 doi:10.20944/preprints202105.0415.v1

Article

Low-Cost Algorithms for Metabolic Pathway pairwise comparison

Esteban Arias-Méndez [1,2] and Diego Barquero-Morera [3] and Francisco J. Torres-Rojas [1]

[1] Escuela de Computación, Instituto Tecnológico de Costa Rica, Cartago 30101, Costa Rica; esteban.arias@tec.ac.cr
[2] PaRMa Group, Instituto Tecnológico de Costa Rica; parma@tec.ac.cr
[3] Ingeniería en Biotecnología, Instituto Tecnológico de Costa Rica, Cartago 30101, Costa Rica; dbarquero@ic-itcr.ac.cr

Version May 13, 2021 submitted to Biomimetics

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. Metabolic pathway comparison has been defined as a computationally complex task [1,2]. In a previous work [3], two different approaches that simplify the problem of comparing pathways represented as graphs were introduced. The first algorithm consists of the transformation of a two-dimensional graph structure, representing a metabolic pathway, to a one-dimensional structure and thus aligning the corresponding data using a reduced 1 dimension string. The second algorithm consists of performing a paired analysis between reactions in pathways and thus eliminating all similarities, finally, showing these differences to the user. The suggestion is to use the information provided by these algorithms as a previous analysis to a deeper, more expensive, comparison tool use. Here we provide an extension of this work with more data and deeper analysis. These methods have shown to be an effective way to treat the problem of metabolic pathway comparison as listed in the discussion and results section. Our results show evidence of a quick, simple and effective way to resolve the described problem.

Keywords: metabolic pathways, graph comparison; graph alignment; graph depth-first traversal; graph breadth-first traversal; global alignment; local alignment; semiglobal alignment

1. Introduction

Metabolomics is the study of the physio-chemical and biochemical reactions at the cellular level; an ordered series of reacting substrates or metabolites to be transformed into other by catalytic enzyme reactions [4], [5], called metabolic-pathways, which provide information to better understand those living processes.

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

2. Metabolic Pathways and Graphs

A metabolic pathway is an ordered sequence of biochemical reactions between metabolites, which are chemical compounds that act as substrates to get transformed into other compounds (in this case, products), through a series of reactions catalyzed by enzymes. [4], [5]. A big number of metabolic

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

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 by the previously described elements. Some of them have curated data that has been verified by several collaborating laboratories.

Metabolic databases store the specifications of the metabolic pathways. Data has been stored in a way that is like that of a directed graph data structure, which is used in computer science to shape relationships and to describe networked processes. Inquiries can be made through protein, metabolite, related gene or gene abbreviation; it will depend on the target and the organization of every base. KEGG (www.genome.jp/kegg/) [6] and MetaCyc (part of BioCyc https://biocyc.org) [7], [8] are examples of the best and most important repositories of such data, providing access to metabolic pathways of several organisms from all kinds of taxa.

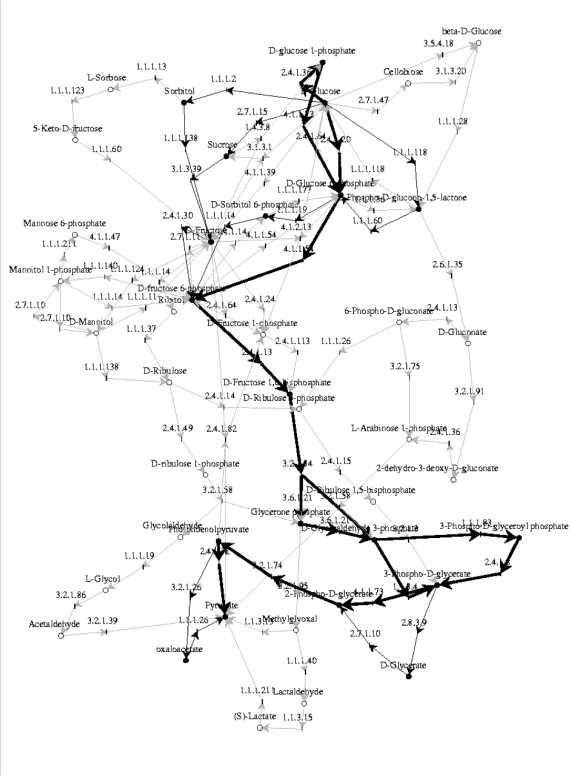


Figure 1. In the background, a network of pathways as the result of work done by [9], where computational prediction methods were used to combine data concerning the relationships between different compounds obtained and between different organisms, from the databases KEGG, ENZYME, and BRENDA. This allowed to obtain a network of reactions related, in this case, to the Glycolysis process, which is essential in many organisms, as it is used to obtain energy (in the form of pyruvate, an energy-rich intermediary) from sugars (mainly glucose). Highlighted, the common elements (in bolded lines); a common pathway is shown between two organisms, yeast and MG (Mycoplasma genitalium) for the calculated network.

To study metabolic processes, it is necessary to span 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 reactions between them, the metabolic pathway. These pathways can be organized into bigger and more complex processes that make up a network of metabolic pathways when several of these interact. Then, the interest is usually on a specific set of points that forms a pathway (see figure 1).

Biologically, there is not one way only to obtain a product, there could be a lot of different procedures that lead to the same result; modifying in a greater or lesser way the procedure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

One of the most important areas of research on metabolomics is the *pairwise comparison* of metabolic pathways of processes of agronomic, pharmaceutical, medical and commercial importance, also know as *pathway alignment*, *graph alignment* or simply *alignment*. The reasons are abundant and relevant. For example, knowing these processes can give us tools to modify said processes with the goal of producing more or better food, learning how to control different viruses or diseases at a cellular level for a better bio-control, as well as improving the development of medicines or more effective treatments, since the acknowledgment of 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 in plants, knowing a metabolic network can be used as a tool to improve growing techniques or used to extract components that are important in the improvement of the human food supply by maximizing food production.

A better understanding of phylogenetic evolution, speciation and reconstruction [10], [11] and the discovery of more effective drugs [12] may be possible thanks to the comparative analysis of different organism's pathways.

In aid of representing all kinds of processes and relationships, graphs and other dynamic data structures have been used through the years. Recently, Biology has been using a lot of well-known data structures to represent all kinds of data. When it comes to the digital representation of metabolic-pathways, they are frequently modeled as directed graphs. Many different techniques have been developed for the alignment or pairwise comparison of these graphs and interesting pathways. As explained in [2] and [1], most of these comparisons can be represented as problems in the class of NP-Complete, which are very complex to solve, even for computers.

Complex algorithm solutions have been then applied, which generally use heuristic techniques that seek to reduce the time of graph-alignment. This problem is much more complex when looking for a comparison between multiple pathways. Some tools like PathVisio [13], MetDraw [14] or NetCoffe [15], provide basic information about pathways, their components, graphs images, and other information, but not analysis tools such as a direct pairwise pathway comparison.

When comparing graphs associated with metabolic-pathways, the difference between homologous paths and similar paths must be taken into consideration. The similarity is often considered to be a measurable and tangible evaluation of some properties of the graph or pathway, while homology is more intuitive. We can see people look homologous because we can find some similarities: a head, arms, legs, eyes, ears, etc. They are not necessarily similar, even if they are homologous. In the case of metabolic pathways, multiple paths can have the same number of interactions or reactions, which would likely mean that they have a homologous shape, but the reacting components or nodes may be very different.

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 are considered common practices. There are a few 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, alignment techniques are valid for a step-by-step comparison of each stage of the metabolic pathway, but, an efficient comparison mechanism at the computational level, which can then be used with different sources of information for the proper study of the metabolic pathways of interest and their subsequent analysis, is still required.

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

The main tools to analysis related pathways are based on the idea of metabolic-pathway comparison, using graph data structures. Computer scientists have proposed several mechanisms for effective comparison. To this regard, Ay & Kahveci [1], proposed SubMAP (Alignment of Pathways with Subnetwork Mappings) which focuses on finding common sub-parts between different pathways. The algorithm CAMPways from Abaka et. al [2] promise to be efficient at run-time; they made a review of tools developed and described the NP costs associated to align graph associated to metabolic pathways.

In some works, like [16] and [17], some heuristic techniques have been applied to reduce the time taken by the graph-alignment algorithms. This causes some loss of generality but makes the data easier to process. The complexity of the problem is bigger when looking to compare multiple pathways or graphs at the same time.

Another approach, more general to graphs, was used by [16], in which the M-GRAAL algorithm was used. This method relies on the calculation of edge-correctness, which represents the percentage in which a graph is topographically like another graph. It is defined as the ratio of edges in graph 1 that are aligned to edges in graph 2 [18]. The goal of the M-GRAAL algorithm is to align two different networks in such a way that edge-correctness is maximized. This task has great computational complexity, but M-GRAAL provides a very good approximation, however on metabolic pathways the topology is not as important as the order of the reacting metabolites. Pinter et. al [17] proposed a bottom-up dynamic programming method to align pathways of different graphs; however, this implementation requires a transformation from the original graph to a multi-source tree (which is a directed acyclic graph).

