Escuela de Computación
Bachillerato en Ingeniería en Computación

TEC | Tecnológico de Costa Rica

Proyecto #2

Comparación de Rutas Metabólicas

IC-8022 Introducción a la Biología Molecular Computacional

Tecnológico de Costa Rica, Sede Central Cartago

Escuela de Computación, Ingeniería en Computación

I Semestre 2017

Prof. Ing. Esteban Arias Méndez

La comparación de rutas metabólicas es una operación importante hoy en día para el estudio de metabolismo aplicable en múltiples áreas de interés como medicina, agricultura, farmacia, biotecnología y otros.

INTRODUCCIÓN

Propósitos

- Aplicar conocimientos aprendidos sobre alineamiento y comparación en estructuras de datos más complejas a las lineales como las usadas hasta ahora.
- Poner en práctica la evaluación de algoritmos para su validación, uso y pruebas.

DESARROLLO

Utilizando 2 algoritmos propuestos por el profesor deberá evaluarlos para verificar su utilidad como mecanismo para comparar rutas metabólicas.

Se trabajará sólo sobre el Sistema Operativo Linux, usando solamente lenguajes C o Python. Se recomienda trabajar su código en módulos, funciones o procedimientos, incluso en varios archivos para facilitar el trabajo. Debe hacer uso del git de la Escuela de Computación para llevar todo el desarrollo de su proyecto: git.ec.tec.ac.cr

Este trabajo es para ser desarrollado por parejas de 2 personas máximo.

Para su proyecto deberá trabajar realizando las siguientes actividades:

1. Investigar sobre el problema de comparación de rutas metabólicas, indicar porqué el mismo es importante. Refiérase con fuentes documentadas al costo computacional que se ha descrito y herramientas que se han usado hasta ahora para tal fin.
2. Se deben consultar Bases de Datos de rutas metabólicas y documentar los formatos en que se pueda accesar dicha información de forma automática (textual).
3. Buscar rutas metabólicas de ejemplo, al menos 15 rutas, obtener su diagrama gráfico y su formato descrito en texto (las reacciones).
4. Para cada ruta de trabajo deberá escribirla en un formato de archivo de texto tipo json, como en los ejemplos provistos.
5. Aplicar los 2 algoritmos provistos a sus archivos en formato .json a sus 15 rutas para obtener similitudes y diferencias entre todas las rutas.
6. Investigar sobre herramientas actuales de comparación y hacer pruebas de comparación de sus 15 rutas en alguna herramienta de software existente
7. Compare los datos obtenidos de similitud entre las herramientas encontradas y los algoritmos propuestos.
8. Luego de las pruebas realizadas, deberá tabular la información y generar conclusiones al respecto de los algoritmos dados.
9. Brinde recomendaciones de mejora sobre el código, los algoritmos y el texto explicativo que se incluye al final en este documento.

Las rutas de trabajo pueden ser de procesos similares entre organismos distintos por ejemplo e incluso de procesos diferentes para realizar las comparaciones.

Cada pareja de trabajo debe usar 15 rutas diferentes entre sí. Pueden ponerse de acuerdo entre parejas si quieren compartir datos entre rutas similares corridas entre cada pareja, por ejemplo para trabajar una ruta en común o un organismo en particular.

Se provee un documento de presentación usado en clase que describe los algoritmos: 20170509 - comparación de rutas metabólicas.pptx. Además, a continuación, un texto que describe los algoritmos usados y los mecanismos usados para su implementación.

Se provee en código Python la implementación de dichos algoritmos para pruebas sobre documentos con rutas tipo .json como los ejemplos provistos.

No se proveen los algoritmos de alineamiento. Deberá coordinar con sus compañeros de curso, para que todos usen el mismo código de alineamientos global, local y semiglobal en Python o C. Todos deben usar el mismo código para que las pruebas y resultados obtenidos entre todos sean homólogas.

El código dado en Python, toma los archivos del directorio de rutas para listarlos al usuario e indicar cuales archivos desea trabajar. Posteriormente siga las instrucciones. Puede usar recorridos de los grafos por anchura o por profundidad, modificando el código. Luego a los datos obtenidos deberá aplicarles los alineamientos: global, local y semiglobal y tabular los datos obtenidos para cada experimento realizado. Debe hacer pruebas de alineamiento con los 3 métodos y con ambos tipos de recorrido, para cada posible combinación de las 15 rutas de trabajo, para validar la información. El programa puede mostrar los datos usando los nombres originales en las rutas o bien nombres codificados usando letras para simplificar el alineamiento.

EVALUACIÓN

1. Rúbrica de evaluación:

    - El proyecto se calificará con los siguientes criterios:

        i. 80% - Pruebas completas de sus 15 rutas, comparación entre todas ellas usando ambos recorridos de grafos y los 3 alineamientos para cada par de rutas comparadas. Proveer conclusiones relevantes sobre el proceso y recomendaciones sobre los algoritmos, resultados, implementación y el texto dado.
        ii. 20% - Documentación completa del trabajo.

2. El proyecto debe resolverse, implementándolo de la mejor manera.

3. De forma global, se evaluará la presentación del trabajo según los parámetros solicitados, estrategias empleadas y la calidad, la entrega a tiempo del trabajo y la documentación completa correspondiente.

