Capítulo 10

Aprendizaje Dinámico de Redes Bayesianas, Aplicado a Redes Regulatorias Genéticas

Beatriz C. Luna Olivera[1]
Eduado Ortiz Hernández[2]
Eduardo Sánchez Soto[3]
Jessica M. Hernández Martínez[4]

Abstract: In the last decades genetic regulatory networks have been widely studied, these are built considering the complex interactions between genes and proteins; one of the problems in this area is to decide which network correspond to the obtained data, this is known as network learning. There are various models used to represent a regulatory network, Bayesian networks being one of them. Bayesian networks are of two types, static or dynamic, the dynamic ones, in comparison to the static, have been studied very little in the learning of gene expression data. The information theory score function, known as mutual information introduce dynamics in Bayesian network learning. This measure has the possibility of building a statistical test of independence based on the distribution X^2 that penalizes in the dynamic Bayesian network the degrees of interaction between each variable and their parents. In this paper we use the test mutual information for a dynamic Bayesian network by means of an algorithm. In particular it is applied to the Arabidopsis thaliana flower

[1]bcluna@unpa.edu.mx. Laboratorio de Cómputo Científico y Matemáticas Aplicadas, Universidad del Papaloapan.

[2]eortix@unpa.edu.mx. Laboratorio de Cómputo Científico y Matemáticas Aplicadas, Universidad del Papaloapan.

[3]esanchez@unpa.edu.mx. Laboratorio de Cómputo Científico y Matemáticas Aplicadas, Universidad del Papaloapan.

[4]jmhernandez@unpa.edu.mx. Laboratorio de Cómputo Científico y Matemáticas Aplicadas, Universidad del Papaloapan.

regulatory network, responsible for morphogenesis genetic control of it. At the end is presented the analysis of information obtained.

Keywords: Genetic regulatory networks, network learning, dinamic bayesian networks, mutual information test.

Resumen: En las últimas décadas las redes regulatorias genéticas han sido ampliamente estudiadas, éstas se construyen considerando las interacciones complejas entre los genes y las proteínas; uno de los problemas en esta área es poder decidir a qué red corresponden los datos obtenidos, esto se conoce como aprendizaje de redes. Existen diversos modelos utilizados para representar una red regulatoria, siendo las redes bayesianas uno de ellos. Las redes bayesianas son de dos tipos, estáticas o dinámicas, las dinámicas a comparación de las estáticas han sido muy poco estudiadas dentro del aprendizaje de los datos de expresión de los genes. La función de puntuación teórica conocida como prueba de información mutua introduce la dinámica en el aprendizaje de una red bayesiana. Esta medida tiene la posibilidad de construir una prueba estadística de independencia basada en la distribución X^2 que sirve para penalizar dentro de la red bayesiana dinámica los grados de interacción entre cada variable y sus variables padres. En este trabajo usamos la prueba de información mutua para obtener una red bayesiana dinámica mediante un algoritmo. En particular se aplica a la red regulatoria de la flor Arabidopsis Thaliana encargada del control genético de la morfogénesis de la flor. Se presenta al final el análisis de la información obtenida.

Palabras clave: redes regulatorias genéticas, aprendizaje de redes, redes bayesianas dinámicas, prueba de información mutua.

10.1 Introducción

La biología de sistemas utiliza las áreas de biología clásica, matemáticas y computación para crear y entender el modelo de un sistema biológico mediante las técnicas y conceptos de esas áreas. La complejidad de los procesos celulares y la gran cantidad de datos de un sistema biológico obtenidos de los experimentos realizados en el laboratorio, hacen difíciles las tareas de modelar, simular y analizar matemáticamente o computacionalmente el sistema [Ideker2001]. El desarrollo de un modelo para un sistema biológico sirve para facilitar el estudio de la dinámica y las propiedades de procesos celulares que posee, mediante algún método o función de aprendizaje.

