BMC Bioinformatics

SOFTWARE Open Access

Metadag: a web tool to generate and analyse metabolic networks

Pere Palmer-Rodríguez[1*], Ricardo Alberich[1], Mariana Reyes-Prieto[2], José A. Castro[3] and Mercè Llabrés[1*]

*Correspondence: pere.palmer@uib.es; merce.llabres@uib.es

[1] Mathematics and Computer Science Department, University of the Balearic Islands, Ctra Valldemossa, Km 7.5, Palma 07122, Balearic Islands, Spain
[2] Sequencing and Bioinformatics Service, Foundation for the Promotion of Health and Biomedical Research of the Valencian Community (FISABIO), Avda. de Catalunya, 21, 46020 Valencia, Valencia, Spain
[3] Biology Department, University of the Balearic Islands, Ctra Valldemossa, Km 7.5, 07122 Palma, Balearic Islands, Spain

Abstract

Background: MetaDAG is a web-based tool developed to address challenges posed by big data from omics technologies, particularly in metabolic network reconstruction and analysis. The tool is capable of constructing metabolic networks for specific organisms, sets of organisms, reactions, enzymes, or KEGG Orthology (KO) identifiers. By retrieving data from the KEGG database, MetaDAG helps users visualize and analyze complex metabolic interactions efficiently.

Results: MetaDAG computes two models: a reaction graph and a metabolic directed acyclic graph (m-DAG). The reaction graph represents reactions as nodes and metabolite flow between them as edges. The m-DAG simplifies the reaction graph by collapsing strongly connected components, significantly reducing the number of nodes while maintaining connectivity. MetaDAG can generate metabolic networks from various inputs, including KEGG organisms or custom data (e.g., reactions, enzymes, KOs). The tool displays these models on an interactive web page and provides downloadable files, including network visualizations. MetaDAG was tested using two datasets. In an eukaryotic analysis, it successfully classified organisms from the KEGG database at the kingdom and phylum levels. In a microbiome study, MetaDAG accurately distinguished between Western and Korean diets and categorized individuals by weight loss outcomes based on dietary interventions.

Conclusion: MetaDAG offers an effective and versatile solution for metabolic network reconstruction from diverse data sources, enabling large-scale biological comparisons. Its ability to generate synthetic metabolisms and its broad application, from taxonomy classification to diet analysis, make it a valuable tool for biological research. MetaDAG is available online, with user support provided via a comprehensive guide. MetaDAG: https://bioinfo.uib.es/metadag/ User guide: https://biocom-uib.github.io/MetaDag/

Keywords: Metabolic networks construction, Analysis of metabolic networks, Comparison of metabolic networks

Background

In the current omics era, the focus of metabolic function analysis has shifted from individual organisms to entire microbial communities, as well as the intricate metabolic interactions among community members. The vast quantity of sequence data generated from metagenomic and metatranscriptomic samples requires specialized tools for the taxonomic and functional annotation of genomes, genes, and proteins, as well as their subsequent integration and analysis. In this context, the reconstruction and analysis of metabolic networks are crucial for understanding the metabolic profiles and their interactions within microbial communities. These networks can be extensive, often encompassing hundreds or even thousands of interrelated reactions. As a result, the development of bioinformatic tools for their visualization and comprehensive analysis has become a necessity.

Several methodologies for metabolic reconstruction have been developed, including machine learning-based approaches that predict and reconstruct metabolic pathways (see [1] and [2] for an extensive review of these methods). Nevertheless, the constantly increasing amount of metabolic information stored and characterized in several public repositories according to their functions, such as KEGG [3–5], BioCyc [6], MetaCyc [7], among others, strengthens the need for automated metabolic reconstruction methods based on curated metabolic data information. In this line of research, various approaches to metabolic network reconstruction, analysis, and comparison can be found in the literature [8, 9](see [10–12] for surveys on different approaches and tools). Each approach selects a representation of metabolic networks that models information of interest, proposes a similarity or a distance measurement, and possibly supplies a tool. Models of metabolic networks range from a high-level abstraction of metabolic information, such as abstract metabolic networks, to a lower level of abstraction, such as reaction networks, metabolic hypergraphs, or stoichiometric matrices.

In [13] a novel methodology for the analysis of metabolic networks was presented. This methodology incorporates the concept of strongly connected components in a reaction graph, called *metabolic building blocks*. Additionally, a new model of metabolic networks, known as a *metabolic DAG*, was defined to combine the reaction graph information with the network's topology. This methodology has been used in different contexts, and proven to be successful in comparing, analysing, and visualising metabolic networks, as shown in previous studies [14, 15]. As a result, we present here MetaDAG, a web-based tool that implements the metabolic DAG methodology. MetaDAG is a continuation of the protocol introduced in [13], and its innovation lies in the tool that makes the methodology accessible and easier to use for any researcher.

MetaDAG automates the metabolic network reconstruction using several different identifiers of data stored in the KEGG database, which we choose as our source of metabolic information due to its curated nature and standardized presentation.

MetaDAG enables metabolic network reconstructions using a variety of inputs, such as a single organism, groups of organisms, specific reactions, enzymes, or KEGG Orthology (KO) identifiers. This flexibility supports reconstruction across a range of sample types-from individual microbial samples to consortia and complex metagenomic samples-making it a powerful tool for diverse analytical needs. Additionally, MetaDAG can generate "synthetic metabolisms" independent of taxonomic classification. By identifying interactions within the metabolic data, it creates artificial metabolic networks, addressing research gaps that often focus exclusively on established model organisms.

Metabolic networks are generated by retrieving the reactions associated with user-specified queries from the KEGG database. Initially, it computes a metabolic network as a reaction graph. From the reaction graph, a directed acyclic graph called a metabolic

DAG (m-DAG for short) is computed by collapsing all strongly connected components of the reaction graph into single nodes, called metabolic building blocks (MBB for short). In the m-DAG, two MBBs are connected through an edge if there is at least one pair of reactions (one in each MBB) connected by an edge in the reaction graph. As a result, the m-DAG representation reduces considerably the number of nodes while keeping the network's connectivity. Hence, at first glance, this m-DAG representation offers an easy-to-interpret topological analysis of the reconstructed metabolic network. Both models of metabolic networks, the reaction graphs and the m-DAGs, are displayed in an interactive web to aid users in visualising and analysing the networks, as well as to retrieve the node's information linked to KEGG. Therefore, with MetaDAG, reaction graphs, and m-DAGs can be generated effortlessly and automatically.

Furthermore, when examining different groups of organisms, experiments, or samples, MetaDAG also calculates the core and pan metabolism associated with each group. It then provides the results of a comparative analysis of their respective m-DAGs. This comparative analysis offers valuable insights into the shared and unique metabolic features across different groups, experiments, or samples, as shown in the results of a Eukaryotes test and a gut microbiome analysis, conducted for this work to assess the tool's performance and functionality. Given the vast amount of information MetaDAG generates for each query, a comprehensive user guide in R is provided to assist users in understanding, handling, and analysing all the results. (See https://biocom-uib.github.io/MetaDag/).

Implementation

MetaDAG is an innovative web-based tool designed to offer users a robust and effective method for analysing and visualising complex DAGs that model metabolic networks. The client-side, or front-end, of MetaDAG has been crafted using Angular and TypeScript, ensuring a robust and responsive user interface.

The Angular framework enables consistent integration of components, bringing a dynamic and interactive user experience. TypeScript, a superset of JavaScript, further enhances the front-end development process, promoting better code organization, early error detection, and improved maintainability.

The server-side, the back-end of MetaDAG, has been engineered using Java 19, a versatile and highly reliable programming language, and deployed as a self-contained JAR file to ensure stability over time, in combination with the popular Spring framework. The Spring framework simplifies the development process by providing a comprehensive set of tools and libraries for building scalable and performant applications. With Spring, the back-end of MetaDAG benefits from features such as dependency injection and aspect-oriented programming, making it a solid foundation for data-intensive operations.