On the other hand, the edge-correctness does not reflect, whether the correctly aligned edges are near each other and form a connected graph. Therefore, often the size of the largest common connected subgraph that is preserved under the alignment is also used as an indicator of the alignment's quality. Formally, it may be defined as the number of nodes or edges in the largest connected subgraphs [19].

In a previous work [3], two different low-cost algorithms were developed as simple mechanisms for the comparison of two metabolic pathways that can be used as a previous step to a deeper and more time-consuming analysis to be applied for the graph comparison associated to the pathways.

In this work we will be using some simple but important definitions that are important to clarify:

- *Node label*: for any given node in a graph, we refer to "label" as the associated string used to identify each node (which corresponds to the compound or metabolite name of a metabolic pathway represented by said node). Each label is unique within a pathway, meaning that two nodes in the same graph can't have the same label name.
- *Equivalent nodes*: refers to any given labeled node that is present in both graphs being compared, meaning that both pathways involve the use of same compound described by the associated nodes.
- *Analogous order*: for two aligned sequences S and T, the elements with analogous order between both sequences are those that conform the largest possible sub-sequence of both S and T.

It is not the intention of this work to give a definitive answer to the result of the metabolic pathway pairwise comparison problem, or to indicate that one pathway is better than another one, 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 metabolic pathway, only to provide reference information for the interested party. This work is a detailed extension of the ideas introduced in 2017, with a deeper analysis and discussion.

3. Algorithms

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

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

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 pathway has two key elements: at least one starting *origin* point as substrate and at least one final product or *destiny* as output. If the pathway is then viewed as a graph, this graph will have at least a root and an important target product or leaf node (using common nomenclature for trees in data structures).

Concerning graphs, we have mentioned some algorithms; in the case of a graph which represents a metabolic pathway, when applying a traversal algorithm to the graph (which visits all the nodes) it becomes trivial to obtain the list of elements that conform said graph. This would be a 2D to the 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 should be the final product of the route as such.

In the following example, it can be observed that both metabolic pathways have elements in common, equivalent nodes and reactions, that are easily homologous, but it is of interest to quantitatively measure their similarity. The first step for this analysis is to visualize both pathways as graphs (figure 2) and to label the nodes by its corresponding metabolites (figure 3). This means that all the nodes of the graphs are distinguished nodes. Equivalent nodes or equivalent reactions means that the same metabolite or same reacting metabolites are present in both graphs.

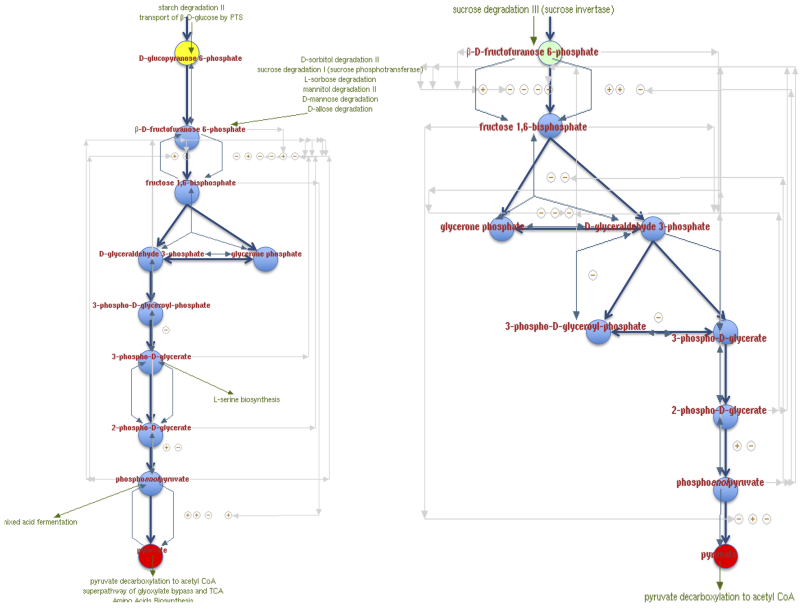


Figure 2. Glycolysis I and Glycolysis IV model metabolic pathways as graphs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

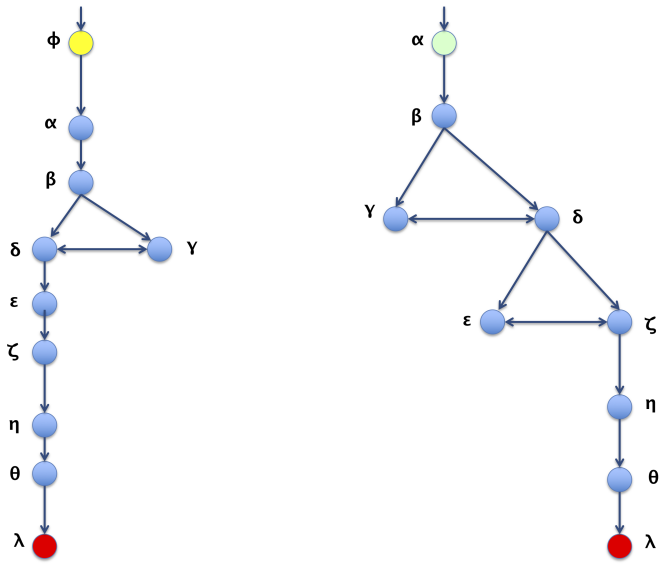


Figure 3. Nodes relabeled according to their corresponding metabolites to simplify processing.

For graphs that represent a metabolic pathway, a traversal lecture (which visits all the nodes) is helpful to get the series of elements using a selected root and, sometimes, the desired end. This is the required transformation from 2D to 1D. Common graph-traversal algorithms are, depth-first [20] and breadth-first [21]; more about this on [22], [23], [24]. It has been observed that when applying a depth-first algorithm the information obtained is not relatively proportional and relevant to the route because the product may 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 4. When performing a breadth-first traversal, the nodes are visited by levels, which corresponds more closely to how the metabolites reactions occur until the expected product is reached. Breadth-first traversal for the routes is shown in figure 5.

ϕ	α	β	γ	δ	ε	ζ	η	θ	λ
α	β	δ	ζ	η	θ	λ	γ	ε	

Figure 4. Depth-first traversals of pathways for graphs in figure 3.

ϕ	α	β	δ	γ	ε	ζ	η	θ	λ
α	β	γ	δ	ε	ζ	η	θ	λ	

Figure 5. Breadth-first traversals of pathways for graphs in figure 3.

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

It should be noted that there will be a loss of information in such a transformation. Figure 6 shows this fact, mainly on the order of the elements and their original relationships. We look to demonstrate that such loss of information during the process is tolerable and acceptable for a correct pairwise comparison result.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

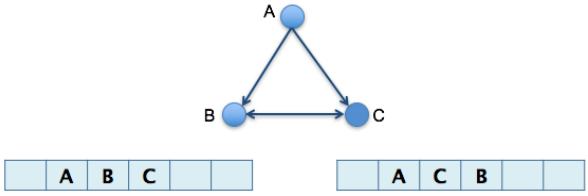


Figure 6. Possible loss of information due to 2D to 1D transformation.

Once the pathway data is raised to obtain the traversal in a 1D format we proceed to apply traditional sequence alignment techniques: global (GA) [25], local (LA) [18], and semi-global (SGA). With this, we get numeric values comparison of the sequences from the graphs.

A sample of the results of this process is then summarized on figures 7 and 8.

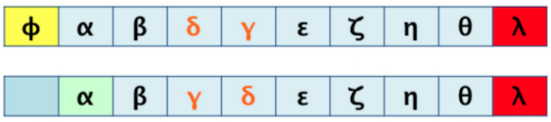


Figure 7. Global alignment generated for transformed graphs, optimal value reached: +3.

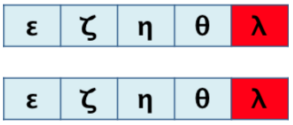


Figure 8. Local alignment generated for transformed graphs, optimal value reached: +5.

3.2. Algorithm 2: Differentiation by pairs

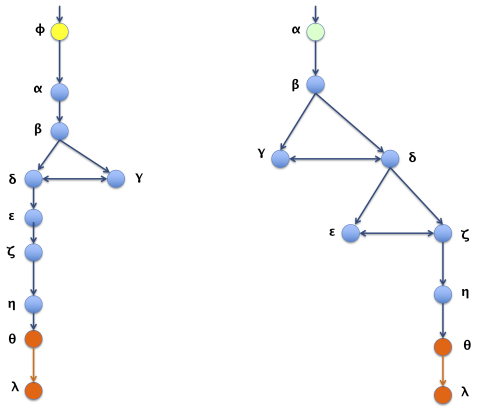
In many situations, when we need to compare objects, the similarities may be evident. In such cases, the differences between both objects become more relevant to the observer than the coincidences. This second algorithm intends to ignore the common pairs of objects and focus on the difference between the elements of both of the graphs that are being evaluated. This algorithm consists of the elimination of equal pairing edges and nodes from both graphs, remaining then only with the differences in the comparing structures.

