Computación Evolutiva y Aprendizaje Automático para la Inferencia, Modelado y Simulación de Redes Regulatorias de Genes

Carballido Jessica Andrea, Ponzoni Ignacio, Gallo Cristian Andrés
Laboratorio de Investigación y Desarrollo en Computación Científica (LIDeCC)
Dto. de Cs. e Ing. de la Computación (DCIC) - Planta Piloto de Ingeniería Química (PLAPIQUI)
Universidad Nacional del Sur

CONTEXTO

La línea de investigación aquí presentada constituye una de las tres grandes líneas de investigación abarcadas por el laboratorio de investigación LIDeCC (Laboratorio de Investigación y Desarrollo en Computación Científica), dirigido por la Dra. Nélida Beatriz Brignole. Actualmente, los doctores Ignacio Ponzoni, Jessica Carballido y el becario doctoral Cristian Gallo están trabajando en estos temas. Es una línea de investigación que se encuadra en el área de Bioinformática, y que cuenta con el soporte de varios proyectos de investigación[1].

RESUMEN

Los alcances principales de esta línea de I/D consisten en diseñar técnicas computacionales que asistan a expertos en bioinformática en la obtención de nuevos conocimientos sobre el funcionamiento de los mecanismos de regulación existentes a nivel molecular en los organismos biológicos. Más específicamente, se busca desarrollar sistemas de software que asistan en la reconstrucción (o descubrimiento) de la estructura relacional presente en las redes regulatorias de genes.

Palabras clave: *bioinformática, computación evolutiva, aprendizaje automático, redes regulatorias de genes*

INTRODUCCION

El advenimiento de nuevas tecnologías en el área de genómica, tales como los *microarrays* [Sohler, 2006], han posibilitado la obtención de grandes volúmenes de información relativas al funcionamiento molecular de seres vivos. Esto ha derivado en una verdadera revolución del conocimiento en el campo de la Biología Molecular, la cual ha dado origen a una nueva disciplina, la Bioinformática. En la actualidad, gran parte de los esfuerzos llevados adelante en esta área apuntan a generar modelos computacionales que permitan extraer conocimiento a partir de la gran cantidad de datos biológicos disponibles actualmente [Schlitt y Brazma, 2007]. En este contexto, el uso de técnicas de aprendizaje automático [Bishop, 2006] y computación evolutiva [De Jong, 2006] está teniendo gran impacto en el diseño de modelos predictivos orientados a facilitar el descubrimiento automatizado de nuevo conocimiento biológico [Cios *et al.*, 2007; Handl *et al.*, 2007].

En el caso particular de la línea de I/D aquí presentada, el principal objetivo consiste en el desarrollo de software para la inferencia, validación y simulación de redes regulatorias de genes, tema de vital y actual relevancia en el campo de la genómica funcional [Schlitt y Brazma, 2007]. Una red regulatoria de genes (GRN, por su sigla en inglés) representa las relaciones de alto nivel que gobiernan las tasas de transcripción (regulación) de los genes en ARNm (ácido ribonucleico mensajero). Esta interacción puede representarse mediante una red (grafo), donde los nodos identifican genes cuyos niveles de expresión son regulados por la acción (activadora o inhibitoria) ejercida por otros genes de la misma red. Dichas interacciones son denotadas mediante aristas que interconectan los nodos del grafo. Descubrir y estudiar la estructura de las GRNs de diferentes organismos resulta de fundamental importancia para avanzar en la comprensión del funcionamiento biológico de los seres vivos, y posee innumerables aplicaciones tanto en

[1] **PICT 2006 Categoría B**. Tema: Desarrollo teórico de técnicas de Computación Evolutiva basadas en dominancia de Pareto para Optimización Combinatoria Multiobjetivo. Código: 1652. Director: Dra. Jessica A. Carballido. **PGI-UNS** Tema: *Aplicaciones de Computación Científica*. **PGI-UNS** Tema: *Técnicas de Aprendizaje Automático y Computación Evolutiva aplicadas al Diseño de Modelos Predictivos en Bioinformática*.

medicina (por ejemplo, el desarrollo de nuevos medicamentos [Huang, 2002]), como en biotecnología (por ejemplo, el diseño de nuevas variantes de cultivos [Yamaguchi-Shinozaki y Shinozaki, 2006]). Esta última aplicación se encuadra dentro de las áreas de investigación prioritarias establecidas en el **Plan Estratégico Nacional de Ciencia, Tecnología e Innovación "Bicentenario" (2006-2010)**, más específicamente en el área temática denominada ***"Competitividad y Diversificación Sustentable de la Producción Agropecuaria"***, que promueve el desarrollo de nuevas variantes de cultivos con resistencias genéticas a factores adversos (bióticos y abióticos) y mayor eficiencia en la captación de nutrientes, agua, radiación, etc [MinCYT, 2006].

