Escuela de Computación

Ingeniería en Computación

Informe Final de Proyecto de Investigación

Algoritmos alternos de bajo coste para la comparación de rutas metabólicas en plantas

Investigadores

Esteban Meneses Rojas

Esteban Arias Méndez

Junio 2017

Ficha del Proyecto

Código y Título del proyecto

Algoritmos alternos de bajo coste para la comparación de rutas metabólicas en plantas

Escuela responsable

Escuela de Computación

Autores y direcciones

Dr. Esteban Meneses Rojas, Ph.D. – Coordinador
esmeneses@itcr.ac.cr

Ing. Esteban Arias Méndez
esteban.arias@tec.ac.cr

Periodo de ejecución

Febrero a Diciembre 2016

Tabla de Contenidos

1. Resumen

Las rutas metabólicas proveen información clave para alcanzar un mejor entendimiento de la vida y sus procesos; esta información es útil para el mejoramiento de la medicina, la agronomía, farmacia y otras áreas similares. La herramienta de análisis principal usada para estudiar estas rutas está basada en la idea de la comparación de rutas, usando estructuras de datos tipo grafos. La comparación de grafos ha sido definida como una tarea computacionalmente compleja.

Se proponen 2 algoritmos con enfoques diferentes que simplifican el problema de comparar rutas representadas como grafos. El primer algoritmo consiste en la transformación de una estructura grafo de 2 dimensiones a una estructura de 1 dimensión, posteriormente alinear los datos correspondientes usando la estructura 1D reducida. El segundo algoritmo consiste en realizar un análisis de pares entre grafos, es decir una relación de 2 nodos iguales presentes en ambos grafos, para eliminar las similitudes, y finalmente mostrar las diferencias al usuario.

Nuestros resultados muestran evidencia de una forma rápida, simple y efectiva para resolver el problema descrito.

El mecanismo propuesto en el algoritmo 1 puede ser usado como un evaluador previo para predecir buenas comparaciones en caso de un análisis más profundo se desee. Se muestra que la pérdida de información o precisión no afecta mucho el resultado, que es dar al usuario un puntaje de similitud entre 2 rutas analizadas.

Para el algoritmo 2 la propuesta es ofrecer al experto un punto de vista adicional para su evaluación de la ruta en cuestión. En este caso no se provee un puntaje, pero si la lista de diferencias.

Palabras clave

grafos, comparación de grafos, recorrido anchura-primero, recorrido profundidad-primero, alineamiento de secuencias

2. Abstract

Metabolic pathways provide key information to achieve a better understanding of life and all its processes; this is useful information for the improvement of medicine, agronomy, pharmacy and other similar areas. The main analysis tool used to study these pathways is based on the idea of pathway comparison, using graph data structures. Graph comparison has been defined as a computationally complex task.

We propose two algorithms with different approaches which simplify the problem of comparing pathways represented as graphs. The first algorithm consists in the transformation of a two-dimensional graph structure to a one-dimensional structure, and thus aligning the corresponding data using a reduced 1D structure. The second algorithm consists in performing a pair analysis between graphs, that is to say a relation of 2 equal nodes present in both graphs, and thus eliminating all similarities, finally, showing these differences to the user.

Our results show evidence of a quick, simple and effective way to resolve the described problem.

The mechanism proposed in algorithm 1 can be used as a prior evaluator to predict good comparisons in case a deeper analysis is desired. We show that the loss of information or precision does not affect much the result, which is to give the user a similarity score between the two analyzed pathways.

For algorithm 2 the proposal is to offer the expert an additional point of view for his evaluation of the pathway in question. In this case, no score is provided but the listed differences.

Keywords

graphs, graph comparison, breadth-first traversal, depth-first traversal, sequence alignments

3. Introducción

Los organismos que consideramos como vivos hoy en día, tienen en común un proceso que conocemos como el ciclo de la vida: nacer, crecer, alimentarse, reproducirse y morir; la célula es considerada la unidad y base fundamental de estas etapas para especímenes formados de 1 sola célula hasta aquellos conformados por miles de millones de células en uno solo.

Luego de los enormes avances logrados en el campo de la ciencia y específicamente en la bioinformática y la biología molecular computacional con el uso de técnicas computacionales para la secuenciación completa de la información genética de un individuo, es decir su genoma, así como el análisis que se ha dado en años recientes para comprender dicha información; se trabaja actualmente en usar dicha información para comprender otros procesos de interés. Áreas como la proteómica, epigenética y la metabolómica tienen gran impacto hoy en día en múltiples campos como medicina, agricultura, salud y otros.

