Comparación de rutas metabólicas mediante algoritmos de evaluación simples

Esteban Arias-Méndez
Escuela de Computación
Tecnológico de Costa Rica
Cartago, Costa Rica
esteban.arias@tec.ac.cr , earias@ic-itcr.ac.cr

Francisco Torres-Rojas
Escuela de Computación
Tecnológico de Costa Rica
San José, Costa Rica
torres@ic-itcr.ac.cr

***Abstract*—Para comprender mejor la vida y sus procesos, las rutas metabólicas proveen información útil para el mejoramiento de la medicina, agronomía, farmacia y otras. Las herramientas principales usadas para estudiar estas rutas están basadas en la idea de la comparación de rutas usando estructuras de datos tipo grafos. La comparación de grafos ha sido definida como computacionalmente costosa. Se proponen dos formas diferentes que simplifican el problema de la comparación de rutas representadas como grafos. El primer algoritmo consiste en la transformación de los grafos a comparar, de estructuras en 2 dimensiones a estructuras lineales, para posteriormente alinear dichas estructuras usando algoritmos clásicos de alineamiento. El segundo algoritmo consiste en realizar una comparación basada en pares de metabolitos reaccionantes, eliminando de los grafos los pares iguales, para mostrar al final las reacciones entre metabolitos diferentes entre cada ruta metabólica. Nuestros resultados muestran buena evidencia de un mecanismo rápido, simple y efectivo para el problema descrito.**

Keywords — grafos, recorrido por anchura primero, recorrido por profundidad primero, alineamiento global, alineamiento local, alineamiento semiglobal.

I. Introducción

Los organismos que consideramos como vivos hoy en día son aquellos que realizan un proceso en común que conocemos como el ciclo de la vida: nacer, crecer, alimentarse, reproducirse y morir. La célula es considerada la unidad más pequeña de organismo vivo y base fundamental de estas etapas para especímenes formados de 1 sola célula hasta aquellos confirmados por miles de millones de células en un solo organismo tal como nosotros los seres humanos.

La biología molecular ha logrado grandes avances con la bioinformática o biología computacional que aplican técnicas computacionales en el desarrollo de la disciplina. Se trabaja actualmente en usar dicha información para comprender procesos de interés. Proteómica, epigenética y metabolómica tienen gran impacto hoy en día en múltiples campos.

En metabolómica interesa comprender los procesos metabólicos que ocurren en los organismos. Esto es, el conjunto de reacciones bioquímicas y físico-químicas que ocurren a nivel celular. Complejos procesos interrelacionados los cuales son la base de la vida a escala molecular.

Hoy en día existen varias bases de datos para metabolismo o rutas metabólicas que almacenan las descripciones de estos procesos. La forma en que han sido anotadas es de forma similar a estructuras de datos tipo grafos dirigidos, usados en computación para modelar muchos tipos de relaciones y describir procesos. En éstas es posible realizar consultas vía proteína, metabolito, vía gen o abreviatura del gen; según el enfoque y organización de cada base en particular. KEGG (www.genome.jp/kegg/) [11] y MetaCyc (parte de BioCyc https://biocyc.org) [5] son ejemplos de las más grandes e importantes usadas hoy en día y proveen acceso a rutas metabólicas de diversos organismos tanto animales como plantas.

II. Rutas metabólicas y Grafos

Una ruta metabólica es una secuencia ordenada de reacciones bioquímicas entre diversos actores denominados metabolitos, que son sustratos que mediante procesos enzimáticos catalizan transformaciones hasta obtener un producto [3], [15]. Esto es un proceso de transformación de moléculas de unos compuestos o metabolitos a otros. Las bases de datos en línea proveen mecanismos para predecir rutas metabólicas según elementos previamente descritos. Para muchas rutas se cuenta con datos curados, que han sido confirmados por varios laboratorios.

En la actualidad, el estudio de procesos metabólicos de interés traspasa varias áreas de conocimiento para el análisis de la información disponible. En el caso particular de los estudios sobre metabolismo interesa conocer, además de los metabolitos involucrados, los pasos o reacciones entre cada paso de un proceso, conocidos como una ruta metabólica. Dichas rutas pueden estar siendo afectadas o reguladas en procesos mayores y más complejos que conforman redes de rutas metabólicas, cuando varias de estas interactúan entre sí. A nivel biológico no existe una única forma de obtener un producto, es decir, pueden existir para algunos casos diversos procedimientos para un mismo producto; con pequeñas o mayores modificaciones.