Durante el último lustro, surgieron diferentes métodos para efectuar la ingeniería inversa de GRNs empleando técnicas de inteligencia artificial [Schlitt y Brazma, 2007]. Estos métodos utilizan datos de expresión genética, los cuales constituyen una medida de la abundancia de ARNm de un subconjunto (o totalidad) de los genes presentes en el genoma de un organismo. Dichas mediciones son usualmente obtenidas mediante experimentos en laboratorios con *microarrays*, los cuales permiten recolectar datos de expresión genética a gran escala. A los fines prácticos, la información de expresión genética obtenida experimentalmente puede ser organizada en una estructura matricial, ***E***, la cual representa una serie de tiempo, donde las filas corresponden a los genes observados y las columnas a las distintas muestras relevadas mediante los *microarrays*. De este modo, el elemento e_{ij} de ***E*** contiene un número real que indica el valor de expresión del gen i-ésimo en la muestra j-ésima. Sobre la base de esta información, los métodos de inteligencia computacional propuestos para reconstruir GRNs infieren potenciales asociaciones de regulación entre genes, mediante la búsqueda sistemática de correlaciones entre los datos presentes en la matriz ***E***. Este proceso usualmente requiere efectuar primero una discretización de los datos, obteniéndose una matriz ***E'*** que representa la evolución del valor de expresión de cada gen en términos de un conjunto finito de estados. La idea básica detrás de esta metodología es que los diferentes estados de un gen constituyen una medida de su grado de actividad dentro de cada muestra. De este modo, la identificación de un patrón de comportamiento similar (o contrapuesto) entre diferentes genes puede estar indicando la presencia de un potencial vínculo de regulación entre los mismos, entendiendo por patrón a una secuencia consecutiva de estados contenida dentro de una o más filas de ***E'***. Las asociaciones identificadas mediante estos métodos pueden luego ser expresadas como potenciales reglas de regulación, las cuales permitirán la posterior reconstrucción del grafo correspondiente a la GRN estudiada.

Varios métodos provenientes de la inteligencia artificial fueron propuestos para inferir GRNs a partir de datos obtenidos mediante *microarrays*. Las redes booleanas fueron una de las primeras técnicas empleadas [Akutsu *et al.*, 1999], y algunas variaciones de este enfoque han sido publicadas más recientemente [Mehra *et al.*, 2004]. Estos algoritmos utilizan variables booleanas para expresar el estado de un gen, activo (1) o inactivo (0), mientras que las interacciones son modeladas mediante funciones booleanas que calculan el estado de un gen a partir de la activación de otros genes. La sencillez de este enfoque constituye su principal ventaja pero también su mayor limitación. Las redes booleanas permiten analizar en forma eficiente redes regulatorias de gran dimensión, dado que se realizan fuertes suposiciones que simplifican la estructura y dinámica de los sistemas regulatorios genéticos. Por ejemplo, se asume que el estado del gen es siempre activo o inactivo, sin permitir el modelado de niveles de expresión intermedios. Además se asume que las transiciones entre estados ocurren sincrónicamente. Por ende, cuando las transiciones no suceden simultáneamente, lo cual es usual en la mayoría de los casos, ciertos comportamientos quedan fuera del modelado computacional. Las redes bayesianas proveen un enfoque probabilístico para abordar el modelado de GRNs. En este caso la representación es un grafo dirigido acíclico, donde cada nodo representa una variable

(usualmente genes) y las aristas representan dependencias. Ejemplos de la aplicación de redes bayesianas al modelado de GRNs pueden encontrarse en [Friedman, 2004; Pe'er, 2005]. En general el uso de esta metodología para estudiar GRNs posee interesantes ventajas, tales como la capacidad de lidiar con aspectos estocásticos y con la presencia de ruido en los datos. Sin embargo, hay situaciones, como las relaciones autoregulatorias, en que las reglas booleanas y las funciones lineales no son suficientes para expresar la lógica de control de una GRNs [Schlitt y Brazma 2007]. En estos casos, se requieren funciones de control más complejas que las brindadas por las redes bayesianas.

