arte 32](#_Toc462407865)

[5.3 Redes complejas 32](#_Toc462407866)

[5.3.1 Introducción a las redes complejas 32](#_Toc462407867)

[5.3.2 Análisis de redes complejas 36](#_Toc462407868)

[5.3.3 Estado del arte 37](#_Toc462407869)

[5.4 Redes de co-expresión de genes 38](#_Toc462407870)

[5.4.1 Introducción a las redes de co-expresión de genes 38](#_Toc462407871)

[5.4.2 Estado del arte 38](#_Toc462407872)

[5.5 Aprendizaje Bayesiano fundamentos y estado de arte 42](#_Toc462407880)

[5.5.1 Definición y representación de una red Bayesiana 42](#_Toc462407881)

[5.5.2 Tipos de razonamiento en redes bayesianas 44](#_Toc462407882)

[5.5.3 Principio de la máxima verosimilitud (maximum likelihood) 45](#_Toc462407883)

[5.5.3.1 Contexto del problema del aprendizaje 45](#_Toc462407884)

[5.5.3.2 La función de verosimilitud (likelihood) 45](#_Toc462407885)

[5.5.3.3 Estimación de máxima verosimilitud (MLE) 46](#_Toc462407886)

[5.5.4 MLE para redes Bayesianas 46](#_Toc462407887)

[5.5.4.1 Un ejemplo simple 46](#_Toc462407888)

[5.5.4.2 Descomposición global de la verosimilitud (likelihood) 48](#_Toc462407889)

[5.5.5 Aprendizaje de la estructura de una red Bayesiana 49](#_Toc462407890)

[5.5.5.1 Definición del problema 49](#_Toc462407891)

[5.5.5.2 Resumen de métodos 49](#_Toc462407892)

[5.5.5.3 Puntajes de estructuras 50](#_Toc462407893)

[5.5.5.4 Búsqueda de estructuras 51](#_Toc462407894)

[5.5.6 Estado del arte 54](#_Toc462407895)

[5.5.6.1 Enfoque basado en restricciones 54](#_Toc462407896)

[5.5.6.2 Enfoque basado en puntaje 55](#_Toc462407897)

[5.5.6.3 Modelo promedio Bayesiano 55](#_Toc462407898)

[5.6 Markov Chain Monte Carlo (MCMC) 57](#_Toc462407899)

[5.6.1 Introducción 57](#_Toc462407901)

[5.6.2 Algoritmos 57](#_Toc462407902)

[5.7 Relaciones entre las redes de co-expresión (CGN) y las redes Bayesianas de interacción de genes (BIN) 58](#_Toc462407903)

[6. Bibliografía 60](#_Toc462407904)

[7. Anexo 64](#_Toc462407905)

**FIGURAS**

pág.

[Figura1 Entender como los rasgos moleculares de diferentes dominios biologicos se conectan en redes es critico para el progreso de la biología del cancer 11](#_Toc462413159)

[Figura 2 Diagrama de flujo de los pasos metodológicos 14](#_Toc462413160)

[Figura 3 El explorador de ontología de genes AmiGO 22](#_Toc462413161)

[Figura 4 Un micro-arreglo puede contener miles de "spots". Cada spot contiene muchas copias de la misma secuencia de ADN que únicamente representa un gen de un organismo 24](#_Toc462413162)

[Figura 5 Esquema del protocolo experimental para estudiar expresión diferencial de genes 25](#_Toc462413163)

[Figura 6 Datos de expresión de genes antes y después del procedimiento de normalización 28](#_Toc462413164)

[Figura 7 Red o grafo 32](#_Toc462413165)

[Figura 8 Red no dirigida 33](#_Toc462413166)

[Figura 9 Red dirigida 33](#_Toc462413167)

[Figura 10 Redes dirigidas y ponderadas 33](#_Toc462413168)

[Figura 11 Estructura de una red 34](#_Toc462413169)

[Figura 12 Propiedad del mundo pequeño(Watts & Strogatz, 1998) 36](#_Toc462413170)

[Figura 13 Red de co-expresión con 7221 genes para 18 pacientes de cáncer gástrico 38](#_Toc462413171)

[Figura 14 Patrón de regulación negativo o invertido 39](#_Toc462413172)

[Figura15 Mezcla de patrones de regulación 40](#_Toc462413173)

[Figura 16 Una red bayesiana para el problema de cáncer de pulmón 42](#_Toc462413174)

[Figura 17 Ejemplo de una búsqueda que requiere eliminación de arista 51](#_Toc462413175)

[Figura 18 Ejemplo de búsqueda que requiere invertir 51](#_Toc462413176)

[Figura 19 Red de co-expresión de genes (CGN) vs. red bayesiana de interacción de genes (BIN) 57](#_Toc462413177)

**TABLAS**

pág.

[Tabla 1 Descripción corta de los conjuntos de datos 13](#_Toc462413190)

[Tabla 2 Cronograma general del proyecto 18](#_Toc462413191)

[Tabla 3 Matriz de expresión de genes que contiene filas para representar genes y columnas para representar condiciones biológicas(M. Babu, 2004), 29](#_Toc462413192)

[Tabla 4 Datos de expresión de genes por medida relativa(M. Babu, 2004) 30](#_Toc462413193)

[Tabla 5 Datos de expresión por(M. Babu, 2004) 30](#_Toc462413194)

[Tabla 6 Datos expresión por valores discretos(M. Babu, 2004) 30](#_Toc462413195)

[Tabla 7 Selección preliminar de nodos y valores para el ejemplo de cáncer de pulmónRef. (Korb & Nicholson, 2010) 41](#_Toc462413196)

**ECUACIONES**

pág.

[Ecuación 1 Tasa de expresión de un gen dado 27](#_Toc462413526)

[Ecuación 2 Distribución conjunta de las variables aleatorias *v* 43](#_Toc462413527)

[Ecuación 3 Función de verosimilitud local 47](#_Toc462413528)

[Ecuación 4 Delta del puntaje 53](#_Toc462413529)

# Presentación del proyecto

## Planteamiento de problema

Los experimentos de **micro-arreglos** de DNA miden simultáneamente los niveles de expresión de miles de genes. **Un problema importante en biología molecular es a partir de las medidas de niveles de expresión de Micro-arreglos**, **descubrir** las **redes de interacción** y las **características biológicas claves** de los **sistemas celulares** (Faming, 2010). Uno de los **enfoques** más usados para este problema es **aprender una red bayesiana desde los datos de expresión** (Pearl, 1988) y (Cowell, 1998).

La propuesta para abordar este problema es desarrollar un marco de trabajo de redes Bayesianas para análisis de conjuntos de genes de interés biológico a partir de datos de expresión generados por micro-arreglos.

Los **micro-arreglos** de DNA han sido la herramienta más utilizada para explorar transcriptomas (conjunto de todas las moléculas de *mRNA* de una célula o población de células). En cáncer de mama, por ejemplo, esta tecnología tuvo éxito mejorando el sistema de clasificación, para proponer nuevos genes importantes en proceso de oncogénesis. Sin embargo, aún **hay margen para mejorar el diagnóstico y la predicción de la enfermedad** (Prestat et al., 2013).

En el **descubrimiento de redes de interacción** a través de la teoría de redes y análisis de redes complejas (**Network sciences**), se han podido modelar eventos complejos de biología molecular con el objetivo de abordar su estudio desde una perspectiva sistémica observando relaciones e interacciones en conjunto en lugar de ver elementos de manera aislada. Usando puntos y líneas se construyen redes que recrean matemáticamente distintas relaciones entre biomoléculas (proteínas o genes), como por ejemplo, interacciones entre proteínas o mecanismos regulatorios y de expresión; de manera que cada nodo corresponde a un gen o una proteína y las aristas representan algún tipo de interacción entre ellos.

En el marco de trabajo de redes Bayesianas que se propone, nos interesa inicialmente, construir un tipo de red de interacción de genes, llamado **redes de co-expresión de Genes** (**GCN**). Las 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. De esta manera, dos genes se conectan si sus perfiles de expresión están asociados entre las perturbaciones estudiadas.

Una vez se tiene la **GCN**,el paso a seguir es construir otro tipo de red de interacción; una red de regulación de genes, por medio de una red Bayesiana. Las **redes de Bayes** son una poderosa herramienta de representación de conocimiento y razonamiento bajo condiciones de incertidumbre que son típicas en aplicaciones del mundo real. Una red bayesiana es un grafo acíclico dirigido, en el cual los nodos representan las variables de un dominio, como, por ejemplo, en una red bayesiana de interacción de genes, los nodos representan genes. Las aristas corresponden a la dirección de dependencias probabilísticas entre ellos, y siguiendo con el ejemplo, las aristas representan relaciones de interacción (co-expresion y/o regulación) entre genes.

En este marco de trabajo Bayesiano, se articulará el análisis de redes complejas con el aprendizaje basado en la simulación de Markov Chain Monte Carlo (**MCMC***)*, para aprender una red Bayesiana de interacción de genes (**BIN**). Muchos enfoques han sido desarrollados para el aprendizaje de redes Bayesianas. Uno de estos enfoques está basado en la **simulación** de **MCMC**. Aunque este enfoque resulta interesante, trabaja bien solo para problemas con un número muy pequeño de variables (Faming, 2010), sin embargo, definiendo una **cadena de Markov** sobre el espacio de posibles estructuras de la red bayesiana*,* es posible generar un **conjunto de estructuras viables probabilísticamente (es decir que la estructura y los parámetros de la red se ajustan a los datos) para el grafo**, haciendo una caminata aleatoria en la cadena de Markov asociada (Koller & Friedman, 2009a).

Finalmente, para darle **coherencia (es decir, que los resultados tengan sentido desde la biología) y significancia (que los resultados sean pertinentes) biológica a los hallazgos del marco Bayesiano propuesto**, con base en la **BIN** aprendida y los **perfiles de expresión de genes** analizados en la **GCN**, se **vinculará información externa** para comprender mejor los procesos biológicos, y poder realizar nuevos descubrimientos (M. Babu, 2004). Gene ontology (**GO**) es la principal iniciativa bioinformática para unificar la representación del gen y los atributos de los productos del gen, en todas las especies (The Gene Ontology, 2008). Las **GO** comprenden diversas bases de datos de anotaciones (por ejemplo, **GeneMANIA** (Warde-Farley et al., 2010)) que permiten caracterizar a nivel funcional los genes, y con esta caracterización se asocian posibles funcionalidades para un conjunto de genes dado, y obtener información de interés biológico para los mismos.

## Propuesta de investigación

En este trabajo, se propone construir en principio una **red de co-expresión de genes** (**GCN**), a partir de datos de expresión de micro-arreglos. Después, ala red de co-expresión de genes se aplicará **análisis de redes complejas** para detectar *módulos* o agrupamientos (*clusters*) de genes que se co-expresan similarmente, para construir entonces una red bayesiana de interacción de genes por simulación de Markov Chain Monte Carlo. Posteriormente, se identificaran conjuntos de genes de interés biológico (**signature gene**); analizando las relaciones entre ambas redes y su importancia biológica a través bases de datos de anotación de ontologías de genes (**GO**). Por último, se someterá el marco de trabajo Bayesiano a experimentación a través de un caso de estudio para explorar diversos transcriptomas de personas (con enfermedades como cáncer, VIH o autismo), algunos animales (como ratas, monos o chimpancés) o algunos cultivos (como levadura, arroz o caña). Es decir, se propone desarrollar un **marco de trabajo bayesiano para análisis de conjuntos de genes de interés biológico a partir de datos de expresión generados por micro-arreglos**.

En el marco mención, después de construir las redes de co-expresión y redes bayesianas de interacción de genes, y de analizar estas redes y sus relaciones, se pretende **inferir en los perfiles de expresión**, **patrones** que **complementados** con conocimiento biológico de las **GO**, aporten en el **descubrimiento de funciones de genes** o **interacciones de genes de importancia biológica**.

Para la construcción de redes de co-expresión se propone emplear el Análisis de Redes Ponderadas de Co-expresión de Genes (**WGCNA** del inglés Weighted Gene Coexpression Network Analysis) como un método para identificar módulos y sus relaciones con otros módulos, con fenotipos externos e identificar genes *hub,* es decir, altamente conectados con módulos (Langfelder & Horvath, 2008). Este método será usado para identificar módulos de co-expresión.

Para la construcción de la red bayesiana de interacción de genes se hará por simulación *MCMC*, usado el algoritmo Metropolis Hasting (***MH****)*. Una ventaja estos algoritmos es la habilidad para muestrear a partir de distribuciones probabilísticas de alta dimensionalidad.

Después, se llevará a cabo el análisis de ambas redes y sus relaciones, aplicando análisis de redes complejas, y además, se incorporarán anotaciones de GO, con el objetivo de descubrir en los perfiles de expresión, funciones y/o interacciones de genes de interés biológico.

Finalmente, se validará este marco de trabajo a través de un caso de estudio (ver numeral 3.2.4).

## Pregunta de investigación e hipótesis

Que pueden las redes de co-expresión y las redes bayesianas de interacción de genes construidas a partir de datos de expresión generados por micro-arreglos, decirnos acerca de genes de interés biológico y las características biológicas claves en los sistemas celulares ?

**La hipótesis** de este trabajo es que **existe una relación** entre **(1)** las **redes** de **co-expresión** y las **redes** bayesiana de **interacción** de genes construidas a partir de datos de expresión generados por micro-arreglos, con respecto al **(2) proceso biológico** que dio origen a los datos de expresión de genes generados por micro-arreglos.

Usando datos reales y simulados es posible capturar **relaciones** entre las variables del **modelo** (nodos de la red) y los **procesos biológicos subyacentes**, entendiendo como procesos biológicos subyacentes, por ejemplo a una variedad de condiciones y topologías de rutas metabólicas; en una ruta bioquímica las reacciones son organizadas en rutas las cuales pueden converger, ramificarse o intersectarse para formar redes complejas(Blair, Kliebenstein, & Churchill, 2012).

## Motivación y justificación del trabajo

“Todos los modelos son incorrectos, pero algunos son útiles”(Box & Draper, 1987).

La motivación principal del trabajo es la de contribuir al entendimiento de algunas características claves de los sistemas celulares y como diferentes dominios biológicos se conectan y dan origen a un fenotipo o enfermedad.

