Navigating This Repo

Jim Brunner

August 17, 2017

1 Parent Folder

The parent folder contains all the python and bash scripts used in the project, as well as the python module of functions. The python module of function is `co_occ_funs.py`. The python scripts are:

- `adding_data.py` - Experiment to test network building with growing data set.
- `cluster_net.py` - Clustering for network that has been built.
- `co_occurrence.py` - Creates Networks.
- `examples.py` - Makes examples for a paper.
- `falsepm.py` - Validation experiments.
- `mc_speed.py` - Test time to build networks.
- `network_stats.py` - Compares various statistics of a network to the null model.
- `rand_samp.py` - Creates a "sample" that can be analyzed.
- `sample_analysis.py` - Analyses a sample, creates a table that can be added to a cytoscape network.

The bash scripts are

- `net_making.sh` - Runs `co_occurrence.py`, `cluster_net.py`, `network_stats.py`, and `falsepm.py`, for genus and species level. Creates a folder with the date and time: `MM_DD_HH_MM_networks`, and subfolders for taxonomic level and network creation method.
- `sample_tests.sh` - runs `rand_samp.py` and `sample_analysis.py`

For detailed code documentation, see `writeups/docs/code_docs.pdf`.
In addition, the parent folder contains the following text files:

- `count_by_type.txt` - Contains a count of the sample types in `merged2.txt`, and the holdouts from `08_14_14_57_networks`
- `merged_assignment.txt` - Small set of GOTTCHA output data (from HMP?)
- `merged2.txt` - Larger set of HMP data.

Also, it contains this file and the various auxiliary files used to create it.

2 pycache

This folder is created by python when a module is created (in this case `co_occ_funs.py` and can be ignored.

3 08_08_17_09_networks & 08_14_14_57_networks

These folders contain network files that can be put into cytoscape. The numbers are date and time in MM_DD_HH_MM format. In each folder there two subfolders corresponding to taxonomic level. In each of these, there is again two subfolders corresponding to the network building method. The bins folder contains networks built using the "binning" method, and the pears folder contains networks built using Pearson correlation.

All network files are .tsv text files, and can be read into a python environment using pandas read_csv('filename', sep = '\
For each network, there are four files, which differ in the end of the filename:

- adj.tsv - This is the adjacency matrix, used for any analysis you want to do in python, not for cytoscape
- list.tsv - This is a list of edges, and can be imported into cytoscape using Import Network From File.
- node_data.tsv - This is a node data table and can be added to the existing cytoscape network using Import Table From File.
- held.tsv - This is a list of the data columns that were not used in creating the network.

The procedure for importing a network into cytoscape is then:

1. Click "Import Network From File" (a down arrow pointing at a cartoon of a 3 node network) and click a the network file ending in list.tsv. Make sure there isn't an index column in the import box that comes up (if there is, click the dropdown above the column and select "not imported").
2. Click "Import Table From File" (a down arrow pointing at a cartoon of a spreadsheet) and select the file ending in node_data.tsv*that has the same beginning of the file name as the* list.tsv *file* . In the resulting import box, at the top dropdown, change the selection next to "Where To Import Table Data" to "Selected Networks Only", and select the correct network. Make sure the first column (with a key above it) is a list of taxa names. If there is an index list, don't import it and put the key above the list of taxa names.
3. If you want colors, go to "style", find the color attribute you want (node fill, for example), select "passthrough mapping", and select a data column containing the word "color".

3.1 pears

In the pears folder there is a couple of extra things. These are: a stats.txt file that contains the results of network_stats.py, and folder called validation_plots. This folder contains the results of falsepm.py.

4 Old Scripts

This folder contains some old code that might not work anymore because of changes to co_occ_funs.py.

- add_gender.py - added the gender column to node data in 08_08_17_09_networks
- cleanup.py - cleaned up network files for easier import into cytoscape. No longer needed, this cleanup has been incorporated into co_occurrence.py.
- color_key.py - creates a plot of colors. Can be used to create a key for the coloring of nodes that is in the node data table.

5 RCode

Pavel's folder - contains metadata for data I don't have

6 Stat Figs

Contains figures made. The files in the folder are made from various experiments using merged_assignment.txt data, while the folder merged2 used data from merged2.txt.

7 Test Samples

This folder contains "samples" that can be put through `sample_analysis.py` to make examples.

8 Writeups

This contains all the documentation. The files are picturs (`.png` files) and one bibliography file (`.bib` file) which is used to add references in a LaTeX document. Documentation was created in LaTeX, meaning numerous auxiliary files are created. The `.tex` writes the document, and the `.pdf` file is the what should be read. The other files can be ignored. To edit a document, open the `.tex` file and write away. Must have LaTeX installed on the computer, and I like TeXstudio as an editor.

8.1 docs

Code documentation.

8.2 notes

All notes from the summer related directly to this project. This document contains some ideas that I haven't explored.

8.3 paper

The paper on the project.

8.4 presentation

Presentation for the project (also created in LaTeXusing beamer).