4. Sobre la documentación y presentación:

    a. 2pts - El subject del correo a ser enviado debe ser:
    [BMC] – Proyecto # 2 – Sus Nombres Completos
    b. 2pts - El correo debe contener de forma separada:
        i. los archivos de texto de los códigos fuentes que permiten la solución y funcionalidad del mismo. Si se hicieron cambios al código provisto se deben documentar los cambios introducidos.
        ii. un archivo PDF con la documentación completa

        No envíe archivos ejecutables o binarios.

    c. La documentación en PDF con el nombre de archivo igual al subject del correo enviado. Esta documentación debe tener un apartado, que indique los pasos a seguir, para poder ejecutar el código (librerías a instalar y otros) en caso de usar herramientas adicionales a las brindadas.
        i. 5pts – La documentación debe incluir una portada con los datos completos: TEC, carrera, sede, curso y código, profesor, periodo, fecha de entrega, número de proyecto, título del proyecto, nombres completos con número de carnet de la pareja de trabajo y un abstract en inglés en la misma portada.

ii. 5pts – La introducción del documento es una descripción breve del trabajo realizado y herramientas usadas. Como mini-marco teórico incluya las referencias a los algoritmos implementados y las herramientas empleadas, así como otras fuentes de consulta utilizadas.
iii. 10pts – Como desarrollo debe explicar, los procedimientos, rutinas, la lógica que utilizo para resolver el problema. indicar los ejemplos de código que ha usado como guía para el desarrollo de los mismos usando las referencias bibliográficas correspondientes. Explicar el uso y funcionamiento.
iv. 20pts – Análisis de resultados, explicando el trabajo implementado, la forma de realización, funcionamiento, general, ejemplos documentados, problemas presentados, estructuras de datos empleadas, algoritmos usados, etc.
v. 10pts – Una sección de conclusiones y/o observaciones sobre el proyecto.
vi. 10pts – En una sección de Apéndices incluya el código fuente documentado del proyecto y su explicación. En caso que haya hecho cambios al código dado. Explique la estructura del código empleada en su proyecto, módulos, etc.

5. La tarea puede realizarse de forma individual o en parejas de 2 máximo.

6. Se debe entregar en digital a más tardar el Miércoles 31 de Mayo, antes de la media noche. Debe hacerlo de forma simultánea a los correos siguientes y copiarse usted mismo y su compañero de trabajo. Cada día de atraso serán 15pts menos de la nota de la tarea:

    a. earias@ic-itcr.ac.cr
    b. kecastro@ic-itcr.ac.cr

7. Cualquier consulta puede hacerla al foro, o personalmente en clase o al correo del profesor con copia al asistente a los correos anteriores.

8. Durante la revisión del proyecto deben estar presentes ambos miembros de la pareja de trabajo, la no presentación les restará 10pts a cada uno de los ausentes.

9. De entregar esta tarea antes de la fecha indicada se reconocerán puntos extra, entre más días antes más puntos extra !!!

INFORMACIÓN

Alternative low cost algorithms for metabolic pathway comparison

Esteban Arias-Méndez
Ingeniería en Computación, Escuela de Computación, Tecnológico de Costa Rica
[Costa Rica Institute of Technology, School of Computing, Computing Engineering]
Member of the PAttern Recognition and Machine Learning (PARMA) Group and The Happy Few.
Cartago, Costa Rica
esteban.arias@tec.ac.cr

***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 are based on the idea of pathway comparison using multiple graph data structures; graph comparison has been defined as a computationally complex task. We propose two 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 that of 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 and thus eliminating all similarities, finally, showing these differences to the user.**

Keywords—metabolic pathway comparison; graph comparison

Introduction

Nowadays, we consider living organisms to be those that go through a common process known as the cycle of life: birth, growth, development and death. The cell is considered the smallest unit of life of a living being, in which many of the processes which support this cycle take place.

Bioinformatics or computational biology are some examples of the great progress that molecular biology has achieved since computer techniques began to be applied towards this discipline. For example, genome sequencing, as well as the process of analysis being used to interpret this information. Proteomics, epigenetics, and metabolomics, are areas that are showing to have great impact in several other fields such as medicine, agriculture, health, among others.

In the case of metabolomics, the main interest is to know and understand the metabolic processes that occur in organisms. It is, the set of biochemical and physiochemical reactions that happen on a cellular level. Complex interrelated processes that function as a base for life at a molecular scale are necessary for every single cell activity, such as: growth, development, structure keeping, and responding to stimuli.

There are a variety of metabolic data bases that stores the specifications of these processes called metabolic pathways. Data has been stored in a way that is similar to that of a directed graph data structure, which is used in computer science to shape relationships and to describe processes. Inquiries can be made through protein, metabolite, gene or gene abbreviation; it will depend on the target and the organization of every base. KEGG (www.genome.jp/kegg/) [11] and MetaCyc (part of BioCyc https://biocyc.org) [5] are examples of the best and most important repositories of such data, providing access to metabolic pathways of several organisms such as animals and plants.

Metabolic Pathways and Graphs

A metabolic path is an ordered sequence of biochemical reactions between various actors named metabolites, these are substrates that through enzymatic actions catalyze

transformations until getting a product [3],[16]. This is a process of molecular transformation of metabolites into to others. A big number of metabolic pathways are at present unknown, and a lot of reactions are missing on the already known pathways. Online databases provide mechanisms to predict metabolic pathways per previously described elements. It is important to automate new reconstructions of metabolic pathways that explain unknown reactions and in turn link missing pieces of the pathways that are being established, or to extrapolate them into other related species (see figures 1 and 2). Finally, some pathways have curated data that has been verified by several collaborating laboratories.

To study metabolic processes, it is necessary to span over several areas of knowledge to analyze all the available information. When studying the metabolism, besides knowing about the metabolites involved, it is important to acknowledge the steps or reaction between them, the metabolic pathway. These pathways can be impacted or divided into bigger and more complex processes that make up a network of metabolic pathways (figure 1) when several of these interact.