En biología del cáncer, la integración de los datos es de vital importancia debido a la compleja interacción entre la genética, la señalización celular y las rutas metabólicas (ver **Figura 1**Entender como los rasgos moleculares de diferentes dominios biologicos se conectan en redes es critico para el progreso de la biología del cancer**)**(Blair, Trichler, & Gaille, 2012).

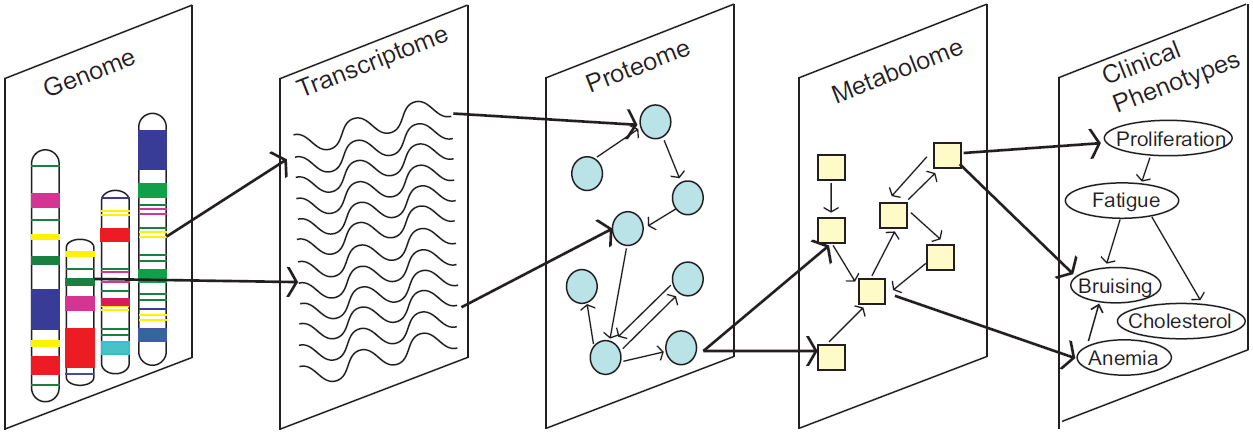


Figura1 Entender como los rasgos moleculares de diferentes dominios biologicos se conectan en redes es critico para el progreso de la biología del cancer

El descubrimiento de redes de interacción desde datos de expresión de genes y las características biológicas claves de los sistemas celulares es en general un problema abierto(Faming, 2010).

El desarrollo de modelos matemáticos que sean consistentes y predictivos de los reales mecanismos biológicos subyacentes es un objetivo central de la biología de sistemas, ya que los modelos en mención pueden dirigir el diseño y ejecución de experimentos y estudios poco probables de éxito, y de paso ahorrar un poco de tiempo y dinero. La evaluación y mejora de la utilidad de tales modelos continuará siendo un área activa de investigación(Blair, Trichler, et al., 2012).

En ese sentido la justificación del trabajo se apoya en contribuir computacionalmente en la solución de problemas abiertos actuales en biología molecular y biología de sistemas.

# Objetivos

## Objetivo General

Desarrollar un marco de trabajo de redes Bayesianas para análisis de conjuntos de genes de interés biológico a partir de datos de expresión generados por micro-arreglos.

## Objetivos Específicos

1. Revisar y adaptar computacionalmente la construcción y visualización de redes de co-expresión de genes.
2. Desarrollar la construcción automática de la arquitectura óptima de una red bayesiana y sus parámetros a partir de los datos de expresión de genes.
3. Conceptualizar, diseñar e implementar las herramientas de análisis de redes complejas y bayesianas de expresión de genes y sus relaciones.
4. Someter 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.

# Materiales, metodología y cronograma

## Materiales

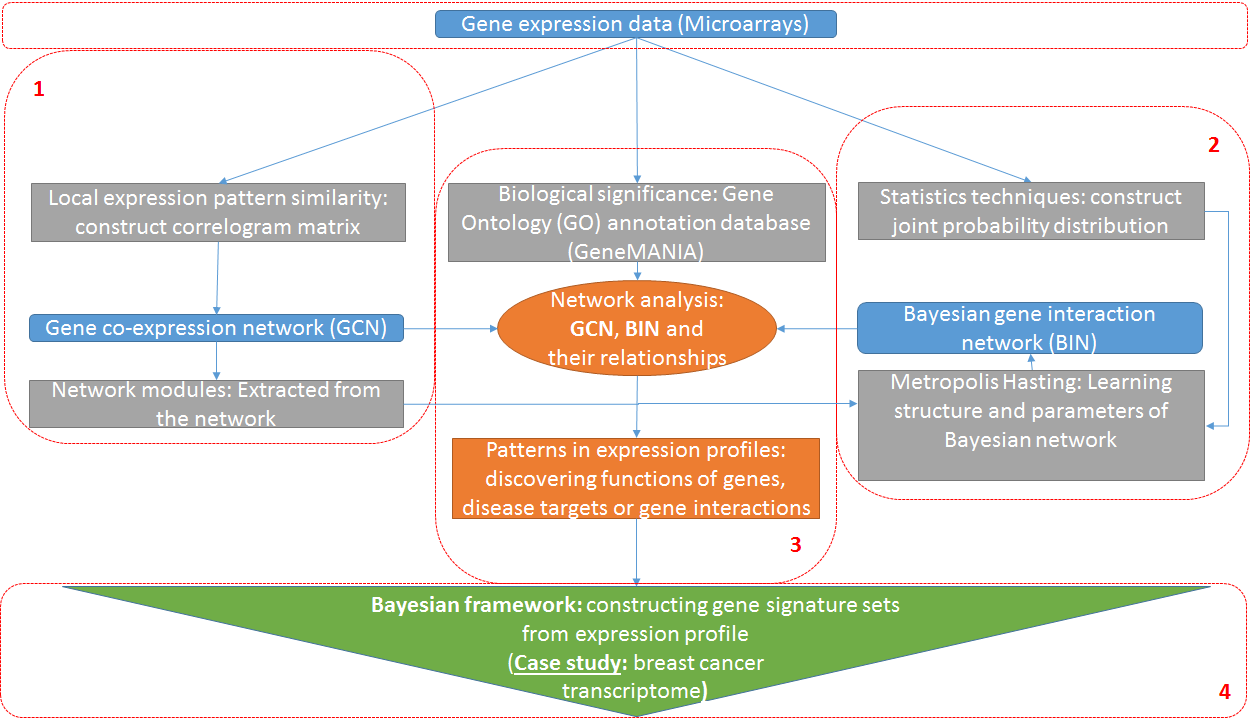
Existen diversas fuentes de conjuntos de datos de expresión de genes (gene expression data) de micro-arreglos; el centro nacional para información y biotecnología (**NCBI** - The National Center for Biotechnology Information advances science and health) provee acceso a información genómica y biomédica.

A continuación, en la **Tabla 1**, se muestran algunas características de diferentes conjuntos de datos de expresión y sus fuentes, los cuales serán usados para apoyar la **experimentación del marco de trabajo bayesiano**.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Organism** | **Dataset** | **No. of genes** | | **No. of samples** | **Source** | |
| Yeast Sporulation | Yeast | | 474 | 7 | http://cmgm.stanford.edu/pbrown/sporulation webcite | |
| Yeast | Yeast KY | | 237 | 18 | http://faculty.washington.edu/kayee/cluster/ webcite | |
| Yeast | Yeast cell cycle | | 384 | 18 | http://faculty.washington.edu/kayee/cluster webcite | |
| Human | GDS825 | | 277 | 8 | NCBI |
| Mouse | GDS958 (Subset) | | 4000 | 12 | NCBI |
| Rat | GDS3702 (Subset) | | 3000 | 12 | NCBI |
| Rice | Thaliana | | 517 | 13 | http://homes.esat.kuleuven.be/~sistawww/bioi/thijs/Work/Clustering.html webcite |
| **Ref.** *Roy et al. BMC Bioinformatics* 2014 15(Suppl 7):S10 doi:10.1186/1471-2105-15-S7-S10.  **Tabla 1** Descripción corta de los conjuntos de datos | | | | | |

## Metodología

Se sugieren cuatro pasos metodológicos para el desarrollo del proyecto. A continuación, se muestra el respectivo diagrama de flujo en la **Figura 2**, y seguidamente se detalla cada paso**:**



**Figura 2** Diagrama de flujo de los pasos metodológicos

### Construcción y visualización de redes de co-expresión de genes

Como primer paso se construirá una red de co-expresión de Genes (**GCN**), a partir de los datos de expresión de genes generados por micro-arreglos (tomados de los conjuntos de datos descritos en la **Tabla 1**), para mayor detalle de este proceso **ver** numeral **5.4**.

### 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 realizará el desarrollo de los métodos de aprendizaje de la arquitectura óptima de una red Bayesiana y sus parámetros, a partir de un conjunto de datos de entrenamiento, y dicho desarrollo se empleará para el aprendizaje Bayesiano, a partir de datos de expresión de genes.

En este paso metodológico se analizaran los módulos (o clusters) obtenidos a partir de la red de co-expresión (CGN) obtenida anteriormente, para filtrar las estructuras optimas de la red bayesiana de interacción (BIN) que se ajustan a los datos, y por medio de simulación de Markov Chain Monte Carlo; definiendo una cadena de Markov sobre el espacio de posibles estructuras de la red bayesiana*,* se generará un conjunto de estructuras viables probabilísticamente para el grafo, haciendo una caminata aleatoria en la cadena de Markov asociada (Koller & Friedman, 2009a) para más detalle **ver** numeral **5.6**.

#### Procedimiento de búsqueda de la arquitectura optima de una red Bayesiana

##### Calculo del espacio de búsqueda

El espacio de búsqueda se determinará en cada iteración de la simulación MCMC (ver numeral 5.6), a través de operaciones de transformación de la estructura (por ejemplo: agregar una arista entre dos nodos, eliminar una arista del grafo o invertir la dirección de una arista del grafo) ver numeral 5.5.5.4.1

##### Adopción del algoritmo Metropolis Hasting

Para nuestro procedimiento de búsqueda de la estructura de la red Bayesiana a partir de los datos adoptaremos en el algoritmo *Metropolis Hasting* (Hastings, 1970),correspondiente a la técnica de simulación MCMC (ver numeral 5.6)

#### Metodología de utilización del componente de aprendizaje de una arquitectura de red Bayesiana

La metodología de uso del aplicativo que se desarrollará para la construcción de la arquitectura óptima de una red bayesiana a partir los datos, se compondrá de las siguientes tareas:

* *Ejecutar la optimización por puntaje (según función de verosimilitud):* se realizará la ejecución del aplicativo con un número elevado (100.000 por ejemplo) de simulaciones, sin guardar información sobre la exploración del espacio de búsqueda, a excepción del **último modelo encontrado**. El **objetivo** aquí será calcular todos los puntajes para poder **analizar el comportamiento de convergencia de la simulación**.
* *Determinar el puntaje promedio en el nivel óptimo y su desviación*: El **objetivo** de este paso será solo **determinar los parámetros claves del comportamiento de convergencia**.
* *Ejecutar un número reducido de simulaciones,* ***iniciando con el modelo óptimo*** *encontrado en el primer paso*, con base en los resultados obtenidos en los pasos anteriores, se **reducirá el número de iteraciones** de la simulación, cambiando el criterio de aceptación (en el algoritmo de Hasting): se aceptará cualquier solución con puntaje mayor o igual a: . Se guardarán las **matrices de adyacencia** de los distintos mejores puntajes; y se repetirá hasta que no haya nuevos candidatos (por ejemplo, 100 intentos seguidos sin que salga una nueva matriz de adyacencia). El **objetivo** será **obtener los posibles grafos del nivel óptimo**.

En el siguiente paso metodológico **Análisis de redes**, se analizaran los grafos obtenidos del nivel óptimo, con base en el conjunto de matrices de adyacencia distintos, y entonces se filtrarán los grafos sobre-entrenados (overfitting, es decir, grafos con muchas aristas o estructuras muy complejas) empleando análisis de redes complejas (ver numeral 5.3.2).

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

Con la construcciónde las redes de co-expresión (**CGN**) y las redes bayesianas de interacción de genes (**BIN**), el paso a seguir es la selección de genes de interés biológico usando las redes mencionadaspara realizar análisis de redes complejas sobre los módulos o agrupamientos (*clusters*) de genes en dichas redes, y emplear bases de datos de anotaciones de ontologías de los genes (**GO**) como GeneMANIA(Warde-Farley et al., 2010).

Esto consistirá en identificar genes relevantes dentro de procesos biológicos de interés mediante la caracterización de los módulos a nivel funcional y de su forma (topología). Los módulos se anotarán funcionalmente y se determinarán las categorías biológicas más relevantes mediante análisis de rutas (*ORA – Over Representation Analysis*) con el fin de encontrar los más llamativos relacionados con los procesos de interés.

Posteriormente cada módulo será caracterizado topológicamente midiendo propiedades que indican la relevancia biológica de los genes al interior de ellos (Ver numerales **5.3** y **5.6**).

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

En el último paso, se someterá 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 serán los siguientes:

* **Selección del problema**: Se seleccionará para la experimentación del marco de trabajo bayesiano la exploración de transcriptomas de cáncer. Los micro-arreglos de DNA han sido la herramienta más utilizada para explorar transcriptomas de personas (con enfermedades como cáncer, VIH o autismo), algunos animales (como ratas, monos o chimpancés) o algunos cultivos (como levadura, arroz o caña).Por ejemplo, en cáncer de mama, esta tecnología tuvo éxito mejorando el sistema de clasificación, para proponer nuevos genes importantes en proceso de oncogénesis(Prestat et al., 2013).
* **Selección de los datos**: Se seleccionará el análisis de un conjunto de datos de expresión de genes de acuerdo a la problemática de aplicación escogida:

<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1378>

breast cáncer / tamoxifen monotherapy (microdissected tummos biopsies); es un conjunto de datos (dataset) de 60 pacientes con cáncer de mama (ER-positive primary) y tratados con tamoxifen (monotherapy) por 5 años (Data weregeneratedfromLCMed cáncer cells).