10.2 Redes Regulatorias Genéticas

El cromosoma de los organismos sirve para transportar biológicamente características hereditarias; es una macromolécula consistente de cadenas de ácido desoxirribonucleico (ADN). Las cadenas de ADN contienen información codificada para producir cada proteína o molécula de ácido ribonucléico (ARN) presente en un organismo. Cuando una cadena está dividida en pequeños fragmentos (conocidos como genes), su función dentro de un organismo consiste en obtener la información necesaria para construir una molécula de ARN o una proteína. Una proteína sintetizada contiene una región codificada de ADN, la cual puede funcionar como un factor de transcripción para algunos sitios de ADN.

Cuando la célula de un organismo reconoce el inicio de un gen o varios genes para construir una molécula de ARN o proteína, la sección de ADN llamada promotor indica que se debe iniciar una transcripción de ADN a ARN; a este proceso se conoce como regulación de transcripción de ADN. Posteriormente cuando se reconoce el inicio de un gen éste se copia en una molécula de ARN, lo que resulta de este proceso de regulación del ARN y transporte se conoce como ARNm, el ARNm también es conocido como ARN mensajero. El proceso de regulación de translación de ARN convierte el código de ARN en una proteína. Para que se pueda tener una proteína modificada es necesario pasar por el proceso de regulación de modificación de proteína. Todo este conjunto de procesos es importante dentro de la biología molecular [Hidde2000].

La Red Regulatoria Genética sirve para representar las interacciones de los genes en una estructura, que se puede modelar como un grafo dirigido compuesto por nodos (genes) y arcos (interacciones). Cada interacción entre los genes se comporta como una activación (induce la transcripción con otros genes) o inhibición (reprime la actividad de transcripción).

10.3 Modelos gráficos

Un modelo gráfico es una herramienta que se utiliza frecuentemente para resolver problemas en matemáticas aplicadas e ingeniería [Murphy1998]. La representación de los modelos gráficos probabilísticos se realiza por medio de grafos constituidos por nodos (las variables aleatorias) y arcos (la independencia condicional entre las variables). Los modelos gráficos dirigidos, a comparación de los no dirigidos tienen una representación que puede interpretarse como causalidad. También, y debido a su aplicación en la genética, se dice que un nodo es padre de otro llamado hijo si existe una arco dirigido partiendo del primero y terminando en el segundo. Además de la estructura del modelo es necesario especificar la distribución de probabilidades, ya sea continua o discreta.

10.3.1 Redes bayesianas

Una red bayesiana consta de tres elementos: un conjunto de *variables aleatorias* $X = \{x_1, ..., x_n\}$, un *grafo dirigido acíclico* (GDA) $GDA = (X, A)$ (donde A son los arcos (o aristas) entre cada variable X_i) y una *distribución de probabilidad* sobre X, representada en la red bayesiana como $P(X)$ [Murphy1999], que gracias a las independencias condicionales puede ser factorizada de la siguiente forma:

$$P(X) = \prod_i P(x_i \mid pa(X_i)),$$

donde las P(x_i | pa(X_i)) son las probabilidades condicionales de P(X) y pa(X_i) representa todos los padres que puede tener la variable x_i.

10.3.2 Redes bayesianas dinámicas

Las redes bayesianas tal como se han presentado hasta este momento carecen de la capacidad de representar aspectos temporales en series de tiempo. Una red bayesiana dinámica (RBD) es un modelo gráfico que describe un sistema que cambia dinámicamente o evoluciona a través del tiempo [Mihajlovic2001]. Los espacios de tiempo de una RBD son interconectados con relaciones representadas por arcos (punteados en este trabajo) entre variables específicas en los cuales se observa la dinámica de la red. Las variables del modelo son denotadas como los estados de tiempo de la RBD, los estados del tiempo t dependen tanto de su estado anterior, conocido como el tiempo t_1, como de los tiempos anteriores t_2, t_3, ...,t_n. Cabe mencionar que los nodos, arcos y la probabilidad de la interpretación estática de la red bayesiana también se ven reflejados en la RBD. Como el modelo de la red bayesiana es "expandido" mediante una RBD, nosotros sólo podemos conocer algunos valores de probabilidad condicional de las dependencias e independencias entre las variables para el primer tiempo, siempre y cuando tengamos algunas observaciones del modelo estático, pero para encontrar las relaciones de independencia entre las demás variables se deben medir o calcular los estados anteriores. Las relaciones temporales son incorporadas en tablas de probabilidad condicional entre cada variable para cada espacio de tiempo.