Biologically there is not one only way to obtain a result, there could be a lot of different procedures which lead to the exact same result; modifying in a bigger or lower way the procedure.

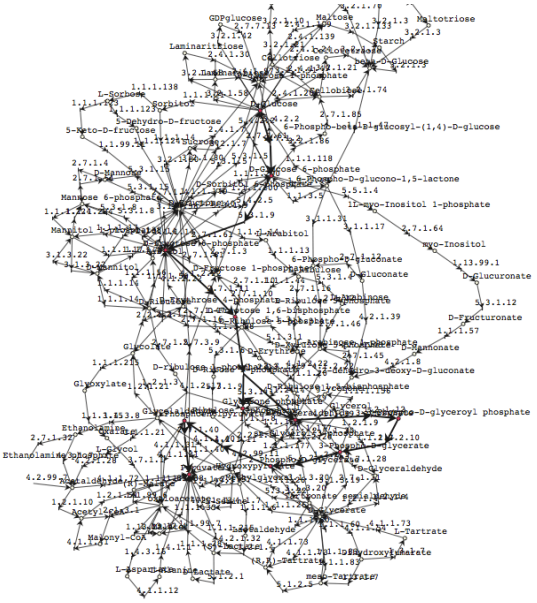


Fig. 1. Shows a network of pathways, which is the result of work done by [13] using computational prediction methods to combine data concerning the relationships between different compounds obtained and between different organisms in the databases: KEGG, ENZYME and BRENDA and as such obtain a network of reactions related to the Glycolysis process which is an essential process in Many organisms (used to obtain energy (pyruvates) from sugars (glucose).

One of the most important areas of research, is the comparison of metabolic pathways of processes of interest at an agronomic, pharmaceutical, medical, and commercial level.

Reasons are abundant and important. For example, knowing these processes can give us tools to modify said processes with the goal to produce more or better food, learn how to control different viruses or diseases at a cellular level for a better bio control, as well as to improve the development of medicines or more effective treatments. Since the acknowledge about pathways and metabolic networks is important when dealing with medicines in clinical studies concerning the action-reaction effect of drugs on organisms.

A clear example, would be plants, knowing a metabolic network can be used as a tool to improve cultivation techniques to extract components that are important in to the topic of improving human food supply by maximizing food production.

A better understanding of phylogenetic evolution, speciation and reconstruction [17], [8] and the discovery of more effective drugs [7] may be possible thanks to the comparative analysis of different organism's pathways.

To facilitate this analysis of metabolic pathways, graphs are used and directed graph structures are used to describe the metabolites involved in each process and their interactions as reactions. Some tools like PathVisio [14], MetDraw [10] or NetCoffe [9], provide basic information about pathways (routes), their components, graphs and other information, but not analysis tools such as a direct route comparison.

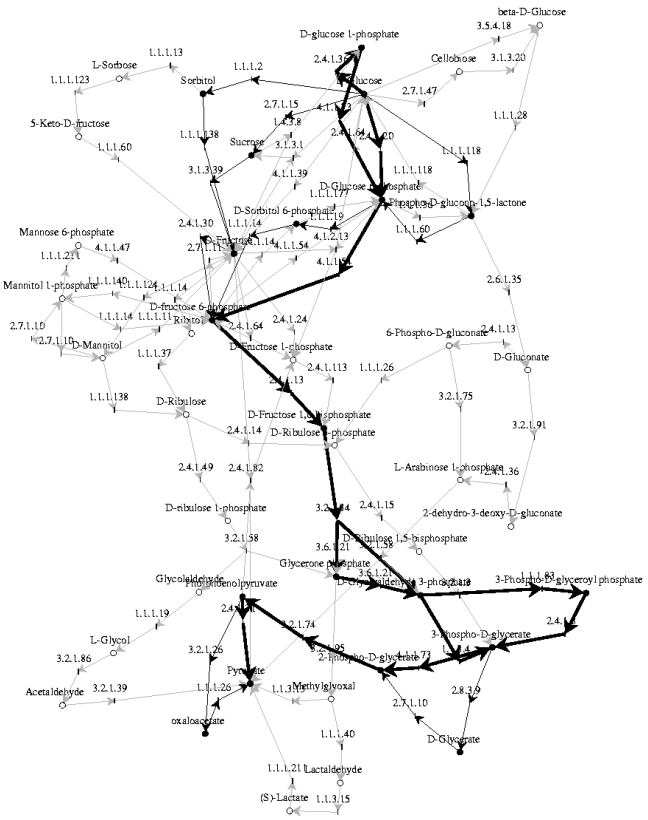


Fig. 2. Shows the common elements (in bolded lines) between two particular organisms, yeast and MG (mycoplasma genitalium) for the calculated network resulting from Figure 1, [13].

Computer Scientists have proposed several mechanisms for effective comparison of these routes within a species or between species. To this regard Abaka et. al [1] has done a review of the most important tools developed until now, including NP costs associated with the alignment of two routes treated as graphs. Tasks that have been described as computationally complex, as mentioned also in Ay & Kahveci [2] who makes a proposal called SubMAP (Subnetwork Mappings in Alignment of Pathways) which does not provide a comparison of 1 to 1 or 1 to many as in the first approaches, but rather focuses on finding common sub-parts between different routes, or similar subnets. The CAMPways algorithm of Abaka et. al [1] promises to be more efficient at runtime than do called "state of the art" algorithms. However, this algorithm refers to two evaluations or measures that can be self-conflicting: similarity and topological similarity of the given routes. The analysis of the information obtained from routes, as well as the analysis of the previous tools, is a complex mechanism dependent upon interpretation and the processing of existing information

Here we present a different approach to the mechanisms used so far for the comparison of metabolic pathways and propose two simpler alternatives that can be used as a precede to a deeper, costlier, and more time-consuming analysis.