For human beings, it is common to detect the obvious equalities but could be difficult to detect all of them. The main objective of this algorithm is to remove the common elements in both graphs, that is, equal reactions between two pathways or entire graphs, and without these, find all the differences between given pathways. With this algorithm denoting the distinct reacting metabolites we can see all the remaining differences for the proper analysis by the expert.

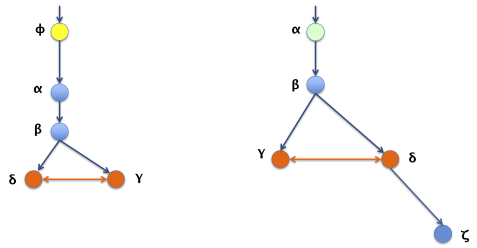
This method differs from a traditional numeric alignment of paths, in which the coincidences take a more relevant role than the differences between a given pair of routes. We also calculate a numeric value, using a relation between the number of differences and the total number of reactions, which is therefore called "Numerical Differentiation by Pairs". So far, no similar approach has been found to this one, so the resulting data is more an intuitive homology for the user than a value of similarity to be used. The results are correlated later to the numerical values of the first proposal to validate the differences found. The process is thoroughly explained in [3]. However on figure 9 a summary of the process is shown.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

a.: Sample graphs.



b.: Progressively eliminating the common reactions between both graphs



c.: Differences found.

Figure 9. Peer differentiation process. a. shows the original sample graphs been compared. b. shows one of the intermediary steps. c. shows the resulting differences found

4. Materials and Methods

4.1. Obtaining the data

We first considered two options, KEGG and MetaCyc, both are self announced as public repositories. However KEGG now has the KEGG FTP Academic Subscription available as a paid service by Pathway Solutions for academic users who wish to bulk-download. For our needs to get all

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

the available data for testing purposes, we did not consider to pay for this service as an option. Then, the database chosen for obtaining the pathways dataset was MetaCyc.

Also, the MetaCyc database focuses on individualizing the pathways into biologically meaningful units (occasionally distinguishable for individual organisms), instead of combining reaction and pathways from multiple species into a single chimeric pathway, as KEGG. Nevertheless, MetaCyc does have super-pathways, which are pathways that comprise multiple sub-pathways, but they usually do occur in a single organism, as explained in the MetaCyc User Guide [1]. The characteristic individualization of these pathways allow for a good random sample pool of graphs, as there are expected to be all kind of comparisons, between contrasting or similar pathways, whereas the chimeric nature of data from other sources implies the unification of the would-be-compared graphs into single large units, making harder the process of randomly sampling for comparisons.

MetaCyc categorizes data in several ways, one of which is the organism the data belongs to. It does so by assigning an ID to each organism. For obtaining the pathways dataset, the organism ID "*META*" was selected, which refers to the multi-organism Pathway Database that contains general metabolic data and is not restricted to a single organism, as explained in the MetaCyc website [2]. The database also categorizes metabolic pathways according to their biological functions and classes of metabolites involved in the reactions. It does so with a hierarchy of pathway classes, where each class composes a large group of sub-classes and pathways, first grouping them by general characteristics (for example, "*Biosynthesis*"), and further down the hierarchy, forming detailed classes containing only pathways (for example, "*Glycogen and Starch Biosynthesis*") [3]. In this work, we refer to those categories as "families" of metabolic pathways.

The websites with the hierarchy information for the organism ID "*META*" was automatically traversed, storing the ID of each pathway when found and the most detailed subclass ("family") containing it. Using the ID of each pathway, the XML file for each pathway was requested to the database, and the data was refined to store the described nodes and edges into JSON files on *ReactionLayout* (RNL) and *DictionaryPathWay* (DPW) formats. Basic pathways (those with no sub-pathways referenced in their XML file) were prioritized; afterwards, super pathways were assembled, using both the instructions for nodes and edges contained within the respective XML file, as well as the already processed instructions for the reference sub-pathways (stored in DPW files). This way, a dataset of **3241 basic metabolic pathways and super pathways** was obtained.

DictionaryPathWay or DPW format, a JSON file, consists of a directed graph data structure based on a dictionary where the keys are strings with the label of a node, and the values are lists with the label of other nodes that each key-node directs to. As a dictionary, this structure doesn't allow duplicate nodes, and will merge the edges of duplicate nodes into a single key (merging also the associated lists). This is allowable for a metabolic pathway that occurs into a single chemical background, where the nodes represent chemical compounds that are dispersed in a theoretically ubiquitous manner across the system, as long as no compartmentalisation of the reactions is involved. This is the case indeed, as metabolic databases tend to miss compartmentalisation when representing metabolic pathways [26].

ReactionLayout or RNL format, a JSON file, is another way of storing a directed graph data structure based on a dictionary, where this time, the keys are strings with the database identifier for the reactions, and the values are lists that contain two internal lists: the first one stores the nodes' label for the substrates of the reaction, while the second one stores the labels for the products.

---

1 https://metacyc.org/MetaCycUserGuide.shtml
2 https://metacyc.org/PToolsWebsiteHowto.shtml#dbselect
3 https://metacyc.org/META/new-image?object=Pathways

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

4.2. Selection of matching candidates for the comparisons

After gathering the metabolic pathway dataset, an automatic selection was conducted for choosing candidates for the experimental comparisons. Meaningful scenarios, close to what would be an actual practical use of the tool, were desired, so the selection process consisted of two criteria. The first one indicates that, for a comparison between two similar pathways, both graphs must contain the same pair of "*origin*" and "*destiny*" nodes; an *origin* being a node that directs to one or more nodes, but none other node directs to it (i.e. analogous to the root of a tree data structure), while a *destiny* node is one such that at least one node directs to it, but it itself doesn't direct to any other node (i.e. a leaf on a tree data structure). Under this simplified biological context, the origin nodes would represent initial substrates for the pathway, while the destiny nodes would be the expected products of interest.

The second criterion for the selection process aims to detect which graphs, from each pathway, can generate valid lectures or traversals between the given *origin* and *destiny* nodes. It takes into account that it is possible to start the traversal of the graph from a given origin node, but never reach a desired destiny node without starting another traversal from a different origin. Such a case can be seen in 10 [4], where the destiny node *succinate* can only be reached by starting at the origin node *phosphoenolpyruvate*. Therefore, for this test, to consider a pair of pathways as a candidate for a given comparison, there must be an actual valid traversal lecture between the same given origin and destiny nodes. It is worth mentioning that this also implies there can be more than one valid traversal or lecture for a given pair of pathways, as long as they meet both selection criteria.

A third argument applied to further filter the candidates for the experimental comparisons was to select only the data with *full-coverage traversals*. This refers to consider those graphs in which all nodes are covered in a single traversal or lecture, we call this a fair comparison.

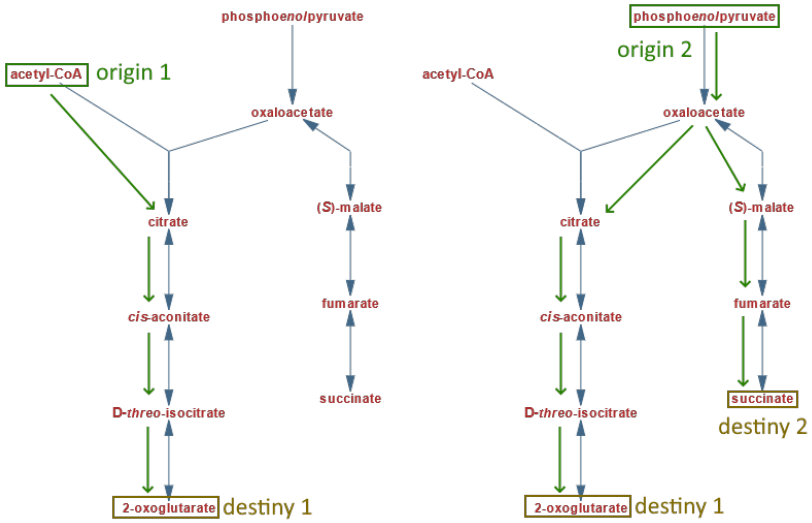


Figure 10. Possible traversals or lectures from a single pathway. Original figure from MetaCyc

4.3. Graphs vs Pathways

The characteristics of the graphs associated to each pathway and the pairwise comparison are taken into account, as factors that could possibly influence the results. As so, these factors are also annotated for each comparison:

- *Size ratio between graphs*: For a given graph, we refer to the "size" as the amount of nodes described within the graph. The size ratio between graphs consists of dividing the smaller

[4] https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5913&detail-level=1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

graph's size (graph with the least amount of nodes) over the bigger graph's size (graph with the largest number of nodes), to obtain a value between 0 and 1 that represents how different are the sizes of the compared graphs. This value can also be interpreted as how much of the biggest graph's size can be covered by the smallest graph's size.