Un aspecto que en general ha sido escasamente tratado por los métodos antes mencionados es el caso de la ***regulación diferida en el tiempo*** (*time-lagged regulation*). Esta forma de regulación sucede cuando la relación causa-efecto de una acción regulatoria requiere varias unidades de tiempo para que efectivamente ocurra. En términos de la matriz ***E'***, esto implica que un cambio en el estado de un gen i-ésimo en la muestra k puede implicar una modificación en el estado del gen j-ésimo recién en la muestra k+m, siendo m un número entero positivo. Este fenómeno empezó a ser tratado recientemente por ***algoritmos de reconstrucción de GRNs basados en inferencia de potenciales reglas de regulación***. Dentro de esta categoría se encuentran los algoritmos propuestos por Soinov *et al.* [2003] y Li *et al.* [2006]. Ambos tratan la reingeniería de GRNs como un problema de clasificación mediante el uso de árboles de decisión, y aplican el algoritmo C4.5 [Quinlan, 1992] para inferir estas estructuras arbóreas que conducen a la identificación de ***potenciales reglas de regulación***. Lamentablemente, ambos métodos heredan las limitaciones del algoritmo C4.5 que pierde precisión en problemas de clasificación que requieren la evaluación de atributos con valores comprendidos dentro de subrangos de los reales [Ruggieri, 2002].

RESULTADOS OBTENIDOS

En este contexto propusimos un nuevo método, denominado GRNCOP [Ponzoni *et al.*, 2007], que emplea un proceso de discretización adaptiva y optimización combinatoria que supera las limitaciones de los métodos anteriores. No obstante, GRNCOP sólo infiere relaciones con ciertos patrones específicos de regulación diferida en el tiempo, por lo cual su aplicabilidad sufre de restricciones en la práctica.

Otra cuestión importante que ha comenzado a ser estudiada recientemente es el desarrollo de métodos computacionales que faciliten la validación y simulación de las GRNs reconstruidas mediante técnicas inteligentes [Mendes *et al.*, 2003; Batt *et al.*, 2005]. Al respecto, la generación de ***redes regulatorias artificiales*** es de fundamental importancia para los procesos de validación de los algoritmos de inferencia de GRNs [Mendes *et al.*, 2003; Carballido y Ponzoni, 2008]. En particular, el modelo evolutivo basado en el Genoma Artificial de Reil [1999] propuesto por Quayle y Bullock [2006] constituye una alternativa prometedora. También la verificación automática de propiedades está empezando a ser utilizada en el estudio de GRNs [Antoniotti *et al.*, 2003; Batt *et al.*, 2005; Siebert *et al.*, 2008]. El objetivo aquí consiste en poder simular el comportamiento dinámico de una GRN, partiendo de modelos computacionales de GRNs. En este contexto, las ***técnicas de Model-Checking***, provenientes del área de ingeniería de software, resultan muy atractivas para su empleo en este campo, dado que las GRNs pueden ser modeladas como autómatas finitos [Bérard *et al.*, 1998]. Luego, las técnicas de Model-Checking pueden ser utilizadas para razonar sobre el funcionamiento de una GRN mediante el uso de sistemas deductivos y lógica temporal proposicional. Por ejemplo, una herramienta de Model-Checking puede detectar inconsistencias entre las relaciones inferidas para una GRN, ayudando a descartar asociaciones entre genes obtenidas por casualidad (*by chance*). Más aún, estas herramientas pueden ser utilizadas para extraer nuevo conocimiento a partir de la corroboración de nuevas hipótesis biológicas,

las cuales pueden ser expresadas, y evaluadas por el sistema deductivo, mediante consultas lógicas (*logical queries*). De este modo, resulta factible elaborar una metodología integral para la inferencia y simulación de GRNs que facilite el descubrimiento de conocimiento en este tópico, lo cual constituye precisamente el objetivo general, a largo plazo, de estas investigaciones.