El estudio de procesos metabólicos de interés traspasa varias áreas de conocimiento para el análisis de la información disponible. En este caso interesa conocer, además de los metabolitos involucrados, los pasos o reacciones entre cada paso de un proceso, conocidos como una ruta metabólica.

Las rutas metabólicas consisten de metabolitos y reacciones bioquímicas o enzimáticas que transforman éstos en otros similares y enzimas catalizando estas reacciones proveen información valiosa sobre los centros de procesamiento de una célula funcional y del metabolismo celular en general.

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

Se considera entonces el problema de proveer un alineamiento de reacciones de 1-a-muchos en un par de rutas metabólicas. Dichas rutas se han modelado tradicionalmente como grafos, un tipo de estructura de datos dinámica de uso frecuente para modelar redes en computación. Sin embargo, el costo de un alineamiento de grafos es computacionalmente costoso [12], [13], [14], no imposible, pero costoso.

Se planteó como Objetivo General de este trabajo: Implementar y evaluar 2 algoritmos alternos de comparación de rutas metabólicas propuestos y determinar su eficiencia.

Para cumplir este objetivo se organizó el trabajo en 4 objetivos específicos:

1. Recolección bibliográfica
2. Obtener datos de referencia de algoritmos actuales
3. Programar algoritmos propuestos
4. Evaluar efectividad de algoritmos y comparar

Se presentan acá los 2 algoritmos originales para proveer un alineamiento efectivo de bajo costo. Los algoritmos alternos propuestos son:

1. Transformación de grafo: variar la forma en que se procesa la ruta metabólica, transformando los datos a una representación lineal "equivalente" para un posterior alineamiento.

2. Diferenciación por pares: mediante la búsqueda de pares de metabolitos reaccionantes y equivalentes en un par de rutas, proveer una lista de diferencias.

4. Marco Teórico

Metabolismo es un conjunto de interacciones químicas que construyen todos los componentes esenciales para la vida de un organismo. Gran cantidad de componentes de importancia biológica, o metabolitos, tales como azúcares, lípidos, amino ácidos, ácidos nucleicos, y otros componentes secundarios son generados en la célula a través de series de reacciones bioquímicas llamadas rutas metabólicas. Cada ruta metabólica es un patrón, una serie de reacciones, catalizadas por enzimas, las cuales son producidas por la célula y regulan la ruta metabólica [1], [2]. Dependiendo de las enzimas que cada especie produce hay rutas diferentes posibles que la producción o degradación de un producto puede tomar, proveyendo diferentes patrones; esta es la razón por la cual las rutas metabólicas entre organismos varían. Las rutas metabólicas están interconectadas entre ellas, creando redes enormes de interacciones entre metabolitos y enzimas.

Las rutas metabólicas se modelan en computación como estructuras de datos tipo grafos. Dadas un par de rutas metabólicas, un alineamiento de dichas rutas corresponde a un mapeo entre subestructuras similares del par para proveer un valor de comparación o alineamiento. El propósito de este alineamiento es poder determinar un grado de comparación entre este par de rutas. Alineamientos satisfactorios pueden proveer aplicaciones útiles en reconstrucción de árboles filogenéticos, diseño de medicamentos y por sobretodo mejorar nuestro entendimiento del metabolismo celular [8], [9].

En computación se han propuesto varios mecanismos para la comparación efectiva de dichas rutas para una especie o entre especies; sin embargo, los costos asociados para dichas tareas se han descrito como computacionalmente complejos o NP duros [14]. El costo de un alineamiento de grafos es computacionalmente costoso [12], [13], no imposible, pero costoso.

Herramientas tales como NetCoffee [12] permiten encontrar un alineamiento global de múltiples redes de interacción proteína-proteína maximizando una función objetivo mediante un recocido simulado, cristalización simulada o enfriamiento simulado, que es un algoritmo de búsqueda meta-heurística para problemas de optimización global; el objetivo general de este tipo de algoritmos es encontrar una buena aproximación al valor óptimo de una función en un espacio de búsqueda grande. Sin embargo, NetCoffee no es exclusivo para información de redes metabólicas sino interacciones entre proteínas en general.