* **Aplicación del marco de trabajo bayesiano**: Una vez elaborado el marco de trabajo bayesiano se aplicará el conjunto de datos seleccionado.
* **Verificación de los resultados**: Finalmente se hará el análisis y discusión de los resultados obtenidos.

## Cronograma

El proyecto se plantea para realizar en dos años, desglosado por componentes semestrales que podrán cuantificar el avance y contribuir al desarrollo global del proyecto. La siguiente tabla muestra el cronograma general para el proyecto.



**Tabla 2** Cronograma general del proyecto

# Resultados esperados

El desarrollo del presente trabajo de grado generará los siguientes entregables:

* Interfaces de programas de aplicación (API) para la adaptación computacional de la construcción y visualización de redes de co-expresión de genes.
* Aplicación computacional de la construcción automática de la arquitectura óptima de una red bayesiana y sus parámetros, a partir de los datos de expresión de genes y los módulos de la red de co-expresión de genes obtenida por el entregable anterior.
* Herramientas computacionales de análisis de redes complejas y bayesianas de expresión de genes y sus relaciones.
* Documentación de la experimentación del marco de trabajo bayesiano a través del caso de estudio.

# Marco teórico y estado del arte

Como preparación para el desarrollo del proyecto se ha elaborado un cubrimiento de ontologías de los genes (GO), datos de expresión de genes de micro-arreglos, redes de co-expresión, redes bayesianas para datos de expresión de genes y sus relaciones. Se busca con este marco teórico y estado del arte, preparar las bases para caracterizar los procesos de la construcción de una red de co-expresión y redes bayesianas de interacción de genes desde datos de expresión de genes generados por micro-arreglos desde lo que se encuentra en la literatura científica, y que serán aplicados en la exploración de transcriptomas.

Ya que no hace parte del alcance de este trabajo revisar la biología del gen, se recomienda **ver**(Watson et al., 2013), (Nguyen, Bulak Arpat, Wang, & Carroll, 2002) y (Tözeren & Byers, 2004) para mayor información.



## Ontología de genes (Gene Ontology - GO)

### Introducción

Los perfiles de expresión de genes pueden vincularse con información externa para comprender mejor los procesos biológicos y poder realizar nuevos descubrimientos (M. Babu, 2004). Gene ontology (**GO**) es la principal iniciativa bioinformática para unificar la representación del gen y los atributos de los productos del gen, en todas las especies (The Gene Ontology, 2008).

Los objetivos principales de GO son:

1. Mantener y desarrollar su vocabulario controlado del gen los atributos de los productos del gen (RNA - proteínas).
2. Anotar los genes y los productos del gen, y asimilar y diseminar sobre los datos anotados.
3. Proveer herramientas para fácil acceso a todos los aspectos de los datos provistos por GO, y permitir la interpretación funcional de datos experimentales usando GO, por ejemplo, mediante el análisis de enriquecimiento (enrichment analysis)

Aunque la nomenclatura del gen en sí misma pretende mantener y desarrollar vocabulario controlado del gen y sus productos, GO extiende el esfuerzo usando **lenguaje etiquetaado** para que los datos sean legibles por una máquina, y hacerlo de manera unificada para todas las especies.

Pragmáticamente, una ontología es una representación de algo, acerca de lo cual, tenemos interés en saber. Así, la ontología representa algo que es detectable o directamente observable y sus relaciones.

La ontología cubre tres dominios:

1. **Componente celular**: Partes de la célula o su ambiente extra-celular.
2. **Función molecular:** Actividades elementales de un producto de un gen a nivel molecular, por ejemplo, la catálisis.
3. **Proceso biológico**: Operaciones o conjunto de eventos moleculares con un comienzo y fin definido, pertinente para el funcionamiento de unidades vivas integradas: células, tejidos, órganos y organismos.

Cada término GO en la ontología tiene un **nombre**, el cual puede ser una **palabra** o varias palabras, un **identificador** alfanumérico único, una **definición** con fuentes citadas y un **espacio de nombres** que indica el dominio al cual pertenece.Un término también puede tener sinónimos.

La ontología GO está estructurada como un **grafo a-cíclico dirigido**, y **cada término** tiene **relaciones** con **uno o más términos** en el mismo dominio, y algunas veces de otros dominios. El vocabulario GO está diseñado para ser **neutral-especies** e incluir términos aplicables a eucariotas, procariotas y organismos unicelulares y multicelulares.

Los **archivos de ontologías GO** están disponibles libremente, por ejemplo, un término podría verse así(Consortium, 2009b):

**id**: GO:0000016

**name**: lactase activity

**namespace**: molecular\_function

**def**: "Catalysis of the reaction: lactose + H2O = D-glucose + D-galactose." [EC:3.2.1.108]

**synonym**: "lactase-phlorizin hydrolase activity" BROAD [EC:3.2.1.108]

**synonym**: "lactose galactohydrolase activity" EXACT [EC:3.2.1.108]

**xref**: EC:3.2.1.108

**xref**: MetaCyc:LACTASE-RXN

**xref**: Reactome:20536

**is\_a**: GO:0004553 ! hydrolase activity, hydrolyzing O-glycosyl compounds

La **anotación del genoma** es la práctica de capturar datos sobre el producto de un gen, y las **anotaciones de GO** emplean términos de la **ontología de GO** para hacer esto.

Los miembros del consorcio GO envían sus anotaciones para la integración y diseminación en el *web site GO*, donde se pueden acceder en línea. Además del identificador del producto del gen y el término relevante de GO, las anotaciones GO tienen los siguientes datos: La **referencia** usada para la anotación, la fecha y el creador de la anotación.

El siguiente es un ejemplo de anotación tomado de(Consortium, 2009a)

**Gene product:** Actin, alpha cardiac muscle 1, UniProtKB:P68032

**GO term:** heart contraction ; GO:0060047 (biological process)

**Evidence code:** Inferred from Mutant Phenotype (IMP)

**Reference:** PMID 17611253

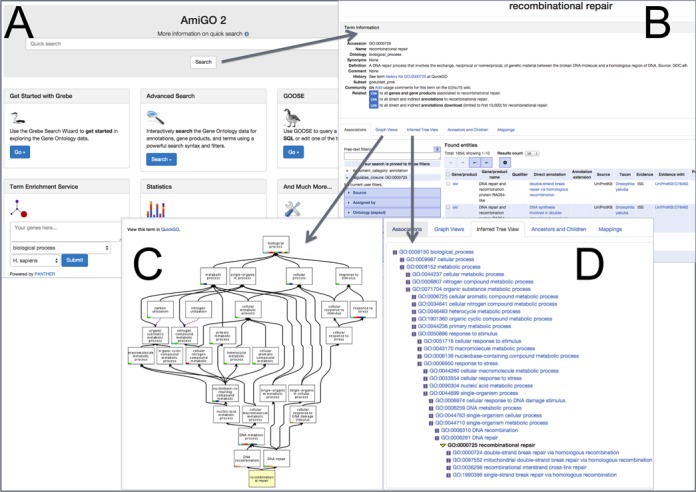
**Assigned by:**UniProtKB, June 6, 2008

### Estado del arte

Recientemente, muchos algoritmos de aprendizaje de máquinas (*machine learning*) han sido diseñados e implementados para predecir anotaciones GO(Pinoli, Chicco, & Masseroli, 2015).

Existe un gran número de herramientas disponibles en línea y para descargar que usan los datos proporcionados por GO, siendo la gran mayoría proporcionada por terceros. El consorcio GO desarrolló y soporta dos herramientas:

1. **AmiGO**: Es una aplicación basada en web que permite a los usuarios consultar, explorar y visualizar ontologías y datos de anotación de producto de genes. Adicionalmente, tiene una herramienta **BLAST**, que permite análisis de conjuntos de grandes conjuntos de datos y una interfaz para consultar directamente la base de datos GO.



Ref. Nucleic Acids Res. 2015 January 28;43(Database issue):D1049-D1056.

Figura 3 El explorador de ontología de genes AmiGO

A – Entrada a la página del portal.

B – Muestra de salida de una búsqueda rápida de un término GO.

C – Despliegue gráfico.

D – Vista de árbol del término GO y su lugar en la ontología.

AmiGo también se puede descargar e instalar para uso local, en cualquier base de datos empleando el esquema de base de datos GO. La herramienta es de código abierto (*open source*) y está disponible como parte de la distribución de software **go-dev**(Consortium, 2015).

1. **EBO-Edit**: Es un editor de ontologías de código abierto independiente de la plataforma, desarrollado y mantenido por el consorcio GO. Esta implementado en Java, y usa un **enfoque orientado a grafos** para desplegar y editar ontologías. Incluye búsqueda exhaustiva e interfaz de filtros con la opción de procesar subconjuntos de términos para hacer una visualización distinta. La herramienta también proporciona, un analizador (*reasoner*) que puede inferir enlaces que no se han definido explícitamente, basándose en las relaciones existentes y sus propiedades. A pesar de que fue desarrollada para ontologías biomédicas, está puede usarse para ver, buscar y editar cualquier ontología. Se encuentra disponible para descargar en(Consortium, 2014).

## Datos de expresión de genes en micro-arreglos

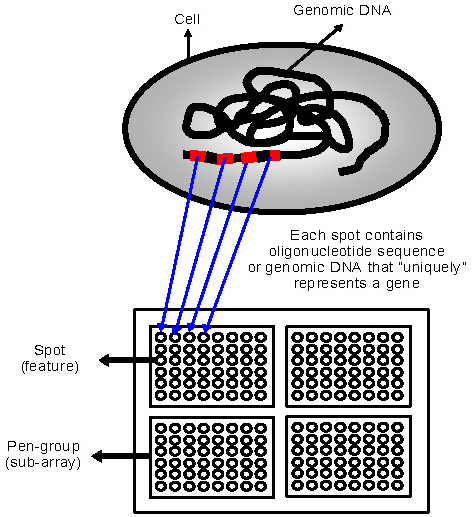
A continuación, se presenta un breve resumen de los conceptos básicos sobre experimentos de micro-arreglos; dando una percepción de lo que los datos representan realmente, y en el numeral 5.4, se proporciona información de varios métodos computacionales que se pueden emplear para derivar resultados significativos de tales experimentos.

### Definición y función

La tecnología de micro-arreglos se ha convertido en una herramienta indispensable que muchos biólogos usan para **monitorear amplios niveles de expresión de genes** del genoma en un organismo dado.

Un micro-arreglo es frecuentemente una lámina o platina de vidrio (*glass slide*) en la cual se fijan de manera ordenada, moléculas de DNA a una localización específica llamada *spots* (o *features*).

Un micro-arreglo puede contener **miles de *spots*** y **cada *spot****,* puede contener unos pocos millones de copias de moléculas idénticas de DNA, que corresponden únicamente a un gen (ver **Figura 4 Un micro-arreglo** puede contener miles de "spots". Cada spot contiene muchas copias de la misma secuencia de ADN que únicamente representa un gen de un organismo).

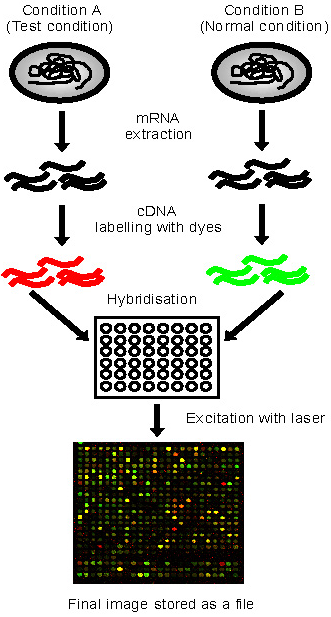


**Ref.**http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/

Figura 4 Un micro-arreglo puede contener miles de "spots". Cada spot contiene muchas copias de la misma secuencia de ADN que únicamente representa un gen de un organismo

El *DNA* en un *spot* puede ser genómico (*gDNA*) o un tramo corto de hebras oligonucleótido que corresponden a un gen.

Los micro-arreglos pueden ser usados para medir expresión de genes de muchas maneras, pero una de las aplicaciones más populares es para **comparar expresión de un conjunto de genes** de una célula en una condición particular A (***condition A***) con respecto al mismo conjunto de genes de la célula referencia bajo condiciones normales (***condition B***)(ver Figura 5).



**Ref.**http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/

Figura 5 Esquema del protocolo experimental para estudiar expresión diferencial de genes

Primero el *RNA* es extraído de las células. Paso a seguir, las moléculas de *RNA* se transcriben de forma inversa en *cDNA* usando una enzima transcriptasa inversa y los nucleótidos son marcados con tintas fluorescentes diferentes. Por ejemplo, el *cDNA* del cultivo de células en *condición A* puede ser etiquetado con tinta **roja**, y el cultivo de células en *condición B* con tinta **verde**.

Los *spots* son pintados en la platina de vidrio (*glass slide*) por un robot o sintetizados por el proceso de fotolitografía.

Una vez las muestras han sido marcadas, se les permite hibridar en la misma lámina de vidrio (*glassslide*). En este punto, cualquier secuencia cDNAen la muestra hibridará con *spots* específicos en la lámina de vidrio que contengan su **secuencia complementaria.**

Como premisa se tiene que la **cantidad** de *cDNA* ligado a *spots* será directamente proporcional al **número inicial** de moléculas *RNA* presentes para el gen en **ambas muestras**.

El siguiente al paso a la hibridación, una vez los *spots* en el micro-arreglo han hibridado, **son excitados por láser y escaneados** a longitudes de onda apropiados para detectar los colorantes rojo y verde.

**La cantidad de fluorescente emitido bajo excitación corresponde a la cantidad de ácido nucleico ligado**.

Por ejemplo, si el *cDNA* de la *condición A* para un gen particular está en mayor abundancia que el de *condición B*, se podría encontrar el *spot* en rojo. Si el gen fue expresado en el mismo grado en ambas condiciones, el spot podría estar amarillo, y si el gen no fue expresado en ninguna condición, el *spot* podría estar en negro.

Así, lo que se ve al final del experimento es una **imagen del micro-arreglo**, en la cual cada spot que corresponde a un gen tiene un valor fluorescente asociado que representa el nivel relativo de expresión de dicho gen.

