Manual (under construction)

Contents

massoc modules

massoc consists of several modules that can be used independently given the correct *settings.json* file. This manual contains an explanation of all *massoc* parameters, for both the command line interface (CLI) and the graphical user interface (GUI).

Currently, *massoc* contains five modules, with corresponding GUI tabs:

- `massoc input` **1,2**
- `massoc network` **3**
- `massoc database` **4**
- `massoc netstats` **5**
- `massoc metastats` **5**

The software can be run through the CLI by first specifying a module and then the set of input parameters. For example, running both CoNet and SPIEC-EASI with a custom CoNet script can be done as follows:

```
massoc network –tools conet spiec-easi –conet_bash
../conet_script.sh –conet ../CoNet3 –fp ../filepath
```

Note that the `–fp` parameter is required for all modules, as this specifies the location of the *settings.json* file that was generated by previous modules or by the user.

massoc input

This module sets several initial parameters, such as the default file path and filenames. Moreover, it imports BIOM files and tab-delimited files and preprocesses these.

`massoc input -fp` **1**

All files will be exported to the file path specified here. Additionally, *massoc* will save some parameters to a *settings.json* file stored in this location.

`massoc input -biom` **2**

Supply the complete file paths to BIOM files, or select the appropriate files through the explorer window. BIOM files are imported through the BIOM API and therefore need to meet the standards described by http://biom-format.org/.

`massoc input -net` **3**

The listbox first enables either the *Network inference* tab or the *Open network file* button. The `-net` parameter or button can then be used to select weighted edge lists. Weighted edge lists will be compared to imported BIOM files to ensure the taxon identifiers match those in the BIOM files.

GUI-only **4**

Enables selection of tab-delimited files below.

`massoc input -otu` **5**

Supply the complete file path to the tab-delimited count table. More than one count table can be provided, given that taxonomy files and sample data files are provided in the same order. Note that the number of column headers needs to be equal to the number of columns (row names should also have a header).
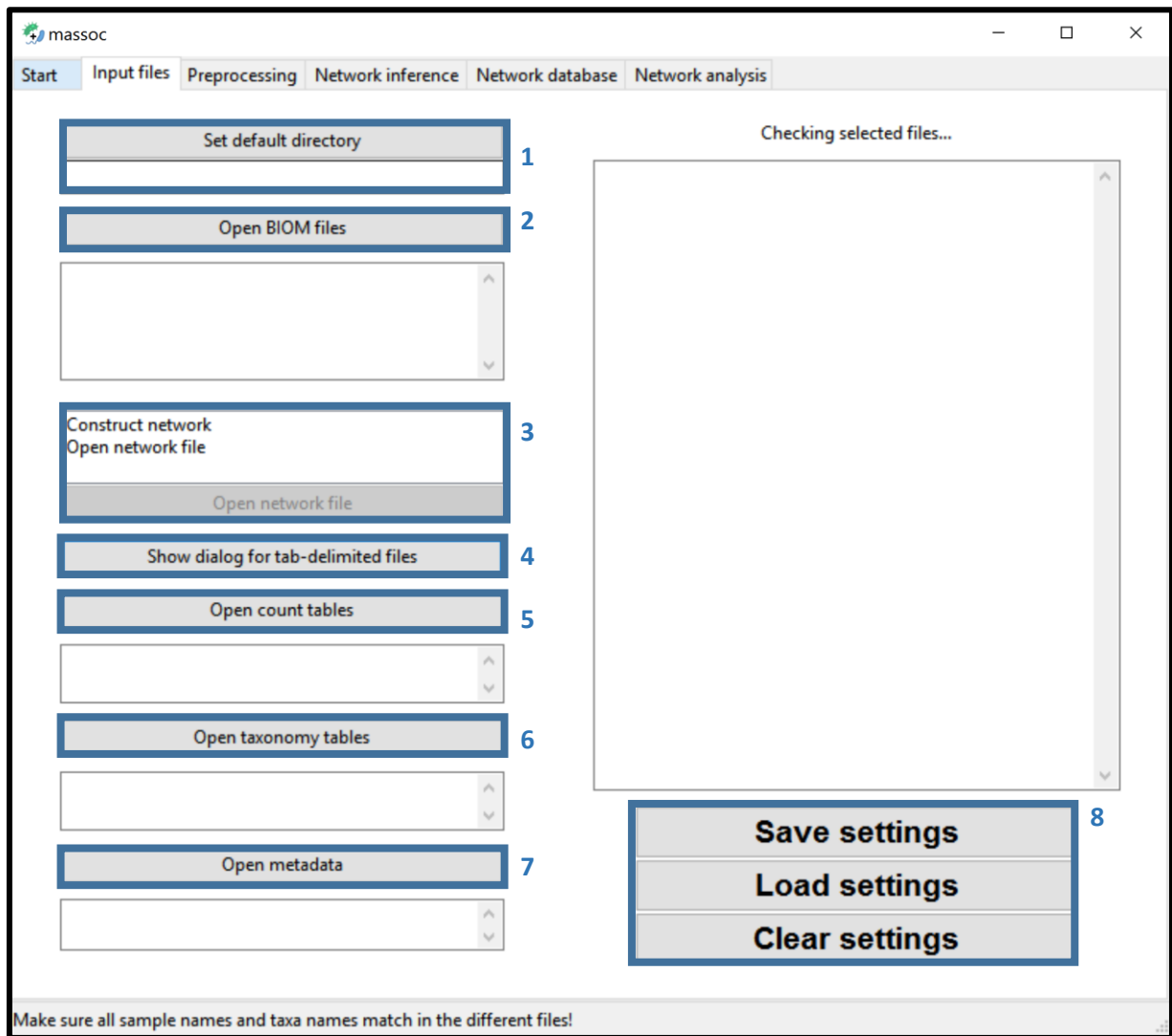
`massoc input -tax` **6**