4.1 Rutas metabólicas

Una ruta metabólica es una secuencia ordenada de reacciones bioquímicas entre diversos actores denominados metabolitos, estos son sustratos que mediante procesos enzimáticos

catalizan transformaciones hasta obtener un producto [1], [2]. Esto es un proceso de transformación de moléculas de unos compuestos o metabolitos a otros. Una gran cantidad de rutas metabólicas son aún desconocidas y/o muchas reacciones hacen falta en las rutas ya conocidas. Las bases de datos en línea proveen mecanismos para predecir rutas metabólicas según elementos previamente descritos. Se considera importante poder automatizar reconstrucciones de novo de rutas metabólicas que incluye elucidar reacciones desconocidas para unir los faltantes en las rutas que se establecen o realizar extrapolaciones en otras especies relacionadas o no. 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. Una de las mejores maneras de procesar esta información de rutas es como estructuras de datos tipo grafos.

A nivel biológico no existe una única forma de obtener un producto, es decir pueden existir para algunos casos diversos procedimientos o rutas metabólicas 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 intervenirlos 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 [5], [6] 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 [10], MetDraw [11] o NetCoffee [12] 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 [13] 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 [14] 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 [13] 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.

4.2 Grafos y alineamientos

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 dentro de 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 [13], 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 pues se parecen en su forma general: 1 cabeza, 2 brazos, 1 tronco, 2 piernas, 2 ojos, etc. Pero, aunque 2 personas sean homólogas 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.

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/) [3] y MetaCyc (parte de BioCyc https://biocyc.org) [4] 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.

Por otro lado, el alineamiento, es un procedimiento ampliamente utilizado en los últimos años como mecanismo principal de comparación de secuencias. Los algoritmos tradicionales tales como Smith-Waterman [16], para alineamiento local, y Needleman-Wunsch [15], para alineamiento global, han sido lo principal para la comparación de secuencias genómicas y su análisis. Como se ha indicado, en el caso de las rutas metabólicas se han propuesto algoritmos similares para alineamiento de rutas metabólicas, basados en grafos dirigidos cuyos costos computacionales han probado ser elevados [12], [13] y [14].

En este trabajo, se busca probar la efectividad de al menos 2 algoritmos originales nuevos propuestos, los cuales mediante diferentes formas de tratamiento de la información se espera que puedan dar un valor de comparación similar al que se usa con el alineamiento estándar. La originalidad de la propuesta radica en algoritmos alternos de bajo costo. Acá se sacrificaría precisión en beneficio de menor costo y mejor rendimiento donde la pérdida de información sea tolerable. De esta manera se pueden evaluar datos de una manera más rápida antes de decidir si se aplican métodos más precisos pero costosos.

Los algoritmos alternos propuestos son:

1. Transformación de grafo: variar la forma en que se procesa la ruta metabólica, transformando los datos en una representación lineal “equivalente” para un posterior alineamiento.

2. Diferenciación por pares: mediante la búsqueda de pares de metabolitos reaccionantes y equivalentes en un par de rutas, proveer una lista de diferencias.

5. Metodología

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.

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.

Mediante un mecanismo alterno de comparación de rutas metabólicas se busca ampliar el espectro de resultados para su posterior análisis que permita establecer nuevas relaciones o conexiones no descritas previamente entre rutas u organismos.

Con un tratamiento diferente dado a la información expresada en los digrafos asociados a las rutas metabólicas se pueden obtener resultados relevantes con un menor costo computacional.

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.

Para cada uno de los objetivos planteados se propuso la metodología de desarrollo para cada uno:

1. Recolección bibliográfica

En este primer punto se procedió con la revisión de la literatura relacionada con los algoritmos de comparación tradicional de grafos. De la misma manera se tabuló la información concerniente a las propuestas actuales de alineamiento de grafos y las propuestas específicas de comparación automática entre rutas y entre redes metabólicas.

Por otro lado, se tabuló la información de referencia sobre las bases de datos metabólicas, era necesario procesar y registrar los formatos y los tipos de datos usados para almacenar la información en estas bases de datos.

Con la información de formatos de representación de estas bases de datos se decidió sobre cual modelo se podría trabajar, esto pues los formatos vistos hasta ahora no cuentan con un estándar de definición, sino que cada base de datos utiliza su propio formato para almacenar datos.