### Introducción al procesamiento de imágenes, transformación y normalización

#### Procesamiento de imágenes y análisis

El primer paso en el análisis de datos de micro-arreglos es procesar la imagen que almacena el nivel relativo de expresión de cada gen, resultado del proceso que se mencionó en el numeral anterior.

El procesamiento de imágenes comprende los siguientes pasos, para mayor detalle ver (M. Babu, 2004):

1. Identificación de los *spots* y diferenciación de los mismos contra falsas señales
2. Determinación del área de *spot* a ser estudiado y determinación de la región local para estimar el fondo de la hibridación.
3. Informe de resumen estadístico y asignación de intensidad de *spot* después de restar la intensidad de fondo.

#### Transformaciones de la taza de expresión

La taza de expresión (*expression ratios*) es una métrica frecuentemente usada para medir el nivel de expresión relativo para un gen, el cual puede ser medido como la cantidad de luz roja o verde emitida durante la excitación.

La tasa de expresión es definida como:

Ecuación 1 Tasa de expresión de un gen dado

Para cada gen en el arreglo, donde representa a métrica de intensidad de *spot* para la muestra de prueba y representa la métrica de intensidad de *spot* para la muestra de referencia.

Aunque la tasa de expresión puede revelar patrones inherentes en los datos, también, puede eliminar toda la información acerca de los niveles de expresión absoluta de los genes, y los problemas asociados aparecerán cuando se intente identificar genes regulados diferencialmente. Algunas **técnicas para mitigar la eliminación** (perdida de niveles de expresión absoluta) que se acaba de mencionar se pueden ver en(M. Babu, 2004).

#### La normalización de datos

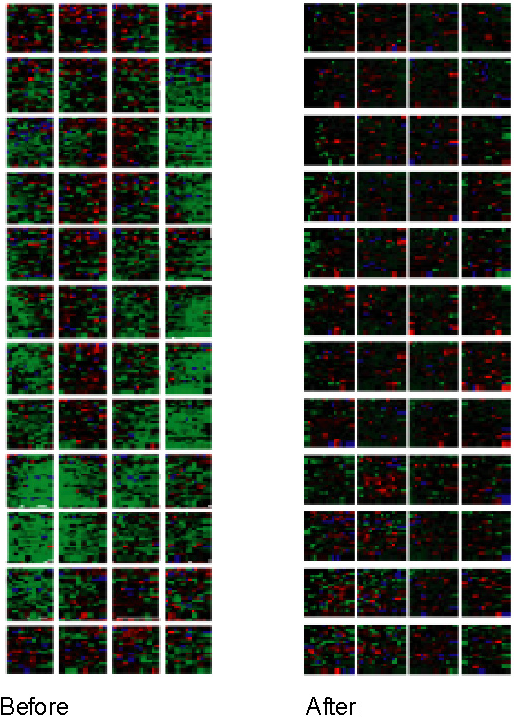
La normalización es un término usado para describir la eliminación de variaciones causadas por la eficiencia en el marcado diferencial de los dos colorantes fluorescentes o variaciones causadas por diferentes cantidades de material *mRNA* en las dos muestras. Así, la normalización permite la comparación apropiada de los datos obtenidos de las muestras.

El primer paso en un procedimiento de normalización es elegir un conjunto de genes (*gene-set*); el cual consiste de genes para los cuales el nivel de expresión podría no cambiar bajo las condiciones estudiadas, es decir, la taza de expresión para todos los genes del conjunto se espera sea igual a 1.

Desde el conjunto de genes seleccionado, se calcula un factor de normalización (*normalization factor*), el cual es un número que cuenta la variabilidad observada en dicho conjunto.

Este procedimiento se aplica para los otros genes de una condición biológica (condición A o B) en el experimento de micro-arreglos.

El procedimiento de normalización genera un cambio de los datos. La Figura 6**,** presenta datos de expresión antes y después de un procedimiento de normalizaciónpara mayor detalle ver (M. Babu, 2004).



**Ref.**http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/

Figura 6 Datos de expresión de genes antes y después del procedimiento de normalización

### Análisis de datos de expresión de genes

Los datos procesados, después del procedimiento de normalización, pueden ser representados en forma de matriz, por medio de una *matriz de expresión de genes* (VerTabla 3)

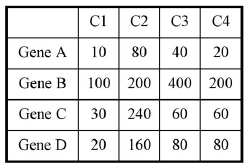


Tabla 3 Matriz de expresión de genes que contiene filas para representar genes y columnas para representar condiciones biológicas(M. Babu, 2004),

Cada fila de la matriz corresponde a un gen y cada columna podría corresponder a una condición experimental o punto de tiempo específico en el cual la expresión del gen correspondiente ha sido medido.

El perfil de expresión de un gen (*gene expression profile*) es el nivel acumulado de expresión de dicho gen, en diferentes condiciones experimentales. El perfil de expresión de la muestra es el nivel acumulado de expresión para todos los genes bajo una condición experimental.

Adicional a los niveles, se puede agregar anotación al gen o a la muestra. Por ejemplo, se puede proporcionar la función del gen o el detalle biológico de la muestra (estado “normal” y estado “enfermo”).

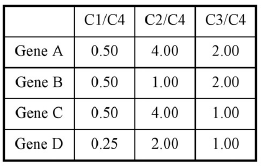
Dependiendo de si se utiliza anotaciones o no, el análisis de datos de expresión de genes puede ser clasificado en dos tipos: (1) aprendizaje supervisado y (2) aprendizaje no-supervisado.

En el caso del aprendizaje no-supervisado, los datos de expresión son analizados para identificar patrones que pueden agrupar genes o muestras en grupos (*clusters*) sin emplear anotaciones. Por ejemplo: los genes con **perfil de expresión similar** pueden ser agrupados sin usar anotaciones, y posteriormente, la información de anotaciones puede ser tomada en cuenta después para dar hacer inferencias biológicas significativas, para mayor detalle **ver** (M. Babu, 2004) o el numeral 5.4.

#### Representación de datos de expresión de genes

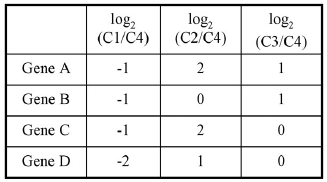
Para hacer útil cualquier comparación o análisis biológico debemos saber lo que representan los datos de la matriz de expresión de genes. Estos datos pueden ser representados de cinco diferentes maneras(para más detalle ver (M. Babu, 2004)):

1. **Medida absoluta**: Cada celda de la matriz representa el nivel de expresión del gen en unidades abstractas,**ver**Tabla 3.
2. **Medida relativa o tasa de expresión**: El nivel del gen en unidades abstractas es normalizado con respecto a su expresión en una condición biológica de referencia, como se muestra en la Tabla 4.



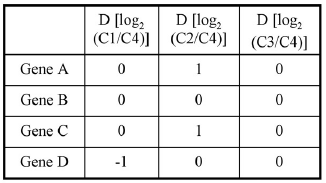
**Tabla 4** Datos de expresión de genes por medida relativa**(M. Babu, 2004)**

1. **(**tasa de expresión**):** La medida relativa o taza de expresión es mapeada de manera simétrica por la función logaritmo de base 2, como puede verse en la Tabla 5.



**Tabla 5** Datos de expresión por**(M. Babu, 2004)**

1. **Valores discretos:** La representación se convierte a números discretos. Por ejemplo, se puede usar una matriz binaria de 1 y 0, donde 1 significa que el gen correspondiente se expresó por encima de un umbral y 0 por debajo, como puede observarse en la Tabla 6.



**Tabla 6** Datos expresión por valores discretos**(M. Babu, 2004)**

1. Representación de perfiles de expresión como vectores: Un perfil de expresión (de un gen o de una muestra) puede ser pensado como un vector y representado en el espacio del vector.

Así, el perfil de expresión de un gen puede ser representado como el vector fila:

Y, el perfil de expresión de una muestra puede ser representado como el vector columna:

### Estado del arte

La Secuenciación de Alto Rendimiento de ARN (RNA-Seq del inglés RNA Sequencing)es una tecnología mucho más reciente que los micro-arreglos, para detectar miles de genes expresados (transcritos) en un solo experimento, y se basa en las nuevas técnicas de Secuenciación de Segunda Generación (NGS del inglés Next Generation Sequencing) que ofrecen la ventaja de producción de grandes cantidades de secuencias a un bajo costo(Metzker, 2010) y (Mardis, 2013).

En la técnica RNA-Seq las secuencias producidas son proporcionales a la expresión, hecho que permite la detección y medición de esta última. En comparación con los micro-arreglos de ADN, el RNA-Seq permite determinar la expresión de manera más precisa(Young et al., 2012), y además no está limitada a la detección de transcritos basado en un genoma conocido lo que permite su aplicación en el estudio de organismos no modelo.

## Redes complejas

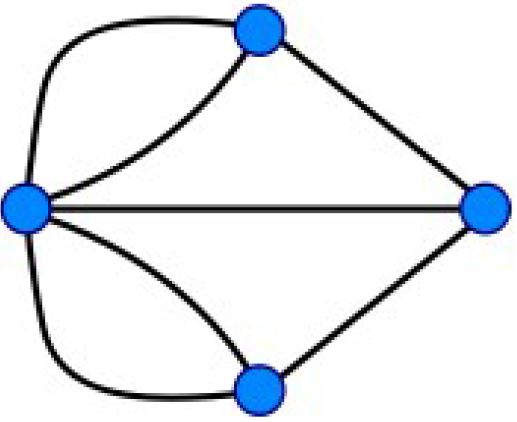
### Introducción a las redes complejas

El análisis de redes complejas (network science) es un campo académico interdisciplinario que estudia redes complejas (complex networks) como redes de telecomunicaciones, redes de computadores, redes biológicas, redes cognitivas y semánticas y redes sociales. Este campo se apoya en teorías y métodos incluyendo teoría de grafos desde la matemática, mecánica estadística desde la física, minería de datos y visualización de información desde las ciencias de la computación, modelamiento inferencial desde la estadística y estructura social desde la sociología. El concejo de investigación nacional de los Estados Unidos, lo define como “el estudio de las redes que representan fenómenos físicos, biológicos y sociales, dirigido a la construcción de modelos predictivos de dichos fenómenos”

Recientemente a través de la teoría de redes se han podido modelar eventos complejos de biología molecular con el objetivo de abordar su estudio desde una perspectiva sistémica observando relaciones e interacciones en conjunto en lugar de ver elementos de manera aislada.

Usando puntos y líneas se construyen redes que recrean matemáticamente distintas relaciones entre biomoléculas (proteínas o genes) como por ejemplo interacciones entre proteínas o mecanismos regulatorios y de expresión.Las redes de co-expresión de genes (**NCG**) son un ejemplo de aplicación de este enfoque para estudiar la complejidad de las relaciones entre perfiles de expresión génica obtenidos masivamente mediante tecnologías como los micro-arreglos de ADN o técnicas NGS (**ver** numeral **5.4**).

Una red o grafo es un conjunto de **nodos** (vértices), con conexiones entre ellos, llamados **arcos** (aristas).



**Figura 7** Red o grafo

Matemáticamente, un grafo H es un par ordenado ***G= (V, E)*,** donde *V* es un conjunto de vértices o nodos y *E* es un conjunto de aristas o arcos, que relacionan estos nodos.

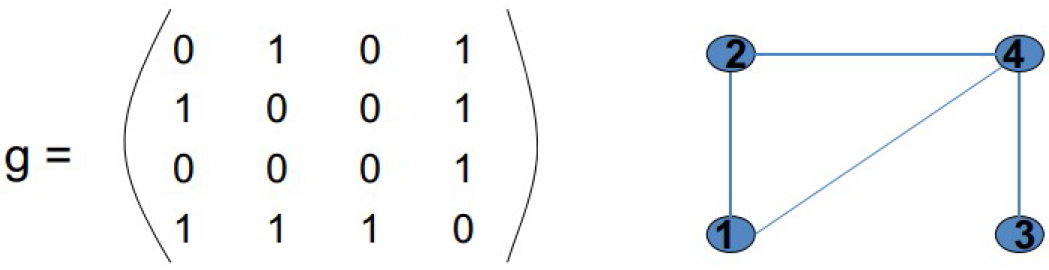
Normalmente V suele ser finito, y se llama **ordinal** del grafo *G* a su número de vértices *|V|.*

El **grado** de un vértice o nodo en*V* es igual al número de arcos *E* que se encuentran en él. Un bucle es una arista que relaciona al mismo nodo.

Existen diferentes tipos de redes:

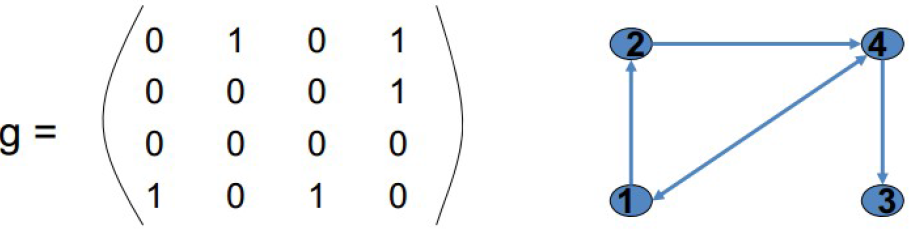
1. **Redes No dirigidas**: En las cuales no hay un origen y un destino en la relación, este tipo de relaciones son simétricas

Como puede observarse en siguiente figura, las redes se pueden representar por medio de una **matriz de adyacencia**, indicando para cada posición , si el nodo *i* está relacionado con el nodo *j* por medio de una arista.



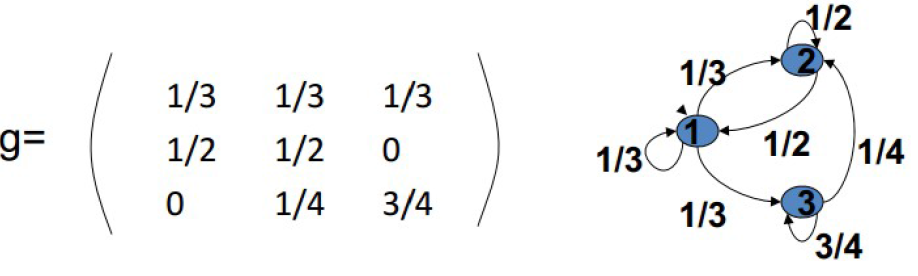
**Figura 8** Red no dirigida

1. **Redes dirigidas**: En estas redes para cada arista o arco, puede haber una dirección en un sentido, pero no necesariamente en el otro sentido.



