Case study: soil data from the Arctic

In this case study, we are going to perform an analysis of data collected from arctic soils. This data can be found here: https://qiita.ucsd.edu/study/description/104

Chu H, Fierer N, Lauber CL, *et al.*: Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ Microbiol.* 2010; **12**(11): 2998–3006.

In the original BIOM file, the mapping file is not included. The BIOM file in the Github repository has already had the mapping file added:
https://github.com/ramellose/massoc/raw/master/data/demo.biom

The *massoc* CLI is built up from several modules. All are accessible through the *massoc* script. For information on the different modules, run:

```
massoc -h
```

The first step is to import data and specify some variables. These will be saved to a *settings.json* file in the specified location. The *settings.json* file will be used by downstream applications later on.

```
massoc input -biom ../demo.biom -fp ../test
-name test -min 10 -rar True -prev 10 -levels otu
```

massoc input specifies the module that is being used. The complete filename for the BIOM file can be specified after `-biom`; the filepath where all intermediate and the final output files will be stored, should be specified after `-fp`. `-name` is used as a prefix for stored files.

All other supplied variables control how the biom file is processed. In this case, we specify that samples should have at least 10 counts, otherwise they are filtered. Afterwards, they are rarefied to even depth through the `-rar True` parameter. To rarefy to another level, specify that level instead of True. The `-prev` parameter controls the minimum prevalence level; in this case, taxa with a prevalence below 10% are filtered. Finally, we are only going to construct networks at the lowest taxonomic level. Other levels could also be specified through the `-levels` parameter.

The next step is to infer networks. Select CoNet and make sure to specify the location of your CoNet3 folder. Here, we are just going to run CoNet with the settings specified in the CoNet.sh script included with *massoc*. If you write your own shell script, you can use the `-conet_bash` parameter. Because *massoc* will use the *settings.json* file written previously, you do not need to specify any other parameters.

```
massoc network -fp ../test -tools conet -conet ../CoNet3
```

If everything ran correctly, you can load the CoNet network into Cytoscape using the *conet_otu_test.txt* edge list. This list does not contain any additional metadata; the edge weight is converted to -1 and 1 so it is comparable to outputs from other network inference tools. If you want to do additional analysis and export the taxonomic information as well, you can export a rich GraphML from the Neo4j database. This demo assumes that you have already installed Neo4j Server on your machine and carried out an initial configuration. If not, please refer to https://neo4j.com/docs/operationsmanual/current/installation/.

You can start up the database and upload the BIOM and network files with the following command:

```
massoc neo4j -fp ../test -n ../neo4j –u neo4j -p test -a
bolt://localhost:7687
```

If you need to clear or quit the database, simply append with:

```
-j clear    -j quit
```

Note: it is possible that there is still a lingering Java process running if you previously opened the database, but did not close it safely. This will prevent *massoc* from uploading your data. If this happens, open your task manager and look for a process called “Java™ Platform SE Binary”. This process places a lock on the database to make sure it cannot be accidentally overwritten. Killing the process clears the lock so you can start a new database.

