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

September 15, 2018

Resumen

Se escribirá al final

Introducción

Sintesis del problema que se resuelve, estado del arte la solución propuesta y los resultados y conclusiones. Breve resumen de los capitulos y su organización.

1 Marco teórico y estado del arte

1.1 Redes de co-expresión génica

1.1.1 Introducción

- Definición de gén, transcriptomas y proteomas (dogma central de la biología).
- Micro-arreglos (definición y caracterización del experimento de medición de niveles de expresión usando micr-arreglos)
- Definición de red o grafo
- Que son las redes de co-expresión génica.

1.1.2 Representación de las redes de co-expresión génica

Matrices de adyacencia simétrica, representación gráfica (grafos no-dirigidos)

1.1.3 Métodos de construcción de redes de co-expresión génica

Coeficiente de pearson, Coeficiente de Kendall, umbral de significancia, normalización de la matriz de adyacencia

1.1.4 Análisis de redes complejas

- Principales propiedades y metricas para el anàlisis de redes complejas (grado, conectividad, intermediación y coeficiente de clustering [8]).
- Definición de clusters de nodos y nodos HUB

1.1.5 Trabajos relacionados

[9], et. al. Definición dendogramas y mapas de calor

1.2 Redes Bayesianas

1.2.1 Introducción

- Defición de variable aleatoria y función de distribución de probabilidad (caso discreto).
- Definición de probabilidad condicional y teorema de Bayes.
- Definición de prior, posterior y verosimilitud.
- Modelo estadístico (multinomial-dirichlet)
- Simulación MCMC
- Que son las redes Bayesianas

1.2.2 Representación de las redes Bayesianas.

Matrices de adyacencia no simétrica, representación gráfica (grafos dirigidos)

1.2.3 Métodos de aprendizaje de redes Bayesianas y sus parámetros

Método basado en puntaje, método basado en restricciones, método de simulación MCMC.

1.2.4 Trabajos relacionados

Simulación MCMC [7],[2], et. al.

2 Adopción y caracterización de un método para la construcción y visualización de redes de co-expresión génica.

Tomar como ejemplo la aplicación de WGCNA con datos de expresión de E. coli para ilustrar los conceptos

2.1 Caracterización de construcción de una red de co-expresión génica usando WGCNA

- Definición de las funciones de similaridad y no-similaridad y sus parámetros
- Construcción de la red de co-expresión génica
- Identificación de clusters de nodos
- Identifiación de nodos HUB

2.2 Visualización de una red de co-expressión génica

Ilustrar la red usando el paquete de software Cytoscape

2.3 Discusión

3 Caracterización y desarrollo de la construcción de un método de aprendizaje de redes Bayesianas y sus parámetros a partir de datos de expresión génica.

3.1 Caracterización de un método de aprendizaje de redes Bayesianas y sus parámetros

Tomar como ejemplo el problema de cancer de pulmon para ilustrar el método.

- Caracterización del algoritmo de Metropolis Hasting [4]
- Criterio de aceptación basado en verosimilitud (distribución polinomial)
- Cálculo y definición del espacio de busqueda para la caminata aleatoria (formula matematica de potencias de la matriz de adyacencia)
- Análisis de resultados alrededor del sobre-entrenamiento
- Estrategias para evitar el sobre-entrenamiento (restricción del grado en los nodos y definición de casi-independencia)

- Promediando las redes optimas y sus puntajes por medio de la función de verosimilitud (representación por matriz de adyacencia binaria)
- Discusión de resultados

3.2 Desarrollo de la construcción de un método de aprendizaje de redes Bayesianas

Especificación de los algoritmos, estructuras y tecnologias empleadas en el desarrollo (lenguajes de programación etc.)

3.3 Discusión

4 Análisis comparativo entre las redes de co-expresión génica y las redes Bayesianas de interacción génica para obtener un conjunto de genes de interes biológico.