10.3.3 Aprendizaje en redes bayesianas

El aprendizaje en redes bayesianas puede ser una alternativa para el modelado de la red regulatoria genética. Para poder saber cuáles son los genes que interactúan en la representación del modelo de una red regulatoria genética es necesario llevar a cabo un aprendizaje basado en los datos disponibles. El proceso de aprendizaje para obtener una red bayesiana a partir de los datos, conlleva elegir uno o ambos de los siguientes aspectos:

- **Aprendizaje estructural:** A partir de los datos se establecen las independencias condicionales entre los genes.
- **Aprendizaje paramétrico:** Dada la estructura, se obtienen las probabilidades asociadas.

10.4 Prueba de información mutua

La mayoría de los trabajos dentro del área del aprendizaje de modelos gráficos utilizan medidas Bayesianas para calcular la calidad de un red. Estas medidas tienen la desventaja de no poder medir correctamente los ciclos y sobre todo la dirección de las interacciones. Como alternativa se puede utilizar la información mutua, que se aplica en este trabajo y se prueba en las redes regulatorias genéticas. Formalmente, la prueba de información mutua $MI_D(X_i,Pa_G(X_i))$ sirve para medir el grado de interacción entre cada variable aleatoria X_i y sus padres $Pa_G(X_i)$ si y sólo si ellos son independientes, pero penalizando su valor por medio de un término relativo a la distribución X^2. El desarrollo de la prueba de información mutua se basa en el tratamiento de la información mutua proporcionada por cada nodo de la red, con la finalidad de indagar qué información comparte un nodo con otro. Verifica si la información extraída de la red es relevante para el problema que se esté modelando [Campos2006], [Xuan2012]. Las funciones de puntaje representan una opción para medir los grados de conveniencia de un GDA con un conjunto de datos; es decir, se encargan de reducir el número de variables necesarias para representar una RBD dependiendo de la probabilidad de las variables que influyen en el GDA. La prueba de información mutua ha sido considerada para el aprendizaje de redes bayesianas pero no en su totalidad para el aprendizaje de redes bayesianas dinámicas [Campos2006], [Nguyen2011] y [Xuan2012].

La siguiente ecuación representa la prueba de información mutua MIT para redes bayesianas estáticas:

$$SS_{MIT}(G:D) = \sum_{i=1,Pa_i\neq 0}^{n} \{2N, I(X_i, Pa_i) - \sum_{j=1}^{s_i} X_\alpha, l_i\sigma i(j)\}$$

La prueba de información mutua indica que tan "bueno" o "malo" es el modelo que nosotros proponemos respecto a los datos con los que contamos. La forma en que la prueba de información mutua trabaja el aprendizaje del modelo de una RBD, consiste en discretizar el tiempo y crear una réplica de cada variable aleatoria para cada punto temporal de ese modelo. Para realizar la discretización del tiempo se utiliza un método condicionado por el fenómeno en estudio y a las condiciones prácticas de laboratorio. Las muestras se tomaron cada hora debido

a que el tiempo de duplicación de los genes es aproximadamente de 20 min y que se requiere de un tiempo similar para llevar a cabo la toma de las muestras y realizar las mediciones correspondientes. El modelo dinámico debe tener al menos dos muestras en instantes de tiempo consecutivos.

10.4.1 Ji cuadrada o X^2

X^2 es una prueba usada en problemas estadísticos, aplicable a variables aleatorias (o estocásticas) discretas o continuas dependiendo del contexto en el que se esté trabajando, para evaluar hipótesis acerca de dos variables categóricas.

Cuando se hace uso de esta prueba estadística se requiere un nivel de significación, el cual indica que el resultado de éste va a ser satisfactorio o no satisfactorio dependiendo del número de redes que propongamos. El nivel de significación es comúnmente representado por el símbolo griego α (alfa). Para nuestras pruebas usaremos un valor de α de 0.05, ya que es un valor aceptable para realizar pruebas.