Graphs

To represent metabolic pathways, the use of graphs and dynamic data structures has been used to model various relationships between processes of all kinds. Generally, a route is represented as a directed graph in a computer. From there, different techniques have been developed to align and compare a directed graph, corresponding to the routes of interest whose associated complexity cost is NP. Complex algorithm solutions are then applied, which generally use heuristic techniques that seek to limit the time of alignment of a graph or route against another. This problem is much more complex when looking for a comparison between a single route and multiple routes.

The difference between two homologous routes and two similar routes must be considered. Homology can be described as a high-level comparison, while similarity is defined as a measurable and tangible valuation. We can say that two people are homologous because their general form is similar: 1 head, 2 arms, 1 trunk, 2 legs, 2 eyes, etc. However, although 2 people are homologous they might not be similar. In the case of routes, multiple route may possess the same amount of interactions or reactions and thus have a homologous form, but the reactant metabolites differ.

Simple low-cost comparison methods

Pathways, when viewed as graph-type data structures, allow the application of a wide variety of existing algorithms. In traditional literature concerning graphs it is not common to explore this type of comparative algorithms, but the traversal of all nodes within a graph or the exercise of finding the shortest path between two nodes is considered a common practice. That is, traditional algorithms such as the minimum spanning tree, minimum distances or shortest paths either between all nodes or a pair of given nodes.

In bioinformatics, the alignment techniques are valid for a step-by-step comparison of each stage of the metabolic pathway. An efficient comparison mechanism at the computational level is still required, which can then be used with different sources of information for the proper study of the metabolic routes of interest and their subsequent analysis.

By means of an alternative mechanism for the comparison of metabolic routes, it is sought to broaden the spectrum of results for subsequent analysis to establish new relationships or connections not previously described between routes or organisms. With different treatment of the given information, expressed in the digraph associated with the metabolic pathway, relevant results can be obtained while achieving lower computational cost.

We propose a different comparison approach for metabolic pathways by means of two alternative algorithms: first, we do not display the metabolic path strictly as a digraph, we do an initial transformation of said two-dimensional structure into a linear structure that is more simple and computationally cheap to align; another alternative would be to do pair analysis of reactions in the graphs, that is to say to analyze the relations 1 to 1 of reaction between two metabolites within the graphs, and thus obtain a common denominator of both structures, finally we then simplify said graphs so that it is easier to determine the points of divergence between the routes.

It is not the intention to give a definitive answer to the result of comparing two routes or to indicate that one method is better than another, rather we seek to provide an additional point of view as support, to be considered by an expert in the matter at the time of making their observations, evaluations and conclusions about the process they are studying. It is not sought to give a "correct" answer on which is the best route, only to provide reference information for the interested party.

Next, we explain the two algorithms applied to the problem of alignment or comparison of digraphs that correspond to metabolic pathways.

Algorithm 1: Transformation of 2D graph to a 1D or linear structure for later alignment and evaluation.

When analyzing a graph against another to obtain some similarity value, two aspects must be considered, which are relevant to the case of the metabolic pathways and existing graph algorithms.

In the case of metabolic pathways, it is common to observe in the description of the raw data obtained from the various databases that, although they are modeled as a graph with different relations between them and even containing internal cycles, it is characteristic that every route has two key elements: a starting point substrate and a final product as output. If the path is then viewed as a graph, this graph will have a root and an important target leaf node (using common nomenclature for trees in data structures).

Concerning graphs, we have described several algorithms that allow us to obtain all nodes in an efficient manner, a method for generating optimal routes between any pair of

nodes, etc. In the case of a graph which represents a metabolic path, when applying a traversal algorithm to the graph (which visits all the nodes) it becomes trivial obtain the list of elements that conform said graph. This would be a 2D to 1D transformation of the graph. If we take the starting point of the route as the root of the graph, then all the nodes must be visited until arriving at the node of interest that would be the final product of the route as such.

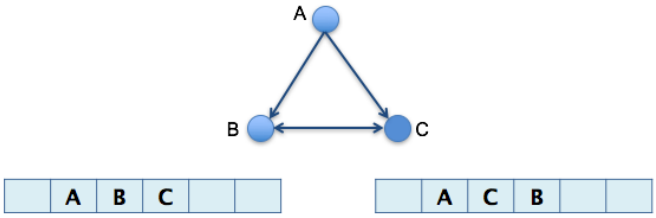


Fig. 3. Loss of information due to 2D to 1D transformation.

It should be noted that there will be a loss of information in such a transformation. Figure 3 shows this fact, mainly on the order of the elements and their original relationships. It seeks to demonstrate that such loss of information during the process is tolerable and acceptable for a correct comparison result.

After the transformation of the graphs to be analyzed, the linear information obtained is used to apply conventional alignment algorithms: global [18] and local [19]; with which we can obtain comparison values of said linear sequences.

Taking as reference the previously mentioned Glycolysis process, two examples from the MetaCyc database (metacyc.org), two glycolysis processes that provide the same product from two different sources can be observed in Figures 4 and 5.