The back-end has been designed in two layers. The front layer is responsible for serving data requests from the front-end (e.g.: serving the files requested or preparing the data the user wants to download). Furthermore, when a user makes a new query, the front layer prepares the data needed to fulfill the request and passes them to the calculation layer, wherein an independent process of the calculations is performed. This way, the front layer can serve customer requests quickly. The calculations are made independently, and they can last a significant amount of time, so a contact email for the user has to be provided. When the calculations are finished, an email giving access to the results is sent.

For guidance purposes, during the development, several tests were conducted, and we measured the response times for these tests. The results are shown in Table 1. The calculations were performed on a server equipped with two AMD 7282 processors and 512 GB of DDR4/3200 MHz memory.

In the simplest case, a specific pathway of an organism, the result can be generated in just a few seconds with an average response time of 1.07 s per test. However, when dealing with larger datasets, the required time increases significantly. This is true for the global metabolic network of a list of organisms, where it is possible to use the entire set of available eukaryotes and prokaryotes (8,935 species), leading to response times that can exceed 40 h. Moreover, if m-DAG similarities need to be calculated, the response time increases further. Storage space required for the resulting data also varies depending on the type of experiment and the data involved. For instance, in the case of a specific pathway of an organism, only a few megabytes are needed. However, for the previously mentioned global metabolic network of a list of organisms, which involves the complete set of available organisms, the resulting data can require slightly more than 70 GB.

MetaDAG makes use of the fundamental biological information available on the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. It is important to denote, since the metabolic reconstructions by MetaDAG are constructed in a presence/absence scenario only, and KEGG is a widely recognized and highly curated database that contains an extensive collection of biological pathways and networks. It provides standardized nomenclature and annotations for genes, proteins, enzymes, orthologs, and pathways. We consider the KEGG database since it is an extensive, reliable, and widely used resource explicitly designed to present data in a standardized way.

MetaDAG's design

We first recall the methodology of metabolic networks and the construction of metabolic DAGs. Next, we present its implementation in the proposed tool *MetaDAG* to generate, analyse, and compare metabolic networks.

Table 1 Execution times with MetaDAG

Query	Mean (s)	Std. Dev. (s)	Number of Tests
1	1.07	1.00	179
2	24.61	16.19	97
3	1239.31	80.00	80
4	12237.17	132.00	132
5	75.28	96.00	96
6	9905.37	10169.47	51

This table shows, for every query available in MetaDAG, the mean and standard deviation of the execution times of different performed tests

Reaction graphs & Metabolic DAG models

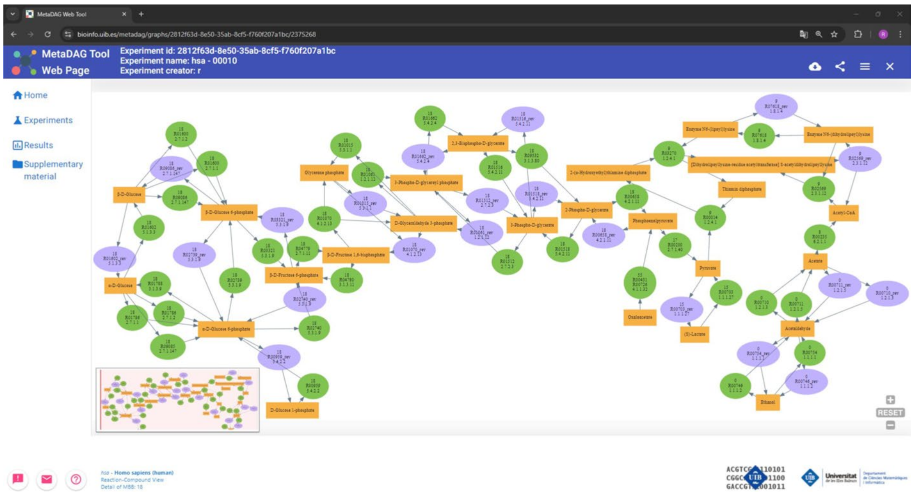
A *reaction graph model* of metabolism is a directed graph $G_R = (R, E)$ where R is the set of chemical reactions present in the metabolism and E is the corresponding set of edges. In this model, there is an edge from R_i to R_j if, and only if, there is at least one metabolite produced by R_i that R_j consumes. In the case of a reversible reaction, two different nodes are considered, one for the forward reaction and the other for the backward reaction. The set of chemical reactions is not necessarily from one single species; it may include reactions coming from multiple species or synthetic organisms.

A *path* from node u to node v in a directed graph G is a sequence of nodes $\{u_0, u_1, ...u_k\}$ such that $u_0 = u$, $u_k = v$, and (u_i, u_{i+1}) is an edge in G for $i = 0, ..., k-1$. Two nodes u, v are said to be biconnected if there is a path in each direction between them. It turns out that biconnectivity is an equivalence relation; therefore, it creates a partition (or cluster) of the set of nodes of G. Every cluster, called a *strongly connected component*, is a subgraph such that every pair of nodes in it are biconnected, and it is maximal under inclusion with this property [16, 17]. In addition, the quotient graph by the biconnectivity relation, called the *condensation* of G, is defined such that each strongly connected component is contracted to a single node. There is an arc from a strongly connected component s_i to a strongly connected component s_j if, and only if, there is an arc in G from a node $u \in s_i$ to a node $v \in s_j$.

A *directed acyclic graph*, DAG for short, is a directed graph with no cycles. The condensation of a directed graph is always a DAG. Thus, for every reaction graph G_R, we can consider its collection of strongly connected components and compute its condensation, which will become a DAG. We call *metabolic DAG* (m-DAG for short) the condensation of a reaction graph G_R, and we call a *metabolic building block* (MBB for short) each strongly connected component in the reaction graph G_R. Those MBBs with only one reaction are called *essential MBBs*. Notice that essential MBBs are the nodes in the m-DAG such that if one were to be deleted, it would increase the number of connected components in the m-DAG.

Given a set of reaction graphs, its *pan-metabolism* is the reaction graph obtained when considering the reactions that belong to one or more reaction graphs in the set. The *core-metabolism* is the reaction graph obtained when considering the reactions belonging to all the set's reaction graphs. Intuitively, the pan-metabolism represents the joint metabolism, while the core-metabolism represents the shared metabolism. These terms were inspired by the pangenome concept [21].

Figure 1 shows the MetaDAG interface used to visualize the reaction graph generated by the first query, specifically the reaction graph of the glycolysis pathway in *Homo sapiens*. Notice that we show the reaction graph of a single pathway, which is a very reduced set of reactions compared to the complete metabolism of *Homo sapiens* (35 reactions versus 1860). The MetaDAG interface used to visualize the metabolic DAG associated with this reaction graph is shown in Fig. 2. We refer the reader to Fig. 4 in [13] to visualize the connection between the reaction graph in Fig. 1 and its corresponding m-DAG in Fig. 2. This m-DAG contains 7 MBBs: 4 MBBs with more than one reaction (depicted in yellow) and 3 MBBs (depicted in green), each consisting of only one reaction. Note that removing the green MBB at the top of the m-DAG does not increase the number of connected components, which remains one. However, if we remove the other green MBBs,