- *Complexity ratio between graphs*: For a given graph, we refer to the *complexity* as the amount of edges described within the graph. The complexity ratio between the graphs consists of dividing the complexity number of the less complex graph (graph with the least amount of edges) over complexity of the more complex graph (graph with the largest number of edges),to obtain a value between 0 and 1 that represents how different are the complexity amounts of the compared graphs. This value can also be interpreted as how much of the more complex graph can be covered by the less complex graph.
- *Equivalent nodes ratio*: This is obtained by dividing the amount of equivalent nodes over the size of the larger graph. It produces a value between 0 and 1 that can be interpreted as how much percentage of equivalent-nodes present in the larger graph can be found also within the smaller graph. Equivalent nodes means that the same metabolite is present in both graphs.

4.4. Formulas

For the evaluation process in the next section, we define then some formulas as the basis for our metrics of analysis.

- *Absolute Score*: $S = xm + yn + zg$, where S is score, x is number of matches, m is value of a match, y is number of mismatches, n is value of a mismatch, z is number of gaps, g is value of a gap.
- *Relative Global*: $r_G = x_G \ / max(|S|, |T|)$, where r_G is the relative global score, x_G is the number of matches of the respective global alignment and S and T are the aligned sequences.
- *Relative Local*: $r_L = x_L \ / min(|S|, |T|)$, where r_L is the relative local score, x_L is the number of matches of the respective local alignment and S and T are the aligned sequences.
- *Relative Semiglobal*: $r_{Sg} = x_{Sg} \ / min(|S|, |T|)$, where r_{Sg} is the relative semiglobal score, x_{Sg} is the number of matches of the respective semiglobal alignment and S and T are the aligned sequences.

4.5. Executing pairwise comparisons

The pairwise comparison generally is any process of comparing entities in pairs to judge which of each entity is preferred, or has a greater amount of some quantitative property, or whether or not the two entities are identical. In psychology literature, for example, it is often referred to as paired comparison.

For each pairwise comparison, both graphs representing each one a different pathway traversed in a breadth-first manner, starting each at the selected "*origin*" node (which is taken as the root of the tree data structure), and concluding when all accessible leaves are reached, wherein a *destination* node can be found. As noted before, it is possible that some nodes get excluded from a particular traversal, if they are located in a path only accessible by traversing from another "*origin*" node.

Each traversal lecture yield a linear sequence of the traversed nodes, in the order they were visited. The first node of the sequence will always be the chosen *origin*, whereas the *destiny* node can be found anywhere in the sequence. The sequences obtained from the traversals are later aligned with global (GA) [25], local (LA) [18], and semi-global (SGA) sequence alignment algorithms. On the other hand, the Difference by Pairs algorithm is performed directly on the graph structures.

Then, we define different metrics for each pairwise comparison executed:

- *Global Score*: score generated by the absolute score formula for the optimal global sequence alignment between two traversals (algorithm 1). The base values used were: match = 1 and mismatch = -1; for gap value, the comparisons were performed using 3 different values: -2, -1 and 0 (explained why later).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

- *Relative Global Score*: value between 0 and 1 obtained from the relative global formula, a percentage value p%, interpreted as "there is a p% similarity between both traversals" or "at least p% elements of the small traversal is present in analogous order on the long traversal".

- *Local Score*: score generated by the absolute score formula for the optimal local alignment between two traversals (algorithm 1). The base values used were: match = 1, mismatch = -1, gap = -2.
- *Relative Local Score*: value between 0 and 1 obtained from the relative local formula, a percentage value p%, interpreted as "at least p% elements of the small traversal is present in analogous order on the long traversal".

- *Semiglobal Score*: Score generated by the absolute score formula for the optimal semiglobal alignment between two traversals (algorithm 1). The base values used were: match = 1, mismatch = -1, gap = -2.
- *Relative Semiglobal Score*: Value between 0 and 1 obtained from the relative semiglobal formula, a percentage value p%, interpreted as "at least p% elements of the small traversal is present in analogous order on the long traversal".

- *Differentiation by Pairs*: For each one of the two pathways, a list of distinguished reactions present on said pathway but absent in the other one was obtained as a result. Each reaction, or edge in the graph, is represented as a string of the form "node_0 -> node_1 ", meaning a *metabolite*0 is being transformed into *metabolite*1.
- *Numerical Differentiation by Pairs*: For a given pairwise comparison between two pathways, consists of the total amount of elements of the previous differentiation by pairs results, divided by the sum of complexity of both graphs. This provides a value between 0 and 1 that represents a percentage of the distinguished reactions (edges) constituted between the two graphs are unique in each graph. Complexity of the graph should be understood as the number of reactions in a single pathway.

Its also important to consider further interpretation of relative global scores in particular. Let's consider figure 11:

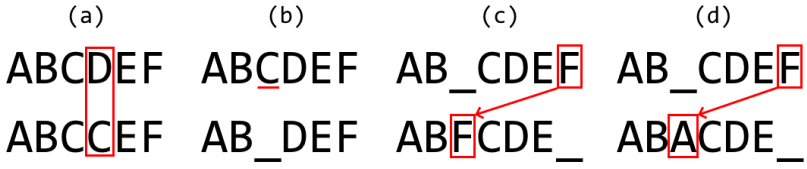


Figure 11. Relative global scores for different scenarios

All shown cases on figure 11 have 5 matches over a maximum sequence length of 6, so all would be 83% (5/6) similar according to the Relative Global: $r_G = x_G\ /max(|S|, |T|)$, formula. Therefore, a difference under this standard could mean a substitution (case a), a deletion or addition (case b), a transposition (case c) or a deletion and addition (case d).

5. Results and Discussion

First, we 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 an accurate and useful result of the comparison.

For the procedure of the first algorithm *Transformation of the 2D pathway graph to a 1D or linear structure for later alignment and evaluation*, we make use of graph traversal by breadth, as previously indicated, using a depth-first traversal may not provide information in order like the one described by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

a pathway and the results for different graphs can be seemingly random. In the case of breadth-first search, a level crossing is performed, like the way a metabolic route works in nature. So, depth-first traversal is not relevant to the proposed process. The cost of this first algorithm approaches in the order of $O(|V| + |E|)$, where V: is the set of vertices or nodes of the graph and $E = VxV$: is the set of edges or arcs.

For the second algorithm *Differentiation by pairs*, it must be considered that for each reaction that exists in the first path or graph $G1$, it must be found in the second path or graph $G2$. That is, if $R1$ is the number of reactions counted for $G1$ and $R2$ the quantity for $G2$, there will be a maximum $R1xR2$ comparisons when it is common for a half-time on average to perform such comparisons. Thus, we can establish a worse case in the order of $O(R1xR2)$.

5.1. Execution tests

From the 3241 pathways available on the dataset, we look that each comparison "match" comprises two distinct corresponding graphs with an origin node selected, and a destiny node expected. Let's remember that each pathway may have different traversals from the same pathway, producing extra comparisons for a single pathway.

For analysis purposes, we considered pathways with valid traversal between an origin (root node) and destiny (leave node) for both graphs and, for a fair comparison criteria, we also considered full-coverage traversals only. It means that we are considering pathways as connected graphs.

Taking into account all the data could mean several million comparisons and many hours of computer work. So, to simplify the process, we considered a random statistic sample with a selection criteria and in the same proportion of elements presented in the total population data, they were: number of nodes (size) and 1 origin node (root) only (the latter guarantees full-coverage traversals).