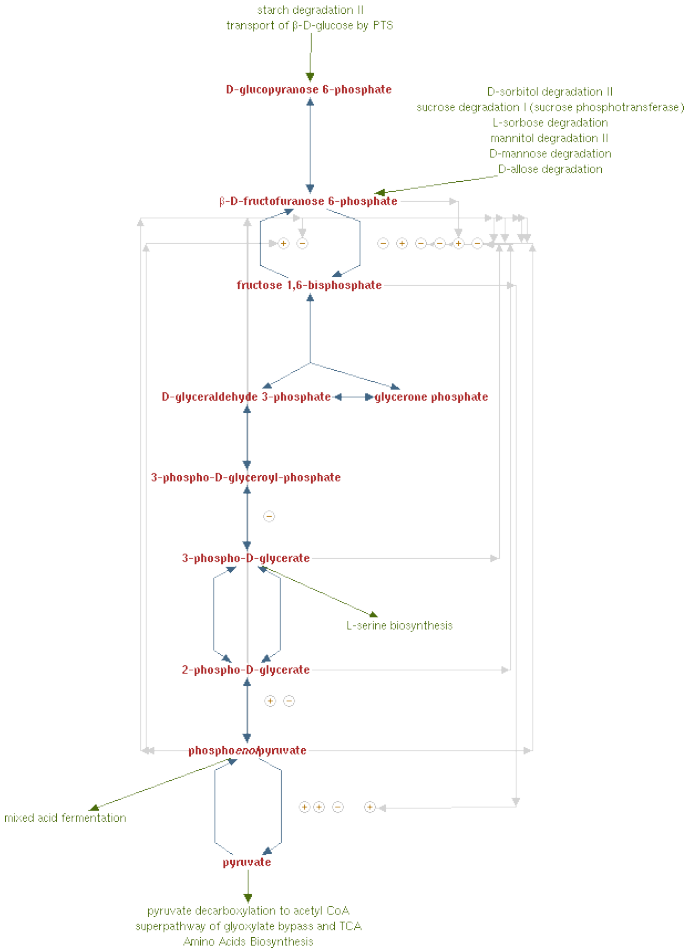


Fig. 4. MetaCyc Pathway: glycolysis I (from glucose-6P)[1]

It can be observed that both routes have elements in common and are easily homologous, but it is interesting to measure their similarity. The first step for this analysis is to visualize both routes as graphs and to label the nodes per the corresponding metabolites, see Figures 6 and 7.

Once the nodes have been simply labeled per their corresponding metabolite we proceed to perform graph traversal. There are two common graph traversal strategies, by depth [20] or by width [4], (see also [6],[12],[15]). When applying a depth first algorithm the information obtained is not relatively proportional and relevant to the route because the product will appear in the middle of the 1D row and not at the end of said row (as one might expect in a series of reactions which hold said product at the end). For example, depth first traversal would give a result as shown in figure 8. When performing a breadth first traversal, the nodes are visited by levels, which corresponds more closely to the way in which the metabolites reactions occur until the expected product is reached. Breadth first traversal for the routes shown is shown in figure 9.

1 From: http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=GLYCOLYSIS&detail-level=1&detail-level=0

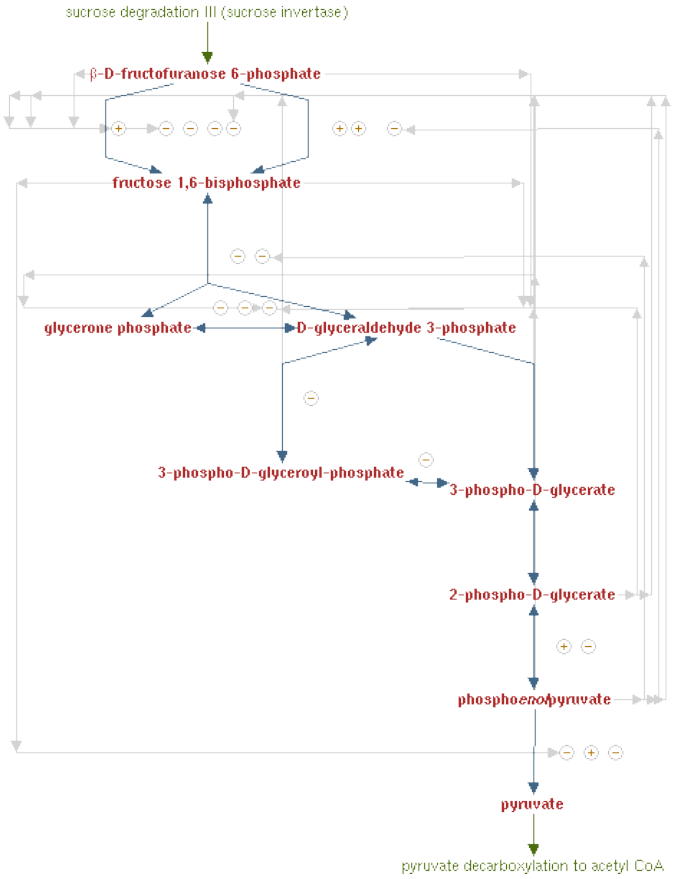


Fig. 5. MetaCyc Pathway: glycolysis IV (plant cytosol)[2]

Per this observation, useful data corresponds mainly to that generated by the breadth first traversal algorithm.

Once the route data is raised to obtain the routes in a 1D format we proceed to apply the traditional alignment algorithms, see figure 10. In this example the comparison value reached by the overall alignment is +3 which (in this case) indicates a good degree of similarity. Also, applying a local alignment between the same data gives us a value of +5, which represents the section of the metabolic path that most closely resembles the two.