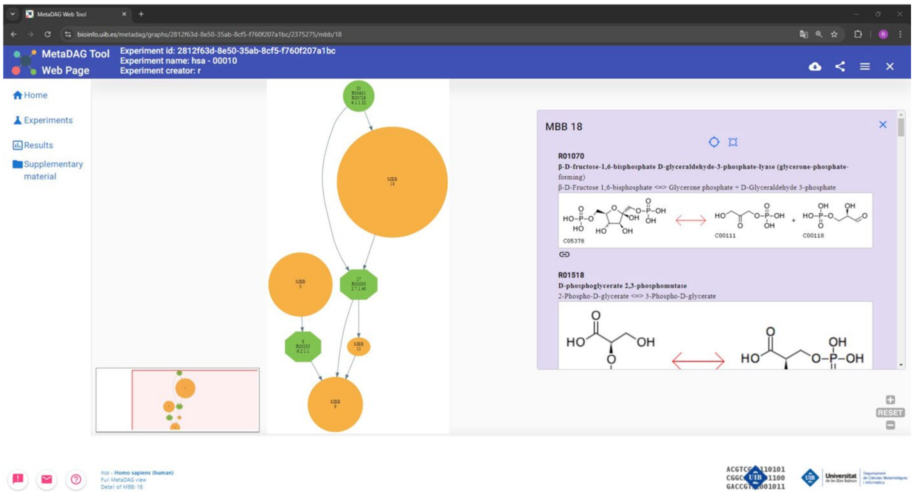


bioinfo.uib.es/metadag/graphs/2812f63d-8e50-35ab-8cf5-f760f207a1bc/2375268

Experiment id: 2812f63d-8e50-35ab-8cf5-f760f207a1bc
Experiment name: hsa - 00010
Experiment creator: r

hsa - Homo sapiens (human)
Reaction-Compound View
Detail of MBB: 18

Fig. 1 Reaction graph of the glycolysis pathway in *Homo sapiens*, computed and displayed with the tool MetaDAG



bioinfo.uib.es/metadag/graphs/2812f63d-8e50-35ab-8cf5-f760f207a1bc/2375275/mbb/18

Experiment id: 2812f63d-8e50-35ab-8cf5-f760f207a1bc
Experiment name: hsa - 00010
Experiment creator: r

MBB 18

R01070
β-D-fructose-1,6-bisphosphate D-glyceraldehyde-3-phosphate-lyase (glycerone-phosphate-forming)
β-D-Fructose 1,6-bisphosphate <=> Glycerone phosphate + D-Glyceraldehyde 3-phosphate

R01518
D-phosphoglycerate 2,3-phosphomutase
2-Phospho-D-glycerate <=> 3-Phospho-D-glycerate

hsa - Homo sapiens (human)
Full MetaDAG view
Detail of MBB: 18

Fig. 2 Metabolic Directed Acyclic Graph (m-DAG) corresponding to the reaction graph of the glycolysis pathway in *Homo sapiens* computed and displayed with the tool MetaDAG

the m-DAG splits into two connected components. Therefore, only these two MBBs are considered essential MBBs. To easily explore the reactions that belong to an MBB, users can select it, and a small window on the right presents information for all the reactions within the MBB. This includes the KEEG ID of the reaction, the graphical representation of the molecular structures of each substrate and product, and a hyperlink pointing to the reaction information on the KEGG webpage.

MetaDAG's calculation and outcome scope

In this section, we explain all the experiments implemented in the tool MetaDAG. Queries 1 to 3 are devoted to computing the metabolic graphs of a single pathway, a single organism, or a single pathway in all organisms from the KEGG catalog. Queries 4 to

6, compute the metabolic graphs of a list of organisms, KOs, compounds, enzymes, or reactions, as well as the construction of synthetic organisms.

Although all metabolic information is retrieved from the KEGG database, we are not restricted to model organisms in general, only in queries 1 to 4 specifically designed to obtain the metabolic networks of the selected organisms and pathways from the KEGG catalog. The other queries, on the contrary, are designed to obtain the metabolic networks of a metagenome sample or transcriptomics, metabolomics, and proteomics data which can be considered as synthetic or new organisms. Therefore, it allows obtaining the metabolic networks for non-model organisms, and even communities.

MetaDAG provides the reaction graphs and the corresponding metabolic DAG models under the following queries:

1. *A specific pathway of an organism*: to generate a reaction graph and m-DAG, users are required to choose one pathway and an organism from the list of metabolic pathways and organisms available in the KEGG database. Then, all reactions present in the selected pathway for the chosen organism are taken into account to construct the reaction graph and m-DAG.
2. *Global metabolic network of one organism*: this query generates the reaction graph and m-DAG for the entire metabolism of a single organism. In this query, the focus is on selecting a singular organism, and the metabolic models are constructed by incorporating all reactions present across all pathways of the chosen organism.
3. *Specific pathway of all organisms*: in this query, a single pathway from the catalog is chosen. The metabolic models are then constructed by considering all reactions present in that pathway for at least one organism. Essentially, this query provides the reaction graph and m-DAG for a reference pathway sourced from KEGG.
4. *Global metabolic network of a list of organisms*: the user has to provide a list of organisms using their KEGG identifier, available at https://www.kegg.jp/brite/br08610. Then, the reaction graph and m-DAG of each organism are constructed in the same manner as in the second experiment. Additionally, the pan-metabolism and core-metabolism of all organisms in the provided list are generated. Users also have the flexibility to select specific groups of organisms from the list, and for each group, their pan and core metabolisms are constructed as well. Furthermore, users can request similarities among all m-DAGs as explained in the next section.
5. *Synthetic metabolism of a list of compounds, reactions, enzymes, or KOs*: this query generates the reaction graph and m-DAG for a given set of compounds, reactions, enzymes, or KO identifiers. When a set of enzymes is provided through the EC number or a set of KO identifiers, or when a list of reactions is specified, even if they are not necessarily from the same organism, metaDAG constructs the reaction graph by retrieving all relevant information from KEGG. Namely, from the list of enzymes, we compile the list of reactions each enzyme catalyses. Using this information, we generate a reaction-compound network as the foundation to generate the metabolic reaction graph. Similarly, when a list of compounds is provided, the models are constructed by considering all reactions where a compound from the list is present in either its product or substrate. Additionally, users can request the similarities among all m-DAGs.

6. *Comparison of several experiments*: this query allows joining the previous queries. Namely, the user may compare a set of organisms and a set of reactions, enzymes, KOs, or compounds, and obtain the analysis of the whole experiment. That is the pan and core metabolisms. In addition, the metabolisms of a synthetic organism can also be constructed. Finally, users can request the similarities among all m-DAGs.

Pairwise similarity of m-DAGs

For every pair of m-DAGs, the tool provides two similarity measures based on the similarity of their MBBs: the *Munkres-similarity* and the *MSA-similarity*.

The Munkres-similarity is the similarity defined in [13] which we briefly recall here. Given two nodes, MBB_1 and MBB_2, their similarity is based on the reactions' similarity score as it was defined in [18]. Then, the similarity score, $S_{mbb}(MBB_1, MBB_2)$, is computed by:

- defining a complete bipartite graph in which the reactions in MBB_1 and MBB_2 are nodes and the weight of each edge $(R_i, R_j) \in MBB_1 \times MBB_2$ is the similarity of R_i and R_j;
- applying the maximum weighted bipartite matching algorithm to the resulting graph to obtain the best match between MBB_1 and MBB_2;
- summing the scores of the best match and dividing it by $\max \{|MBB_1|, |MBB_2|\}$.

Finally, the similarity measure between two m-DAGs, $Sim(mD_1, mD_2)$ is computed by:

- defining a complete bipartite graph in which the MBBs in mD_1 and mD_2 are nodes and the weight of each edge $(MBB_i, MBB_j) \in mD_1 \times mD_2$ is $S_{mbb}(MBB_1, MBB_2)$;
- applying the maximum weighted bipartite matching algorithm to the resulting graph to obtain the best match between mD_1 and mD_2;
- summing the scores of the best match and dividing it by $\max \{|mD_1|, |mD_2|\}$.

