MetaDAG Tool

Summary of experiment data files and formats

Depending on the type of experiment, the generated data will vary. Here, you will find a description of all the possible content that can be produced.

Basic information files

These files contain general information and are all located in the root directory of the downloadable data.

Experiment.json

This file contains a brief experiment description.

Report.pdf

This file contains a detailed description of the calculations performed.

Different_MBB.csv

This is a comma-separated file containing information related to Metabolic Building Blocks (MBB). One of the objectives of the MetaDAG Tool is to identify the MBBs of the considered samples or organisms. In this file, each MBB is represented in a row, and the information for each one is organized into columns. The contents are as follows:

- `MBB id`: a unique identifier for the MBB.
- `natural`: a binary value (0/1). A value of 1 indicates that the MBB has been observed in a metaDAG of at least one sample or organism, while a value of 0 means that the MBB has been generated due to its organization into groups or categories.
- `#pathways`: the number of KEGG pathways in which the MBB appear.
- If the experiment includes classification by taxonomic categories, the file provides, for each category, the count of samples or organisms belonging to that category that also have the MBB in their respective metaDAGs.
- `#groups`: if the experiment includes user-defined groups, this column displays the count of groups in which the MBB is present.
- For each considered group, the column provides the count of samples or organisms belonging to that group which also possess the MBB in their respective metaDAGs.
- `#Reactions`: for the entire set of reactions appearing in the metabolism of at least one sample or organism, if the MBB contains the reaction, a value of 1 is assigned, and if not, a value of 0 is recorded.
- `Reactions`: text string listing the set of reactions forming the MBB.

Different_mDAG.csv

This is also a comma-separated file containing general information related to the Metabolic DAGs (mDAGs). Similar to the previous case, each mDAG is represented in a row, and the information is organized into columns, including:

- `mDAG Id`: a unique identifier for each metabolic DAG.
- `#Categories`: the number of taxonomic categories containing species with this mDAG (present only when taxonomic category classification is applied).
- For each taxonomic category considered, a column indicating the number of samples or organisms with the mDAG (present only when taxonomic classification is used).
- `#Groups`: the number of user-defined groups containing species with this mDAG (present when grouping is applied).

- For each group considered, a column with the number of organisms or mDAG's belonging to the group. Those columns are present only if the experiment has been configured with some grouping.
- `#MBB`: the quantity of Metabolic Building Blocks (MBB) in the current mDAG.
- For the entire set of identified MBBs, a 1 or 0 indicating whether the MBB belongs to the mDAG or not.
- `#Members`: the number of samples or organisms with the mDAG.
- `Members`: a text string listing the samples or organisms with the mDAG.

Results.csv

This is a comma-separated file containing general information about the samples or organisms included in the experiment, as well as for each grouping or taxonomic level under consideration. For each entry, the following columns are included:

- `Organism`: The KEGG acronym for the organism, or if it's a sample or a simulated/synthetic organism, the user given name.
- `Categories`: The list of taxonomic categories associated with the organism or sample. In the case of a synthetic organism or a sample, the value in this column is "`Unknown`."
- `Groups`: The list of groups to which the current sample or organism belongs.
- `mDAG Id`: The associated metabolic DAG identifier.
- `Full Name`: The name of the organism or the identification provided for the sample.

For each reaction, it is represented as '`NA`' if the reaction is not present in the organism's metabolism, or as the MBB identifier if the reaction is part of the organism's metabolic processes.

Similarities_MBB_MSAMethod.csv

The document contains a similarity matrix for MBBs. These distances are calculated using the MSA (Multiple Sequence Alignment) method. The data is organized in a comma-separated values format, with the MBB ids located in both the first row and the first column

Similarities_MBB_MunkresMethod.csv

Similar to the previous scenario, this document contains a similarity matrix for MBBs. However, in this case, the distances are computed using the Munkres (Hungarian) method. The data is organized in a comma-separated values format, with the MBB ids present in both the first row and the first column. In both cases, the similarities are calculated based on the similarities between the reactions associated with each MBB.

Similarities_mDAG_MSAMethod.csv

This document includes a similarity matrix for mDAGs. The distances are computed using the MSA (Multiple Sequence Alignment) method. The data is formatted in comma-separated values, with the mDAG ids listed in both the first row and the first column.

Similarities_MBB_MunkresMethod.csv

This document also includes a similarity matrix for mDAGs. In this case, the distances are computed using the Munkres (Hungarian) method. The data is organized in a comma-separated values format, with the MBB ids present in both the first row and the first column.

The similarities are determined based on the relationships between the MBBs associated with each mDAG.

Metabolic DAG specific files

Depending on the selected experiment, it is possible that the resulting data is organized into various directories, each grouping information related to specific categories, groups, samples, or organisms involved in the experiment. Within each category, several files are generated.