Supply the complete file path to the tab-delimited taxonomy table. Note that the number of column headers needs to be equal to the number of columns (row names should also have a header).

`massoc input -sample` **7**

Supply the complete file path to the tab-delimited sample metadata table. Note that the number of column headers needs to be equal to the number of columns (row names should also have a header).

GUI-only **8**

Saves, loads or clears current settings in the GUI. The name and location of the settings file can be specified after clicking the *Save settings* button.

massoc

Start | Input files | Preprocessing | Network inference | Network database | Network analysis

Set default directory — 1

Open BIOM files — 2

Construct network
Open network file — 3

Open network file

Show dialog for tab-delimited files — 4

Open count tables — 5

Open taxonomy tables — 6

Open metadata — 7

Checking selected files...

Save settings
Load settings
Clear settings — 8

Make sure all sample names and taxa names match in the different files!

massoc input

This part of the *input* module carries out preprocessing steps.

`massoc input –name` **9**
The specified prefix is added to all intermediate files.

GUI-only **10**
Visualizes taxon prevalence and sample counts for the selected file.

`massoc input –min` **11**
Minimum mean abundance required. If taxa have a mean abundance lower than the specified value, their counts are added to a bin and the taxa removed from the file.

`massoc input –rar` **12**
Rarefaction step. If 'To even depth' is selected in the GUI or `–rar True` for the CLI, all samples are rarefied to even depth. If 'To count number' is selected, a box appears that can be used to specify the minimum count number. Samples with lower counts are removed and the remaining samples are then rarefied to even depth.

`massoc input –prev` **13**
Prevalence filter; specify as a percentage. Taxa with a prevalence below the specified threshold have their counts added to a bin and are then removed from the file.

`massoc input –split` **14**
Once the biom file has been read into *massoc*, a list of metadata variables will appear in this box. One of these (qualitative) variables can be selected to split the data into multiple count tables and run network inference separately.

`massoc input –levels` **15**
Taxonomic levels used for network inference. Taxon counts are binned depending on their taxonomic assignments. This binning currently occurs after the other preprocessing steps.