2. Obtener datos de referencia de algoritmos actuales.

Utilizando como base la información del punto anterior se procedió a analizar cuantitativamente el comportamiento y rendimiento de los algoritmos que se han propuesto hasta ahora.

Se recolectaron y analizaron los datos de ejecución de al menos 2 algoritmos vigentes para la comparación de rutas metabólicas. Se utilizaron como referente los datos estadísticos provistos por las propias investigaciones publicadas al respecto y una evaluación propia de algunas herramientas relevantes en su enfoque a los algoritmos acá propuestos.

La información procesada se tomó como insumo para evaluar el estado del arte actual y como referente para la evaluación propia de los algoritmos propuestos.

3. Programar algoritmos propuestos.

Este objetivo llevó la propuesta de una idea a una realidad mediante la programación y prueba de los algoritmos propuestos en un lenguaje de programación de alto nivel que permitió poder evaluar la efectividad de los algoritmos propuestos mediante la aplicación de datos reales tomados de bases de datos de rutas metabólicas reales.

Acá se debió trabajar y seleccionar alguno de los formatos de datos del primer objetivo y con información puntual de alguna base de datos metabólica de importancia.

4. Evaluar efectividad de algoritmos y comparar.

Este objetivo se logró mediante la comparación de los resultados obtenidos y la información accesible por las herramientas disponible. Debido a la diferencia de los datos, de los algoritmos y su comportamiento y enfoque no fue posible realizar un posterior Análisis de Varianza de los datos obtenidos con los nuevos algoritmos podrá determinar si hay mejora en el costo computacional asociado a los algoritmos propuestos en contraste con los algoritmos existentes. Sin embargo, los resultados obtenidos, mostrados más adelante, elucidan evidencia de la validez de los algoritmos propuestos.

Acá se debe tomar como referente los datos del análisis realizado en el punto de objetivo número 2 y compararlos con los datos de pruebas realizadas a los algoritmos propuestos programados en el punto de objetivo número 3.

La efectividad esperada se refiere a una mejora asociada al costo computacional en el tiempo de ejecución de los algoritmos, se muestra que el costo computacional es bajo, al mismo nivel que algoritmos clásicos de grafos y mucho menor a los costos NP en contraste con los algoritmos vigentes.

6. Resultados

Para referencia del lector a continuación se explican acá brevemente los 2 algoritmos para tratar el problema de alineamiento o comparación de dígrafos que corresponden a rutas metabólicas propuestos. La explicación ampliada e ilustrada se puede revisar en el artículo publicado en el iWOBI 2017 que se adjunta en los anexos. Para cada algoritmo se presentan los resultados relevantes obtenidos.

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

Al analizar un grafo contra otro para compararlos y obtener algún valor de similitud se deben considerar dos aspectos relevantes para el caso particular de las rutas metabólicas y algoritmos sobre grafos ya existentes.

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: un sustrato 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.

Sobre grafos se han descrito diversos algoritmos que permiten obtener todos sus nodos de forma eficiente, generar rutas óptimas entre cualquier par de nodos, etc. 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.

Se debe tener en cuenta que habrá una pérdida de información en dicha transformación. Se demuestra acá que dicha pérdida de información durante el proceso es tolerable y aceptable para un resultado certero de comparación.

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 [16], semiglobal, and local [15]; para obtener valores de comparación de dichas secuencias lineales.

En grafos es común que se recurra a hacer dos tipos de recorridos, por profundidad o por anchura. 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. Los valores alcanzados como valor de comparación por el alineamiento global fueron positivos cuando es evidente que se cuenta con rutas similares y el alineamiento local en los mismos casos brinda un valor aún mayor para la sección de ruta metabólica que más se asemeja entre ambas.

6.2 Algoritmo 2: Diferenciación por pares

Muchas veces 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 una comparación global para resaltar en su lugar los puntos divergentes entre un par de rutas dadas. Tomando como base la información alcanzada que se muestra en la figura 7 sobre las dos rutas de ejemplo. Se propone buscar los pares de reacciones comunes entre ambos grafos y eliminarlos de los mismos. Este proceso se muestra en la serie de pasos en el artículo publicado.