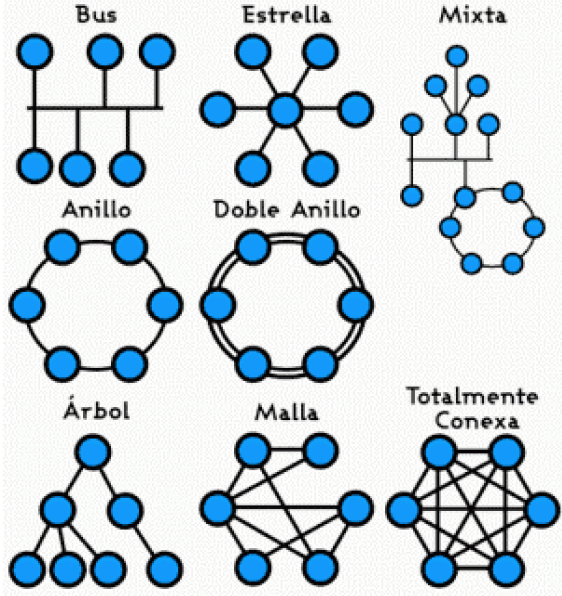
**Figura 9** Red dirigida

1. **Redes dirigidas y ponderadas**: Es una red dirigida que además tiene asociada un peso en la relación representada por el arco o arista.



**Figura 10** Redes dirigidas y ponderadas

La **estructura de una red** es la forma o topología de la misma. Existen diversas topologías para una red como se muestra en la figura a continuación:



**Figura 11** Estructura de una red

Un **camino** es ir desde el nodo : compuesto de una secuencia de nodos ( y una secuencia de aristas ( de tal forma que existe en *G*para cada *k.*

Un **componente** es el máximo sub-grafo conectado ***(V’, E’)***subconjunto de ***(V, E)*,**

Algunas **propiedades** de las redes se describen a continuación:

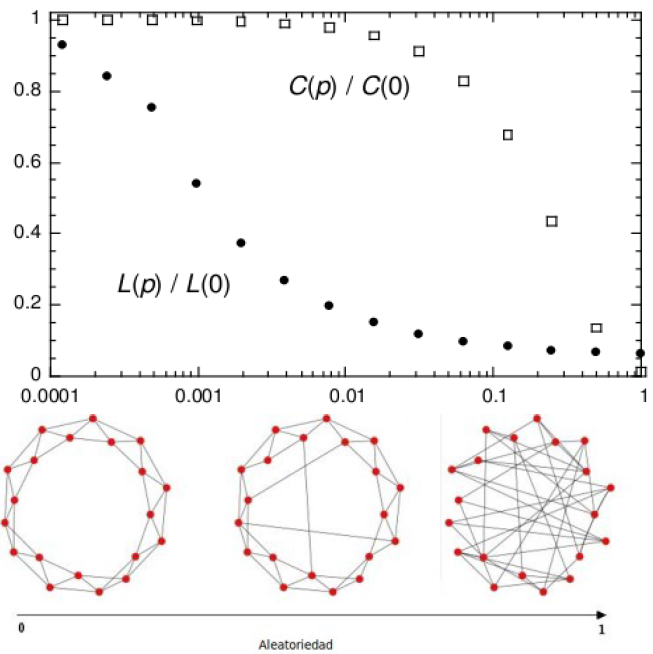
1. **Densidad**: Proporción de aristas presentes en la red sobre el número máximo posible.
2. **Reciprocidad**: Proporción de aristas o arcos que son reciprocos en la red (aplica para redes dirigidas).
3. **Distancia**: La distancia entre un par de nodos, es el número de nodos que debe recorrerse para unirlos.
4. **Distancia geodésica**: Es el la distancia del camino más corto entre un par de nodos.
5. **Diámetro**: Es la longitud de la distancia geodésica más larga entre cualquier par de nodos de la red.

Algunas **medidas de centralidad y prestigio** de las redes, son las siguientes:

1. **Grado de entrada**: Número de aristas totales que un nodo recibe de otros en la red.
2. **Grado de salida**: Número de aristas totales que un nodo envía a otros en la red.
3. **Intermediación (betweeness)**: Medida de frecuencia con la que un nodo aparece en el camino más corto entre dos nodos
4. **Cercanía (closeness)**: Que tan cerca está un nodo de los demás nodos en la red basados en la distancia geodésica (camino de longitud más corta)
5. **Excentricidad (eccentricity)**: La distancia de un nodo al nodo más alejado de él en la red.
6. **Coeficiente de agrupamiento (*clustering*) de cada nodo:** Número de conexiones entre los amigos (nodos que se relacionan) de un nodo sobre el total de posibles conexiones entre los amigos del nodo (es decir mide la transitividad).

### Análisis de redes complejas

Un análisis común en redes complejas es validar la propiedad del mundo pequeño. Esta propiedad se da cuando la red presenta un elevado promedio (*C*) de agrupamiento (*clustering*) y un bajo promedio de la distancia geodésica (*L*) al mismo tiempo. Una interpretación intuitiva de la propiedad del mundo pequeño indicaría que fracción de mis amigos (nodos que se relacionan) son amigos unos de otros?



**Figura 12** Propiedad del mundo pequeño**(Watts & Strogatz, 1998)**

Como puede verse en la figura anterior el análisis de una red que presenta propiedad de mundo pequeño, revela que pocos nodos (*Hubs*) tienen muchas relaciones (aristas) con otros nodos, sin embargo, la mayoría de nodos están pobremente relacionados con respecto al total de nodos de la red.

Con las redes complejas se disminuye la complejidad de conjuntos grandes de datos de expresión, y mediante su análisis topológico se pueden detectar patrones de expresión subyacentes, inferir funcionalidades biológicas e identificar genes relevantes en procesos biológicos de interés como enfermedades o estímulos ambientales.

En el análisis topológico de las redes y módulos, se pueden identificar genes sobresalientes por su alta **conectividad** y/o **centralidad de intermediación**, los cuales podrían ser relevantes en los procesos de los sistemas celulares.

### Estado del arte

Por ejemplo, la **conectividad**, que denota el **grado de conexión** de los genes, y la **centralidad de intermediación** (betweenness centrality), que indica el grado en que un nodo actúa como puente entre otros, son propiedades importantes biológicamente(Prifti, Zucker, Clément, & Henegar, 2010) pues permiten detectar **hubs** (genes altamente conectados) y **bottlenecks** (genes con alta intermediación),respectivamente, los cuales tienden a estar relacionados con la esencialidad en procesosfuncionales. De esta manera se pueden elegir los genes más llamativos desde el punto de vista topológico y que a la vez tengan relevancia biológica. Dong y Horvath (Dong & Horvath, 2007) y Horvath y Dong (Horvath & Dong, 2008)explican varias propiedades de las redes de co-expresión que tienen un significado biológico importante.

## Redes de co-expresión de genes

### Introducción a las redes de co-expresión de genes

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.De esta manera, 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: sin pesos, en las que sus aristas denotan si hay una asociación entre un par de nodos, y con pesos, 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].

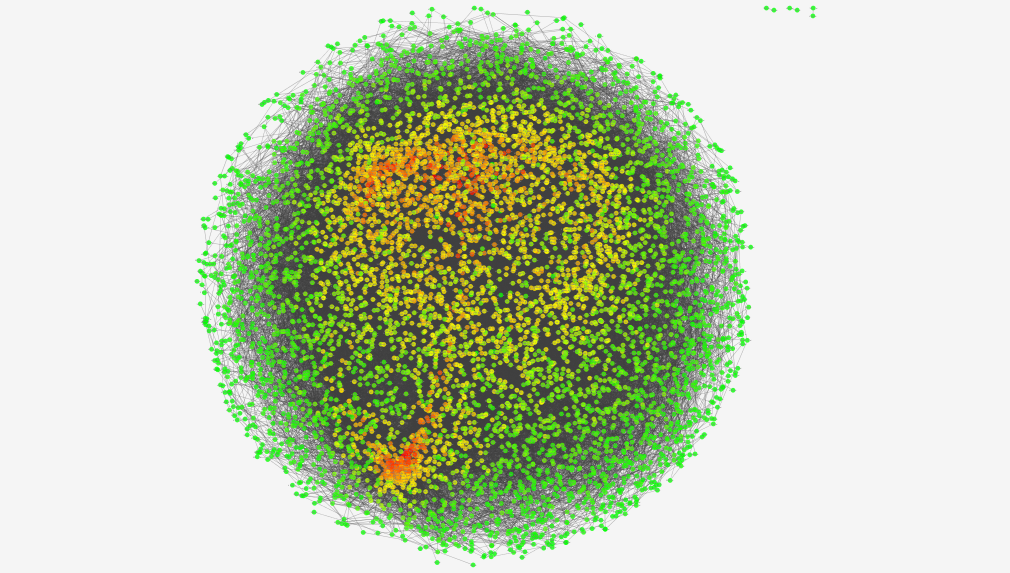
### Estado del arte

Una de las aplicaciones de estas redes es la identificación de grupos de nodos altamente co-expresados denominados **módulos**, los cuales indican **funcionalidades comunes** entre los genes que los conforman. **Zhang y Horvath**(Zhang & Horvath, 2005) definen 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).

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 comprende los siguientes pasos: (1) preparación de los datos de expresión, (2) definición de una medida de similitud entre nodos, (3) definición de una función de adyacencia entre nodos, (4) definición de los parámetros de adyacencia, (5) definición de una medida de disimilitud entre los nodos, y finalmente, (6) identificación de los módulos. Los autores del método han desarrollado WGCNA, 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/), Gephi (http://gephi.github.io/), o gViz ([http://urbm-cluster.urbm.fundp.ac.be/webapps/gviz/)(**Ver**](http://urbm-cluster.urbm.fundp.ac.be/webapps/gviz/)(Ver)Figura13).

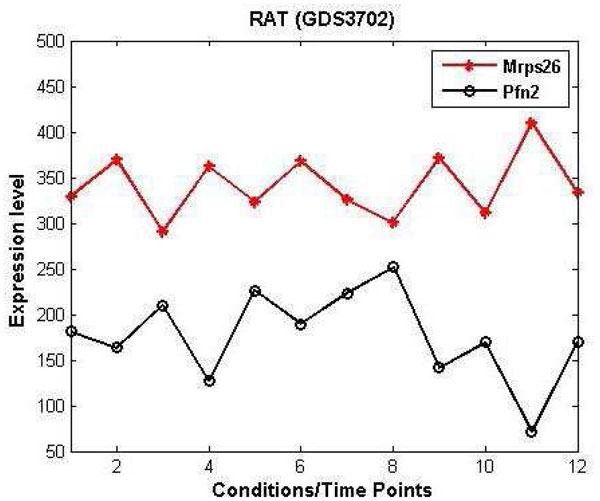


**Ref.***S. Mohammad H. Oloomi -* I have created the network by programming in R and then I have created the network image using Cystoscape.

Figura 13 Red de co-expresión con 7221 genes para 18 pacientes de cáncer gástrico

Por ejemplo, enXue et al.(Xue et al., 2013) usan las CGN para identificar y seleccionar genes candidatos claves en el proceso de desarrollo embrionario en humanos y ratones. Por su parte Davidson et al.(Davidson et al., 2011) identifican genes candidatos para estudios sobre desarrollo reproductivo y mejoramiento del potencial de producción en maíz. Finalmente se puede mencionar a Hollender et al.(Hollender et al., 2014) , quienes a través del análisis de las CGN, identifican genes y **rutas metabólicas** involucrados en la floración y desarrollo de frutos en fresa silvestre.

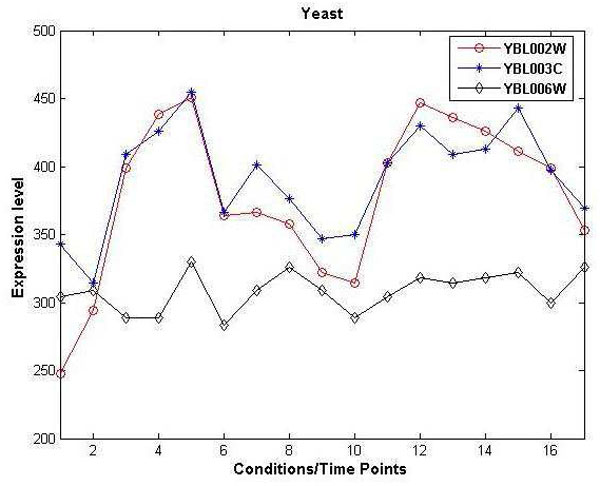
Por otra parte, dos genes pueden estar relacionados el uno con el otro aun cuando sus patrones de expresión muestren comportamiento negativo o invertido (Yu, Luscombe, Qian, & Gerstein, 2003). En la Figura 14**,** los patrones de expresión de genes de rata **Mrps26** y **Pfn2**, tomados del conjunto de datos de NCBI, GDS3702, claramente muestran comportamiento negativo. Las ontologías de los genes sugieren que ambos son responsables de la regulación de producción de *interferón-beta*.



Ref.*Roy et al. BMC Bioinformatics* 2014 15(Suppl 7):S10 doi:10.1186/1471-2105-15-S7-S10.

Figura 14 Patrón de regulación negativo o invertido

También, se han observado en los conjuntos de datos de levadura dados en (Cheng & Church, 2000), los genes *YBL002W*y*YBL003C*con patrón similar y el gen *YBL006W* con patrón invertido con respecto a los otros dos genes. Como se observa en la **Figura15,** los patrones de expresión también comparten regulación mezclada.



Ref.*Roy et al. BMC Bioinformatics* 2014 15(Suppl 7):S10 doi:10.1186/1471-2105-15-S7-S10.