4.1 Relación entre las redes las redes Bayesianas y las redes de co-expresión génica

- Significado e interpretación de co-relación en redes co-reladas y co-expresión en redes de co-expresión génica.
- Significado e interpretación de dependencia entre nodos en redes Bayesianas y redes Bayesianas de interacción de genes.
- Análisis comparativo entre un cluster de nodos co-expresados en una red de co-expresión génica y las dependencias del mismo cluster de nodos en una red Bayesiana de interacción de genes.
- Enriquesimiento de la red Bayesiana análizada empleando bases de datos de Gene Onlotoly (GO) para agregar caracteristicas (features) y funciones genicas a los nodos o genes de estudio modelados por la red Bayesiana.
- Constucción del conjunto de los genes de interes biológico seleccionando los genes correspondientes a los nodos en las raices del grafo correspondiente a la red Bayesiana empleada en el análisis.
- Método de promedio de redes Bayesianas optimas empoleado para el análisis comporatico con las redes de co-expressión génica.

4.2 Que pueden las redes Bayesianas y las redes de co-expresión génica decirnos acerca de genes de interes biológico

Selección de los genes de interes biológico tomando como criterio los nodos padre o raiz, a partir de las redes Bayesiansa promediadas.

4.3 Comentarios conlcuyentes

5 Caso de estudio | Case of study : Application to Lactose Transport in E. coli

5.1 Introducción

- Sistema regulatorio de genes y la red de regulación de E. coli (Atlas bacteriano de UNAM [5])
- El cluster de transporte de lactosa de E. coli

5.2 Flujo de trabajo

5.2.1 Selección de la base de datos de expresión de micro-arreglos

GEO, ArrayExpress, Colombos, M3D, et. al.

5.2.2 Construcción de la red de co-expresión génica para el genoma completo de E. coli

Se aplica al dataset seleccionado de micro-aareglo el WGCNA (Emplear Librerias R).

5.2.3 Método de selección del conjunto de genes de estudio

Especifico (juicio de experto) y general (selección de un cluster por análisis de redes complejas aplicado a una red de co-expresión génica de E. coli modelada por WGCNA)

5.2.4 Construcción de la red Bayesiana de interacción de genes para el conjunto de genes del cluster X,Y,Z de E. coli seleccionado.

Se aplica el aprendizaje de redes bayesianas al conjunto datos de expresión de genes del clusster seleccionado posterior a la discretización empleando el método de quantiles (libreria R) y graficación de las redes usando Cytoscape y RCytoscape.

5.2.5 Análisis comparativos entre las relaciones de co-expresión en la red de co-expresión génica y las relaciones de dependencia en la red Bayesian de interacció de genes.

Selección del conjunto de genes de interes biologico usando el criterio de las raices del grafo correspondiente a la red Bayesiana empleada en el análisis.

5.2.6 Discusión de resultados

5.3 Caso de estudio

5.3.1 Aplicación del flujo de trabajo al cluster de lactosa en e. coli

Utilizar la base de datos M^{3D}[3], y corroborar los resultados con las bases NCBI GEO [1] y ArrayExpress [6]

5.3.2 Aplicación del flujo de trabajo a otros cluster de lactosa en e. coli: Zinc, Sodium, Phosphorelay signal transduction system y Response to arsenic-containing substance

Utilizar la base de datos M^{3D}[3]

5.4 Resultados concluyentes

Discusión de los resultados obtenidos con la experimentación de datos de los genes de E. coli (semejanzas y disparidad entre modulos)

Conclusiones

Relacionar con los resultados del caso de estudio (aplicar referencia cruzada)

- Las redes Bayesianas de interacción de genes empleadas para modelar componentes de una red regulatoria bacteriana pueden mostrarnos en sus nodos raiz (por su caracter jerarquico) genes reguladores de interes biológico responsables de la activación o inhibición de un conjunto de genes que en una red de co-expresión génica solo se aprecian co-expresados. Es decir, que una red Bayesiana comparada con la red de co-expresión génica puede explicar (con cierto nivel de incertidubre, es decir, con una probabilidad) con un análisis causal de la red (por las dependencias modeladas entre sus nodos, soportadas en el concepto de probabilidad condicional) aspectos topologicos de una red de co-expresión como por qué un conjunto de genes se estan co-expresando. Por ejemplo, el cluster de lactosa en E. coli esta compuesto por los genes *LacA*, *LacY* y *LacZ* los cuales aparecen co-relados en una red de co-expresión, sin embargo, en una red Bayesiana podemos ver ademas el gen Lacl como nodo padre de *LacA*, *LacY* y *LacZ*; al consultar la función génica de *Lacl* podemos corroborar que se trata de un gen regulador de lactosa en E. coli.