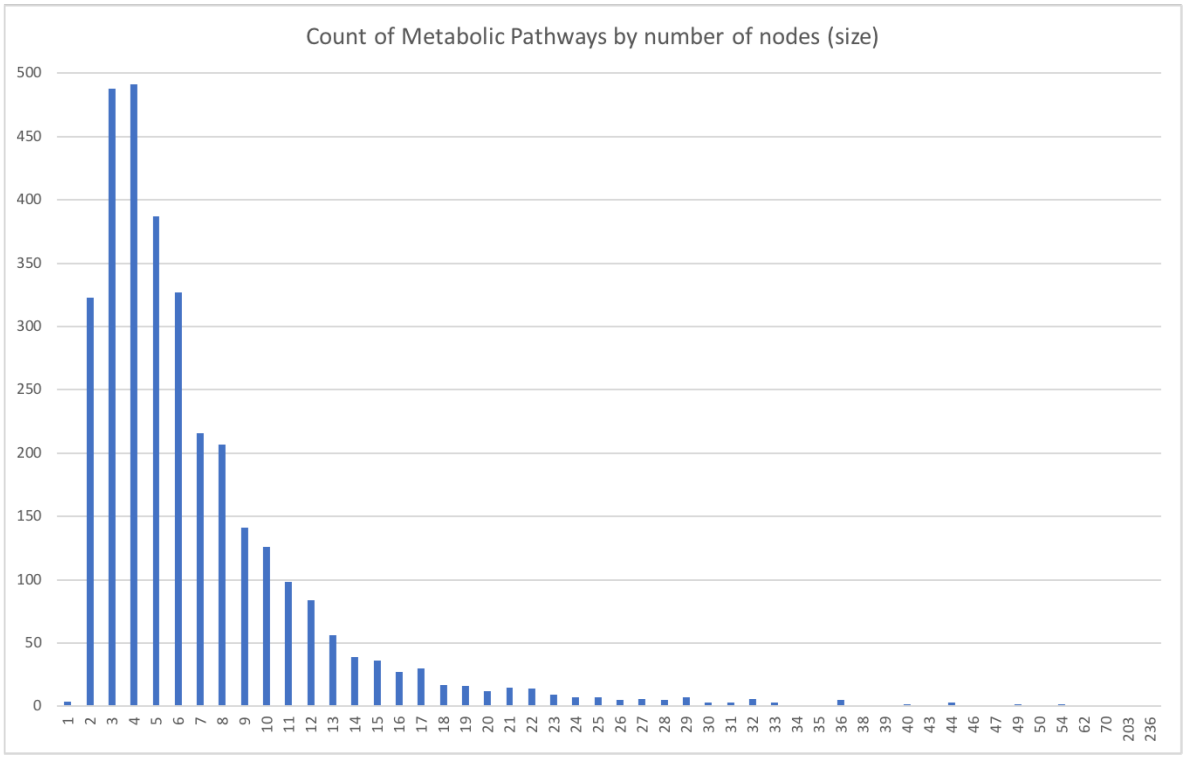


Figure 12. Quantity of Metabolic Pathways ordered by size

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

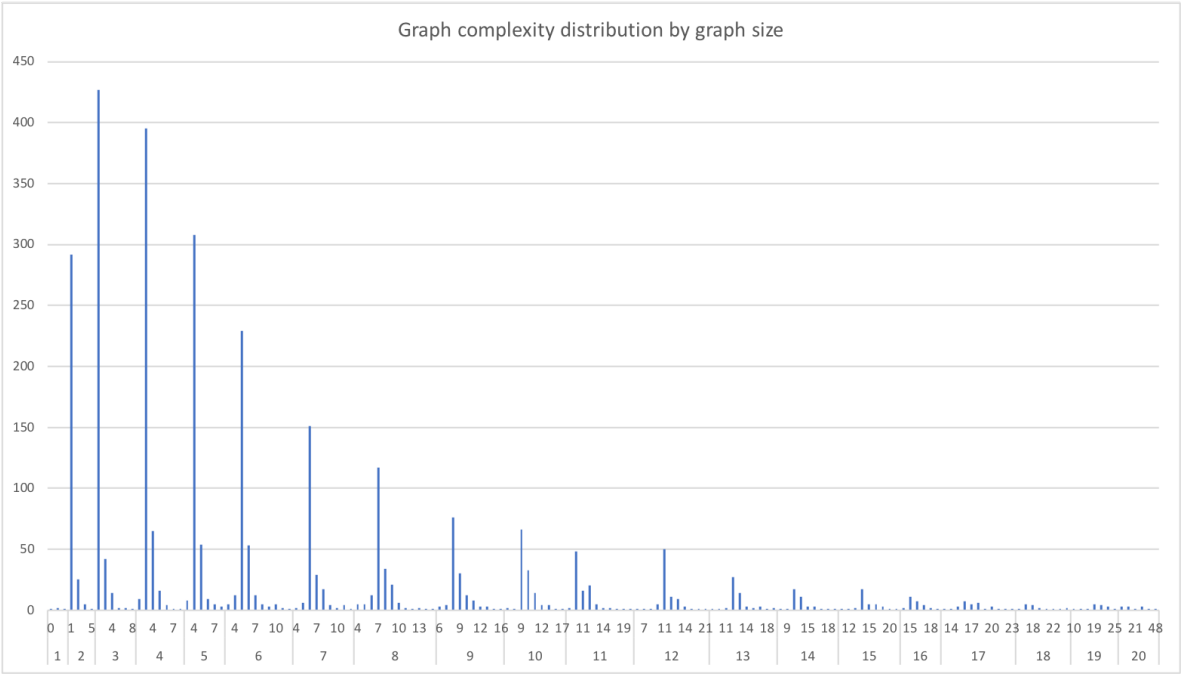


Figure 13. Metabolic Pathways ordered by size and its associated complexities

On figures 12 and 13 we can see the distribution of the data categorized by size and complexity, representing the amount of pathways with each characteristic.

The most representative values selected were: size from 2 to 20 for 3125 pathways (96.4%), selecting then pathways with 1 origin only provides a total of 2340 pathways, distributed as shown by the bars' height on figure 14.

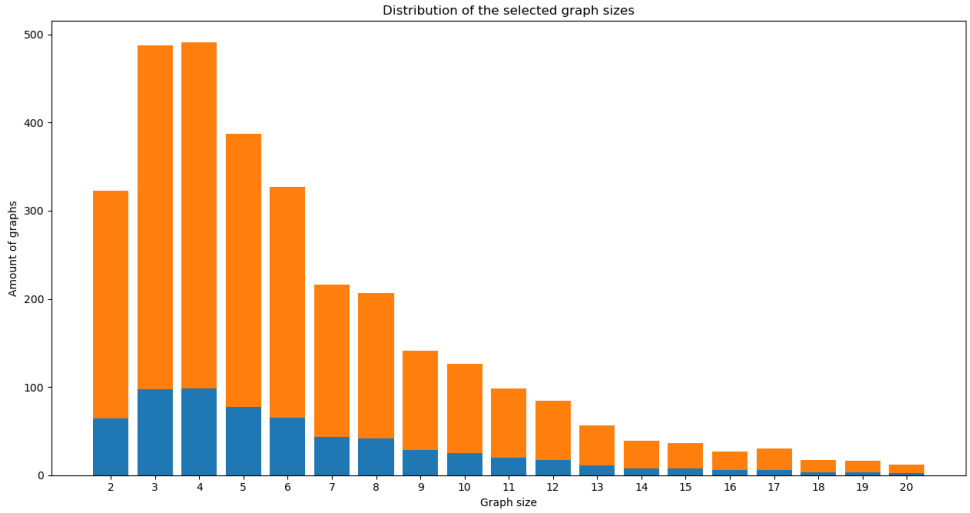


Figure 14. Size distribution quantity of metabolic pathways selected. The height of each bar represents the total pathways selected using the defined criteria of size and origin. Blue represents the random statistic sample of 20 % of the dataset in a scale proportion of sizes

Then we selected a random statistic sample of 20% (468 pathways) in a scaled proportion of the selected size criteria (2 to 20), meaning 20% of each one of these sizes, as shown on figure 14 in blue color, for a total of 109 278 pairwise comparisons.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

All pairwise comparisons were measured using the proposed algorithms with their variants. Also each matching pair was tested using the third-party external tool. We reviewed many previous works to evaluate our results, considering similarity in the outputs, not all of them were available, updated, open-source, accessible, etc. We selected a tool that was available and also provides a pairwise comparison with a 1 to 1 score in a scale 0.0 to 1.0. This tool is called TM&MPAlign (in this work called simply "TMPAlign"), a newer version of the tool MPAlign introduced in 2014 [27].

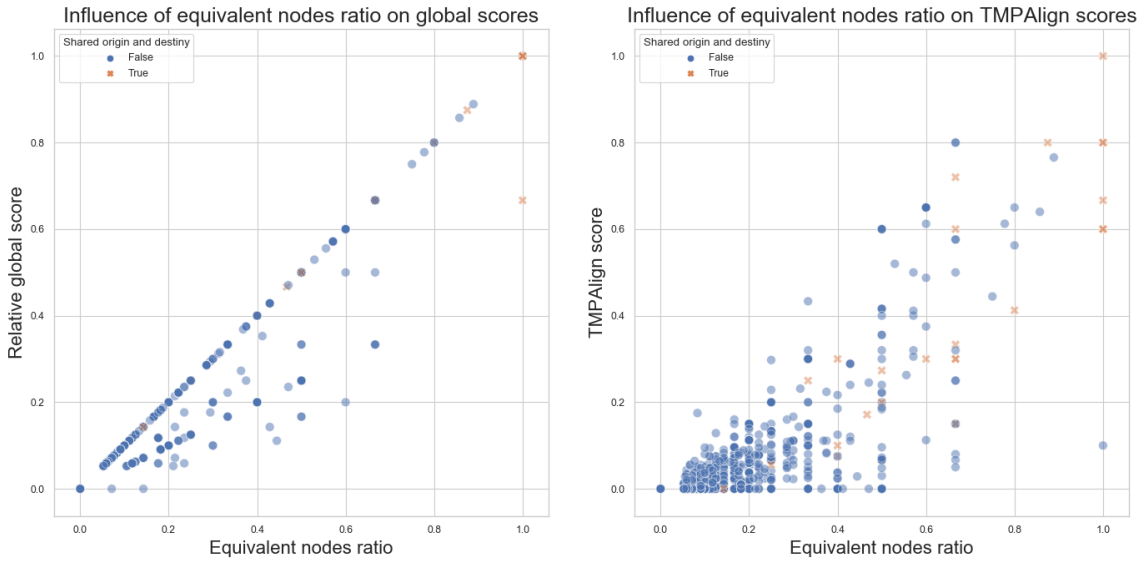


Figure 15. Equivalent nodes ratio vs Global scores and TMPAlign scores, with a random statistic sample of 20%. Pairwise comparisons where both origin and destiny are equivalent are denoted with an orange X marker.