**Figura15** Mezcla de patrones de regulación

La ontología de los genes sugiere que los tres genes en mención están involucrados en la organización de nucleosomas, DNA, cromatina y cromosoma y sub-unidades macro-moleculares celulares. UN grupo de genes puede compartir una combinación de co-regulación positiva y negativa en pocas condiciones o en pocos puntos de tiempo. Computacionalmente, muchas técnicas no consideran los patrones de regulación positivos y negativos para presentar co-expresión o interacción de genes con importancia biológica asociada (Roy, Bhattacharyya, & Kalita, 2014).



## Aprendizaje Bayesiano fundamentos y estado de arte

### Definición y representación de una red Bayesiana

Un modelo de red bayesiana puede ser definido como un par **.** Aquí es un **grafo a-cíclico dirigido** que representa la estructura de la red, denota un conjunto de nodos y un conjunto de aristas. Igualmente, cada elemento del vector representa una probabilidad condicional (Faming, 2010).

**Ejemplo**: problema de **cáncer de pulmón** tomado de (Korb & Nicholson, 2010), un paciente ha estado sufriendo de dificultad para respirar (***dyspnoea***), y visita al doctor preocupado de tener cáncer de pulmón. El doctor sabe que otras enfermedades como tuberculosis, bronquitis son posibles causas del síntoma, además del cáncer de pulmón. También, el doctor sabe qué otra información relevante incluir: si el paciente es o no **fumador** (incrementa la probabilidad de cáncer o bronquitis), y que tanta exposición a **aire contaminado** ha tenido el paciente. Un examen de **rayos X** positivo podría también indicar TB o cáncer.

Representando la problemática del ejemplo, por medio de una red bayesiana, el **conjunto de nodos** del **grafo** podrían ser los que se muestran en la siguiente tabla:

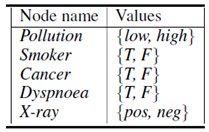
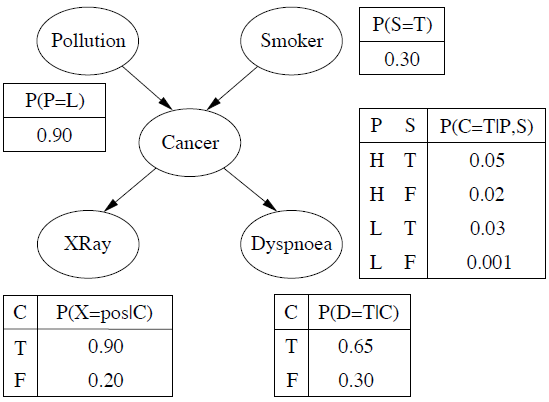


Tabla 7 Selección preliminar de nodos y valores para el ejemplo de cáncer de pulmónRef. (Korb & Nicholson, 2010)

La estructura o topología de la red captura relaciones entre los nodos. Continuando con el ejemplo, nos preguntamos qué factores afectan la probabilidad de que el paciente tenga cáncer? Si la respuesta es contaminación y cigarrillo (pollution and smoking), entonces deberíamos tener aristas desde estos nodos a cáncer. Igualmente, tener cáncer produce dificultad respiratoria en el paciente (dyspnoea) y aumenta la probabilidad de un examen de rayos X (*X-ray*) positivo, por lo tanto, debemos tener aristas desde el nodo cáncer a los nodos *Dyspnoea* y *X-ray*. El conjunto de aristas se muestra a continuación, en la Figura 16.

****

**Figura 16**Una red bayesiana para el problema de cáncer de pulmón

**Ref.**(Korb & Nicholson, 2010)

Para un nodo , un padre de es un nodo desde el cual existe una arista que va dirigida a . El conjunto de padres de se denota . Así, de acuerdo a la **Figura 16,** si hacemos , entonces

En el ejemplo (*caso discreto*) es una variable aleatoria con distribución categórica que toma valores en un conjunto finito , Según la **Tabla 7,** si tomamos , entonces, el conjunto finito es

Existenposibles valores conjuntos (*joint states*) de los padres de .Haciendo de nuevo , y , (ver **Figura 16**), entonces

Es decir, que en la tabla asociada al nodo , existen cuatro posibles valores conjuntos (*joint states*), a saber:

(1) ,

(2) ,

(3) y

(4) .

Volviendo de nuevo al par que define la red Bayesiana**,** dijimos que cada elemento del vector representa una probabilidad condicional; es decir, es la probabilidad de que la variable sea igual a j dado que su padre es igual a . Así, haciendo (ver **Figura 16**).

Naturalmente, tiene las siguientes restricciones  **y** .

En una red Bayesiana, la **distribución de probabilidad conjunta** de las variables puede ser especificada de la siguiente forma:

**Ecuación 2** Distribución conjunta de las variables aleatorias *v*

Siguiendo con el ejemplo de cáncer de pulmón, podríamos calcular la probabilidad conjunta, para un caso particular, así (ver **Figura 16**):

Ahora, sea un conjunto de muestras independientes e idénticamente distribuidas (***iid***), obtenidas de **,** y sea el número de muestras para las cuales, la variable es igual a j dado que su padre es igual a , entonces, los contadores de muestras () siguen una **distribución multinomial**. Es decir,

(

### Tipos de razonamiento en redes bayesianas

Las redes Bayesianas proveen representaciones completas de distribuciones de probabilidad sobre sus variables. Esto implica que ellas pueden ser condicionadas sobre cualquier conjunto de sus variables, soportando cualquier dirección de razonamiento:

* Por ejemplo, uno puede realizar **diagnóstico,** razonando desde el síntoma a la causa, como cuando el doctor (en el ejemplo del cáncer de pulmón) observó disnea y entonces, cambio su creencia sobre el cáncer, si el paciente es fumador. Note que este razonamiento ocurre en la dirección opuesta de la arista de la red.
* También, podemos realizar **predicción**, razonando desde nueva información acerca de la causa hacia nuevas creencias acerca del efecto, siguiendo la dirección de las aristas. Por ejemplo, el paciente puede decir a su médico que es fumador, aun antes de que cualquier síntoma haya sido juzgado, el doctor sabe que aumentará la probabilidad de que el paciente tenga cáncer **(Korb & Nicholson, 2010)**.

### Principio de la máxima verosimilitud (maximum likelihood)

#### Contexto del problema del aprendizaje

Supongamos que observamos un conjunto de muestras independientes e idénticamente distribuidas (***iid***) de un conjunto de variables aleatorias a partir de una distribución desconocida . Asumimos que conocemos el espacio muestral con el que estamos tratando.

Ahora, definimos el conjunto de muestras para **entrenamiento** como y consideramos que este consiste de instancias de : .

Lo siguiente que necesitamos considerar es exactamente qué deseamos aprender?

Supongamos ahora que se nos da un modelo paramétrico para el cual deseamos estimar los parámetros. Formalmente un **modelo paramétrico** es definido por una función , especificado en términos de un conjunto de **parámetros**.

Dado un conjunto particular de valores de parámetros y una instancia de , el modelo asigna una probabilidad a . Adicionalmente, para cada elección de parámetros , debe ser una distribución valida; es decir, esta debe ser positiva y

En general, para cada modelo, no todos los valores de parámetros son válidos. Así, necesitamos definir el **espacio de parámetros**, el cual será el conjunto de parámetros permitidos.

#### La función de verosimilitud

El siguiente paso para la estimación de máxima verosimilitud es definir la función de verosimilitud. Esta función, para una elección de parámetroses la probabilidad que el modelo instancie los datos de entrenamiento:

#### Estimación de máxima verosimilitud (MLE)

Una vez definida la función likelihood, podemos usar la estimación de máxima verosimilitud para elegir los valores de los parámetros. Dado un conjunto de datos , elegir los parámetros que satisfaga

### MLE para redes Bayesianas

En un plano más general del problema, ahora revisamos la estimación de parámetros para una red Bayesiana. Para hacerlo posible, la estructura de las redes Bayesianas nos permite reducir el problema de estimación de parámetros a un conjunto de sub-problemas independientes que se pueden resolver usando MLE.

#### Un ejemplo simple

El ejemplo más simple es imaginar una red consistente de dos variables binarias y , con una arista :

* El objetivo aplicando MLE es maximizar la función de verosimilitud o su logaritmo (log-likelihood).
* En este caso, la red en mención es parametrizada por un vector de parámetros, el cual define el conjunto de parámetros para todas las **distribuciones de probabilidad condicional** (**CPD**s) en dicha red.
* En este ejemplo, la parametrización podría consistir de los siguientes parámetros:
  + y especifican la probabilidad de los dos valores de
  + y especifican la probabilidad de dado
  + y especifican la probabilidad de dado
  + Para abreviar usaremos para referirnos al conjunto y para referirnos a
* En este ejemplo, cada instancia de entrenamiento es una tupla que describe una asignación particular a y .
* Así, la función de verosimilitud se define así:
* El modelo de red especifica que tiene forma de producto, Así, podemos escribir
* Cambiando el orden de la multiplicación, podemos escribir equivalentemente el termino como

Es decir, la función de **verosimilitud** sufrió una **descomposición** en dos términos, uno por variable. Además, cada término es una **función de verosimilitud local** que mide el grado de predicción de una variable dado sus padres.

* Considerando los términos individuales de la ecuación, podemos descomponer el segundo término:

así, **la función de verosimilitud se descompone en un producto de términos**, uno por cada grupo de parámetros en . Esta propiedad es la capacidad de descomponer la función de verosimilitud.

* De la anterior expresión, consideremos uno de los términos:
  + Cada uno de los términos individuales puede tomar uno de dos valores, dependiendo del valor de .
  + Si , el termino individual toma
  + Si , el termino individual toma
  + ¿Cuántos casos de cada tipo podemos tener?
  + Fijemos primero aquellos casos donde Estos a su vez se particionan en dos categorías:
    - Obtenemos en cuyos casos de datos donde and , usaremos  **para denotar el número de casos.**
    - Obtenemos en cuyos casos de datos donde and , usaremos  **para denotar el número de casos.**
  + Así, el termino es igual a:

* + Maximizaremos por medio de:

similarmente para

En conclusión, **podemos encontrar la máxima verosimilitud de los parámetros en esta distribución de probabilidad condicional (CPD), simplemente contando cuantas veces cada una de las posibles asignaciones de and aparecen en los datos de entrenamiento**.

#### Descomposición global de la verosimilitud

Iniciemos examinando la función de verosimilitud de una red Bayesiana:

* Supongamos que queremos aprender los parámetros para una red Bayesiana con estructura , y parámetros .
* Esto significa que acordamos el tipo de CPDs que vamos a aprender.
* Sea también, un conjunto dado de muestras
* Escribiendo la función de verosimilitud y repitiendo los pasos realizados en el ejemplo anterior, obtenemos:

Observemos que cada uno de los términos entre corchetes se refiere la **verosimilitud condicional** de una variable dado sus padres en la red.

Usaremos para denotar el subconjunto de parámetros que determinan en nuestro modelo. Entonces podemos escribir

Donde la **función de verosimilitud local** para es:

**Ecuación 3** Función de verosimilitud local

Esta forma es particularmente útil cuando los conjuntos de parámetros son disjuntos. Es decir, cada CPD es parametrizada por un conjunto separado de parametros que no se solapa.

Por lo tanto, este análisis muestra que la verosimilitud (likelihood) se descompone como un producto de términos independientes, uno para cada CPD en la red. Esta propiedad es llamada **descomposición global de la función de verosimilitud.**

### Aprendizaje de la estructura de una red Bayesiana

En esta sección, consideramos la tarea de aprendizaje donde no conocemos la estructura de la red Bayesiana y que nuestros datos son completamente observados.

#### Definición del problema

Supongamos que los datos son generados IID de una distribución subyacente . También, asumimos que es inducido por alguna red Bayesiana over .

#### Resumen de métodos

Existen tres enfoques para aprender sin una estructura pre-especificada:

* Un enfoque utiliza aprendizaje de estructuras basado en restricciones **(constraint-based)**. Este enfoque ve una red Bayesiana como una representación de independencias. Se realizan pruebas de dependencia e independencia condicional a los datos, y entonces se encuentra una red que explica dichas dependencias e independencias. Este enfoque no hace parte del alcance de este trabajo. Se puede encontrar información relacionada en el numeral 18.2 de (Koller & Friedman, 2009b), así mismo, algunas aplicaciones en (Prestat et al., 2013).
* El segundo enfoque es el aprendizaje basado en puntaje (**score-based**). Los métodos basados en puntaje ven una red Bayesiana como una especificación de un modelo estadístico y entonces direcciona el aprendizaje como un problema de selección del modelo. Todo en este enfoque opera con el mismo principio:
  + Definir un espacio hipótesis de potenciales modelos.
  + Definir una función de puntuación (scoring), que mide que tan bien el modelo encaja con los datos.
  + Nuestra tarea computacional es entonces encontrar la más alta puntuación de estructuras de red.
* Por último, el tercer enfoque no intenta aprender una sola estructura; por el contrario, este genera un conjunto de posibles estructuras. Estos métodos de promedio de modelo Bayesiano (**Bayesian model-averaging**), extienden el razonamiento Bayesiano e intentan promediar la predicción de todas las posibles estructuras.

#### Puntajes de estructuras (scores)

##### Puntajes por verosimilitud

Una elección natural para la **función de puntaje** es la **función de verosimilitud**, la cual revisamos en secciones anteriores para la estimación de parámetros. Esta función **mide la probabilidad de los datos dado un modelo**. Así, esta función intuitivamente encuentra un modelo podría hacer los datos tan probables como sea posible.

Asumimos que queremos maximizar la verosimilitud del modelo. En este caso, nuestro modelo es una pareja . El objetivo es encontrar un grafo y los parámetros que maximicen la verosimilitud. Usaremos los parámetros con la máxima verosimilitud para este grafo.

Para encontrar la pareja que maximiza la verosimilitud, debemos encontrar la estructura del grafo que alcanza la más alta verosimilitud cuando usamos los parámetros MLE para .

Definimos

donde es la **función logaritmo de la verosimilitud** (esto es usual, ya que se facilitan los cálculos con el logaritmo se la verosimilitud), y son los parámetros que maximizan la verosimilitud para

##### Puntajes Bayesianos

Como una alternativa la función de puntuación con puntaje Bayesiano está basada en una perspectiva Bayesiana. Siempre que se tiene incertidumbre sobre algo, deberíamos poner una distribución ahí.