6.3 Pruebas y corrida de los algoritmos

Para mostrar una prueba realizada al software que implementa los algoritmos presentados se seleccionó la ruta "Steroid degradation" (ko00984) de la base de datos KEGG. Se utilizaron los organismos Mycobacterium tuberculosis (mtu00984) y Pseudomonas putida (pput00984) para pruebas. La figura 1 muestra las rutas, con a) para mtu00984 y b) pput00984. Los archivos KGML (XML) para cada ruta y organismo fueron descargados e introducidos en la herramienta desarrollada.

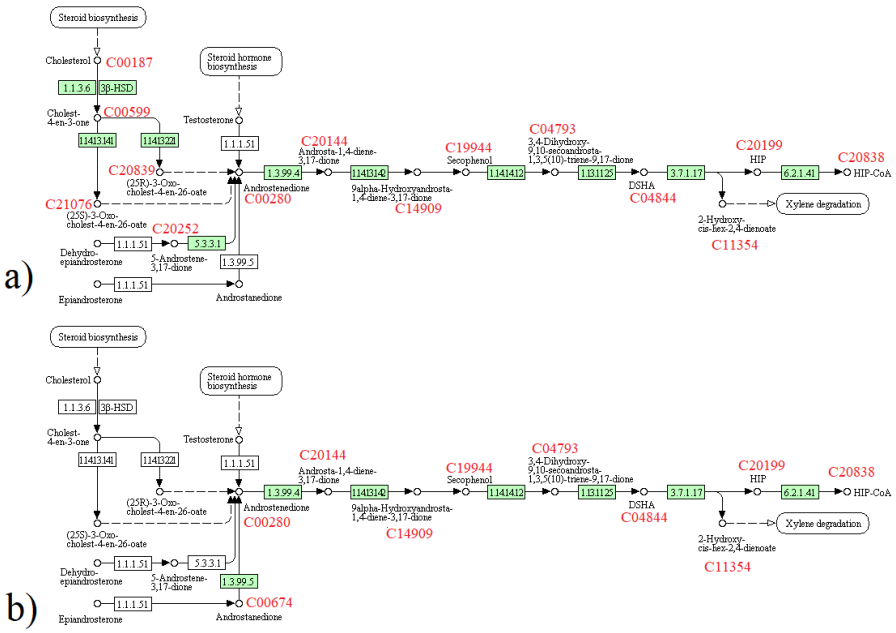

Figura. 1. Salida gráfica de la base de datos KEGG, para la ruta "steroid biosynthesis". Los metabolites están representados con puntos blancos, enzimas representadas con cajas, con su respective número EC. Las cajas coloreadas representan enzimas presents en el organism seleccionado. a) es la ruta para Mycobacterium tuberculosis, b) es la ruta Pseudomonas putida. Fuente: KEGG Pathway [3]. Las etiquetas rojas están en número C-code usados por KEGG.