With the random statistic sample of 20% of the data we can see on figure 15 an interesting observation. All of the pairwise comparisons reporting 0% equivalent nodes generate a score of 0 for all scores, for our Global scores and even for the TMPAlign tool. So, for the rest of the comparisons we are avoiding these comparisons generating 0, since they are not providing significant values and consume an important amount of computation time in our batch runs. This allows a bigger statistic sample for subsequent runs, with more significant results.

5.2. Analysis of algorithms for pairwise comparisons

Several execution metrics were conducted to evaluate the tests for each pairwise comparison previously defined. After the first tests with the sample of 20%, and considering that the comparisons without equivalent nodes always generate the same scores (0, as described before), we increased our random statistic sample to 50% while simultaneously only performing the comparisons with at least one equivalent node, so we can test a broader diversity of metabolic pathways. This new selection represents 1169 pathways, that means 682696 possible comparisons.

For the first algorithm we mainly rely on the score provided by the global, local or semiglobal alignments, however, to provide a better meaning to this, some extra metrics were developed in order to adjust the relationship between the scores and the data, like the coverage of one pathway with others, specially when they are of different sizes. So, the values are indicated in a ratio relationship from 0.0 to 1.0.

Lets consider what we are looking for with each general alignment scores:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

- The global score seeks to analyze the traversal or complete lecture of the pathway and take into account the percentage of similar elements, as a whole. In the tests carried out, it was observed that applying a negative gap assessment such as the -2 standard does not generate any meaning in a metabolic process as such. Hence, the best values have been obtained using a gap value of 0 in the global score. This was the gap value selected for more detailed analysis: 0.
- The local alignment seeks to obtain the best conserved internal fragments (the most similar subgraphs between both routes), this has been achieved well using a standard evaluation of gap = -2, miss = -1 and match = + 1.
- Semiglobal mainly looks for overlaps between the extremes of metabolic pathways to look for similarities in these areas (prefixes against suffixes). The standard values were also used here.

Figure 16. Influence of the graph size in the global comparison scores

For the first algorithm, figure 16 shows the relation between the graphs' size and relative Global pairwise comparison scores. By this, we can see that there is no direct correlation between the size of the graphs and the Global scores. We found that this disordered behaviour is also present in all algorithms.

It is worth noting that, for the relative global score, the "highest possible score" is delimited to the size ratio of the graphs. For example, let's consider two different pathways, one with 3 nodes and the other one with 10 nodes. The best chance of a good comparison here is that the nodes of the smaller pathway are all in the bigger one, in the same order; the highest possible score in this case would be of 30%. This can also be seen in figure 16 as a "diagonal" that bounds the dispersion of the points across the graph.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

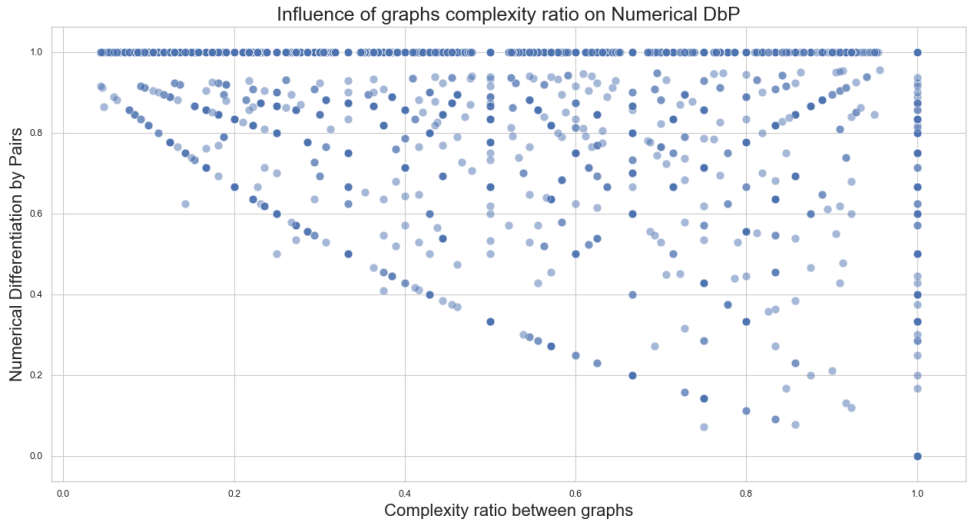


Figure 17. Influence of the graph complexity in the DbP pairwise comparison scores

Then, for the second algorithm, figure 17 shows the relation between the complexity ratio of the graphs and its influence in Numerical DbP scores. The numerical evaluation of the second algorithm seeks for differences between the graphs, according to the difference in edges. We can see that there is no direct correlation between the complexity of the graphs and the numerical DbP scores either.

Also, on figure 18 we can see the relation between the number of equivalent nodes and the Global scores of algorithm 1 in the left side and for TMPAlign tool on the right side. Similar to the threshold observed for the global scores according to the size ratio, a similar behaviour can be seen with the equivalent nodes ratio, with more points getting aligned in the central diagonal than before. This also occurs at a lesser degree for the comparison values obtained with TMPAlign, where we can observe that the scores are less correlated to the equivalent nodes ratio; this implies that TMPAlign could be considering other factors when generating the scores. Nevertheless, there is an important observation here: when the graphs should not be similar (i.e. at low equivalent nodes ratio), both tools tend to show this.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

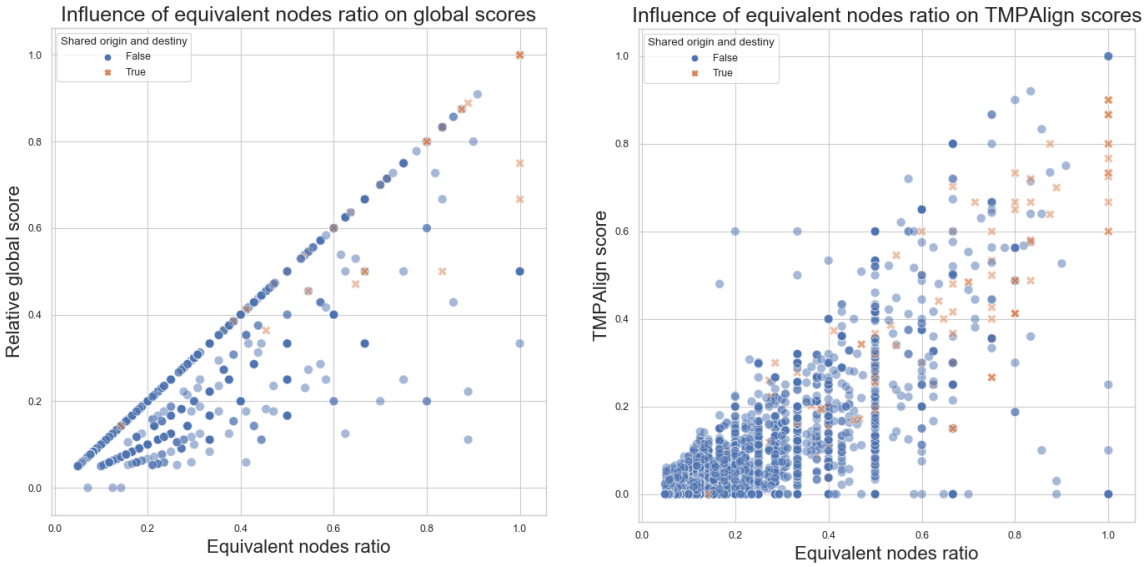


Figure 18. Equivalent nodes ratio vs Global scores and TMPAlign scores, with a random statistic sample of 50% and excluding comparisons without equivalent nodes. Pairwise comparisons where both origin and destiny are equivalent are denoted with an orange X marker.

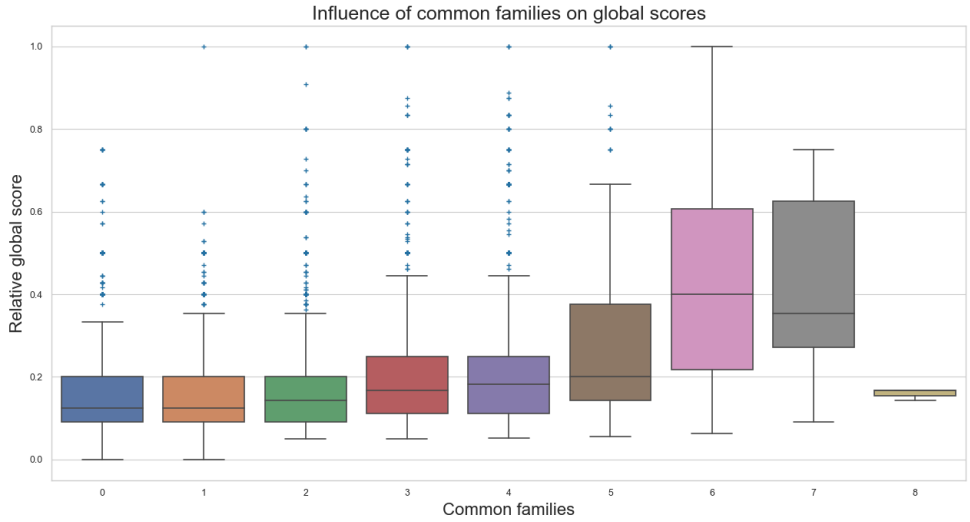


Figure 19. Common families on Global scores.