Una de las actividades más importantes para estos análisis consiste en la comparación de rutas metabólicas de procesos de interés a nivel agronómico, farmacéutico, medicinal, comercial y otros.

Las razones son varias e importantes, el conocer estos procesos puede darnos herramientas para intervenir con la finalidad de copiarlos para producir más o mejores alimentos, aprender a controlar el posible accionar de diversos virus o enfermedades a nivel celular para un mejor biocontrol, así como mejorar el desarrollo de fármacos o tratamientos más efectivos; ya que el conocimiento de las rutas y redes metabólicas es importante para el manejo de medicamentos en estudios clínicos sobre cadenas de acción/reacción de los fármacos en los organismos. Por ejemplo, en el caso de las plantas, el conocer una red metabólica puede ser usado como herramienta para técnicas de mejoramiento de cultivos para extraer componentes que son importantes en los procesos de consumo humano, para maximizar su producción.

Comprender mejor la evolución, especiación y reconstrucción filogenética [16], [8] así como el descubrimiento de fármacos más efectivos [7] puede ser posible gracias al análisis comparativo de rutas de diferentes organismos.

Para facilitar este análisis sobre las rutas metabólicas se utilizan gráficos y estructuras de datos tipo grafos dirigidos para poder describir los metabolitos involucrados en cada proceso y sus interacciones como reacciones. Algunas herramientas como PathVisio [13] MetDraw [10] o NetCoffee [9]. brindan información básica sobre las rutas, sus componentes, gráficas y otra información, pero no herramientas de análisis como lo sería un proceso de comparación de rutas.

En computación se han propuesto varios mecanismos para la comparación efectiva de dichas rutas para una especie o entre especies. Al respecto Abaka et. al [1] han realizado un repaso por las más importantes herramientas desarrolladas hasta ahora. Además, brindan pruebas de los costos NP asociados al alineamiento de 2 rutas tratadas como grafos. Tareas que se han descrito como computacionalmente complejos, como se menciona también en Ay & Kahveci [2] quienes hacen una propuesta llamada SubMAP (Subnetwork Mappings in Alignment of Pathways) la cual brinda una comparación no 1 a 1, o 1 a muchos, como en los primeros enfoques propuestos, sino más bien en buscar las subpartes comunes entre diferentes rutas, o subredes similares. El algoritmo CAMPways de Abaka et. al [1] promete ser más eficiente en tiempo de ejecución que los algoritmos definidos como el estado del arte. Sin embargo, este algoritmo hace referencia a dos evaluaciones o medidas que pueden ser conflictivas: similaridad hómologa y similaridad topológica de las rutas dadas. El análisis que se brinda de las rutas al igual que las otras herramientas anteriores es un mecanismo complejo de interpretación y procesamiento de la información existente

Acá se presenta un enfoque diferente a los mecanismos usados hasta ahora para la comparación de rutas metabólicas y se proponen 2 alternativas más simples que pueden ser usadas como una evaluación previa a un análisis más profundo y de más tiempo y costo involucrados.

A. Grafos

Para tratar las rutas metabólicas se ha recurrido al uso de Grafos, estructuras de datos dinámicas, que sirven para modelar de forma práctica diversas relaciones entre procesos de toda índole. La forma general que describe una ruta cualquiera la podemos representar como un grafo dirigido en una computadora. A partir de ahí, se han desarrollado diversas formas de alinear y comparar los digrafos correspondientes a las rutas de interés cuyos costos asociados son NP [1], [2], o bien, tratamientos complejos de algoritmos mediante técnicas heurísticas que buscan acotar el tiempo de alineamiento de un grafo o ruta contra otra. Este problema es mucho más complejo cuando se busca hacer una comparación entre una ruta y múltiples.

Se debe tener presente la diferencia entre 2 rutas homólogas y 2 rutas similares. La homología puede describirse como una comparación de alto nivel, mientras que la similitud se define como una valoración medible y tangible. Podemos decir que dos personas son homólogas, pero podrían no ser similares. En el caso de las rutas podrían conformar la misma cantidad de interacciones o reacciones, tener una forma homóloga, pero los metabolitos reaccionantes diferir. Para este trabajo interesa la similitud como valor o referencia de comparación.