Para las rutas mostradas en la figura 1, la salida obtenida desde la herramienta implementada es la mostrada a continuación. Se hace uso de un diccionario de datos para acortar los nombres largos de los metabolitos y facilitar la visualización de los alineamientos.

```
Breadth First Traversal (BFT)
  Pathway 1: C00187 C00599 C21076 C20839 C00280 C20144 C14909 C19944 C04793
  C04844 C20199 C11354 C20838
  Pathway 2: C00674 C00280 C20144 C14909 C19944 C04793 C04844 C20199 C11354
  C20838

Depth First Traversal (DFT)
  Pathway1: C00187 C00599 C20839 C00280 C20144 C14909 C19944 C04793 C04844
  C11354 C20199 C20838 C21076
  Pathway2: C00674 C00280 C20144 C14909 C19944 C04793 C04844 C11354 C20199
  C20838

Dictionary
 A0 -> C00187          B0 -> C00599          C0 -> C20839
 D0 -> C00280          E0 -> C20144          F0 -> C14909
 G0 -> C19944          H0 -> C04793          I0 -> C04844
 J0 -> C11354          K0 -> C20199          L0 -> C20838
 M0 -> C21076          N0 -> C00674

Low Detail Output
Breadth First Search (BFT)
 Pathway 1: A0 B0 C0 D0 E0 F0 G0 H0 I0 J0 K0 L0 M0
 Pathway 2: N0 E0 F0 G0 H0 I0 J0 K0 L0 M0
```

```
Depth First Search (DFT)
 Pathway 1: A0 B0 C0 D0 E0 F0 G0 H0 I0 J0 K0 L0 M0
 Pathway 2: N0 D0 E0 F0 G0 H0 I0 J0 K0 L0

Alignment Algorithms
Global Alignment (BFT):
 A0B0C0D0E0F0G0H0I0J0K0L0M0
 ------N0E0F0G0H0I0J0K0L0M0
Score: 2
Global Alignment (DFT):
 A0B0C0D0E0F0G0H0I0J0K0L0M0
 ----N0D0E0F0G0H0I0J0K0L0--
Score: 2

Local Alignment (BFT):
 E0F0G0H0I0J0K0L0M0
 E0F0G0H0I0J0K0L0M0
Score: 9
Local Alignment (DFT):
 D0E0F0G0H0I0J0K0L0
 D0E0F0G0H0I0J0K0L0
Score: 9

Semiglobal Alignment (BFT):
 A0B0C0D0E0F0G0H0I0J0K0L0M0
 ------N0E0F0G0H0I0J0K0L0M0
Score: 2
Semiglobal Alignment (DFT):
 A0B0C0D0E0F0G0H0I0J0K0L0
 ----N0D0E0F0G0H0I0J0K0L0
Score: 4

Equality
 C00280 -> C20144
 C20144 -> C14909
 C14909 -> C19944
 C19944 -> C04793
 C04793 -> C04844
 C04844 -> C20199
 C04844 -> C11354
 C20199 -> C20838

Differences Identified
Pathway 1:
 C00187 -> C00599
 C00599 -> C21076
 C00599 -> C20839
 C21076 -> C00280
 C20839 -> C00280
 C20252 -> C00280
Pathway 2:
 C00674 -> C00280
```

7. Discusión y conclusiones

Primeramente, se debe valorar el costo de los algoritmos utilizados para mostrar que son más baratos que los usados hasta ahora. El segundo paso es demostrar que el procedimiento brinda un resultado certero y útil con respecto a la comparación en sí.

Para el procedimiento del primer algoritmo se hace uso del recorrido de grafos o búsqueda por anchura o por profundidad. Como se indicó previamente el usar un algoritmo por profundidad no brinda información similar a la que describe una ruta y los resultados para diversos grafos pueden ser casi aleatorios. En el caso de la búsqueda por anchura se va realizando un recorrido por niveles, similar a la forma en que funciona una ruta metabólica como tal. El costo de estos algoritmos se aproxima en el orden de 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 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 en el orden de O(R_1 x R_2).

Se busca alcanzar un buen valor de precisión en la comparación si sacrificar exactitud en el proceso. El resultado alcanzado es ganar tiempo en un procedimiento simple y sin perder veracidad.

Al aplicar los algoritmos propuestos se obtuvieron valores de comparación efectivos positivos para el alineamiento global y aún mayores para el alineamiento local. Es fácil de constatar la similitud evidente entre las rutas analizadas hasta ahora y además se presenta un mecanismo de evaluación que nos brinda un valor de comparación.

Al realizar pruebas con una ruta diferente en forma y contenido como la que se observa luego de aplicar la transformación mediante el algoritmo 1 y su posterior alineamiento los resultados varían. En el caso del alineamiento de las rutas, los valores alcanzados fueron valores negativos muy bajos, lo que indica que el valor de comparación evidentemente es muy bajo tanto para el alineamiento global como para el alineamiento local. Nuevamente, si se realiza una comparación entre estas se puede observar que ambas son bastante disímiles.

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. Luego de aplicar al algoritmo mostrado a las mismas rutas de trabajo se obtuvieron las diferencias listadas como se mostró

antes. Para cada ruta metabólica se listan las reacciones presentes en ella misma que no están presentes en la ruta contraria. 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.

Al aplicar el algoritmo 2 a las rutas que varían mucho es evidente que hay una gran cantidad de reacciones diferentes entre estas rutas donde se obtiene una cantidad muy elevada de diferencias.

Conclusiones

Se puede establecer que el mecanismo propuesto en el algoritmo 1: “Transformación de Grafo 2D a 1D para posterior alineamiento y evaluación” puede ser usado como evaluador previo y rápido para predecir buenas comparaciones en caso que se desee un análisis más profundo.

En el caso del algoritmo 2: “Diferenciación por pares” la propuesta es brindar al experto, un punto de vista adicional para sus valoraciones sobre las rutas que estudia. No se brinda un valor numérico pero si información que permita al investigador discriminar entre los datos.