If we also consider the common families of the pathway, from a range of about 638 different families in the total population of 3241 pathways, it is easy to denote a great diversity of metabolic pathways. If we consider to group the pathways in a common families criteria, we can observe in figure 19 that most of the scores related to each category remains very close. Also, as the compared pathways have more families in common, the scores tend to be higher.

As a validation for the scores obtained from the algorithms we applied a simple Design of Experiments (DoE) and One-way ANOVA tests. Figure 20 shows the resulting ANOVA for the Global scores. The DoE, in this case, considered the sizes of the graphs as factors, using 19 levels for each one, sizes from 2 to 20, as selected in the statistic sample. The design suggested 361 runs, we executed it with 5 replicas. All selected data for each run and replica was randomly selected from the sample data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

As we can see on the figure, all the replicas shows a normal variation, it means that different data with similar characteristics will provide results in a similar scope.

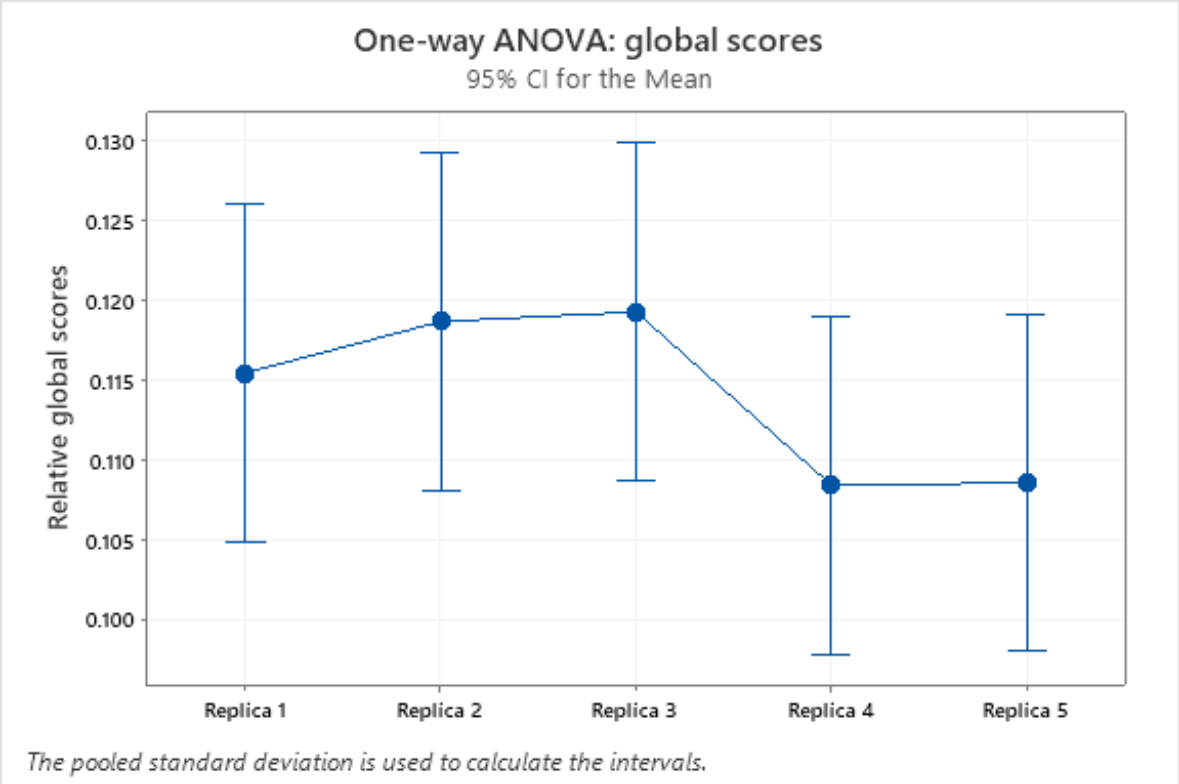


Figure 20. One-way ANOVA: global scores

5.3. Comparison with other tool of reference

The TMPAlign tool used was outdated, it was made available in 2017, written in a python version 2.7 using services of the KEGG database that are not available today as it was expected. Documentation of the tool points out it can work with any database, so we change it to work with the same data files from MetaCyc we are using. Also, TMPAlign is not using the data about enzymes (i.e. when comparing two reactions, it only takes into account the reactions' id when generating the score), since that service is inspired on how KEGG handles this information, and the data obtained from MetaCyc does not fulfill the same criteria. Furthermore, it is worth noting that, for some pairwise comparisons, the tool TMPAlign raised errors during the execution; these were excluded from the subsequent analysis for all algorithms.

It is important to remark that the main goal of using a tool of reference is not to generate the exact same values, but instead prove that, if two pathways are significantly different, both tools can denote it, and the contrary for similar pathways.

Using the previous DoE runs for global scores, we also obtained the scores for the TMPAlign algorithm. On figure 21 we tested the difference between the scores of the Global and TMPAlign algorithm as a way to test its similitude, but mainly to confirm that for non similar given pair of metabolic pathways we get low scores as expected and for similar pathways higher scores. Runs and replicas are the same. All selected data for each run and replica was randomly selected from the sample data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 May 2021 doi:10.20944/preprints202105.0415.v1

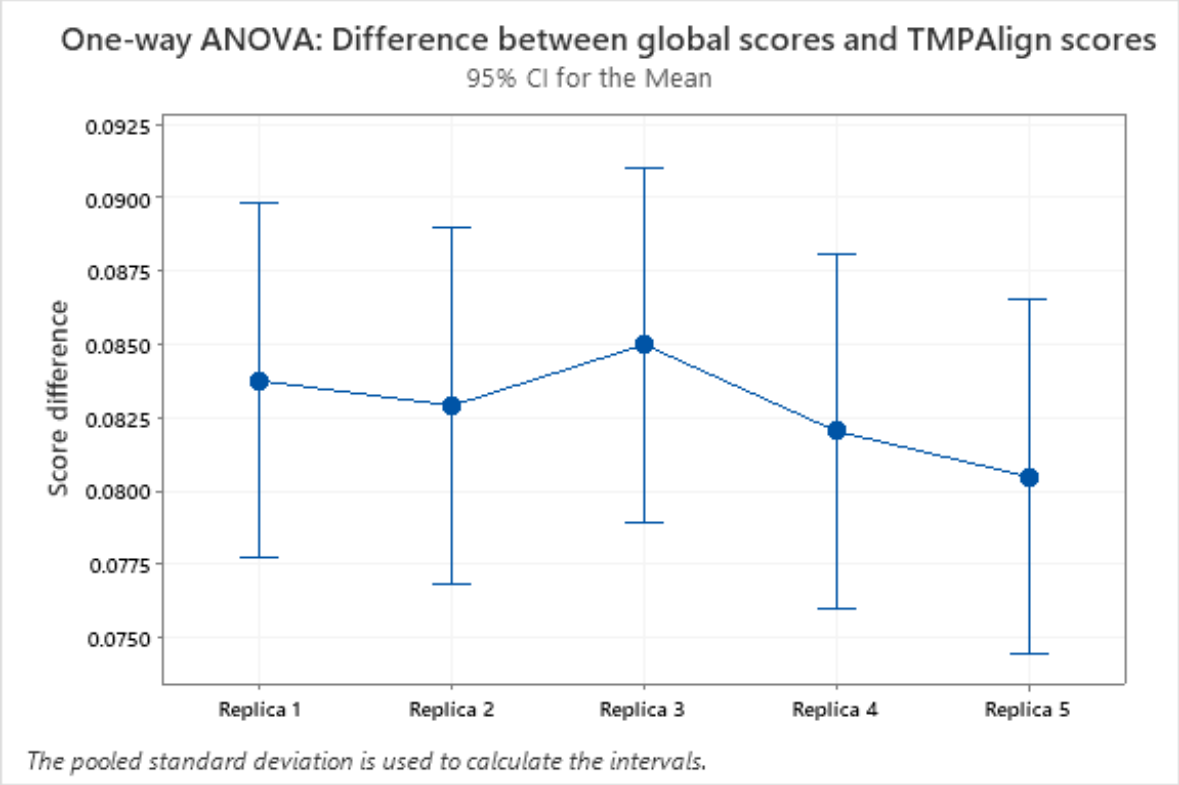


Figure 21. One-way ANOVA: Global scores vs TMPAlign scores

5.4. Execution and timing evaluations

About the cost related to the execution time of the algorithms, this is one of the most important gains obtained with the proposed algorithms. Figure 22 shows the time consumption between our algorithms and the tool TMPAlign and its relation with the graph size. For this comparison, consider that we are including the summarized timing of our algorithms and their versions, all at the same time, for each pairwise comparison, and not a single execution at a time. So, we can see here the sum of the execution times of our algorithms versus a single run of TMPAlign. Even with the accumulated times, our algorithms show an improvement, being, in average, at least 10 times faster than TMPAlign.