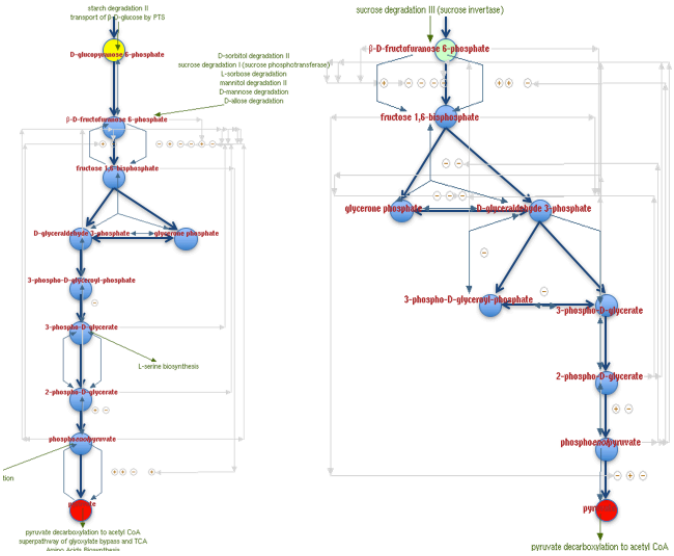


Fig. 6. Model metabolic pathways as graphs.

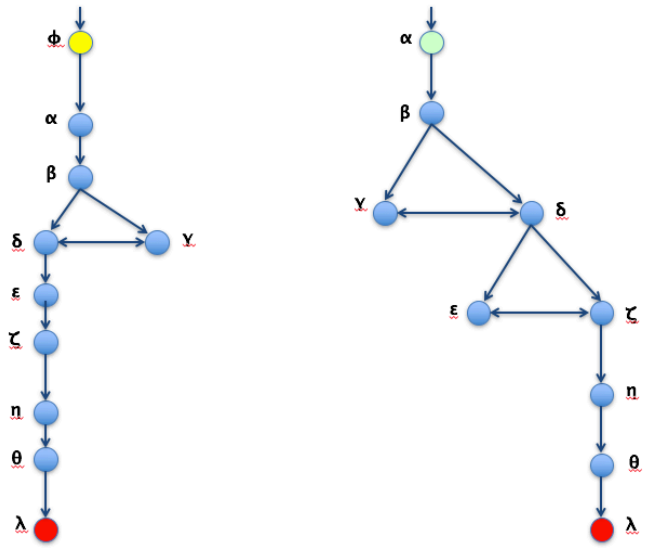


Fig. 7. Re-label nodes according to their corresponding metabolites to simplify processing.

Fig. 8. Depth traversals first of routes as is graphs in figures 6 and 7.

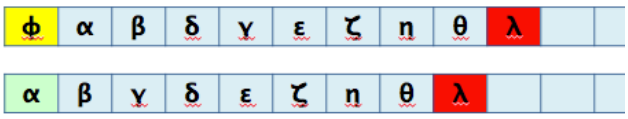
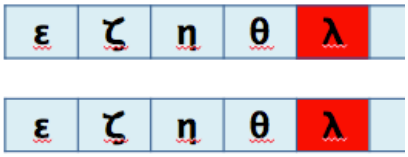


Fig. 9. Width traversals first of routes as is graphs in figures 6 and 7.

2 From: http://metacyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-1042

10.a) Global alignment, optimal value reached: +3



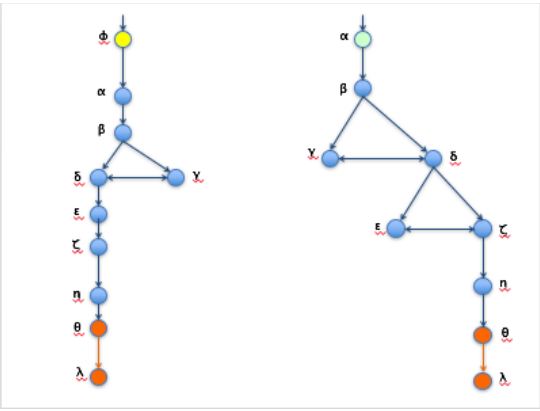
10.b) Local alignment, optimal value reached: +5

Fig. 10. Alignments generated from transformed graphs from 2D to 1D.

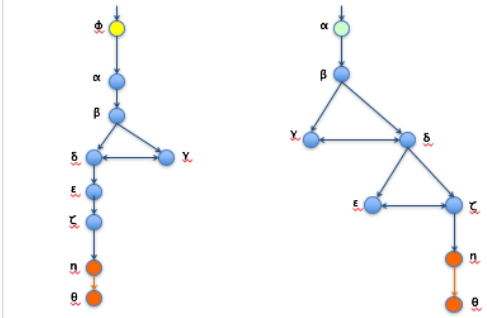
Algorithm 2: Differentiation by pairs

Many times, when we desire to compare two objects, the common elements are evident, it is then in such situations that it becomes more relevant to concentrate on looking for the differences. Based on this idea, this second algorithm seeks to eliminate from the graph the common pairs of objects, that is, equal reactions between the two graphs, to find the differences between the two paths.

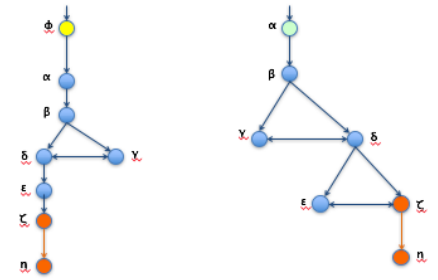
This would be a different approach to traditional alignment that performs a global comparison to highlight instead the divergent points between a given pair of routes. Based on the obtained information shown in figure 7 concerning the two sample pathways. Our proposal is to look for the pairs of reactions common between both graphs and to eliminate them. This process is shown in the series of steps that are described step-by-step in figure 11.