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 entre ambas rutas analizadas.

La herramienta implementada brinda una forma eficiente y de bajo costo para comparar patrones de rutas metabólicas de una forma computacional de muy bajo costo, simplificando los datos, sin comprometer la correctitud de similaridad entre un par de rutas dadas.

Durante la implementación se encontró que uno de los puntos críticos para acertar los valores se debe a la selección correcta de la raíz del grafo a trabajar. Esto sucede ya que al realizar los recorridos de los grafos para su posterior alineamiento, si se hace al azar o en una posición muy diferente del inicio real del proceso biológico el recorrido obtenido no se asemeja a los datos correctos del orden de reacciones del proceso en análisis. El no hacer esto de forma correcta brindaría datos muy variados y no cercanos a una comparación real del proceso. Debido a esto, en la herramienta se ha incluido un control adicional al usuario. El mismo sugiere el punto de partida o raíz para el recorrido, pero el usuario podría seleccionar otro entre la lista de posibles metabolitos. Esto también da paso para que cualquier tipo de grafo se pueda comparar en un futuro.

8. Recomendaciones

Al haber comprobado que los algoritmos propuestos pueden brindar información relevante para el análisis y comparación de rutas metabólicas y una herramienta de software que procesa los datos sería interesante agregar a esta herramienta un servicio que acceda directamente a las bases de datos de metabolismo, extraiga la información de rutas metabólicas de interés y permita aplicar los algoritmos propuestos para su análisis futuro por parte de expertos. El problema que esto implica es que las bases de datos no son compatibles entre sí, varía la información que tienen disponible de rutas y organismos, así como los formatos de archivos. No todas tienen un servicio de consulta automatizado, sino solo procesos manuales de búsqueda y descarga.

Recientemente se han propuesto mecanismos de comparación de secuencias genómicas basado en matrices de valoración, de forma tal que no solo se valoren los elementos que son iguales o diferentes utilizando los mismos valores de alineamiento para todos los elementos sino más bien que se tome en cuenta la afinidad entre los elementos. Por ejemplo, en el caso de proteínas si éstas son de familias similares: hidrofóbicas, azúcares, polaridad positiva o no, etc.; así como aspectos energéticos y de probabilidad de reacciones. Se penaliza menos cuando una proteína se cambia por otra de una misma clase que cuando lo hace por otra. Sería interesante considerar en la comparación de estructuras aspectos como los mencionados para brindar al investigador una fuente adicional de información tanto para metabolitos como para enzimas que participan en el proceso.

Se debe considerar luego las interacciones de unas rutas con otras. Los metabolitos pueden ser el producto final de una ruta o bien un producto intermedio que puede ser a la vez precursor para otras vías metabólicas. Se debe ampliar el análisis a combinar estas rutas para ser tratadas también en su contexto como redes metabólicas.

Al igual que se cuenta con algoritmos de alineamiento múltiple para varias secuencias genéticas se debe continuar trabajando en la comparación de múltiples rutas para encontrar por ejemplo factores similares entre distintas especies, tomando en cuenta aspectos como los mencionados acá.

Un paso siguiente en la comparación de rutas metabólicas puede ser la comparación de enzimas en sí. Diferentes enzimas en diferentes organismos pueden catalizar la misma reacción, pero algunas pueden tener una mejor actividad en algunas condiciones; por lo que la comparación a nivel enzimático podría brindar una gran cantidad de información.

9. Agradecimientos

Se agradece el gran apoyo de la Escuela de Computación, de la Vicerrectoría de Investigación y Extensión y de la Vicerrectoría de Docencia para este proyecto. A la comisión de Becas de Posgrado de la Dirección de Postgrados. Así como al Comité Técnico de la Escuela de Computación y al Centro de Investigaciones en Computación y al Programa de Maestría en Computación.

También se desea agradecer la colaboración y el apoyo en gestiones para este proyecto por parte del profesor Francisco Torres Rojas por su acompañamiento. Y muy especialmente a los estudiantes: Seth Stalley, Kevin Castro Fuentes y Pablo Vargas Rosales. A los estudiantes del curso Electiva Introducción a la Biología Molecular Computacional por tomar su tiempo para la lectura de este trabajo para su publicación y colaboración con pruebas.