En este caso se tiene incertidumbre sobre la estructura y sus parámetros. Por lo tanto, definimos una **distribución a priori**, la cual pone una **probabilidad a priori sobre las diferentes estructuras del grafo**, y un parámetro a priori , que pone una probabilidad a las diferentes elecciones de parámetros una vez el grafo es dado. Aplicando la regla de Bayes, tenemos

donde, como es usual, el denominador es simplemente un factor de normalización que no ayuda a distinguir entre diferentes estructuras.

En ese orden de ideas, definimos el puntaje Bayesiano como:

La habilidad de incluir una distribución a priori sobre las estructuras, nos da una manera de preferir algunas estructuras sobre otras. Observemos que aunque este término parece describir nuestra preferencias por ciertas estructuras, de hecho, este juega un papel relativamente menor.

#### Búsqueda de estructuras

Ahora miraremos como encontrar una estructura con un puntaje alto. Ya tenemos un problema de optimización bien definido con las siguientes entradas:

* Conjunto de entrenamiento
* Función de puntuación (incluyendo distribuciones a priori, si se requieren)
* Un conjunto de posibles estructuras de red (incorporando cualquier conocimiento a prior).

La salida deseada es una estructura de red (tomada del conjunto de estructuras posibles) que maximiza el puntaje.

##### El espacio de búsqueda

Podemos pensar en el espacio de búsqueda como un grafo sobre las posibles soluciones candidatas, conectadas por posibles operadores, que el procedimiento de búsqueda puede ejecutar para moverse entre diferentes soluciones. En el contexto más simple, consideramos el espacio de búsqueda donde cada estado denota una estructura de red completa sobre .

Definimos la conectividad del espacio de búsqueda en términos de los siguientes operadores:

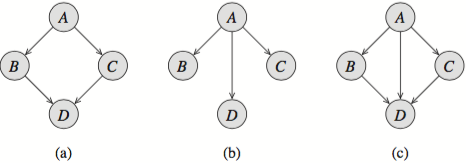
* Adicionar arista
* Eliminar arista
* Invertir arista

Es decir, los estados adyacentes a un estado son aquellos donde cambiamos una arista, ya sea porque adicionamos una, eliminamos una o invertimos la dirección de la arista. Observamos que sólo consideraremos aristas que obtengan redes válidas. O sea, redes acíclicas que satisfagan las restricciones que nosotros hayamos incluido (por ejemplo, restricciones del grado de entrada).

Esta definición del espacio de búsqueda es natural y tiene las siguientes propiedades favorables:

* El diámetro del espacio de búsqueda es en el peor caso . Es decir, que existe una ruta relativamente corta entre cualesquier dos redes que elijamos.
* El puntaje de una red es la suma de puntajes locales. Las operaciones ya resueltas cambian solo en un término de puntaje local.

Supongamos que la red original es la que muestra en la Figura 17a:

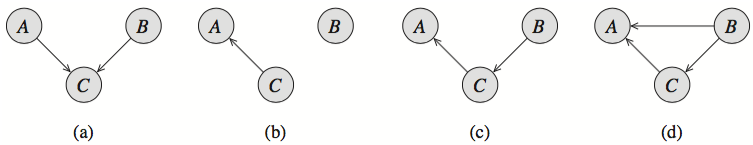


**Figura 17** Ejemplo de una búsqueda que requiere eliminación de arista

**Ref**. (Koller & Friedman, 2009b)

* A es altamente informativa sobre C y B.
* Empezando con una red vacía y adicionando aristas vorazmente, podríamos adicionar las aristas y .
* En algunos conjuntos de datos, podríamos adicionar la arista ; ya que e B y C son los padres de D, también A es informativa sobre D.
* En este punto finalizamos con la red de la Figura 17b.
* Continuando la búsqueda, consideramos diferentes operadores, adicionar la arista , ya que B es un padre de D en la red original, este mejorara la predicción de D cuando se combine con A. Así, la calificación de A and B juntos como padres de D puede ser más grande que la calificación de solo A. Análogamente, adicionaremos también la arista , y alcanzaremos la estructura que se muestra en la Figura 17c.
* Esperamos que elegir B, C como padres de D tendrá una calificación mejor que elegir A, B, C.
* Después, eliminaremos la arista para obtener la estructura original.

Para ver la utilidad del operador de invertir una arista, consideremos el siguiente ejemplo simple:



**Figura 18** Ejemplo de búsqueda que requiere invertir

Ref. (Koller & Friedman, 2009b)

* Suponga la red real que genera los datos tiene la v-estructura mostrada en la Figura 18a.
* Suponga que la independencia entre A y C es más fuerte que entre B y C.
* La arista tiene exactamente la misma calificación que la red con la arista . En este punto no podríamos distinguir entre las dos opciones. Concibamos que tenemos la estructura que se muestra en la Figura 18b.
* El procedimiento voraz procede, y este decide adicionar la arista , dando como resultado la red de la Figura 18c.
* Ahora estamos listos para **invertir** la arista y mejorar la puntuación. Sin embargo, **si no tenemos la operación invertir**, un procedimiento voraz podría no eliminar la arista , ya que disminuye el puntaje.
* Podríamos terminar con la red que se muestra en la Figura 18d. Así, B y C juntos hacen mejor predicción de A que sólo C.

##### El procedimiento de búsqueda

Una vez definimos el espacio de búsqueda, necesitamos diseñar un procedimiento que lo explore y busque la estructura con mayor calificación. La gran mayoría de métodos de búsqueda usados en aprendizaje de estructuras son procedimientos de búsqueda local como por ejemplo el algoritmo “*greedy hill climbing”* (Koller & Friedman, 2009b). En el contexto del aprendizaje de estructuras:

* Tomamos una estructura de red inicial como punto de partida; esta red puede ser vacía , una elección aleatoria, un árbol, o una red obtenida de algún conocimiento previo.
* Calculamos el puntaje de la estructura de red inicial.
* Consideramos todos los vecinos de en el espacio (todas las redes validas obtenidas aplicando un operador individual a ) y calculamos el puntaje de cada uno de ellos.
* Aplicamos el cambio que dirija al mejor incremento del puntaje.
* Continuamos el proceso hasta que ninguna modificación mejore el puntaje.

Existen dos preguntas que podemos hacer:

* Primera, cual es el costo de este proceso.
* Segundo, que podemos decir acerca de la red obtenida del proceso.

##### Descomposición del puntaje (score) y Búsqueda

El factor dominante en el costo del algoritmo de búsqueda es la evaluación de redes vecinas en cada iteración del proceso. Para evaluar cada una de estas necesitamos calificarlas.

Este proceso requiere que hagamos un recorrido de todos los diferentes casos de datos, para calcular las estadísticas suficientes para la nueva estructura. Esta computación puede ser bastante costosa en cualquier algoritmo de aprendizaje de estructuras.

El punto clave es donde la **propiedad de descomposición del puntaje** (score composability property**)** llega a ser **útil**:

* Como vimos en la sección 5.5.4.2, el puntaje que analizamos se descompone en una suma de términos, uno para cada variable .
* Cada una de estas **familias de puntajes** (family scores) es calculada relativamente, solo para las variables de la familia de .
* Un cambio local (adición, eliminación ó invertir una arista) mantiene casi todas las familias de la red sin cambios.

Para las familias cuya composición no cambió, el componente asociado del puntaje tampoco cambia.

Asumamos que la actual red candidata es :

* Para cada operador, calculamos la mejora en el puntaje que podría resultar de este cambio.
* Definimos el **Delta** del puntaje

Ecuación 4 **Delta del puntaje**

Para hacer el cambio del puntaje asociado con aplicar el operador al grafo.

Usando la descomposición del puntaje, podemos calcular esta cantidad de manera relativamente eficiente.

### Estado del arte

Aprender automáticamente la estructura de un grafo de una red Bayesiana (BN) es un problema abierto en reconocimiento de patrones, que ha atraído la atención en los últimos años. Ampliamente hablando, existen dos principales enfoques para aprender estructuras de redes Bayesianas (Neapolitan, 2009) y (Darwiche, 2009).

#### Enfoque basado en restricciones

Los enfoques descritos en (Spirtes, Glymour, & Scheines, 2012), (Wermuth, 1983) y (de Campos & Huete, 2000) pertenecen a esta categoría. Las redes construidas por este enfoque son usualmente asintóticamente correctas, pero como ha sido señalado por (Cooper & Herskovits, 1992) las pruebas de independencia condicional con conjuntos (*condition-sets*) grandes puede ser poco fiable a menos que el volumen de datos sea enorme. Debido a las limitaciones en los recursos de investigación, en los experimentos de micro-arreglos el tamaño de las muestras de datos de expresión de genes es usualmente pequeño (normalmente el volumen de los datos de muestras por paciente, van de 3 a 20), por lo tanto, el enfoque incluido en esta categoría (conditional Independence tests), podría ser inadecuado para datos de expresión de genes.

#### Enfoque basado en puntaje

Las funciones (*scoring)* pueden ser formuladas basándose en diferentes principios, tales como entropía (Herskovits & Cooper, 2013), la mínima longitud de descripción (Lam & Bacchus, 1994) y puntajes bayesianos (*Bayesian scores*) (Heckerman & Chickering) y (Cooper & Herskovits, 1992). Los procedimientos de optimización son usualmente heurísticas, tales como la búsqueda tabú (Bouckaert, 1995) y computación evolutiva (de Campos & Huete, 2000) y (Neiland & Korb, 1999). Desafortunadamente la tarea de encontrar una estructura de red que optimice la función *scoring* es un problema de optimización combinatoria, y es conocido como un problema de complejidad computacional ***NP-hard*** (Chickering, 1996). Por lo tanto, el proceso de optimización frecuentemente se detiene en una estructura óptima local. Por otra parte, para muchos conjuntos de datos de expresión, el tamaño de la muestra es típicamente pequeño con respecto al tamaño de la red. En este caso, la mayoría de las veces existe un número grande de redes con alto *scoring,* sin embargo, sus estructuras son bastante diferentes, así la inferencia de características de la red como presencia o ausencia de una arista particular, basado en una sola red es de alto riesgo (Faming, 2010).

#### Modelo promedio Bayesiano

Una lista no exhaustiva de los trabajos en esta categoría incluyen (Madigan & Raftery, 1994), (Madigan, York, & Allard, 1995), (Giudici & Green, 1999) y (Koller & Friedman, 2009a), donde la simulación es realizada usando el algoritmo de Metropolis-Hasting (**MH**) (Hastings, 1970), y las características de las redes son inferidas promediando sobre un número grande de redes simuladas desde la *distribución posterior*. El promedio sobre diferentes redes puede reducir significativamente el riesgo sufrido por el procedimiento de inferencia basado en un solo modelo. Aunque el enfoque aparece atractivo, trabaja bien solo para los problemas con un número muy pequeño de variables (nodos). Esto es debido al comportamiento de la función de energía de la red bayesiana; la cual presenta altibajos con una cantidad de mínimos locales separados por altas barreras, especialmente cuando el tamaño de la red es grande (aquí la función de energía se refiere a la función de distribución *log-posterior negativa* de la red bayesiana).

Como sabemos por muchos investigadores, el algoritmo MH es propenso a quedar atrapado en mínimos locales indefinidamente en un sistema con altibajos. Para aliviar esta dificultad (Friedman & Koller, 2003) introducen un algoritmo en dos pasos: (1) Utilizar el algoritmo MH para muestrear un orden temporal de los nodos, y muestrear una estructura de red compatible con el orden dado (Para cualquier red bayesiana existe un orden temporal de los nodos, tal que, para cualquier par de nodos X, Y, si existe una arista de X a Y, entonces X debe preceder a Y en dicho orden). Paso a seguir, (2) el algoritmo mejora el mezclado sobre el espacio de estructuras de red, sin embargo, estas estructuras muestreadas no siguen la distribución posterior correcta, porque el orden temporal no induce una partición del espacio de estructuras de red. Es decir, una red puede ser compatible con más de un orden temporal.

(Faming, 2010) basado en (Liang & Zhang, 2009) describe como aprender redes bayesianas usando el algoritmo de aproximación estocástica de Monte Carlo (SAMC) (Liang, Liu, & Carroll, 2007). Una notable característica del algoritmo SAMC es que posee un mecanismo de auto-ajuste, así no queda atrapado en un mínimo local. SAMC pertenece a la clase de algoritmos de ponderación dinámica (dynamic weighting algoritms) (Wong & Liang, 1997) y (Liu, Liang, & Wong, 2001), y las muestras generadas por este algoritmo pueden usarse para evaluar características de la red por medio del estimador de ponderación dinámica. Los estimadores de ponderación dinámica pueden tener mucho más bajo riesgo que el estimador basado en un solo modelo (Faming, 2010).

## Markov Chain Monte Carlo (MCMC)

MCMC corresponde a un área de la estadística, la cual suministra una respuesta al problema de simulación de distribuciones de probabilidad de alta dimensionalidad asociadas a modelos complejos.



### Introducción

Una cadena de Markov (Markov chains) está definida en términos de los estados de un red o grafo, sobre los cuales un proceso estocástico (o Monte Carlo, es decir, no determinístico) realiza una caminata aleatoria. En muchos casos, este proceso tiende a un equilibrio y los estados siguen una distribución probabilística.

Las técnicas MCMC permiten la simulación de una distribución, embebiéndola en una distribución de una cadena de Markov, y entonces, realiza la simulación de dicha cadena hasta que tienda al equilibrio.

Un enfoque muy utilizado en el aprendizaje Bayesiano es la simulación MCMC, donde un conjunto de redes Bayesianas viables son exploradas por una caminata aleatoria que converge a una red optima, que se ajusta a los datos con respecto al puntaje (scoring) de una function de verosimilitud (likelihood) u otra function de evaluación como el puntaje (score) Bayesiano.

### Algoritmos

Metropolis **Hasting** es uno de los algoritmos MCMC el cual está basado en una cadena de Markov, cuya dependencia de los estados anteriores está dividida en dos partes: una distribución propuesta y una aceptación de la propuesta.

La distribución propuesta sugiere un paso siguiente arbitrario en la trayectoria de la cadena. En la aceptación de la propuesta se mantiene la dirección apropiada en la caminata aleatoria, rechazando los movimientos no deseados de la cadena.