GUI-only **16**
Enables selection of settings for clustering the count data.
The following parameters can be set here:
`massoc input –cl`   Choice of clustering algorithm
`massoc input –nclust`   Number of clusters
`massoc input –split`   If set to TRUE, count data is split by cluster assignment

massoc network

This module carries out network inference on the supplied BIOM files.

`massoc network –tools` **1**

Specify tools to run from *massoc*. Currently, only SparCC, CoNet and SPIEC-EASI are supported.

`massoc network –conet` **2**

Location of folder with CoNet Java library; download the latest version here.

`massoc network –spar` **3**

Location of folder with SparCC code, available here.

`massoc network –spiec_settings –conet_bash –spar_pval –spar_boot` **4**

Adjust settings for network inference tools. In the GUI, settings for SPIEC-EASI (algorithm and number of StARS repetitions) can be selected; in the CLI, an alternative R script can be given to run SPIEC-EASI. Similarly, an alternative shell script can be used to run CoNet. Finally, the p-value threshold and number of bootstraps can be specified both through the GUI and through the CLI.

`massoc network –cores` **5**

The number of jobs is equal to the number of files times the number of network inference tools. *massoc* will distribute jobs over the specified number of cores, by default 4.

GUI-only **6**

After all settings have been specified, *massoc* can be executed by clicking this button.
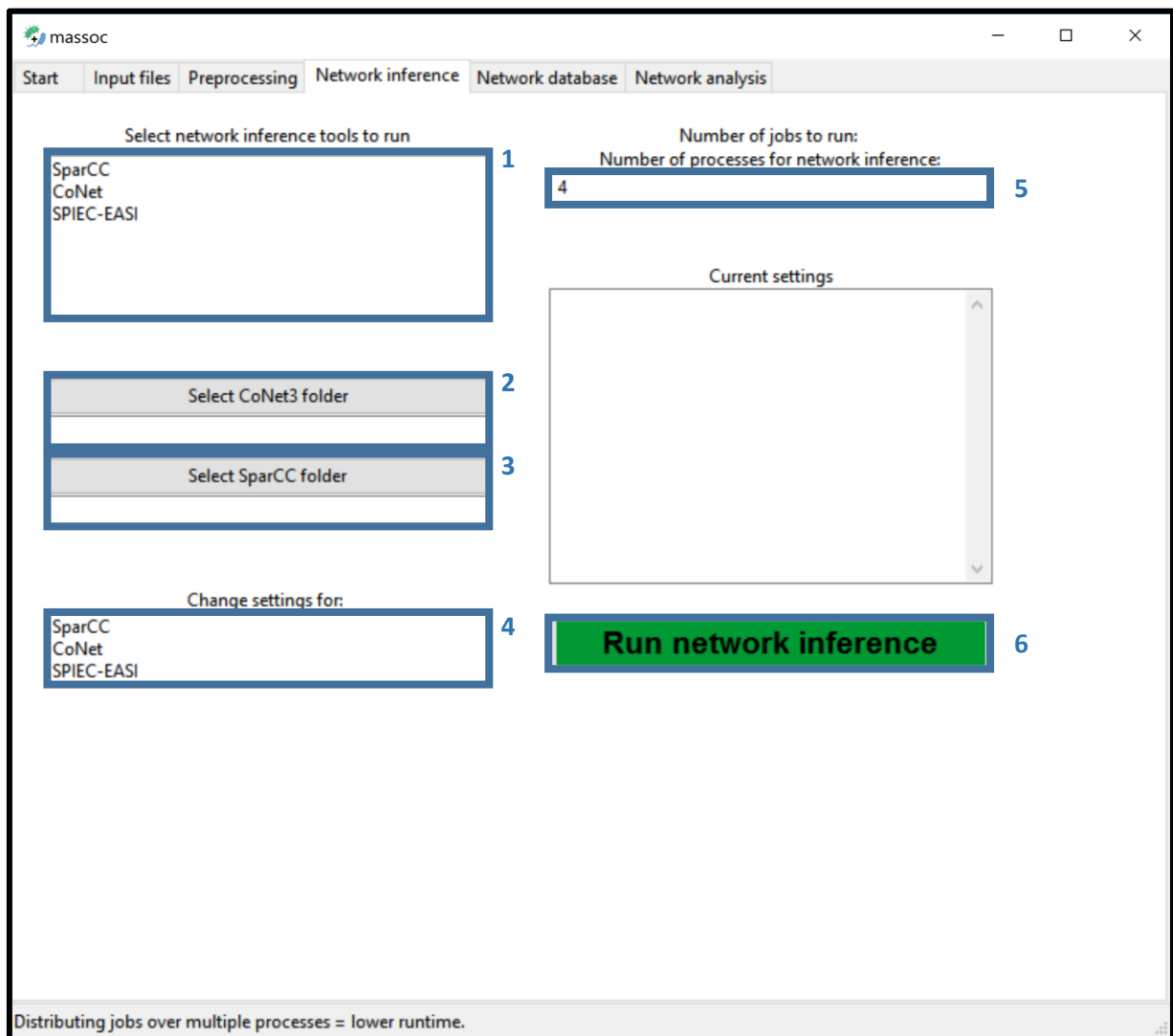
massoc neo4j