B. Métodos simples para comparación

Las rutas al ser vistas como estructuras de datos tipo grafos permiten aplicar una gran variedad de algoritmos ya existentes. En la literatura tradicional sobre grafos no es común explorar esta clase de algoritmos de comparación, pero si es usual hacer recorridos de grafos para obtener todos sus nodos, hacer búsquedas de rutas óptimas entre dos nodos cualesquiera, etc. Es decir, algoritmos tradicionales como el árbol de recubrimiento mínimo, distancias mínimas o rutas más cortas ya sea entre todos los nodos o un par de nodos dados.

Por otra parte, en bioinformática los mecanismos de alineamiento son válidos para una comparación paso a paso de cada una de las etapas de la ruta metabólica. Se requiere aún de un mecanismo de comparación eficiente a nivel computacional, que pueda ser utilizado luego con diversas fuentes de información para el estudio adecuado de las rutas metabólicas de interés y su posterior análisis.

Se propone acá un enfoque de comparación diferente para rutas metabólicas mediante 2 algoritmos alternos: en primera instancia, no visualizar la ruta metabólica estrictamente como un digrafo, haciendo un procesamiento previo para transformar dicha estructura bidimensional en una estructura lineal que sea más sencillo y barato computacionalmente alinear; como otra alternativa se propone hacer una análisis por pares de reacciones en los grafos, es decir analizar las relaciones 1 a 1 de reacción entre 2 metabolitos dentro de los grafos, para sacar las relaciones como un denominador común a ambas estructuras, simplificando luego dichos grafos, de forma que sea más fácil determinar los puntos de divergencia en las rutas para su análisis.

No se busca dar una respuesta definitiva al resultado de la comparación entre dos rutas o indicar que un procedimiento es mejor que otro, más bien se busca aportar un punto de vista adicional como apoyo para que sea considerado por un experto

en la materia a la hora de hacer sus observaciones, evaluaciones y conclusiones sobre el proceso particular que estudia. No se busca dar una respuesta "correcta" sobre cual es una mejor ruta, solo brindar información de referencia para el interesado.

III. Metodología

Se explican los 2 algoritmos para tratar el problema de alineamiento o comparación de dígrafos de rutas metabólicas.

A. Algoritmo 1: Transformación de Grafo 2D a 1D para posterior alineamiento y evaluación.

En el caso particular de las rutas metabólicas es común observar en el detalle que se obtienen de las diversas bases de datos que, si bien se modelan como un grafo con diversas relaciones entre ellas e incluso ciclos internos, es característico que toda ruta tenga dos elementos clave: sustratos de punto de partida y un producto final como resultado. Si se observa entonces la ruta como un grafo, este grafo tendrá definido una raíz particular y una hoja destino de importancia, usando acá nomenclatura de árboles en estructuras de datos.

En el caso particular de un grafo sobre una ruta metabólica, al aplicar un recorrido del grafo que visite todos los nodos se puede obtener de forma simple la lista de elementos que lo conforman. Esto sería transformar de 2D a 1D el grafo. Si tomamos como referencia el punto de inicio de la ruta como la raíz del grafo e inicio del recorrido, se deben visitar entonces todos los nodos hasta llegar como punto final al nodo de interés que sería el producto final de la ruta como tal.

Habrá una pérdida de información en dicha transformación. En la figura 1 se observa este hecho, principalmente sobre el orden de los elementos y sus relaciones originales. Se busca demostrar que dicha pérdida de información durante el proceso es tolerable y aceptable para un resultado certero de comparación.

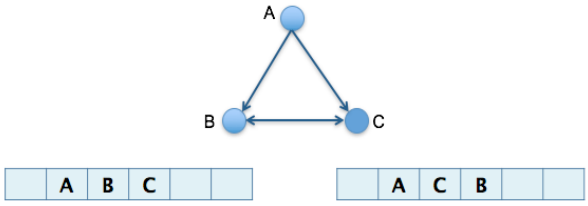

Figura 1. Pérdida de información debido a transformación 2D a 1D

Luego de la transformación de los grafos correspondientes a las rutas en estudio, a la información lineal obtenida se busca aplicar algoritmos de alineamiento convencionales: global [17], semiglobal, and local [18]; para obtener valores de comparación de dichas secuencias lineales.