10.5 Metodología

10.5.1 Estructura del modelo

Partimos de la red regulatoria compuesta por 10 genes que regulan la morfogénesis de la flor Arabidopsis y de los datos reportados en [Mendoza1999] (ver Figura 10.1).

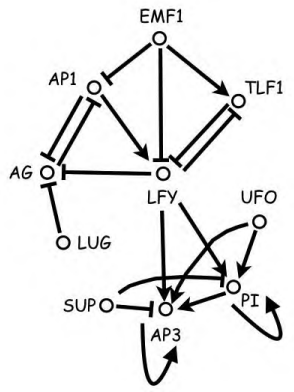

Figura 10.1: Red regulatoria de Arabidopsis.

La construcción de las estructuras gráficas se realizó de la siguiente forma: En la Figura 10.1 se observa que sólo existen 10 nodos (equivalentes a 10 genes

de la flor) que interactúan entre sí. La representación utilizada para los nodos equivale a las variables que van desde x_1 hasta x_{10}, siendo una red bayesiana estática. Para realizar esta estructura como una RBD se representó con otra red bayesiana con mismo número de nodos e interacciones con variables que van desde x_1 hasta x_{10}. Finalmente la parte dinámica se representa uniendo las dos redes bayesianas estáticas. Una parte importante del análisis de la estructura fue tomar algunas decisiones que nos permitieran modelar el conocimiento a posteriori de la RBD. La primera observación que se realizó dentro de la estructura fue que existen relaciones cíclicas, las cuales es necesario eliminar, traspasándolas en el modelo dinámico, y en sus diferentes combinaciones generan un total de 4 redes bayesianas dinámicas (ver Figura 10.2).

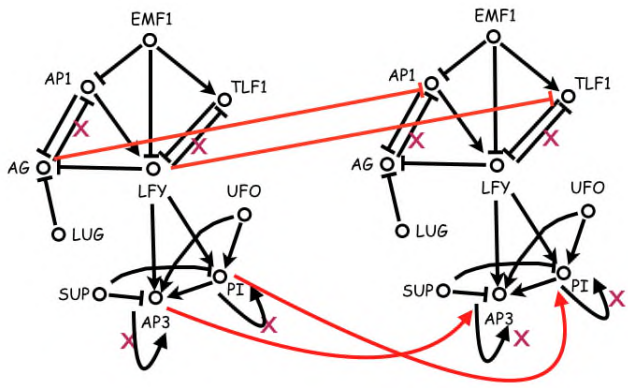

Figura 10.2: Propuesta de red bayesiana dinámica para Arabidopsis.

10.5.2 Prueba de información mutua

A partir de los valores presentados en las tablas de [Mendoza1999], se duplican los valores de tal manera que se establecen una secuencia dinámica en los mismos. Esto es, en cada ranura de tiempo se agrega en una columna extra los datos del instante siguiente. Con estos últimos, se modelarán la dinámica de la gráfica de la derecha en 10.2. Con éstos datos, la prueba fue aplicada a las cuatro redes bayesianas dinámicas posibles. Se calcularon los puntajes aplicando la toolbox de Matlab de [Vinh2011] y [Nguyen2011], cuyos resultados se muestran en la Figura 10.3. El puntaje de la red 1 es de 808.82, seguido muy de cerca por la red 3, con un puntaje de 806.79; con una distancia mayor en puntaje se encuentra la red 4 con 608.76 y la red 2 con 606.76. De forma general, en este trabajo, se probaron el conjunto de redes propuestas y la que obtuvo el mayor puntaje se considera la mejor (red 1).