The MSA-similarity, which stands for Maximum Similarity Assignment, is defined as follows: Let MBB_1 and MBB_2 be two MBBs, its similarity score is

$$\frac{MSA_{mbb}(MBB_1, MBB_2) + MSA_{mbb}(MBB_2, MBB_1)}{2}$$

where $MSA_{mbb}(MBB_1, MBB_2)$ is defined by

$$\frac{\sum\limits_{R_i \in MBB_1} \max\limits_{R_j \in MBB_2} Sim(R_i, R_j)}{|MBB_1|}$$

and, analogously, we define $MSA_{mbb}(MBB_2, MBB_1)$. Again, the pairwise reactions' similarity score is the similarity defined in [18]. Then, the similarity measure between two m-DAGs, mD_1 and mD_2 is

$$\frac{MSA(mD_1, mD_2) + MSA(mD_2, mD_1)}{2}$$

where $MSA(mD_1, mD_2)$ is defined by

$$\frac{\sum\limits_{MBB_i \in mD_1} \max\limits_{MBB_j \in mD_2} MSA_{mbb}(MBB_i, MBB_j)}{|mD_1|}$$

and, analogously, we define $MSA(mD_2, mD_1)$.

We discuss and compare these two similarities in the results section.

MetaDAG's output

For every query, MetaDAG provides the corresponding metabolic models, that is, the reaction graphs and the m-DAGs. In addition, MetaDAG provides the pan and the core reaction graphs and m-DAGs of different experiments or organisms, as well as the pairwise similarities of all m-DAGs. Upon completion of a query, users receive an email containing a job ID, which they can use to access the results on the MetaDAG front page. In the upper-right panel, users have the option to download the results, share them, or display them, through user-friendly icons. By clicking the display icon, a new window opens, allowing users to select any computed metabolic graph for a specific organism or experiment. This includes the m-DAG, reaction graph, and the largest connected component of each m-DAG. Once a metabolic graph is chosen, it is displayed on an interactive webpage to facilitate user exploration and analyses. For detailed instructions on interpreting the results of the MetaDAG tool, please refer to the pipeline available at https://biocom-uib.github.io/MetaDag/.

Metabolic graphs displayed on the webpage

The metabolic graphs displayed on the tool's webpage from the user's query are the following:

- **Reaction Graph.** To help users visualise the reaction graph and contextualise the reactions in it, metaDAG shows not only the reactions, i.e., the nodes in the reaction graph, but also the metabolites associated with each reaction in the KEGG database. Reactions are depicted as green circular nodes, and, for every reversible reaction, a new node depicted by a purple oval is added. The directed edges from a metabolite to a reaction, or vice versa, show if the metabolite is a substrate or a product of the corresponding reaction. Every reaction and compound can be selected, and then, a new window emerges with its information as well as a link pointing to the corresponding information on the KEGG's webpage.
- **m-DAG.** The nodes of the m-DAGs, i.e., the metabolic building blocks (MBBs), are displayed with different types of nodes. Essential MBBs are depicted as green octagons, while green circles are the MBBs with only one reaction. Yellow circles are those MBBs with more than one reaction, and their size refers to the number of reactions inside the MBB. To easily see the reactions that belong to an MBB, the user can select it and a small window appears with the information of all reactions in the MBB. Namely, the ID of the reaction in KEGG, its graphical representation of the molecular structures of each substrate and product, and a link pointing to the reaction information in the KEGG webpage. On the top-right of this window, in the third

icon, the user may contextualize the reactions, since these selected reactions appear highlighted over the reference pathway in the KEGG format.

- **Biggest Connected Component.** When an m-DAG is huge, as in the case of the m-DAG of the entire metabolism of one organism, there are often many isolated nodes. However, as discussed in the results section, these m-DAGs have a considerably large connected component. Therefore, we found it valuable to separately extract the biggest connected component of an m-DAG, which is displayed in the same format as the full m-DAG. In the supplementary material, File S1 presents the core m-DAG for the kingdom of Animalia, whereas File S2 displays its largest connected component.

Downloadable metaDAG data results

All results for a given query can be classified into two categories: On one side, files containing information about the relationships between the m-DAGs, the MBBs, and the various organisms or samples provided. On the other side, several files describing each m-DAG, so that the information can be visually represented. MetaDAG provides a specific viewer to display the m-DAGs, and also the Reaction Graphs, giving users the possibility to carefully analyze those graphs. Additionally, there is an option to download the files containing the data. In this case, users can select which items they wish to download.

In the most extensive case, the available information is organized into the following categories:

- General data: Files with the information describing the whole set of m-DAGs, they consist of; for each m-DAG which MBBs it contains, for each MBB which are the reactions involved, and, to simplify the data handling, a third file having a combination of both views, that is: for each m-DAG which reactions it has and in which MBB.
- Representation data: A detailed representation of each generated metabolic network (m-DAG and reaction graph), including the different format representations of the graphs, such as, graphml, svg, csv, etc.

We refer to File S3 in the supplementary material for a full description of every item that can be downloaded.

Results

The MetaDAG's methodology has been already applied in two different scenarios. First, MetaDAG was successfully applied to obtain the metabolic DAGs of 2,328 symbiotic genomes included in the public database *Symbiotic Genomes Database - SymGenDB* [14] available at http://symbiogenomesdb.uv.es/. All metabolic DAGs as well as the corresponding pan and core metabolisms at the genus level were calculated and stored in the Meta-DAGs section of the database.

Second, to explore the tools' usability regarding the metabolic network's topology, in [15] the reaction graph and the m-DAG of a minimal metabolic network were constructed from the theoretical minimal gene set machinery revised in [19]. The reaction graph of this minimal metabolic network consisted of 80 compounds and 98 reactions, while its m-DAG had 36 MBBs. Additionally, 12 essential reactions were identified in the m-DAG that were critical for maintaining the connectivity of the network. Similarly, the m-DAG of JCVI-syn3.0, and of Candidatus *Nasuia deltocephalinicola* were constructed. JCVI-syn3.0 is an artificially designed and manufactured viable cell whose genome arose by minimizing the one from *Mycoplasma mycoides* JCVI-syn1.0 [20], and Candidatus *Nasuia deltocephalinicola* is the bacteria with the smallest natural genome known to date. The comparison of the m-DAGs derived from a theoretical, an artificial, and a naturally reduced genome, denoted their different lifestyles, with a consistent core metabolism.

Eukaryotes test

To further evaluate our tool's usability, as well as the m-DAG methodology, we ran a test considering all Eukaryotes from the KEGG database. We introduced them as a list of organisms using query number 4 on the metaDAG front-page. Currently, Eukaryotes are distributed in 535 Animals, 154 Fungi, 139 Plants, and 56 Protists (see File S4 in the supplementary material). We obtained the reaction graphs and the m-DAGs of every organism, as well as the pan and core reaction graphs and m-DAGs for every Kingdom group. In addition, the Munkres-similarity and the MSA-similarity for every pair of m-DAGs were calculated. It took 244 minutes to obtain the results from the tool running on an AMD/7282 biprocessor provided with 512 GiB RAM.

Interestingly, we found that the core reaction graph of all Eukaryotes is empty, which means the absence of any common reaction across all Eukaryotic organisms. However, the core reaction graphs within the kingdom taxonomy level were not empty. Table 2 shows the number of common reactions within every kingdom. We can observe that Animals have 133 reactions and 117 MBBs in common, Plants have 303 reactions and 249 MBBs in common, while Fungi and Protists both have 25 reactions and 22 and 21 MBBs, respectively in common.

Regarding the topology of the m-DAGs, we observe that all computed m-DAGs strongly share a topology profile. Namely, they have many isolated nodes, most of them consisting of MBBs with only one reaction, and also, they all have a considerably big connected component. Fig. 3 illustrates the varying sizes of the connected components of each m-DAG, grouped by kingdom. We observe a concentration of connected components ranging from 1 to 20 nodes. Additionally, there is a noticeable gap between the smaller components and those around 250 nodes, indicating that m-DAGs tend to have a significantly large connected component, along with many isolated or very small ones. Indeed, Fig. 4 presents a violin plot showing the size of the largest connected component of each m-DAG. Animals have the largest component, with 640 nodes, followed by plants with 597 nodes. Additionally, we observe significant variability in animals, where the largest component ranges from 200 to 640 nodes, while in plants, the range is more consistent, from 500 to 600 nodes, with only a few exceptions. We refer to the pipeline https://biocom-uib.github.io/MetaDag/for a complete analysis of these graph's topology.