This module starts up the Neo4j database and uploads any required files.

`massoc neo4j -a` **1**
Address of local Neo4j database.

`massoc neo4j -u` **2**
Username for accessing local Neo4j database. Setup through initial Neo4j configuration:
https://neo4j.com/docs/operationsmanual/current/installation/

`massoc neo4j -p` **3**
Password for accessing local Neo4j database.

`massoc neo4j -n` **4**
Location of Neo4j folder.

`massoc neo4j -add` **5**
Uploads an edge list to the Neo4j database. The edge list should contain column headers, as these will be used to create new nodes.

`massoc neo4j -o` **6**
Prefix for output graphml files.

`massoc neo4j -j` **7**
Default job; checks if there is an existing database, then uploads BIOM files that are listed in the *settings.json* file.

`massoc neo4j -j clear` **8**
Deletes all nodes and relationships in current database.

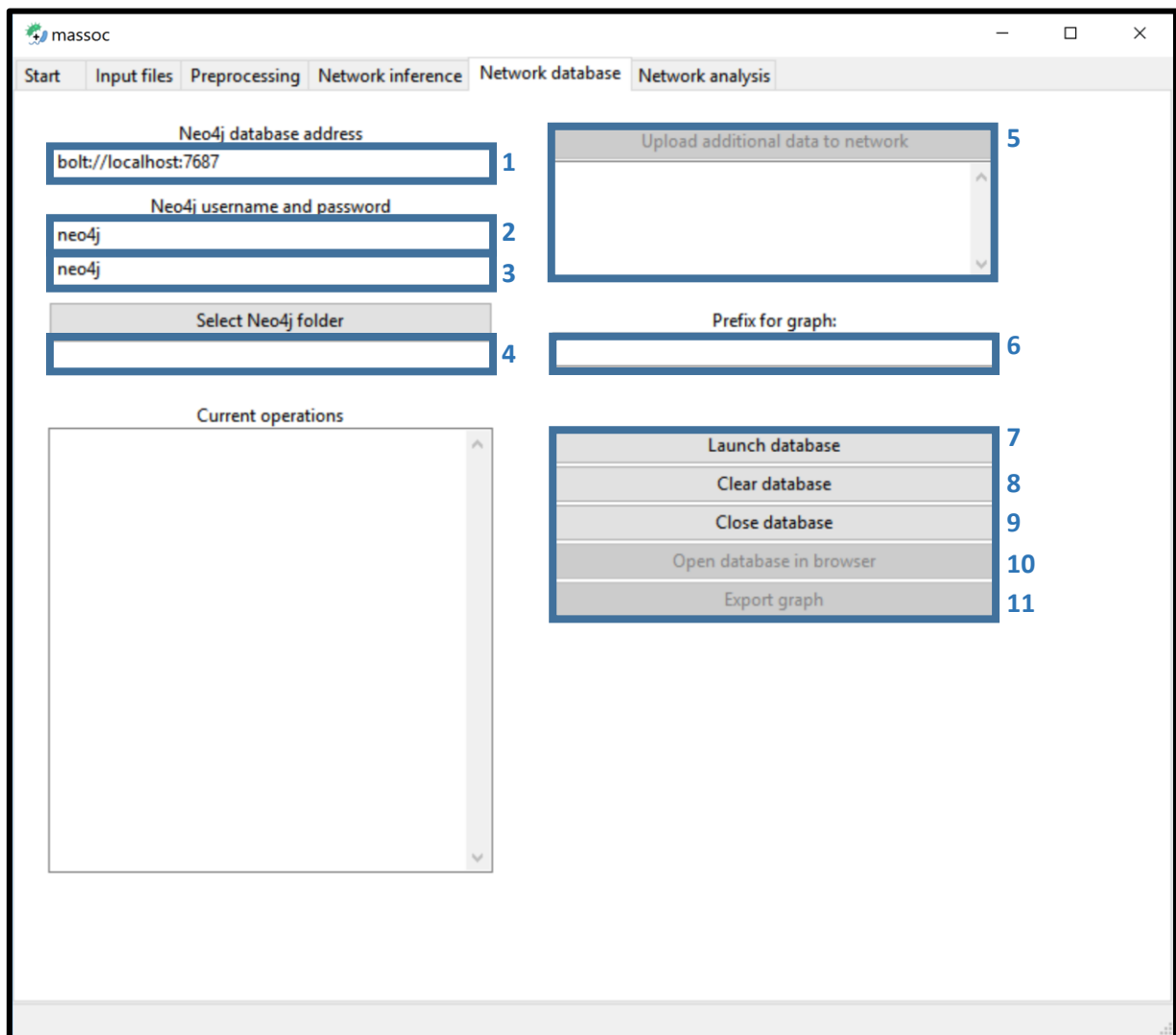`massoc neo4j -j quit` **9**
Safely shuts down current database Java process.

GUI-only **10**
Opens a browser tab with a database interface.

`massoc neo4j -j write` **11**
Writes all existing networks to a graphml file that can be imported in Cytoscape.

massoc netstats and *massoc metastats*

This module carries out analysis on the networks in the Neo4j database.

`massoc metastats –tax` **1**

Merges edges if they are between the same taxonomic groups.

`massoc metastats –w` **2**

Only merges edges if they have the same weight. Currently only compatible with weights of 1 and -1.