The generated files are associated with two types of graphs: those related to the mDAG considered as strongly connected components (indicated by the '`_mDAG`' in their names) and those related to the reactions-compounds graph (indicated by the '`_RC`' in their names). In both cases, these files encompass graphical information in various formats, as well as descriptions of the respective graphs. Additionally, there are files representing only the largest strongly connected component of the metabolic DAG, identified by '`_mDAG_biggerDAG`' in their names.

The graphic formats used include:

- Portable Document Format (`.pdf`): This widely used file format is suitable for structured document representation. It's independent of the operating system, hardware, or other software.
- Scalable Vector Graphics (`.svg`): This vector-based format allows images to be scaled up or down without any loss in quality.

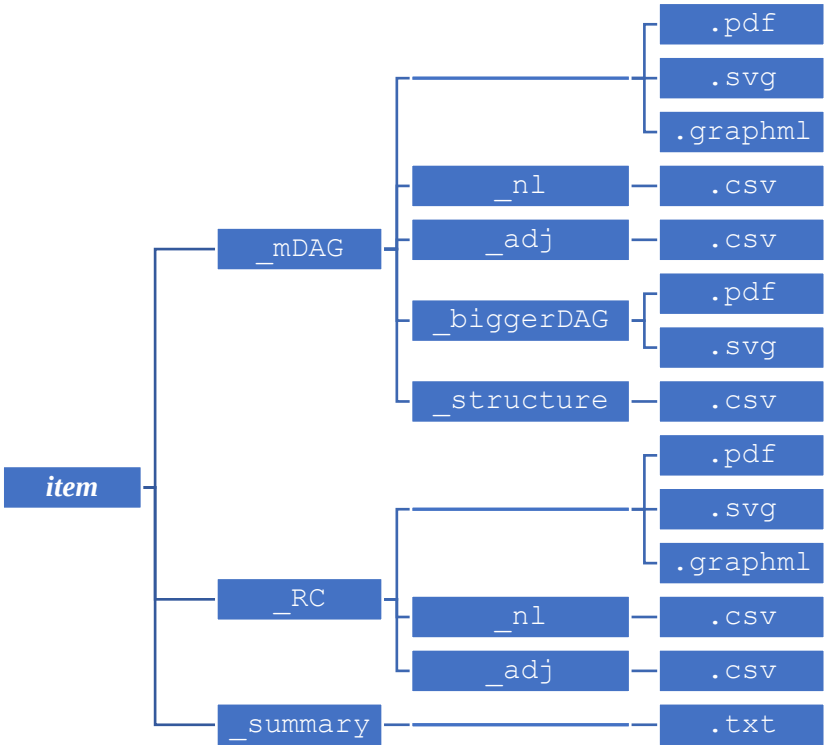
The content description files come in the following formats:

- GraphML format (`.graphml`): This format can be considered a hybrid format, as it includes graphical representation. GraphML is an XML-based description language, allowing the graph to be compatible with a wide range of graph-processing software.
- Node List (`_nl.csv`) and Adjacency List (`_adj.csv`): Both of these files are structured using comma-separated values. The node list represents the nodes of the graph, with the first column containing a unique identifier and the second column containing a label for each node. The adjacency list represents the connections between nodes, i.e., the edges of the graph. The first column specifies the node of origin, the second column designates the destination node, and the third column may either be empty or contain a label for the edge.
- Structure (`_structure.csv`): This file, formatted as comma-separated values, provides a description of strongly connected components, one row for each one forming the metabolic DAG. The first column indicates the number of components, and there are additional columns, each corresponding to a component. A '1' in a column signifies that the component belongs to the strongly connected component, while '0' indicates otherwise.
- Summary (`_summary.txt`): This plain text file provides a detailed description of the connected component. It includes information on the number of different reactions, enzymes, and compounds involved, as well as the MBB and the associated metabolic pathways related to the MBB.

Thus, if the name of the element of interest is ***item*** then the following files are generated:

MetaDAG Tool

Summary of experiment data files and formats



Global directory

This directory contains metabolic DAGs for the entire experiment, including two subdirectories: one for pan metabolism and another for core metabolism.

TaxonomyLevels directory

This directory contains metabolic DAGs for each identified taxonomic classification within the experiment. Each specific taxonomic classification has its own directory, which further contains two subdirectories for core and pan data. Additionally, there is a `taxonomic_classification.txt` file listing the taxonomic classifications used.

Groups directory

If the experiment involves grouping, you'll find a directory within this one dedicated to the relevant data. Inside, there are two subdirectories, for core and pan metabolism data. Additionally, there is a `groups.txt` file listing the represented groups.

Individuals directory

For each sample or organism included in the experiment, there is a dedicated directory containing the respective information. Furthermore, an `individuals.txt` file lists all the samples or organisms.