- Por la naturaleza exploratoria de la caminata aleatoria realiada en la simulación MCMC durante el proceso de aprendizaje de una red Bayesiana, podríamos tener el caso de una red de N nodos, donde un nodo n tiene a los demas $N-1$ como padres en la topológia de esta red. Supongamos que se trata de red con variables binarias, asi que tendremos par el nodo n, 2^{N-1} entradas en la tabla de probabilidad condicional asociada al nodo; para efectos del cálculo del criterio de aceptación del algortimo MH, deberemos evaluar 2^{N-1} funciones de distribución de probabilidad (dirithlet) nos enfrentamos a una complejidad comptacional exponencial. esta situación limita a que las redes Bayesianas solo puedan ser utilizadas para análizar un conjunto de genes de un par de decenas a lo más correspondientes a algún componente módular pequeño del sistema regulatorio bacteriano (el cual puede ser del orden de varios miles de genes). Como trabajo futuro se podria implementar en el marco de trabajo técnicas computacionales como la computación parelela a traves de sistemas distribuidos (MPI-Message Procesing Interface) para parelelizar la evaluación de las PDF dentro del criterio de evaluación del algortimo MH, ya que para el caso del modelo multinomial-dirichlet se trata de una sumatoria apta de ser paralelizada.

- Por otra parte, los sistemas regulatorios en general puden incluir ciclos, dado que un producto del gen como un trascriptoma (RNA) o proteina podría adherirse al citoplasma y posteriormente a una región del ADN que impacte al mismo gen que genero dicho producto (dogma central de la biología). Especialmente en organismos bacterianos este ciclo puede ser mucho mas acelerado que un eukariota. Esta situación limita el modelamiento de componentes de un sistema regulatorio bacteriano al ser este un grafo aciclico. Como trabajo futuro se podría incorporar al marco de trabajo el modelamiento de redes Bayesianas dinamicas para dar tratamiento en este aspecto.

Bibliografia

References

[1] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets update. *Nucleic Acids Research*, 41(D1):D991–D995, nov 2012.

[2] Byron Ellis and Wing Hung Wong. Learning Causal Bayesian Network Structures From Experimental Data. *Journal of the American Statistical Association*, 103(482):778–789, jun 2008.

[3] Jeremiah J. Faith, Michael E. Driscoll, Vincent A. Fusaro, Elissa J. Cosgrove, Boris Hayete, Frank S. Juhn, Stephen J. Schneider, and Timothy S. Gardner. Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Research*, 36(Database):D866–D870, dec 2007.

[4] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, apr 1970.

[5] Miguel A. Ibarra-Arellano, Adrián I. Campos-González, Luis G. Treviño-Quintanilla, Andreas Tauch, and Julio A. Freyre-González. Abasy Atlas: a comprehensive inventory of systems, global network properties and systems-level elements across bacteria. *Database*, 2016:baw089, may 2016.

[6] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y. Amy Tang, Eleanor Williams, Miroslaw Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, Karyn Megy, Ekaterina Pilicheva, Gabriella Rustici, Andrew Tikhonov, Helen Parkinson, Robert Petryszak, Ugis Sarkans, and Alvis Brazma. ArrayExpress update-simplifying data submissions. *Nucleic Acids Research*, 2015.

[7] David Madigan, Jeremy York, and Denis Allard. Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique*, 63(2):215, aug 1995.

[8] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, jun 1998.

[9] Bin Zhang and Steve Horvath. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), jan 2005.