Por último, existe un tercer aspecto relacionado con la capacidad de obtener modelos de GRNs complejas mediante la integración de diferentes métodos de inferencia, e incluso de distintas fuentes de información. Una de las principales falencias, reconocida por la mayoría de los especialistas en el área, es la dificultad de lograr modelos completos de redes regulatorias a partir de un único proceso de inferencia orientado a procesar exclusivamente datos de expresión genética provenientes de series de *microarrays* [Pridgeon y Corne, 2004; Schlitt y Brazma, 2005; Schlitt y Brazma, 2006].

RESULTADOS ESPERADOS

Establecidos los antecedentes, es posible explicitar las metas perseguidas en esta línea de I/D. Principalmente, se busca a futuro lograr el desarrollo completo de una metodología integral de inferencia, validación y simulación de redes regulatorias de genes con las siguientes características específicas:

1. El proceso de reconstrucción de redes regulatorias estará basado en la ***inferencia de reglas*** de regulación. Se elige esta metodología porque promueve un enfoque constructivista del modelo de una red, favoreciendo el descubrimiento de conocimiento que pueda ser explicado en términos de relaciones *causa y efecto*.
2. El proceso de inferencia debe permitir el descubrimiento de ***reglas de regulación diferida en el tiempo***. Esto resulta fundamental para modelar el comportamiento de una red como sistema de eventos discretos que facilite su posterior validación y simulación mediante técnicas de verificación formal de sistemas.
3. El proceso de inferencia central, basado en el descubrimiento de reglas de regulación, debería permitir la ***incorporación de conocimiento biológico adicional*** proveniente de otras fuentes, más allá de los datos de expresión genética obtenidos mediante *microarrays*. Esto permitirá enriquecer los modelos inferidos para redes regulatorias complejas.
4. Además del proceso de inferencia central, basado en el descubrimiento de reglas de regulación, se deberían desarrollar modelos de ***inferencia más específicos orientados*** a la detección grupos de genes co-expresados y sub-topologías (genes *in-hub* y *out-hub*). De este modo, mediante la ***integración del conocimiento*** aportado por cada método de inferencia es posible reconstruir redes regulatorias más complejas. El postulante ya ha efectuado contribuciones en este tema, a través del desarrollo de un algoritmo evolutivo multiobjetivo para detección de biclusters (Gallo *et al.*, 2008).
5. También es necesario estudiar como ***adaptar y aplicar técnicas de Model-Checking*** para simular y evaluar nuevas hipótesis biológicas sobre el comportamiento de estas redes, estableciendo los alcances y limitaciones de estas técnicas, y el tipo de propiedades lógicas que pueden ser evaluadas sobre estos sistemas biológicos.

En conclusión, con el trabajo realizado en el contexto de esta línea de investigación y desarrollo se pretende integrar todas las herramientas y conocimientos desarrollados en estas investigaciones en una arquitectura de software que asista a científicos en el descubrimiento, análisis y simulación de redes regulatorias de genes, con el fin de obtener un producto tangible y práctico.

BIBLIOGRAFIA

1. Akutsu, T.; Miyano, S.; Buhara, S. “Identification of genetic networks from a small number of gene expression patterns under the Boolean network