Concerning the number of reactions in each node of every m-DAG, we also obtain the same pattern as before. Specifically, in all kingdoms, all m-DAGs have a huge

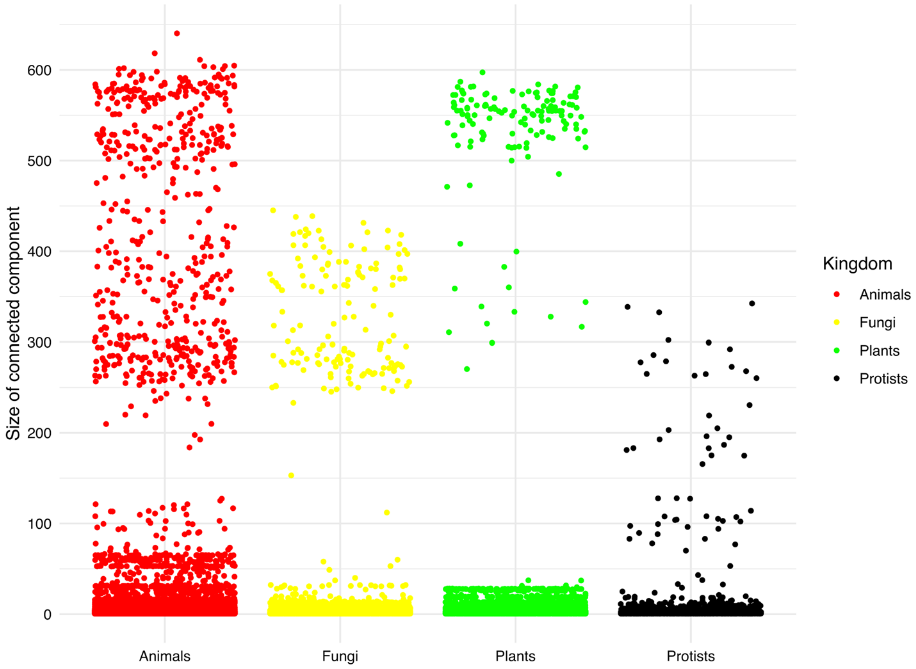


Fig. 3 Size of the connected components of every m-DAG (y-axis) grouped by kingdom (x-axis)

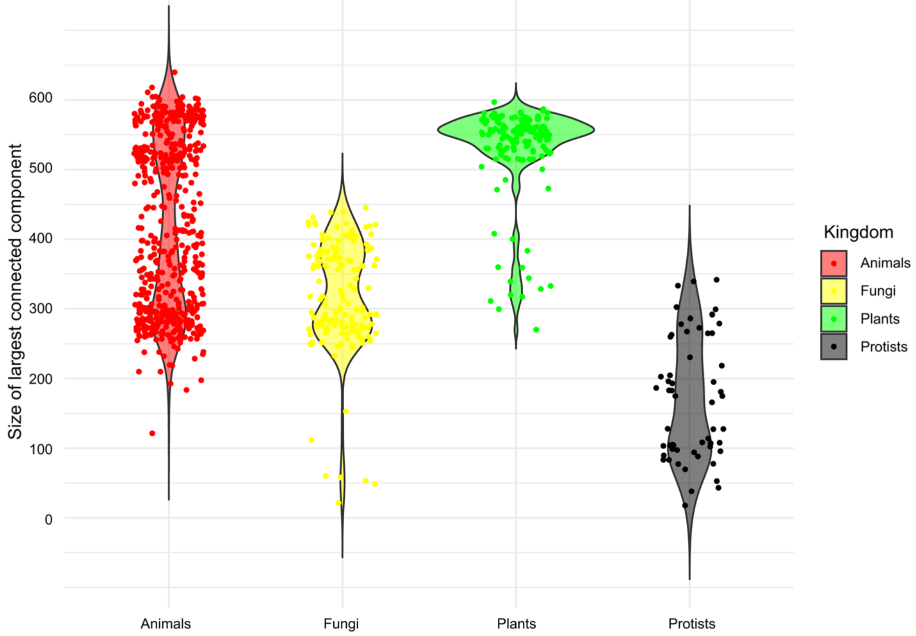


Fig. 4 Size of the largest connected component of every m-DAG (y-axis) grouped by kingdom (x-axis)

node (MBB). And the majority of the MBBs have only one or two reactions. For further details, including all information regarding the metabolic pathways present in the largest MBB of every organism, we refer to File S5 in the Supplementary Material.

Table 2 Reactions and MBBsthat are common by kingdoms

Kingdom	Common reactions	Common MBBs
Animals	133	117
Plants	303	249
Fungi	25	22
Protists	25	21

This table shows, for each kingdom, the number of common reactions present in the metabolism of all organisms within the kingdom, as well as the number of common MBBs

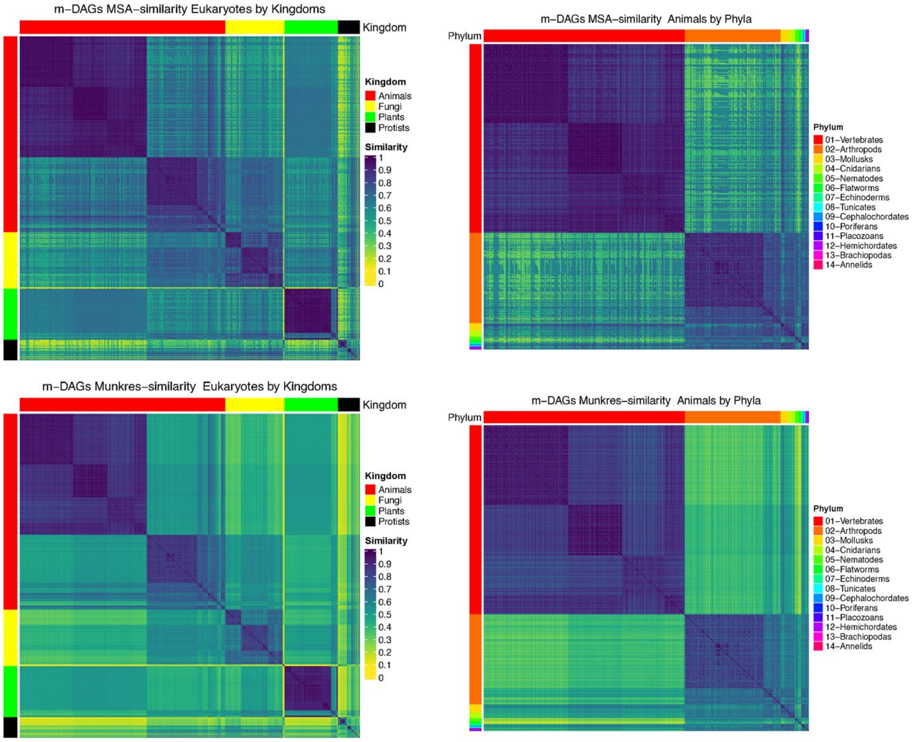


Fig. 5 Heatmaps of the similarity matrices of KEGG Eukaryotes: (a) MSA-similarity at the kingdom level (top-right); (b) Munkres-similarity at the kingdom level (bottom-right); (c) MSA-similarity of animals at the Phylum level (top-left); (d) Munkres-similarity of animals at the Phylum level (bottom-left)

As for the large-scale comparison of all constructed m-DAGs, both, the MSA-similarity and Munkres-similarity measures were calculated with MetaDAG. In Fig. 5 we show the corresponding heatmaps to visualize the results. A heatmap allows for a visual rendering of the similarity matrix produced by each similarity measure. Each cell (i, j) in the heatmap shows the similarity value between the i-th and j-th m-DAGs, colour-coded so that darker colours correspond to a high degree of similarity (from dark blue to yellow). In our case, the heatmaps are symmetric and cells in the main diagonal always show the darkest colour, resulting from comparing an organism with itself. We can observe in this figure that, both similarity measures correctly classify m-DAGs at the kingdom level, and we also clearly distinguish two separate groups within the Animals kingdom. Hence, for each similarity measure, we computed the hierarchical clustering of all Eukaryotes, using the Ward method with 4 clusters. The results are shown in Table 3.