`massoc metastats –v` **3**

Associates taxon abundances to metadata. By default Spearman correlation for quantitative data, or hypergeometric test for qualitative data; no multiple testing correction applied. Use `–v all` if you want to compute associations for all metadata.

GUI-only **4**

Carries out `metastats` operations.

GUI-only **5**

Extracts an overview of the database; this overview is used to present parameters in box 3 and 7.

`massoc netstats –l` **6**

Type of logic operation to carry out. With '*Union'*, all edges are returned; '*Intersection'* only returns edges present in all selected networks, while '*Difference'* only returns edges that are present in one network only.

`massoc netstats –net` **7**

Set of networks to use for logic operations. For example, if there are 4 networks in the database, but only two are selected, the returned intersection will only be for edges present in these two networks. If nothing is specified, all networks are selected.

GUI-only **8**

Carries out `netstats` operations.

massoc

Start | Input files | Preprocessing | Network inference | Network database | Network analysis

Agglomerate edges to:

1
Species
Genus
Family
Order
Class
Phylum

During network agglomeration:

2
Take weight into account
Ignore weight

Associate taxa to:

3
Run database overview first

Run metadata operations **4**

Get database overview **5**

Perform operations:

6
None
Union
Intersection
Difference

Perform operations on:

7
Run database overview first

Run network operations **8**