Manual (under construction)

Input files

In this tab, you can supply files to *massoc*. Moreover, you can save or clear the current settings, and load saved settings from a .xyz settings file. If you supplied the correct inputs, the dialog box on the right will summarize some properties of your files.

Set default directory

This directory contains the filepath to a location where you would like output files to be stored.

Open BIOM files

Specify the complete filepaths of your BIOM files. You can select multiple BIOM files at once, although you can only specify settings once; hence, if these settings are incompatible with some of your BIOM files, this may cause *massoc* to crash.

Construct network or open network file

Selecting one of these options enables you to construct networks from *massoc* or to import an edge list. The IDs specified in these edge lists should match OTU IDs in an uploaded BIOM file or count table, and if available, the weight can be specified in a third column.

Open tab-delimited files

Specify the complete filepaths of your tab-delimited count tables. All tables should be in a format that can be accept by the BIOM-format parser. This implies that they have a header which is preceded by a *#value*, e.g. *#OTU sample sample_1 sample_2.*

Preprocessing

In this tab, you can specify preprocessing steps that *massoc* should carry out before network inference.

Prefix for saved files

Helps identify intermediate files stored by *massoc*.

Show data properties

As multiple files can be supplied, this dialog allows you to select one so that taxon prevalence and sample counts are visualized in the area below.

Remove taxa with mean count below:

Before a prevalence filter or rarefaction, you can remove taxa that have a count number below the supplied number. This can help remove singletons or low-abundance taxa. The removed counts are binned.

Rarefy:

Rarefy to even depth or to a specific count number. If the supplied count number is higher than the sample count number, the sample is removed from the dataset.

Set prevalence filter in %

Filter taxa so only those remain that are present in the specified percentage of samples. The removed counts are binned. Prevalence filters can help remove spurious associations, and the default is therefore set to 20%.

Split samples by metadata variable

Select a metadata column to split the datasets by. For example, half of your datasets were taken from lakes, and the other half from rivers. If this is in the BIOM file, you can use it to split the BIOM file and run network inference on the samples separately (as well as on the original network).

Taxonomic levels to analyze

Select taxonomic levels to perform network inference on. The BIOM file is collapsed to different taxonomic levels, and the BIOM format *collapse* function is adapted to preserve the taxonomy up to the collapsed level. The identifiers of the collapsed taxa are also preserved in these BIOM files.

Clustering dialog

If there are no metadata available that explain microbiome structure, there may still be a structure present that could induce spurious associations (e.g. pH may explain microbiome clustering, but pH was not measured). In that case, *massoc* can run clustering algorithms: it first transforms the data using the centered log-ratio, and then ordinates them in PCA space. The PCA coordinates are used as input for the clustering algorithms. The cluster identities can either be added to the metadata, or used to split the BIOM file in multiple files.

Network inference

In this tab, you can specify which tools should be used to run network inference. For some of these tools, you can also adjust the settings. Finally, an overview of the settings is given, as well as an option to copy these to a command line call.

Select network inference tools to run

Currently, SparCC, CoNet and SPIEC-EASI are supported. When selected, these tools will run network inference using the default settings.
For more details on these tools, please check out the original publications:
SparCC
SPIEC-EASI
CoNet

Select CoNet3 folder

massoc uses the location of your CoNet3 folder (available here) to execute its CoNet.sh script.

Select SparCC folder

Massoc uses the location of the SparCC folder (available here) to execute SparCC functions.

Change settings for SparCC

SparCC bootstraps correlations, and uses the bootstraps to compute pseudo p-values. Lower p-values can result in sparser networks. The number of bootstraps and the p-value threshold can be specified here.

Change settings for SPIEC-EASI

SPIEC-EASI can use two different algorithms: the Meinshausen-Buhlmann algorithm and the Graphical Lasso algorithm. For both of these algorithms, the sparsity of the final network is controlled by the number of StARS repetitions. The number of StARS repetitions and the SPIEC-EASI algorithm can be specified here.

Change settings for CoNet

As CoNet has a multitude of settings, there is no option to adjust them inside *massoc.* However, a future update will contain a utility to supply alternative CoNet.sh scripts to *massoc.*

Number of processes

After all settings have been entered, *massoc* will specify the number of unique networks that need to be inferred. Network inference can be sped up significantly if networks are inferred simultaneously. More cores can be added if a particularly large number of networks needs to be inferred.

At the moment, this functionality has been disabled to accommodate better integration with Pyinstaller.

Export as command line call

While there is an option to generate a command line call, there is no compiled version of the command line utility currently available. However, this is currently under development.

Run network inference

Network inference will start, and a loading bar keeps track of progress. When finished, one or more output edge lists will be written to the previously specified file path.

Network database

In this tab, previously selected files can be uploaded to a Neo4j graph database. You can carry out several operations on the database contents: agglomerate edges, associate taxa to sample metadata and perform logic operations on multiple networks. The resulting networks are exported to a Cytoscape-compatible GraphML format.

Neo4j database access

To access the Neo4j database, you will first need to set it up for your system – check the README for more information. Once it has been set up, *massoc* needs the local address, username and password to carry out database operations.

Launch database

After specifying details, you can launch the database. This will upload all previously selected files (BIOM files, network files and more).

Close database

Shutting down *massoc* does not shut down the database, and the lingering Java process will prevent you from running *massoc* again. Click this button to safely shut down the database.

Agglomerate edges

By selecting this option, you can choose to agglomerate edges. Edges between taxa that share the selected phylogenetic level (e.g. *Roseobacter denitrificans* and *Roseobacter litoralis* at the genus level) will be merged. If you tell *massoc* to take weight into account, only edges that have the same weight are merged. This can reduce the number of edges in your network and can be especially helpful if you have datasets with large numbers of similar tag sequences.

Because the graph database can only store a single graph, this option irreversibly agglomerates edges. Once performed, you will need to upload your data to the database again to restore it.

Associate taxa

Taxa can be associated to sample variables. At the moment, this is done through Spearman's rank correlation coefficient for quantitative variables, and a hypergeometric test for qualitative variables. Because this can take a long time for a complete network, you will need to select metadata variables to test against. There is no multiple-testing correction applied, and this option does not provide a robust statistical test; however, it may indicate which associations could be investigated further.

Logic operations

If you are analyzing multiple networks, it can be helpful to compare those networks. With these operations, you can export the union, intersection or difference of networks.

Currently, there is no option yet to select specific networks. Hence, the *union* option returns all associations in all networks, the *intersection* option returns all associations that are shared between all networks, and the *difference* option returns all associations that are unique for each network.

Run database operations

After running the selected operation, *massoc* will export the resulting network. At the moment, you can only export networks to a rich GraphML format.