model", Pacific Symp. Biocomput. 4, pp. 17-28, 1999.
2. Antoniotti, M.; Policriti, A.; Ugel, N.; Mishra, B. "Model Building and Model Checking for Biochemical Processes", Cell Biochemistry and Biophysics, 38:271-286, 2003.
3. Batt, G.; Ropers, D.; de Jong, H.; Geiselmann, J.; Mateescu, R.; Page, M.; Schneider, D. "Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in Escherichia coli", Bioinformatics, 21 S.1:19-28. 2005.
4. Bérard, B.; Bidoit, M.; Finkel, A.; Laroussinie, F.; Petit, A.; Petrucci, L.; Schnoebelen, P.; McKenzie P. "Systems and Software Verification. Model-Checking Techniques and Tools", Springer. 1998.
5. Bishop, C.M. "Pattern Recognition and Machine Learning", Springer, 1st edition, 2006.
6. Carballido, J., Ponzoni, I. "On Artificial Gene Regulatory Networks", Electronic Journal of SADIO. 8(1):25-34. 2008.
7. Cios, K.; Kurgan, L.; Reformat, M. "Machine learning in the life sciences-How it is used on a wide variety of medical problems and data", IEEE Engineering in Medicine and Biology Magazine, 26(2):14-16. 2007.
8. De Jong, K.A. "Evolutionary Computation. A Unified Approach", MIT Press, 2006.
9. Friedman, N. "Inferring cellular networks using probabilistic graphical models", Science, 303(5659):799-805, 2004.
10. Gallo, C.; Carballido, J.; Ponzoni, I. "A Hybridized Multiobjective Evolutionary Approach for Microarray Biclustering", CLEI 2008 (XXXIV Conferencia Latinoamericana de Informática), 8-12, Septiembre, 2008. En prensa.
11. Handl, J.; Kell, D.B.; Knowles, J. "Multiobjective optimization in bioinformatics and computational biology", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(2):279-292. 2007.
12. Huang, S. "Rational drug discovery: what can we learn from regulatory networks?", Drug Discovery Today, 7(20):S163-S169, 2002.
13. Li, X.; Rao, S.; Jiang, W.; Li, C.; Xiao, Y.; Guo, Z.; Zhang, Q.; Wang, L.; Du, L.; Li, J.; Li, L.; Zhang, T.; Wang, Q.K. "Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling", BMC Bioinformatics, 7:26, 2006.
14. Mehra, S.; Hu, W-S.; Karypis, G. "A Boolean algorithm for reconstructing the structure of regulatory networks", Metabolic Engineering, 6:326-339, 2004.
15. Mendes, P.; Sha, W.; Ye, K. "Artificial gene networks for objective comparison of analysis algorithms", Bioinformatics, 19(Suppl. 2):ii122-ii129, 2003.
16. MinCYT, Plan Estratégico Nacional de Ciencia, Tecnología e Innovación ''Bicentenario'' (2006–2010). Anexo - Documentos consensuados con las Secretarías de Estado responsables de las políticas sectoriales. Prioridades en Investigación, Desarrollo e Innovación para el Programa PROTIS. Página 35, 2006.
17. Pe'er, D. "Bayesian network analysis of signaling networks: a primer," STKE 2005, 281:pl4, 2005.
18. Ponzoni, I.; Azuaje, F.; Augusto, J.; Glass, D. "Inferring association rules between genes using a combinatorial optimization learning process and adaptive regulation thresholds", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(4): 624-634, 2007.
19. Pridgeon, C., Corne, D. "Genetic network reverse-engineering and network size; can we identify large GRNs?," Proc. 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB '04: 7-8 October 2004; La Jolla, California, USA; pp. 32–36, 2004.
20. Quayle, A., Bullock, S. "Modelling the evolution of genetic regulatory networks", J. Theor. Biol. 238:737–753. 2006.
21. Quinlan, J.R. "C4.5: Programs for Machine Learning," Morgan Kaufmann, 1992.
22. Reil, T. "Dynamics of gene expression in an artificial genome: Implications for biological and artificial ontogeny". In: Floreano, D., Mondada, F. & Nicoud, J. D. (ed.): Proc. 5th European Conference on Artificial Life. Lecture Notes in Computer Science, Vol. 1674. Springer-Verlag, pp. 457–466, 1999.
23. Ruggieri, S. "Efficient C4.5," IEEE Trans. on Knowledge and Data Engineering 14:438-444, 2002.
24. Schlitt, T.; Brazma, A. "Modelling gene networks at different organisational levels," *FEBS Letters*, vol. 579, pp. 1859-1866, 2005.
25. Schlitt, T.; Brazma, A. "Modelling in molecular biology: describing transcription regulatory networks at different scales," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 361(1467), pp. 483-494, 2006.
26. Schlitt, T.; Brazma, A. "Current approaches to gene regulatory network modelling," BMC Bioinformatics, 8 (Suppl. 6):S9, 2007.
27. Siebert, H.; Bockmayra, A. "Temporal constraints in the logical analysis of regulatory networks", Theoretical Computer Science, 391(3): 258-275, 2008.
28. Sohler, F. "Contextual Analysis of Gene Expression Data". Master Thesis, Germany, 2006.
29. Soinov, L.A.; Krestyaninova, M.A.; Brazma, A. "Towards reconstruction of gene networks from expression data by supervised learning," Genome Biology, 4:R6, 2003.
30. Yamaguchi-Shinozaki, K.; Shinozaki, K. "Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses", Annual Review of Plant Biology, 57:781–803, 2006.