Si analizamos ejemplos extraídos de la base de datos MetaCyc (metacyc.org), tales como: *glycolysis I (from glucose-6P)*[1] y *glycolysis IV (plant cytosol)*[2], dos procesos de glucólisis (extracción de energía desde azúcares) que brindan el mismo producto desde dos fuentes diferentes, se puede determinar que ambas rutas tienen muchos elementos en común y que son fácilmente homólogas, pero interesa medir su similitud, para tal fin se procede a realizar un recorrido de los grafos. Es común que se recurra a hacer dos tipos de recorridos, por profundidad [19] o por anchura [4], ver también [6], [12], [14]. Al aplicar un algoritmo por profundidad la información que se obtiene no es relativamente proporcional y relevante sobre la ruta debido a que el producto aparecerá en medio de la hilera en 1D y no al final de la misma, como se podría esperar en una serie de reacciones que lleven al final a dicho producto. Al realizar un recorrido por anchura se van visitando los nodos por niveles, lo cual corresponde más a la forma en que van reaccionando los metabolitos hasta llegar al producto esperado. Según esta observación, los datos útiles para ser analizados corresponden principalmente a los generados por los recorridos en anchura primero. Una vez alcanzados los datos de los recorridos para obtener las rutas en un formato 1D se procede a aplicar los algoritmos de alineamiento tradicionales.

B. Algoritmo 2: Diferenciación por pares

Cuando se desea comparar dos objetos, los elementos comunes son evidentes, es entonces que se hace más relevante concentrarse en buscar las diferencias como tal. Con base en esta idea, este segundo algoritmo busca eliminar del grafo los pares comunes de objetos, esto es, reacciones iguales entre ambos grafos, para encontrar las diferencias entre las rutas. Este sería un enfoque de análisis diferente al alineamiento tradicional que busca un valor numérico de similitud como tal.

En este algoritmo se va recorriendo la lista de reacciones en una ruta y se busca en la otra, si la misma existe, se elimina de ambos grafos. Al finalizar el recorrido se listan las reacciones que difieren en cada ruta como las diferencias encontradas.

IV. Discusión: Pruebas y Resultados

Se debe valorar el costo de los algoritmos para mostrar que son más baratos que los usados hasta ahora. Luego se debe demostrar que el procedimiento brinda un resultado certero y útil con respecto a la comparación en sí.

En el primer algoritmo se hace uso del recorrido de grafos o búsqueda por anchura que va realizando un recorrido por niveles, similar a la forma en que funciona una ruta metabólica como tal. El costo de este algoritmo se aproxima como $O(|V|+|E|)$, donde V: es el conjunto de vértices o nodos del grafo y $E \subseteq VxV$: es el conjunto de aristas o arcos.

Para el segundo algoritmo se debe considerar que para cada reacción que existe en la primera ruta o grafo G_1 se deberá buscar la misma en la segunda ruta o grafo G_2. Esto es si R_1 es

[1] Disponible en línea en: http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=GLYCOLYSIS&detail-level=1&detail-level=0

[2] Disponible en línea en: http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-1042

la cantidad de reacciones que contabiliza G_1 y R_2 la cantidad para G_2, se harán máximo R_1 x R_2 comparaciones, cuando es común que en tiempo de ejecución se realice en promedio la mitad de dichas comparaciones. Se establece como peor caso del algoritmo $O(R_1xR_2)$.

El valor de comparación alcanzado por el alineamiento global es de +3 lo que indica un buen valor de comparación o similitud. Para ser más precisos, el alineamiento local nos brinda un valor de +5 para la sección de la ruta metabólica que más se asemeja entre ambas.