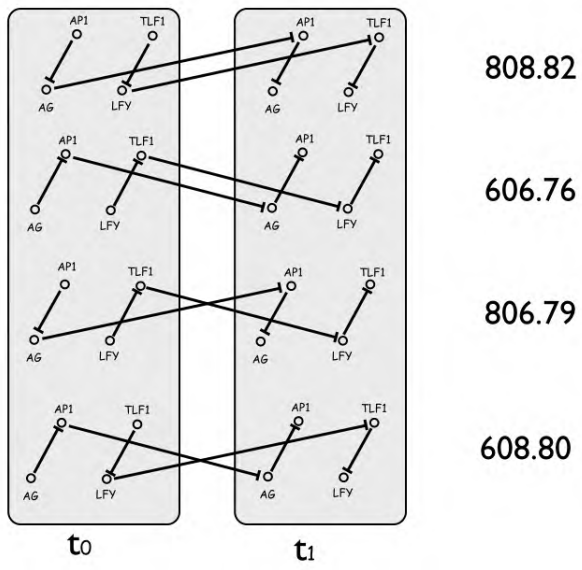

Figura 10.3: Cálculo de MIT para las redes bayesianas dinámicas propuestas.

10.6 Conclusiones

En este trabajo utilizamos las relaciones de independencia condicional en instantes de tiempo, el aprendizaje de redes y la comprobación de éstas. La prueba de información mutua en conjunto a las redes bayesianas permiten utilizar la causalidad en el modelo gráfico, adecuando la estructura de la red bayesiana estática a la RBD para obtener su aprendizaje. El aporte de este trabajo es una propuesta de una estructura dinámica a partir de datos obtenidos de laboratorio de la red regulatoria genética que controla la morfogénesis de la flor Arabidopsis thaliana.

La interpretación de los resultados desde el punto de vista de la dinámica de los procesos biológicos es un trabajo en proceso del cual se publicarán sus resultados en trabajos futuros. Sin embargo, se puede inferir que los cambios en la estructura dinámica podrán modelar de mejor manera los procesos biológicos que son en sí mismos dinámicos. Tiene sentido proponer nuevas conexiones para después ser examinadas con experimentos de laboratorio.

Bibliografía

[Campos2006] De Campos Luis M. **A scoring function for learning bayesian networks based on mutual information and conditional independence tests.** Machine Learning. Vol.7.

[Coen1991] Coen E. S., Meyerowitz E. M. **The war of the whorls: genetic interactions controlling flower development**. Nature, 353, 31-37.

[Hidde2000] Hidde de Jong. 2000. **Modeling and Simulation of Genetic Regulatory.** Institute National de Recherche en informatique et en automatique.

[Ideker2001] Ideker T.et al. 2001. **A new approach to decoding life: systems biology.** Vol.2:343-372

[Mendoza1999] Mendoza Luis, Thieffry Dennis, R. Elena, Alvarez-Buylla (1999). **Genetic Control of Flower Morphogenesis in Arabidopsis Thaliana: a logical analysis**. Bioinformatics, vol 15, nos 7/8 11, pp 593-606,

[Mihajlovic2001] Mihajlovic V. and Petkovic M. 2001. **Dynamic Bayesian Networks: A State of the Art.** Computer Science Departament. University of Twente The Netherlands.

[Murphy1998] A Brief Introduction to Graphical Models and Bayesian Networks *http://www.cs.ubc.ca/ murphyk/Bayes/bnintro.html* (2 de agosto 2014).

[Murphy1999] Murphy K. y Mian S. 1999. **Modelling Gene Expression Data using Dynamic Bayesian Networks.**

[Nguyen2011] Vinh Nguyen.2011. **The GlobalMIT Toolkit for Learning Optimal Dynamic Bayesian Network User Guide.** Monash University, Victoria, Australia.

[Vinh2011] Vinh, N. X., Chetty, M., Coppel, R., y Wangikar, P. P. (2011). **GlobalMIT: Learning Globally Optimal Dynamic Bayesian Network with the Mutual Information Test (MIT) Criterion**, Bioinformatics, DOI: 10.1093/bioinformatics/btr457.

[Xuan2012] Xuan, Nguyen, Chetty, Madhu, Coppel, Ross y Wangikar, Pramod. **Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network**, BMC Bioinformatics, Vol 13, 2012, num 1, pp 131, DOI 10.1186/1471-2105-13-131, ISSN 1471-2105.