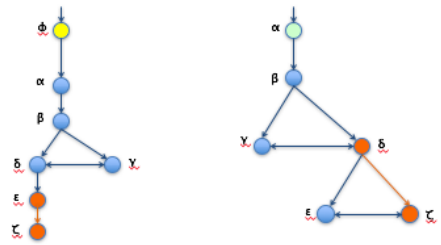
11.a



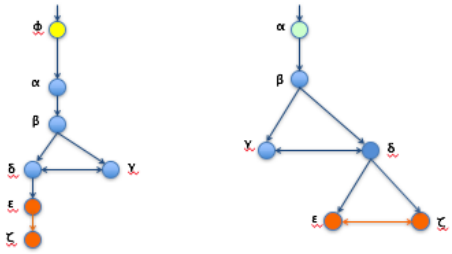
11.b



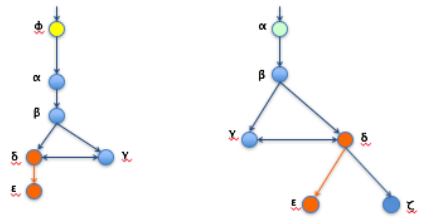
11.c



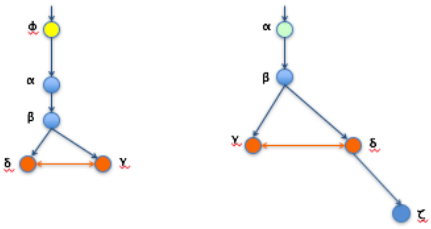
11.d



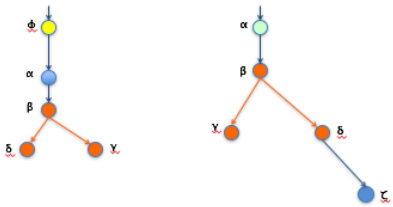
11.e



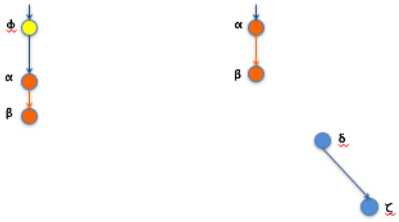
11.f



11.g



11.h



11.i

11.j

v11.k

Figure 11. Peer Differentiation step-by-step process. In 11.k) we observe the differences found between both sample routes after progressively eliminating the common reactions between both graphs.

Discussion: Tests and results

First, we must evaluate the cost of the algorithms used to show that they are less costly than the ones used so far. The second step is to demonstrate that the procedure provides a sure and useful result with respect to the comparison.

For the procedure of the first algorithm, we make use of graph traversal by breadth or by depth. As previously indicated, using a depth first traversal does not provide information like the one described by a pathway and the results for different graphs can be seemingly random. In the case of breadth first search, a level crossing is performed, similar to the way a metabolic route works as such. The cost of these algorithms approaches O (| V | + | E |), where V: is the set of vertices or nodes of the graph and E $\subseteq$ VxV: is the set of edges or arcs.

For the second algorithm, it must be considered that for each reaction that exists in the first G1 path or graph, it must be found in the second G2 path or graph. That is, if R1 is the number of reactions counted by G1 and R2 by the quantity for

G2, there will be a maximum R1 x R2 comparison, when it is common for a half-time on average to perform such comparisons. Thus, we can establish a worse case of O (R1xR2).

We can observe that the routes of Figures 4 and 5 are similar to each other. The goal is to achieve a good value of precision in the comparison without sacrificing accuracy in process. The result achieved is to gain time by means of a simple procedure and without losing truth.

After applying the proposed algorithms, in the above example, we obtained effective comparison values of +3 for the global alignment and of +5 for the local alignment. It is easy to verify the evident similarity between both routes analyzed until now and we present an evaluation mechanism that provides us a similarity score.

When testing with a different pathway, in both form and content, as seen in Figure 12 and after applying the transformation using algorithm 1 and its subsequent alignment we observe that the results vary.

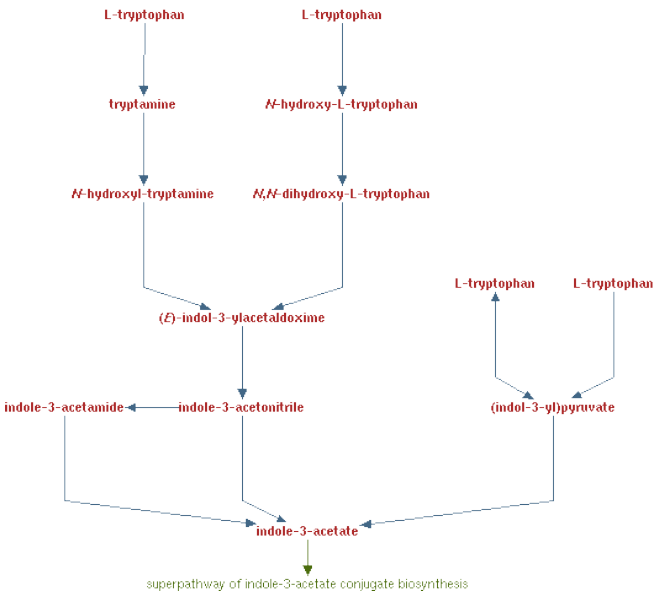


Figure 12. MetaCyc Pathway: indole-3-acetate biosynthesis II [3]

In the case of the alignment of the routes of Figures 4 and 12, the values reached were -10 for the overall alignment, 0 for the local alignment and a value of -20 if we applied a semi global alignment. Again, if a comparison is made between the two in this case both are quite dissimilar.