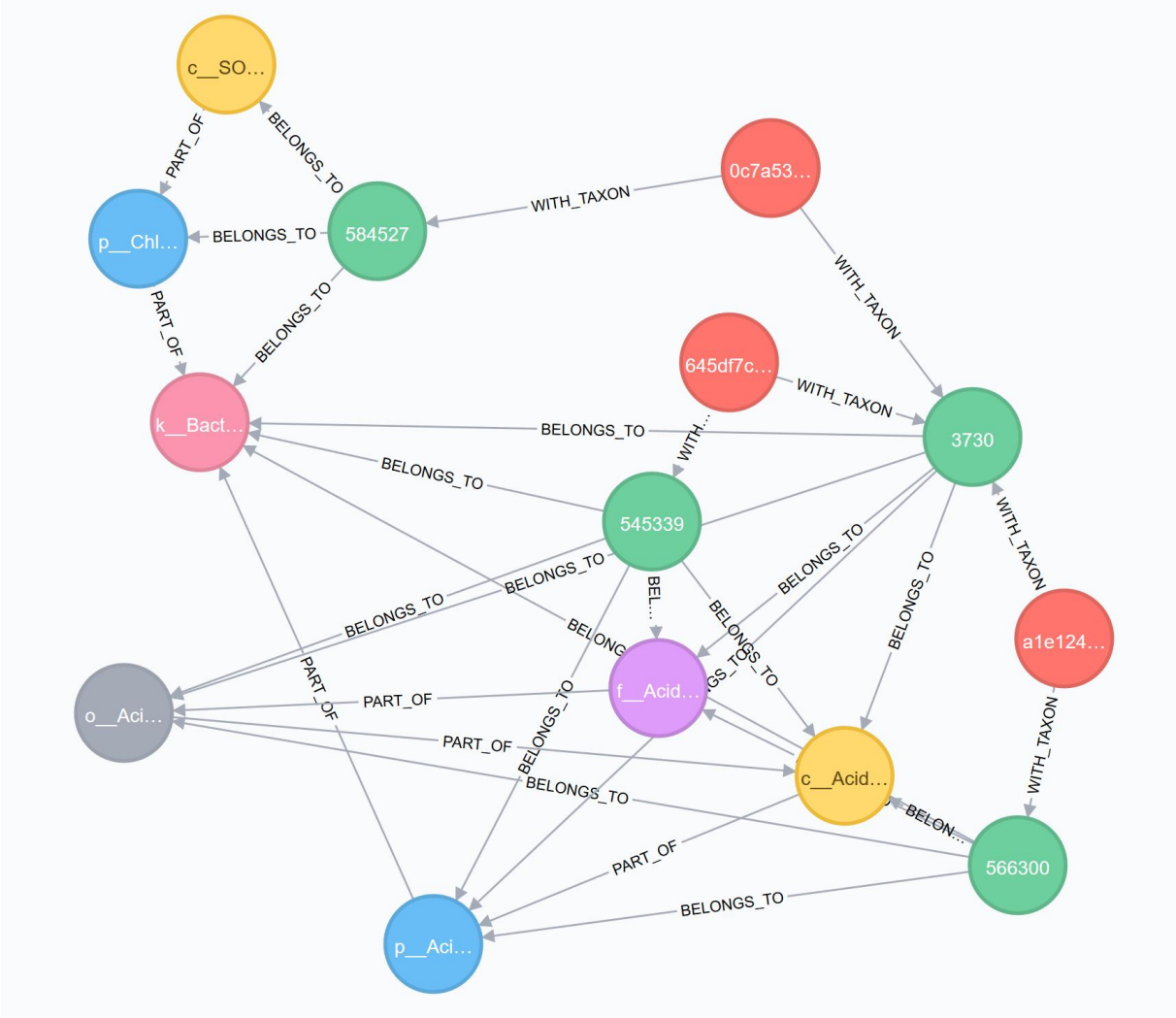
The graph database model used by *massoc* differs from most other microbial network inference tools. If you open the database in your browser, you can explore this database model in more detail.

The Neo4j database can be queried through the Cypher language. If we want to find a taxon and several associated nodes in the databse, we could use the following query in our local environment (by default, available through http://localhost:7474/browser/) :

```
MATCH p=(n)--(:Taxon)--(:Association)--(:Taxon {name: '3730'})--(b)
WHERE NOT n:Sample AND NOT b:Sample AND NOT n:Association AND NOT
b:Association RETURN p LIMIT 60
```

This query will show us other taxa that are connected to taxon 3730. By specifying that the links to those taxa should not be samples or association, we prevent the query from returning all other samples that these taxa were found in. The LIMIT specifies that we get no more than 60 nodes, otherwise the query can clog up the browser rather quickly.