10. Referencias

1. Bruce, A., Dennis. B., Julian, L., Martin, R., Keith, R., James D., W. Molecular Biology Of The Cell third edition. (1994).
2. Lee, J. M., Gianchandani, E. P., Eddy, J. A., & Papin, J. A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS Comput Biol, 4(5), e1000086 (2008).
3. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353-D361. (2017).
4. Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., ... & Krummenacker, M. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*, *42*(D1), D459-D471. (2014).
5. Caetano-Anollés, G., Kim, H. S., & Mittenthal, J. E. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proceedings of the National Academy of Sciences*, *104*(22), 9358-9363. (2007).
6. Baldan, P., Cocco, N., Giummolè, F., & Simeoni, M. Comparing Metabolic Pathways through Reactions and Potential Fluxes. In *Transactions on Petri Nets and Other Models of Concurrency VIII* (pp. 1-23). Springer Berlin Heidelberg. (2013).
7. Guimerà, R., Sales-Pardo, M., & Amaral, L. A. N. A network-based method for target selection in metabolic networks. Bioinformatics, 23(13), 1616-1622. (2007).
8. Mithani, A., Hein, J., & Preston, G. M. Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and non-pathogenic lifestyles in Pseudomonas. Molecular Biology and Evolution, msq213. (2010).
9. Heymans, M., & Singh, A. K. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics, 19(suppl 1), i138-i146. (2003).
10. Kutmon, M., van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R., & Evelo, C. T. PathVisio 3: an extendable pathway analysis toolbox. PLoS Comput Biol, 11(2), e1004085. (2015).
11. Jensen, P. A., & Papin, J. A. MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. Bioinformatics, 30(9), 1327-1328. (2014).
12. Hu, J., Kehr, B., & Reinert, K. NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. Bioinformatics, btt715. (2013).
13. Abaka, G., Bıyıkoğlu, T., & Erten, C. CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways. Bioinformatics, 29(13), i145-i153. (2013).
14. Ay, F., Kellis, M., & Kahveci, T. SubMAP: aligning metabolic pathways with subnetwork mappings. Journal of computational biology, 18(3), 219-235. (2011).
15. Needleman, S. B., & Wunsch, C. D. 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. (1970).
16. Smith, T. F., & Waterman, M. S. Identification of common molecular subsequences. Journal of molecular biology, 147(1), 195-197. (1981).

11. Apéndices

1. Se adjunta el artículo aceptado y presentado en la **5th IEEE International Work Conference on Bioinspired Intelligence (IWOBI 17)** en Funchal, Portugal el 18 de Julio de este año, http://iwobi.ulpgc.es/2017/, el mismo fue publicado en IEEE Xplore: http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=42403

   http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?filter%3DAND%28p_IS_Number%3A7985514%29&rowsPerPage=50&pageNumber=1&resultAction=REFINE&resultAction=ROWS_PER_PAGE

   Arias-Méndez & Torres-Rojas. Alternative low cost algorithms for metabolic pathway comparison. 5th International Conference and Workshop on Bioinspired Intelligence. Funchal, July 2017. Unpublished conference paper. Tecnológico de Costa Rica. (In press).

2. Se adjunta la presentación utilizada para la presentación del artículo del Anexo 1 durante el iWOBI 2017.

3. Se adjunta el artículo enviado al **22nd IBEROAMERICAN CONGRESS ON PATTERN RECOGNITION** que será el 7-10 November 2017, Valparaíso, Chile. http://www.ciarp2017.org/. El mismo aún se encuentra en proceso de revisión.

   Arias-Méndez, E., Castro-Fuentes, K., Stalley, S., Vargas-Rosales, P. A web tool for executing low cost algorithms for metabolic pathway comparison based on graph pattern similarity. CIARP 2017. Manuscript submitted for publication. Tecnológico de Costa Rica.

4. Se adjuntan hojas de datos con información recopilada y tabulada sobre bases de datos, herramientas y algoritmos.

5. Se adjunta el artículo aceptado en las Jornadas Costarricenses de Investigación en Computación e Informática que se realizarán en el TEC el 23 y 24 de Agosto.

   Esteban Arias-Méndez, Francisco Torres-Rojas. "Comparación de rutas metabólicas mediante algoritmos de evaluación simples. JoCICI 2017. August 2017. Unpublished conference paper. Tecnológico de Costa Rica. (In press).