Table 3 MSA and Munkres clusters

Clusters	Animals	Fungi	Plants	Protists
1	331	0	0	0
2	197	0	0	0
3	0	0	125	0
4	7	154	14	56

Clusters obtained at the kingdom level for all m-DAGs of Eukaryotes with the MSA and Munkres-similarity measures

Table 4 Common reactions and MBBs within the different clusters among the total number of them

Clusters	Common reactions	# Reactions	Common MBBs	# MBBs
Cluster 1	525	2037	377	1968
Cluster 2	243	1958	203	1938
Cluster 3	1019	2098	631	1616
Cluster 4	2	2332	2	2842

Munkres and MSA similarity measures

We observe that the Munkres and MSA-similarity separates almost all Animals (except 7 out of 535) into two homogeneous and distinct clusters. Plants are also separated (except 14 from 139) in another homogeneous and distinct cluster, while Fungi and Protists end up within the same cluster together with the 7 animals and 14 plants. These 7 animals are nematodes or flatworms and exhibit a parasitic nature, leading to the development of various diseases. This parasitic condition could result in significant differences in their metabolic characteristics compared to other organisms. Concerning the 14 plants, they are all the green and red algae in the KEGG database. Hence, they are all clustered together. Some of them contain a high lipid content and are suitable for biodiesel production. As in the animal's case, all of them exhibit unique characteristics that can influence their metabolism compared to other plants, and explain their clustering profile.

Similar results are also obtained with the Munkres-similarity. As shown in Table 3, in this case we obtain only 7 of the previous 9 animals clustered together with all Fungi, Protists, and green and red algae. See File S6 in the Supplementary material for a detailed description of these organism's classification.

From this classification, a pertinent question arises: What factors cause these algae and animals to be distinguished from their respective kingdoms? To address this query, we revisited the core metabolism obtained with MetaDAG, but this time for each cluster rather than the kingdom's core metabolism. Table 4 displays the number of reactions and MBBs that are common within each cluster.

We now observe that by dividing animals into two clusters - vertebrates and invertebrates - the number of common reactions increases substantially. Notably, the vertebrates cluster has 377 common MBBs, while the number of common reactions is 525. This indicates that vertebrates share MBBs with more than one reaction. In fact, the largest MBB in this core m-DAG encompasses 1039 reactions. In the plants' cluster, i.e., cluster number 3, we again find that they have 631 MBBs in common from 1019 common reactions. This implies that plants also share MBBs with more than one reaction, where in this case, the largest MBB in this core m-DAG includes 309 reactions. In addition, in cluster number 4 we obtained that only 2 reactions and the corresponding single MBBs are common, namely, R03659 and R04773. Both are catalized by the enzyme 6.1.1.10 which corresponds to the methionine t-RNA ligase. These two common reactions correspond to a tRNA ligase. One of them is related to the selenium metabolism and the other to the methionine metabolism. Both compounds are fundamental to the metabolism of any living organism, indicating that the organisms grouped in this cluster have distinct metabolisms.

In addition, we re-evaluated this classification by increasing the number of clusters to 5 and 6. In this case, we obtained that animals are further divided. Meanwhile, the organisms in cluster number 4, which consists of protists, fungi, algae, and nematodes, continue to be grouped. We refer to the pipeline available at https://biocom-uib.github.io/MetaDag/ for a complete analysis of the hierarchical clustering results.

To further investigate the ability to classify m-DAGs at a deeper taxonomy level, we considered the Animals' classification at the Phylum level. The corresponding heatmaps are displayed in Fig. 5 on the right. We can observe that vertebrates are clearly separated from invertebrates. Also, within the invertebrates in the arthropods' phylum, the class Insects are clearly differentiated from the others. Therefore, we can conclude that both m-DAGs similarities correctly classified the Eukaryotes at different taxonomy ranks. We refer to Files S7 to S12 in the Supplementary material for the heatmaps of the MSA and Munkres similarities at the Phylum level in the Plants, Fungi, and Protists kingdoms.

Lastly, we compared the agreement between the two proposed measures of m-DAG similarity. We calculated the Spearman and Pearson correlation coefficients between the values of MSA-similarity and Munkres-similarity obtained for each pair of m-DAGs. As a matter of fact, we obtained a value of 0.89 for the Spearman correlation coefficient and 0.91 for the Pearson correlation coefficient, indicating that both measures are nearly equivalent.

Figure 6 shows, in a box plot visualization, the similarity values of every pair of m-DAGs. As expected, the MSA-similarity measure obtained higher similarity values with a mean of 0.67 and a standard deviation of 0.18 while the Munkres-similarity measure obtained a mean of 0.55 with a standard deviation of 0.2. Hence, we conclude that both similarity measures almost equally classify the Eukaryotes within the different taxonomy groups.

Gut microbiome analysis

As a final example, we applied the tool MetaDAG to analyse a set of 24 gut microbiome samples. We considered 24 samples from the Study MGYS00000394 in the MGnify database [22]. This study consisted of 12 individuals going through two different diets, Korean (6 individuals) and Western (6 individuals). For every individual, fecal samples were taken before and after three months of diet, and the samples were categorized between individuals who lost a lot of weight (6 individuals) and those who lost little weight (6 individuals). For every sample, we downloaded the KOs functional annotations, also available at Mgnify, resulting from the metagenome analysis they ran from raw Illumina sequence reads. Then, we used query number 6 of the

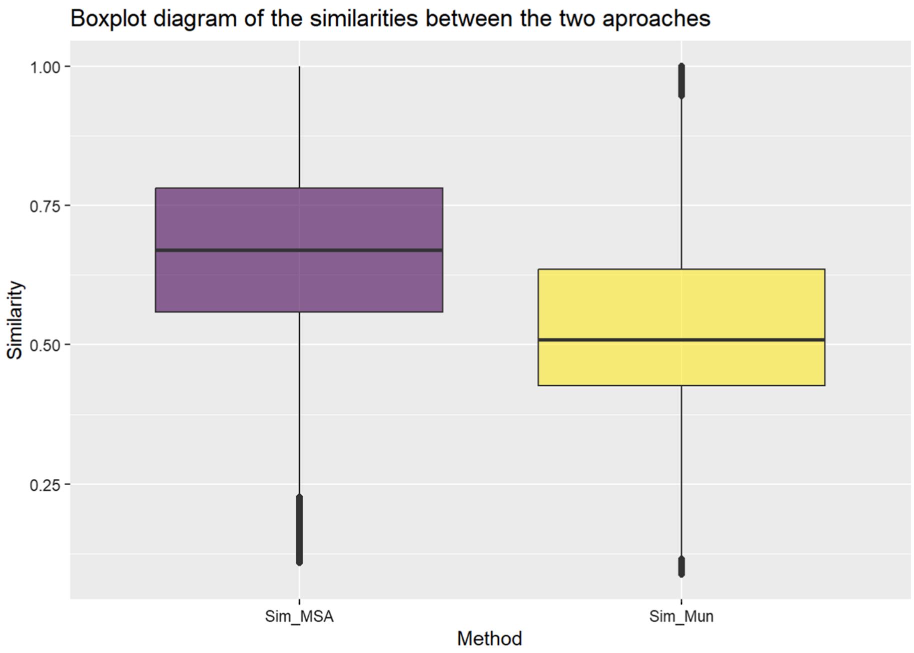


Fig. 6 Boxplot of MSA-similarity values (left) and Munkres-similarity values (right) of the Eukaryotes test

Table 5 This table shows the number of reactions and MBBs that are common in the different groups and the number of reactions and Mbbs in the pan metabolism of each group

Dataset	Common reactions	Common MBBs	Total reactions	Total MBBs
All	3670	678	4423	806
Western	3670	678	4260	773
Korean	4275	782	4423	806
Low All	3670	678	4296	791
High All	4133	764	4423	806
Low Western	3670	678	4114	762
Low Korean	4275	782	4296	791
High Western	4133	764	4260	773
High Korean	4297	792	4423	806

MetaDAG web tool to obtain and compare the metabolism associated with each sample. MetaDAG generated 24 reaction graphs and their corresponding m-DAGs. In addition, we considered different groups by diet (Western/Korean) and subgroups by the amount of weight lost (Low/High).