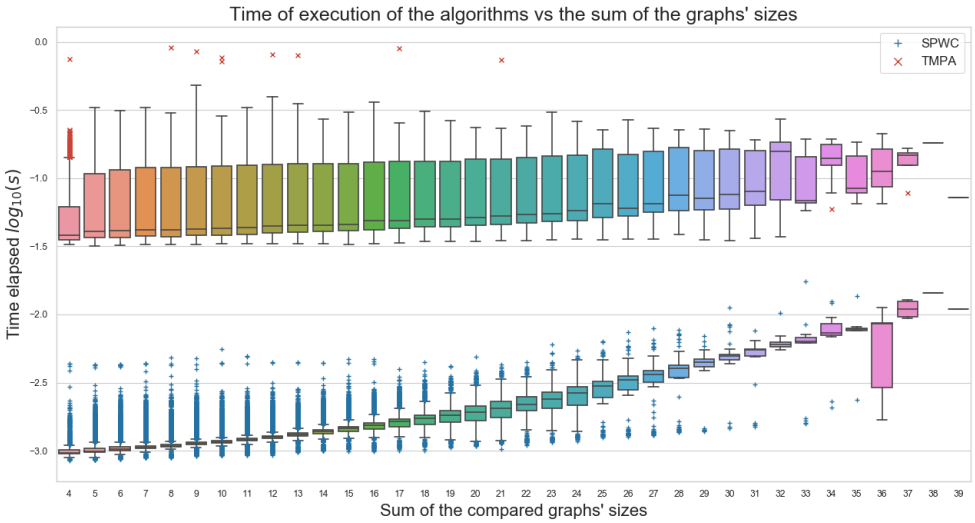


Figure 22. Execution running time comparison between tools and graph size

 doi:10.20944/preprints202105.0415.v1

6. Conclusions

The pairwise comparison of metabolic pathways is an enormous yet interesting problem. In this work, we have procured to provide some insight about it, implementing alternative methods to simplify how the data is used, and by doing so, sacrificing precision on the information representation but not in results.

We can observe that the mechanisms proposed as algorithm 1, with its variants, and algorithm 2 can be used as a prior evaluation process to predict good comparisons in case a deeper analysis is desired. In this case, the analysis could continue by considering extra information like enzymes, chemical similitude between compounds and others.

The comparisons that report 0% equivalent nodes unanimously report a 0% similarity under any of the evaluated algorithms, meaning that, when there are no equivalent nodes between the graphs, it can be safely reported that both pathways are completely different, without the need of executing the comparison algorithms.

It was shown that structural characteristics of the graphs, such as number of nodes (size) or number of edges (complexity), do not bias the comparison results when using the algorithms, as should be for a reliable tool. The only influence that these factors have over the results is that, when comparing graphs with different sizes or complexities, the relative global score or numerical DbP score respectively will be penalized (decreased) because of the structural difference, whether the nodes or edges are actually similar or not. This is naturally to be expected, as these considerations are part of the design for the score system. In other words, the scores obtained from every pairwise comparison are dependent from the inner data nodes; two comparisons of graphs with equals size might produce different scores.

On the other hand, some characteristics that can be obtained from the pairwise comparisons have a slight effect on the comparison results, such as the equivalent nodes ratio between the compared graphs, how many metabolic categories share the graphs, and whether the graphs have equivalent origin and destiny nodes. When the first two have their numerical value increased, the comparison scores tend to increase as well; similarly, when the latter characteristic checks to be true, the comparison scores tend to be higher. This is also to be expected, as said characteristics hint at how related are the pathways. Likewise, it is important to denote that this tendency is not always the case, as there can be comparisons that do not follow this pattern, which is also perfectly normal and shows how diverse can be the metabolic pathways and their comparisons.

The algorithms *Transformation of the 2D pathway graph to a 1D or linear structure for later alignment and evaluation* and *Algorithm 2: Differentiation by pairs* also showed to be extremely resource efficient, surpassing the speed of execution and in a more predictable manner than the tool of reference. This means that the goal of the algorithms (to simplify the graph comparison problem for it to be computationally lighter but still reliable) was achieved. The faster, reliable and more predictable behaviour of the algorithms also means the tool can be successfully employed for batch comparisons, using large datasets of metabolic information, even though this was not the original intended use for the algorithms.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BFT	Breadth-First Traversal
DbP	Differentiation by Pairs
DFT	Depth-First Traversal

References

1. Ay, F.; Kellis, M.; Kahveci, T. SubMAP: aligning metabolic pathways with subnetwork mappings. *Journal of computational biology* **2011**, *18*, 219–235.
2. Abaka, G.; Bıyıkoğlu, T.; Erten, C. CAMPways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways. *Bioinformatics* **2013**, *29*, i145–i153.
3. Arias-Mendez, E.; Torres-Rojas, F. Alternative low cost algorithms for metabolic pathway comparison. 2017 International Conference and Workshop on Bioinspired Intelligence (IWOBI). IEEE, 2017, pp. 1–9.
4. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular biology of the cell*; Garland Science, 2007.
5. Lee, J.M.; Gianchandani, E.P.; Eddy, J.A.; Papin, J.A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS computational biology* **2008**, *4*, e1000086.
6. Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **2011**, *40*, D109–D114.
7. Caspi, R.; Foerster, H.; Fulcher, C.A.; Kaipa, P.; Krummenacker, M.; Latendresse, M.; Paley, S.; Rhee, S.Y.; Shearer, A.G.; Tissier, C.; others. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* **2007**, *36*, D623–D631.
8. Caspi, R.; Billington, R.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Midford, P.E.; Ong, W.K.; Paley, S.; Subhraveti, P.; Karp, P.D. The MetaCyc database of metabolic pathways and enzymes-a 2019 update. *Nucleic acids research* **2020**, *48*, D445–D453.
9. Küffner, R.; Zimmer, R.; Lengauer, T. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* **2000**, *16*, 825–836.
10. Mithani, A.; Hein, J.; Preston, G.M. Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in Pseudomonas. *Molecular Biology and Evolution* **2010**, *28*, 483–499.
11. Heymans, M.; Singh, A.K. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics* **2003**, *19*, i138–i146.
12. Guimerà, R.; Sales-Pardo, M.; Amaral, L.A.N. A network-based method for target selection in metabolic networks. *Bioinformatics* **2007**, *23*, 1616–1622.
13. 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 computational biology* **2015**, *11*, e1004085.
14. Jensen, P.A.; Papin, J.A. MetDraw: automated visualization of genome-scale metabolic network reconstructions and high-throughput data. *Bioinformatics* **2014**, *30*, 1327–1328.
15. Hu, J.; Kehr, B.; Reinert, K. NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. *Bioinformatics* **2013**, *30*, 540–548.
16. Oleksii, K.; Natasa, P. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics* **2010**.
17. Pinter, R.Y.; Rokhlenko, O.; Yeger-Lotem, E.; Ziv-Ukelson, M. Alignment of metabolic pathways. *Bioinformatics* **2005**, *21*, 3401–3408.
18. Smith, T.F.; Waterman, M.S.; others. Identification of common molecular subsequences. *Journal of molecular biology* **1981**, *147*, 195–197.
19. Döpmann, C. Survey on the graph alignment problem and a benchmark of suitable algorithms. *Institut für Informatik* **2013**.
20. Tarjan, R. Depth-first search and linear graph algorithms. *SIAM journal on computing* **1972**, *1*, 146–160.
21. Bundy, A.; Wallen, L. Breadth-first search. In *Catalogue of artificial intelligence tools*; Springer, 1984; pp. 13–13.
22. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Stein, C. *Introduction to algorithms*; MIT press, 2009.
23. Knuth, D.E. The Art of Computer Programming, Vol 1: Fundamental Algorithms, Addisson-Wesley. *Reading, Mass* **1968**, p. 634.
24. Lee, C.Y. An algorithm for path connections and its applications. *IRE transactions on electronic computers* **1961**, *EC-10*, 346–365.
25. 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* **1970**, *48*, 443–453.

 doi:10.20944/preprints202105.0415.v1

26. Wittig, U.; De Beuckelaer, A. Analysis and comparison of metabolic pathway databases. *Briefings in bioinformatics* **2001**, *2*, 126–142.
27. Alberich, R.; Llabrés, M.; Sánchez, D.; Simeoni, M.; Tuduri, M. MP-Align: alignment of metabolic pathways. *BMC Systems Biology* **2014**, *8*, 1–16.