Otro algoritmo muy utilizado en MCMC es el llamado “**Gibbs** sampling”, el cual está basado en una cadena de Markov cuya dependencia de los estados anteriores es gobernada por las distribuciones condicionales (CPD) que surgen a partir del modelo.

## Relaciones entre las redes de co-expresión (CGN) y las redes Bayesianas de interacción de genes (BIN)

La dirección y el tipo de relación de co-expresión no están determinadas en una red de co-expresión de genes. En este paso el tipo de redes que nos interesa construir son las redes bayesianas de interacción de genes (**BIN**), en las cuales una arista dirigida conecta dos genes, representando un proceso bioquímico tal como una reacción, transformación, **interacción**, activación o inhibición.

Comparada con una GCN, una BIN intenta inferir las relaciones de causalidad entre genes, y las aristas representan no solo una co-relación o relación de dependencia entre ellos. Los módulos o sub-grafos altamente conectados en las redes de co-expresión de genes (GCN) corresponden a agrupamientos (*módulos* / *cluster*) de genes que tienen una función similar o se involucran en un proceso biológico común, el cual causa muchas interacciones entre ellos mismos (Roy et al., 2014).

En la **Figura 19,** la dirección de las aristas es obviada en la co-expresión. Mientras los tres genes X, Y y Z están co-expresados, no se determina si X activa Y y Y activa Z o Y activa X y Z u otro gen activa a los tres.

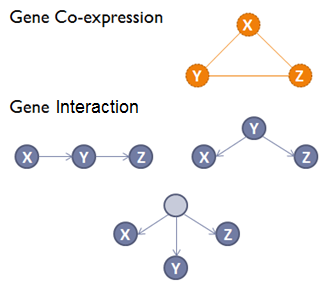


Figura 19 Red de co-expresión de genes (CGN) vs. red bayesiana de interacción de genes (BIN)

Los genes con perfiles de expresión similar normalmente tienen funciones relacionadas. En lugar de analizar pares de genes, se podrían estudiar conjuntos de genes co-expresados que se sabe que interactúan como un módulo funcional envuelto en un proceso biológico particular(M. M. Babu, Luscombe, Aravind, Gerstein, & Teichmann, 2004). Integrando esta información con la conservación evolutiva de las proteínas en más de dos organismos, provee conocimiento de la significancia de los módulos funcionales que han sido conservados en la evolución.

(Neapolitan, 2009) muestra como un módulo de redes (*module networks*) puede ser empleado para analizar datos de expresión de genes, y también los resultados de experimentos de esta utilización.

Los datos de expresión de genes son usados para inferir relaciones regulatorias. Este enfoque es conocido como **ingeniería inversa de redes regulatorias**. A partir de datos de expresión se pueden hacer predicciones de los *reguladores transcripcionales* para un gen o conjunto de genes dados. En(Segal et al., 2003), desarrollaron un modelo probabilístico para identificar módulos de genes co-regulados, sus *reguladores transcripcionales* y condiciones que influencian la regulación, permitiendo la generación hipótesis validables experimentalmente.

En (Gardner, di Bernardo, Lorenz, & Collins, 2003) describieron un método llamado NIR (Network Identification by multiple Regression) para inferir relaciones regulatorias, el cual usa ecuaciones diferenciales no-lineales para modelar redes regulatorias. En este método, se infiere un modelo de conexiones entre genes en una red a partir de mediciones de la dinámica del sistema (por ejemplo, respuesta de genes y proteínas a perturbaciones).

# Bibliografía

Babu, M. (2004). An introduction to Microarray Data Analysis. In R. Grant (Ed.), *Computational Genomics: Theory and Applications*: Horizont Press.

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., & Teichmann, S. A. (2004). {S}tructure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol., 14*(3), 283-291.

Blair, R. H., Kliebenstein, D. J., & Churchill, G. A. (2012). {W}hat can causal networks tell us about metabolic pathways? *PLoS Comput. Biol., 8*(4), e1002458-e1002458.

Blair, R. H., Trichler, D. L., & Gaille, D. P. (2012). {M}athematical and statistical modeling in cancer systems biology. *Front Physiol, 3*, 227-227.

Bouckaert, R. R. (1995). *Bayesian belief networks: from construction to inference*

*Bayesiaanse belief netwerken: van constructie tot inferentie.* Universiteit Utrecht, Faculteit Wiskunde en Informatica, Utrecht.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*: Wiley.

Cheng, Y., & Church, G. M. (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol, 8*, 93-103.

Chickering, D. (1996). Learning Bayesian Networks is NP-Complete. In D. Fisher & H.-J. Lenz (Eds.), *Learning from Data* (Vol. 112, pp. 121-130): Springer New York.

Consortium, T. G. (2009a). ["AmiGO: P68032 Associations"].

Consortium, T. G. (2009b). [&quot;gene\_ontology.1\_2.obo&quot;].

Consortium, T. G. (2014). "Gene Ontology downloads at SourceForge".

Consortium, T. G. (2015). [Download Ontology].

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*(4), 309-347. doi:10.1007/BF00994110

Cowell, R. (1998). *Introduction to inference for Bayesian networks*. Paper presented at the Proceedings of the NATO Advanced Study Institute on Learning in graphical models, Erice, Italy.

Darwiche, P. A. (2009). *Modeling and Reasoning with Bayesian Networks*: Cambridge University Press.

Davidson, R. M., Hansey, C. N., Gowda, M., Childs, K. L., Lin, H., Vaillancourt, B., . . . Buell, C. R. (2011). Utility of RNA Sequencing for Analysis of Maize Reproductive Transcriptomes. *The Plant Genome, 4*(3). doi:10.3835/plantgenome2011.05.0015

de Campos, L. M., & Huete, J. F. (2000). A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning, 24*(1), 11-37. doi:<http://dx.doi.org/10.1016/S0888-613X(99)00042-0>

Dey, D. K., Ghosh, S., & Mallick, B. K. (2010). *Bayesian Modeling in Bioinformatics*: CRC Press.

Dong, J., & Horvath, S. (2007). Understanding network concepts in modules. *BMC Syst Biol, 1*, 24. doi:10.1186/1752-0509-1-24

Faming, L. (2010). Learning Bayesian Networks for Gene Expression Data. In *Bayesian Modeling in Bioinformatics* (pp. 271-291): Chapman and Hall/CRC.

Friedman, N., & Koller, D. (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning, 50*(1-2), 95-125. doi:10.1023/A:1020249912095

Gardner, T. S., di Bernardo, D., Lorenz, D., & Collins, J. J. (2003). {I}nferring genetic networks and identifying compound mode of action via expression profiling. *Science, 301*(5629), 102-105.

Giudici, P., & Green, P. (1999). Decomposable graphical Gaussian model determination. *Biometrika, 86*(4), 785-801. doi:10.1093/biomet/86.4.785

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika, 57*(1), 97-109. doi:10.2307/2334940

Heckerman, D., & Chickering, D. M. (1995). *Learning Bayesian networks: The combination of knowledge and statistical data*.

Herskovits, E., & Cooper, G. F. (2013). Kutato: An Entropy-Driven System for Construction of Probabilistic Expert Systems from Databases. *CoRR, abs/1304.1088*.

Hollender, C. A., Kang, C., Darwish, O., Geretz, A., Matthews, B. F., Slovin, J., . . . Liu, Z. (2014). Floral Transcriptomes in Woodland Strawberry Uncover Developing Receptacle and Anther Gene Networks. *Plant Physiol, 165*(3), 1062-1075. doi:10.1104/pp.114.237529

Horvath, S., & Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol, 4*(8), e1000117. doi:10.1371/journal.pcbi.1000117

Koller, D., & Friedman, N. (2009a). *Probabilistic graphical models : principles and techniques*. Cambridge, MA: MIT Press.

Koller, D., & Friedman, N. (2009b). *Probabilistic graphical models : principles and techniques*. Cambridge, MA: MIT Press.

Korb, K. B., & Nicholson, A. E. (2010). *Bayesian Artificial Intelligence, Second Edition* (2nd ed.). Boca Raton, FL, USA: CRC Press, Inc.

Lam, W., & Bacchus, F. (1994). LEARNING BAYESIAN BELIEF NETWORKS: AN APPROACH BASED ON THE MDL PRINCIPLE. *Computational Intelligence, 10*(3), 269-293. doi:10.1111/j.1467-8640.1994.tb00166.x

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*(1), 1-13. doi:10.1186/1471-2105-9-559

Liang, F., Liu, C., & Carroll, R. J. (2007). Stochastic Approximation in Monte Carlo Computation. *Journal of the American Statistical Association, 102*(477), 305-320. doi:10.1198/016214506000001202

Liang, F., & Zhang, J. (2009). Learning Bayesian networks for discrete data. *Computational Statistics & Data Analysis, 53*(4), 865-876. doi:<http://dx.doi.org/10.1016/j.csda.2008.10.007>

Liu, J. S., Liang, F., & Wong, W. H. (2001). A Theory for Dynamic Weighting in Monte Carlo Computation. *Journal of the American Statistical Association, 96*(454), 561-573. doi:10.1198/016214501753168253

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association, 89*(428), 1535-1546.

Madigan, D., York, J., & Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, 215-232.

Mardis, E. R. (2013). {N}ext-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif), 6*, 287-303.

Metzker, M. L. (2010). {S}equencing technologies - the next generation. *Nat. Rev. Genet., 11*(1), 31-46.

Neapolitan, R. E. (2009). *Probabilistic Methods for Bioinformatics: With an Introduction to Bayesian Networks*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Neiland, J., & Korb, K. (1999). The Evolution of Causal Models: A Comparison of Bayesian Metrics and Structure Priors. In N. Zhong & L. Zhou (Eds.), *Methodologies for Knowledge Discovery and Data Mining* (Vol. 1574, pp. 433-438): Springer Berlin Heidelberg.

Nguyen, D. V., Bulak Arpat, A., Wang, N., & Carroll, R. J. (2002). DNA Microarray Experiments: Biological and Technological Aspects. *Biometrics, 58*(4), 701-717. doi:10.1111/j.0006-341X.2002.00701.x

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Pinoli, P., Chicco, D., & Masseroli, M. (2015). Computational algorithms to predict Gene Ontology annotations. *BMC Bioinformatics, 16*(Suppl 6), S4-S4. doi:10.1186/1471-2105-16-S6-S4

Prestat, E., de Morais, S. R., Vendrell, J. A., Thollet, A., Gautier, C., Cohen, P. A., & Aussem, A. (2013). Learning the local Bayesian network structure around the ZNF217 oncogene in breast tumours. *Computers in Biology & Medicine, 43*(7), 334-341. doi:10.1016/j.compbiomed.2012.12.002

Prifti, E., Zucker, J. D., Clément, K., & Henegar, C. (2010). Interactional and functional centrality in transcriptional co-expression networks. *Bioinformatics, 26*(24), 3083-3089. doi:10.1093/bioinformatics/btq591

Roy, S., Bhattacharyya, D. K., & Kalita, J. K. (2014). Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics, 15 Suppl 7*, S10. doi:10.1186/1471-2105-15-S7-S10

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). {M}odule networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet., 34*(2), 166-176.

Spirtes, P., Glymour, C., & Scheines, R. (2012). *Causation, Prediction, and Search*: Springer New York.

The Gene Ontology, C. (2008). The Gene Ontology project in 2008. *Nucleic Acids Research, 36*(Database issue), D440-D444. doi:10.1093/nar/gkm883

Tözeren, A., & Byers, S. W. (2004). *New Biology for Engineers and Computer Scientists*: Pearson/Prentice Hall.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., . . . Morris, Q. (2010). {T}he {G}ene{M}{A}{N}{I}{A} prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res., 38*(Web Server issue), W214--220.

Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., & Losick, R. (2013). *Molecular Biology of the Gene*: Pearson Education.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of /`small-world/' networks. *Nature, 393*(6684), 440-442.

Wermuth, N. a. L., S. (1983). Graphical and recursive models for contingency tables. *Biometrika, 70*(3), 537-552. doi:10.1093/biomet/70.3.537

Wong, W. H., & Liang, F. (1997). Dynamic weighting in Monte Carlo and optimization. *Proceedings of the National Academy of Sciences, 94*(26), 14220-14224.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., . . . Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature, 500*(7464), 593-597. doi:10.1038/nature12364

Young, M., McCarthy, D., Wakefield, M., Smyth, G., Oshlack, A., & Robinson, M. (2012). Differential Expression for RNA Sequencing (RNA-Seq) Data: Mapping, Summarization, Statistical Analysis, and Experimental Design. In N. Rodríguez-Ezpeleta, M. Hackenberg, & A. M. Aransay (Eds.), *Bioinformatics for High Throughput Sequencing* (pp. 169-190): Springer New York.

Yu, H., Luscombe, N. M., Qian, J., & Gerstein, M. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet, 19*(8), 422-427. doi:10.1016/S0168-9525(03)00175-6

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol, 4*, Article17. doi:10.2202/1544-6115.1128

# Anexo

Producto de los resultados preliminares de experimentación con aprendizaje Bayesiano, se tiene aprobado un artículo en el “9th International Symposium on **Biomathematics** and Ecology Education and Research-2016”, patrocinado por Illinois State University. A continuación, se anexa el Resumen del mismo:

**Strategies to avoid overfitting of MCMC Bayesian learning in some biological applications**

**Abstract**

Model learning from observed data is typically affected by overfitting, because in order to find the model’s best parameter set, all relations between data are used indifferently whether they represent relevant or noisy interactions.

Bayesian networks are widely used in biological modeling (e.g. networks of gene interactions), given that they allow representing graphically and determining statistically the dependence /independence relations between considered variables. A frequent approach in Bayesian learning is Markov Chain Monte Carlo simulation (MCMC), where a set of viable networks are explored by a random walk which converges to a network fitted optimally to data with respect to the likelihood or similar evaluation function.

Here we propose various strategies to mitigate overfitting in Bayesian learning by MCMC in order to reduce the resulting models’ complexity. They either apply constraints inside the MCMC simulation or consider post-optimal operations. We show the effectiveness of these strategies in some biological applications.

**Key words**

Bayesian networks, Bayesian learning, MCMC simulation, overfitting.