For the case of the algorithm 2, we did not find an algorithm with which to compare it since it is a different strategy than the ones proposed so far. But it does provide useful information to the expert who performs an analysis of the observed differences. After applying the algorithm shown step by step in Figure 11 to the same working paths, Figures 4 and 5, the differences listed below were obtained. For each metabolic pathway, the reactions present that are not present in the opposite path are listed.

Differences Identified:

- MetaCyc - glycolysis I (from glucose 6-phosphate)

D-glucopyranose 6-phosphate -> B-D-fructofuranose 6-phosphate
B-D-fructofuranose 6-phosphate -> D-glucopyranose 6-phosphate

- MetaCyc - glycolysis IV (plant cytosol)

D-glyceraldehyde 3-phosphate -> 3-phospho-D-glycerate
3-phospho-D-glycerate -> D-glyceraldehyde 3-phosphate

It should be noted that the reactions are bidirectional in the original routes, which is why the description of each reaction is given in each direction.

After applying the second algorithm to the routes of Figures 4 and 12, it is evident that there are many different reactions between the two routes as presented in the following obtained output. For each metabolic pathway, the reactions that are present and are not present in the opposite path are listed.

Differences Identified:

- MetaCyc - glycolysis I (from glucose 6-phosphate)

D-glucopyranose 6-phosphate -> B-D-fructofuranose 6-phosphate
B-D-fructofuranose 6-phosphate -> D-glucopyranose 6-phosphate
B-D-fructofuranose 6-phosphate -> fructose 1,6-bisphosphate
fructose 1,6-bisphosphate -> B-D-fructofuranose 6-phosphate
fructose 1,6-bisphosphate -> D-glyceraldehyde 3-phosphate
fructose 1,6-bisphosphate -> glycerone phosphate
D-glyceraldehyde 3-phosphate -> fructose 1,6-bisphosphate
D-glyceraldehyde 3-phosphate -> glycerone phosphate
D-glyceraldehyde 3-phosphate -> 3-phospho-D-glyceroyl-phosphate
glycerone phosphate -> fructose 1,6-bisphosphate
glycerone phosphate -> D-glyceraldehyde 3-phosphate
3-phospho-D-glyceroyl-phosphate -> D-glyceraldehyde 3-phosphate
3-phospho-D-glyceroyl-phosphate -> 3-phospho-D-glycerate
3-phospho-D-glycerate -> 3-phospho-D-glyceroyl-phosphate
3-phospho-D-glycerate -> 2-phospho-D-glycerate
2-phospho-D-glycerate -> 3-phospho-D-glycerate
2-phospho-D-glycerate -> phosphoenolpyruvate
phosphoenolpyruvate -> 2-phospho-D-glycerate
phosphoenolpyruvate -> pyruvate
pyruvate -> phosphoenolpyruvate

- MetaCyc - indole-3-acetate biosynthesis II

L-tryptophan -> tryptamine
L-tryptophan -> N-hydroxyl-L-tryptophan
L-tryptophan -> (indol-3-yl)pyruvate
tryptamine -> N-hydroxyl-tryptamine
N-hydroxyl-L-tryptophan -> N,N-dihydroxyl-L-tryptophan
(indol-3-yl)pyruvate -> L-tryptophan
(indol-3-yl)pyruvate -> indole-3-acetate
N-hydroxyl-tryptamine -> (E)-indol-3-ylacetaldoxime
N,N-dihydroxyl-L-tryptophan -> (E)-indol-3-ylacetaldoxime
(E)-indol-3-ylacetaldoxime -> indole-3-acetonitrile
indole-3-acetonitrile -> indole-3-acetamide
indole-3-acetonitrile -> indole-3-acetate
indole-3-acetamide -> indole-3-acetate

Conclusions

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

Parte de este trabajo es generar sus propias conclusiones y recomendaciones!

Future Work

Having verified that the proposed algorithms can provide relevant information for the analysis and comparison of metabolic pathways, it would be useful to implement a complete software tool, capable of: directly accessing metabolic databases, extracting information from metabolic routes of interest and finally applying the proposed algorithms to the benefit of experts.

Recently, techniques for sequence comparison based on valuation matrices have been proposed, so that not only the elements that are the same or different are evaluated using the same alignment values for all elements but rather that the affinity between elements is also considered. For example, in the case of proteins if they are of similar families: hydrophobic, sugars, polarity positive or not, etc.; as well as energy and likelihood aspects of reactions. In this way, it is penalized less when one protein is changed by another belonging to the same class than when it is from another. It would be interesting to consider in the comparison of structures aspects such as those mentioned and as such provide even more information to the researcher.

Next, the interactions between routes should then be considered. Metabolites may be the final product of a route or an intermediate producer which may be a precursor for other metabolic pathways. The analysis should be extended to combine these routes to be treated also in their context as metabolic networks.

Finally, just as there are multiple alignment algorithms for several genetic sequences, it is important to continue working on the problem of comparing of multiple pathways to find, for example, similar factors among different species, considering aspects such as those mentioned previously.

Acknowledgments

This work would not have been possible without the support of professors Francisco Torres and Esteban Meneses. And the valueble help of student assistants Kevin Castro-Fuentes and Seth Stalley.

References

[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] Küffner, R., Zimmer, R., & Lengauer, T. (2000). Pathway analysis in metabolic databases via differential metabolic display (DMD). Bioinformatics, 16(9), 825-836.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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

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