As a first result, we obtain that the core metabolism of the whole set has 3670 reactions distributed in 678 MBBS and the same m-DAGs' topology pattern obtained in the Eukaryotes test, which may support the existence of a core and stable gut microbiota. Table 5 shows the number of common reactions and MBBs when considering the entire dataset, the two diets (Western and Korean), and the amount of weight lost (High and Low). We can observe that there are tiny differences between the different groups and organisms since the corresponding core metabolisms consist of more than an 80% of the total reactions.

Next, to compare the different groups, we considered the Munkres-similarity provided by the tool. Hence, for every pair of m-DAGs we obtained its similarity and we considered the hierarchical clustering methodology to evaluate the results. Figure 7 shows how the samples are clustered. We observe that except for one Western diet individual, the others are split into two homogeneous clusters, those with the Korean diet and those with the Western one. Within each cluster, those who lost a lot of weight are clustered together, as well as those who lost little weight. We also observe that there are no differences before and after diet. In fact, there are 8 clusters that correspond to an individual before and after diet (4 Korean and 4 Western).

Finally, one of the main challenges of working with big data is our capability to correctly identify and interpret the little but important differences between samples. In fact, in research, it is crucial to figure out how to extract the most relevant information. To this end, we made a Partial Least Squares Discriminant Analysis with its sparse variant (sPLS-DA), which enables the selection of the most predictive or discriminative features (MBBs) in data to classify samples [23]. The sPLS-DA was calculated with the R package MixOmics [24].

Figure 8 displays the level of discrimination (coloured from dark blue to dark red) for each MBB and considered groups. For instance, we can observe that for the Korean dietary intervention, the 8 MBBs on the left clusterised this group, while the 6 MBBs on the right clusterised the Western diet. Also, we observe that there is one individual (WLB01 and WLA02) with many unique discriminative MBBs, which corresponds to the outlier we already obtained in the hierarchical clustering. Furthermore, another individual (KHA24) can be considered an outlier for the Korean dietary intervention, since we can also find numerous discriminative MBBs between this sample and the rest. The dendrogram on the top clusterises the MBBs according to their importance to classify the considered groups. In the dendrogram on the left, we can see how those groups (Western/Korean and Low/High) are classified, and they resemble the dendrogram obtained in Fig. 7. On the top left side of the figure, we show the metadata information.

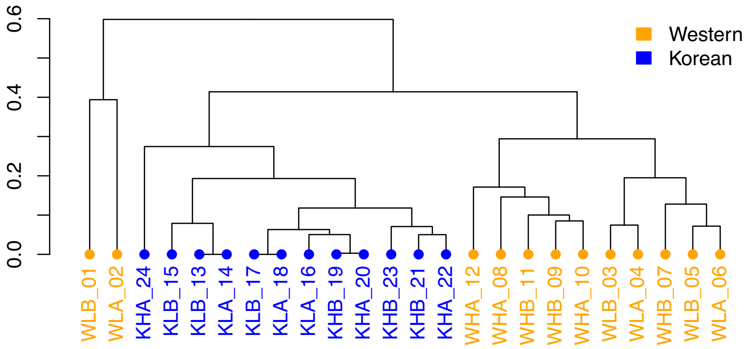


Fig. 7 Hierarchical clustering of the Munkres-similarity of the 24 m-DAGs corresponding to each gut microbiome sample. Labels W/K denote Western and Korean diets, H/L denotes High and Low amounts of weight loss and B/A denotes Before and After diet

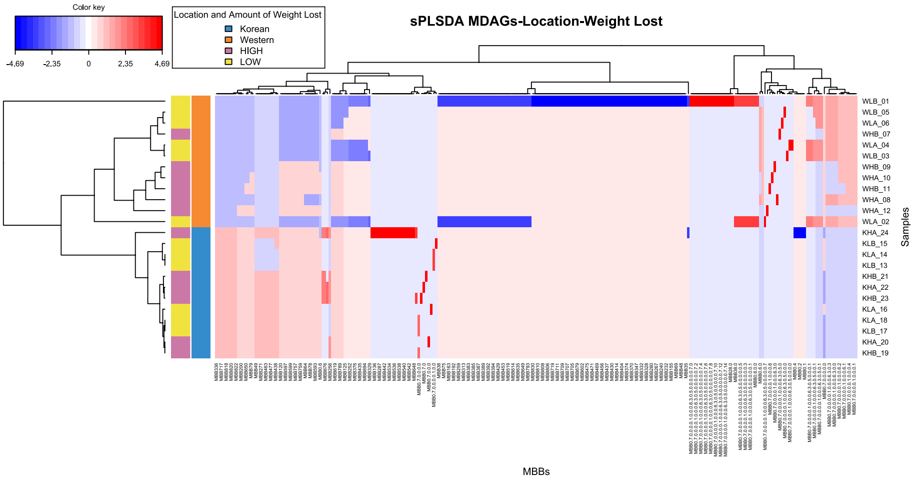


Fig. 8 Heatmap of the partial least Squares discriminant analysis. For each sample corresponding to the labels on the right, the most predictive (dark red) or discriminative (dark blue) MBBs are shown at the bottom. The dendrogram on the top shows the MBBs clusters while the dendrogram on the left shows the m-DAGs classification

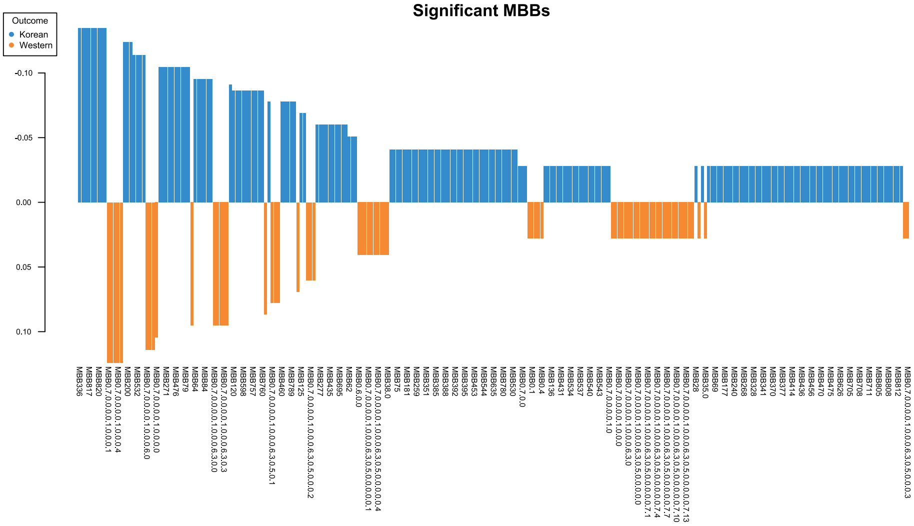


Fig. 9 Loading plot from the sPLS-DA applied to the dataset to discriminate MBBs. Colours indicate the MBBs in which the mean is maximum for each one

Finally, since we wanted to examine those MBBs that appear in the graphic as vertical small lines in dark red, we extracted the most significant MBBs for each group. Figure 9 highlights their distinguishing metabolic capabilities.

Conclusion

This paper introduces a robust implementation of the metabolic networks methodology, enabling the construction of metabolic-Directed Acyclic Graphs (m-DAGs) to address diverse queries. These queries encompass analyzing specific pathways for individual organisms, exploring global metabolic networks of single organisms, studying pathways across all organisms in the KEGG database, investigating global metabolic networks for a list of organisms, examining synthetic metabolism involving various compounds, reactions, or enzymes, and facilitating comparisons between multiple experiments.