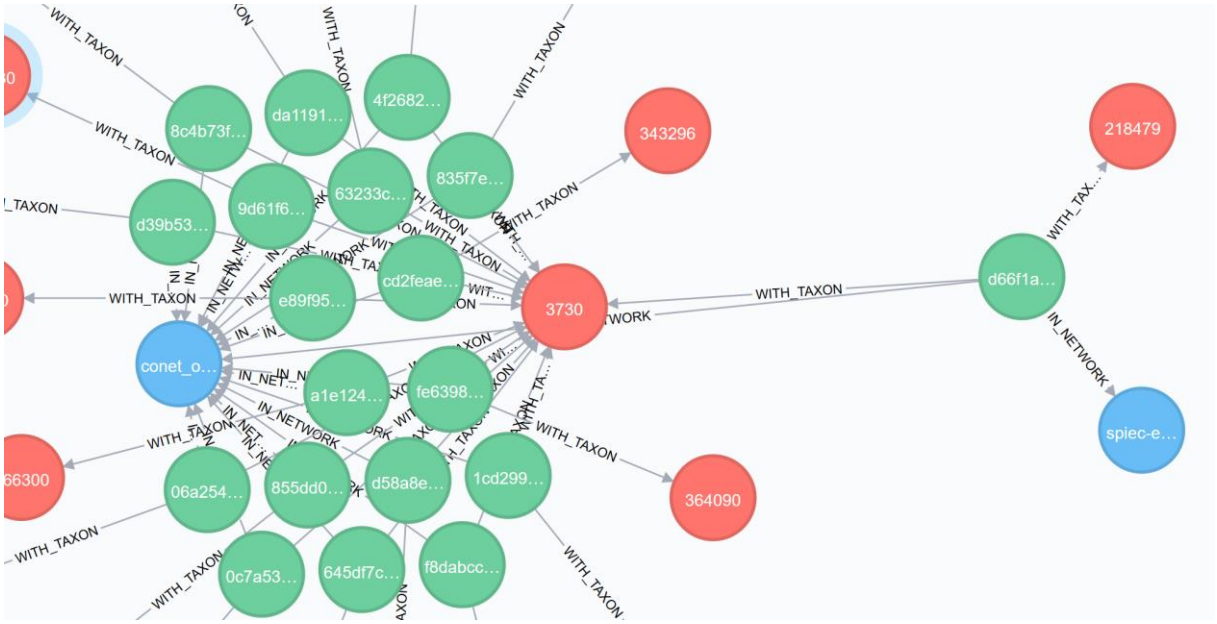
The screenshot below shows how the taxonomy and associations are stored: taxon 3730 has 3 associations (red nodes) with 3 other taxa (green nodes). Two of these taxa are Acidobacteria, while the other taxon belongs to the Chloracidobacteria.



In this case, it is not clear whether the associations were part of the SPIEC-EASI or CoNet networks. We can visualize this in more detail with the following query:

```
MATCH p=()--(:Association)--(:Taxon {name: '3730'}) RETURN p LIMIT 100
```

The screenshot shows that only one of the 19 associations was detected with both CoNet and SPIEC-EASI. This is the association with taxon 218479.



Internally, the Cypher language is used to create new nodes, find specific edges, or check whether a particular edges or node already exists. If you want to manipulate the database in more detail, you can do so through the browser or through a Neo4j driver. *massoc* uses the officially supported Neo4j Python driver, but there are drivers available for many languages.

While many of the operations *massoc* carries out in the analysis steps can also be carried out through a Neo4j driver, *massoc* provides a range utility functions that facilitate analysis of microbial networks. In the last step, we are going to extract subnetworks that only contain edges present in at least 2 networks.

To generate associations to pH, first run:

```
massoc metatastats -fp ../test -var PH
```

Then extract the intersection of SPIEC-EASI and CoNet networks:

```
massoc netstats -fp ../test -l intersection
```

This will generate a graphml file with the intersection of both networks. You can open this file in Cytoscape; it should contain the intersection network, with Spearman correlations of pH to taxon abundance as node properties.

In case there is an issue with the program, *massoc* will write logs to a file called ‘massoc.txt’ one directory above the location of *massoc*. The log file will be refreshed periodically after it reaches a size of 500 KB.