Al realizar pruebas con una ruta muy diferente en forma y contenido como la que se observa en la ruta *indole-3-acetate biosynthesis II*[3] y luego de aplicar la transformación mediante el algoritmo 1 y su posterior alineamiento los resultados varían. Para este caso los valores alcanzados son: -10 para el alinemiento global, 0 para el local y de -20 si aplicamos semiglobal

Para el caso del algoritmo 2 no se encontró un algoritmo con el cual compararlo pues es una estrategia diferente a las propuestas hasta ahora. Pero si se brinda información útil al experto que realiza el análisis sobre las diferencias encontradas. Para las rutas se listan las reacciones presentes en una que no están presentes en la otra.Se debe observar que las reacciones son bidireccionales en las rutas originales, razón por la cual se brinda el detalle de cada reacción en cada dirección.

V. Conclusiones

Se puede establecer que el mecanismo propuesto en el algoritmo 1 puede ser usado como evaluador previo para predecir buenas comparaciones en caso que se desee un análisis más profundo. En el caso del algoritmo 2 la propuesta es brindar al experto, un punto de vista adicional para sus valoraciones sobre las rutas que estudia, sin brindar un valor.

Se demuestra que mediante procedimientos rápidos y bajo costo computacional es posible proveer información relevante para el estudio de comparación de rutas metabólicas de interés y otros análisis. Esto se alcanza al simplificar la información, en el caso de una migración de 2D a 1D, la pérdida de información o precisión no afecta el resultado final para brindarle al usuario un valor de comparación de referencia.

VI. Trabajo Futuro

Desarrollar una herramienta que reciba directamente información de las bases de datos de metabolismo existentes, extraiga la información de interés y permita aplicar los algoritmos propuestos para su análisis por parte de expertos. Proveer mecanismos de comparación de secuencias basado en matrices de valoración que toman en cuenta la afinidad entre los elementos. Proveer alineamientos múltiples.

[3] Obtenido en línea de: https://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-581

Agradecimientos

Este trabajo no hubiera sido posible sin el apoyo de Esteban Meneses, Kevin Castro-Fuentes, Seth Stalley y Pablo Vargas-Rosales. Así como el apoyo de la Escuela de Computación, Vicerrectoría de Investigación y Extensión y Vicerrectoría de Docencia del TEC Costa Rica.

Referencias

[1] Abaka, G., Bıyıkoğlu, T., & Erten, C. (2013). CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways. Bioinformatics, 29(13), i145-i153.

[2] Ay, F., Kellis, M., & Kahveci, T. (2011). SubMAP: aligning metabolic pathways with subnetwork mappings. Journal of computational biology, 18(3), 219-235.

[3] Bruce, A., Dennis. B., Julian, L., Martin, R., Keith, R., James D., W. (1994) Molecular Biology Of The Cell third edition.

[4] Bundy, A., & Wallen, L. (1984). Breadth-First Search. In Catalogue of Artificial Intelligence Tools (pp. 13-13). Springer Berlin Heidelberg.

[5] Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., ... & Walk, T. C. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic acids research, 36(suppl 1), D623-D631.

[6] Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). Introduction to algorithms second edition.

[7] Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. (2007). A network-based method for target selection in metabolic networks. Bioinformatics, 23(13), 1616-1622.

[8] Heymans, M., & Singh, A. K. (2003). Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics, 19(suppl 1), i138-i146.

[9] Hu, J., Kehr, B., & Reinert, K. (2013). NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. Bioinformatics, btt715.

[10] Jensen, P. A., & Papin, J. A. (2014). MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. Bioinformatics, 30(9), 1327-1328.

[11] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research, gkr988.

[12] Knuth, D. (1968). The Art of Computer Programming 1: Fundamental Algorithms 2: Seminumerical Algorithms 3: Sorting and Searching. MA: Addison-Wesley, 30.

[13] Kutmon, M., van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., & Evelo, C. T. (2015). PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol, 11(2), e1004085.

[14] Lee, C. Y. (1961). An algorithm for path connections and its applications. IRE transactions on electronic computers, (3), 346-365.

[15] Lee, J. M., Gianchandani, E. P., Eddy, J. A., & Papin, J. A. (2008). Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS Comput Biol, 4(5), e1000086.

[16] Mithani, A., Hein, J., & Preston, G. M. (2010). Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and non-pathogenic lifestyles in Pseudomonas. Molecular Biology and Evolution, msq213.

[17] Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3), 443-453.

[18] Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197.

[19] Tarjan, R. (1972). Depth-first search and linear graph algorithms. SIAM journal on computing, 1(2), 146-160.