MetaDAG exhibits exceptional efficiency in rapidly computing reaction graphs and m-DAGs across a wide range of query types and data sources. Significantly, the integration with KEGG data empowers researchers to compare metabolisms associated with metagenomic and metatranscriptomic data, further enhancing the tool's utility and versatility.

In addition, MetaDAG facilitates the construction of core and pan metabolisms from selected groups of experiments or organisms. This capability offers valuable insights into shared and distinct metabolic features, which contribute to understanding biological processes.

We are aware that pathway databases like KEGG and model organism knowledge have limitations, including incomplete pathway information, static representations that may not capture dynamic biological processes, and biases toward well-studied organisms, making them less reliable for non-model species. Model organisms offer insights, but generalizing their data to other species can be inaccurate, especially outside controlled lab conditions. Additionally, pathfinding with transcriptomics, metabolomics, and proteomics data is complex due to variability, context-dependence, and data integration challenges. These data types also lack quantitative insights and may suffer from technical inconsistencies, making accurate pathway reconstructions difficult. However, despite these limitations, we are making the most of the available data to enhance our understanding and build the best possible reconstructions with current resources.

MetaDAG not only provides real-time interactive results on its user-friendly webpage but also facilitates further analysis and exploration through downloadable files. In addition, we provide a comprehensive pipeline and guide to analyse the output results effectively. This resource equips researchers with the necessary tools and instructions to make the most of MetaDAG's capabilities. We present here the results of a Eukaryotes test, and a gut microbiome test, as examples of well-known organisms and common uses we believe users can appreciate, and implement in their analysis. Furthermore, we also described MetaDAG's performance and potential across a broad range of applications.

Availability and requirements

Project name: metaDAG Project home page: https://bioinfo.uib.es/metadag/ Operating system(s): Platform independent Programming language: Java, Typescript Other requirements: Angular, Spring Boot, Apache Maven License: End User License Agreement: https://bioinfo.uib.es/metadag/eula This tool is free for academic/non-commercial use. Any restrictions to use by non-academics: This web application is provided for academic, research, and educational purposes. Users are responsible for ensuring compliance with third-party data licenses, including KEGG, and must obtain the necessary permissions for any commercial use of such data. The application outputs may be used in research with proper attribution.

Acknowledgements

We sincerely thank the reviewers for their time and the effort they have invested in reviewing our manuscript and helping us enhance its quality.

Author Contributions

PPR: Tool design and development. RA: Pipeline to analyse MetaDAG results. MRP: Microbiome test analysis. ML,PPR: Conceptualization. JAC,PPR,MRP,ML:Writing-Review & Editing. ML: Supervision and funding acquisition. All authors contributed to the manuscript revision and approved the submitted version

Funding

This work has been supported by the Ministerio de Ciencia e Innovación (MCI), the Agencia Estatal de Investigación (AEI) and the European Regional Development Funds (ERDF) for its support to the project MICROMATES PGC2018-096956-B-C43 and Grant PID2021-126114NB-C44 funded by MCIN/AEI/ 10.13039/501100011033

Availability of Data and Materials

The list of organisms in the dataset, the reaction graphs, and the m-DAGs for each eukaryote from the KEGG database, as well as the pan and core reaction graphs and m-DAGs for each Kingdom group, are available at: https://bioinfo.uib.es/metadag/handleExperiment/0a845f74-826e-3b46-aed9-e7ecf74db262 Materials availability: The supplementary material is available at: https://bioinfo.uib.es/MetaDAG_Supplementary_Material/. The pipeline for analyzing MetaDAG results is available at: https://biocom-uib.github.io/MetaDag/.

Code Availability

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests related to this manuscript.

Received: 15 August 2024 Accepted: 13 January 2025
Published online: 28 January 2025

References

1. Shah HA, Liu J, Yang Z, Feng J. Review of machine learning methods for the prediction and reconstruction of metabolic pathways. Front Mol Biosci. 2021;8:634141. https://doi.org/10.3389/fmolb.2021.634141.
2. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. Genome Biol. 2019;20(1):158. https://doi.org/10.1186/s13059-019-1769-1.
3. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
4. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28(11):1947–51. https://doi.org/10.1002/pro.3715.
5. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587–92. https://doi.org/10.1093/nar/gkac963.
6. Karp PD, Billington R, Caspi R, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 2019;20(4):1085–93. https://doi.org/10.1093/bib/bbx085.
7. Caspi R, Billington R, Keseler IM, Kothari A, Krummenacker M, MidfordPE Ong WK, Paley S, Subhraveti P, Karp PD. The MetaCyc database of metabolic pathways and enzymes - a 2019 update Nucleic. Acids Res. 2020;48(D1):D445–53. https://doi.org/10.1093/nar/gkz862.
8. Ramon C, Stelling J. Functional comparison of metabolic networks across species. Nature Commun. 2023;14(1):1699. https://doi.org/10.1038/s41467-023-37429-5.
9. Cottret L, Frainay C, Chazalviel M, Cabanettes F, Gloaguen Y, Camenen E, et al. MetExplore: manage and Explore metabolic networks. Nucl Acids Res. 2018;46(W1):W495–502.
10. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome?scale metabolic reconstructions. Mol Syst Biol. 2009;5(1):320.
11. Jing LS, Shah FFM, Mohamad MS, Hamran NL, Salleh AHM, Deris S, et al. Database and tools for metabolic network analysis. Biotechn Biopr Eng. 2014;19(4):568–85.
12. Fani R, Fondi M. Origin and evolution of metabolic pathways. Phys Life Rev. 2009;6:23–52.
13. Alberich R, Castro J, Llabrés M, Palmer-Rodríguez P. Metabolomics analysis: finding out metabolic building blocks. PLoS ONE. 2017;12(5):e0177031.
14. Reyes-Prieto M, Vargas-Chávez C, Llabrés M, Palmer P, Latorre A, Moya A. An update on the Symbiotic Genomes Database (SymGenDB): a collection of metadata, genomic, genetic and protein sequences, orthologs and metabolic networks of symbiotic organisms. Database. 2020;2020:160.
15. Reyes-Prieto M, Gil R, Llabrés M, Palmer-Rodríguez P, Moya A. The metabolic building blocks of a minimal cell. Biology. 2021;10(5):1452.
16. Gross JL, Yellen J. Graph theory and its applications (Second Edition). CRC Press; 2011.
17. Clemente JC, Satou K, Valiente G. Phylogenetic reconstruction from non-genomic data. Bioinformatics. 2006;23:110–5.

18. Alberich R, Llabrés M, Sánchez D, Simeoni M, Tuduri M. MP-Align: alignment of metabolic pathways. BMC Syst Biol. 2014;8(1):1–16.
19. Gabaldón T, Peretó J, Montero F, Gil R, Latorre A, Moya A. Structural analyses of a hypothetical minimal metabolism. Philos Trans R Soc B Biol Sci. 2007;362:1751–62.
20. Hutchison CA, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351(6280):aad6253. https://doi.org/10.1126/science.aad6253.
21. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci. 2005;102(39):13950–5.
22. Richardson LJ, Allen B, Baldi G, Beracochea M, Bileschi M, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell L, Curtis T, Escobar-Zepeda A, Gurbich T, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S. Wilkinson D and Finn RD MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res. 2023. https://doi.org/10.1093/nar/gkac1080.
23. LêCao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinform. 2011;12:253. https://doi.org/10.1186/1471-2105-12-253.
24. Kim-Anh Le Cao, Florian Rohart, Ignacio Gonzalez, Sebastien Dejean with key contributors Benoit Gautier, Francois Bartolo, contributions from Pierre Monget, Jeff Coquery, FangZou Yao and Benoit Liquet. mixOmics: Omics Data Integration Project. R package version 6.1.1. 2016; https://CRAN.R-project.org/package=mixOmics

Publisher's Note