Networks Tutorial

Summarise a field of study in genetics by building a literature network

The objective of this exercise is to leverage a systematic approach to data analysis to rapidly gain an overview of a complex field of study. As a bioinformatician, you may find yourself working on projects or for people with whom you share very little domain knowledge. Here is the fastest way to change that!

In addition to learning the top-down structure of a field of study, literature networks can also be incredibly useful for discovering unexplored potential interactions between genes or proteins. For example, if there are several studies looking at the relationship between gene A and gene B, and several studies focused on the interaction between gene B and gene C, but none directly connecting gene A with gene C, perhaps we ought to question whether there is a missing link here? Especially if these studies are all being done at the same time. Maybe it is a result of poor communication between groups, or of competing groups acting secretively so their big finds don't get scooped by someone else. This can be a great way to discover potential research ideas!

There are many ways to build literature networks, and they don't have to focus on genetics. However, here we will look at using co-cited gene symbols from the human genome that have been mentioned in abstracts from a field of study. We will combine this with some field-specific terminology, to hopefully gain a good overview of our area of interest!

N.B. *All python scripts used here have been written for the specific educational purpose of this exercise. Many other fully-fledged network building tools and methods exist, it would be a good idea to see for yourself!*

For example, have a look at:

http://www.citnetexplorer.nl/

https://www.sciencedirect.com/science/article/abs/pii/S1751157715301966

Step 1: Find a reference list of gene symbols

1. Go to the NCBI Taxonomy website
2. Search 'Homo sapiens' select main result
3. On right hand side select 'gene' (click on 224k entries link)
4. On left hand side of results see filtering options
5. Check 'annotated genes', 'protein-coding' to get it down to a sensible number
6. Then go to the 'Send to' link above the table results & send to file
7. Navigate to the file location with the terminal and:
    a. In Linux: `awk '{print $6}' gene_result.txt > gene_list.txt`
    b. In Windows: Open excel and copy the 6th (Symbol) column, create a new sheet, paste it in there & save it as a text file with no header.

Now we have a list of human gene symbols we can use in our literature search!

(this process could also be used if we were researching other model organisms)

Step 2: Download a collection of literature abstracts

1. Go to the NCBI PubMed website
2. Search for a disease name, or any well studied area of genetics
3. Make sure there are enough results (5k+ abstracts)
    a. If there are loads (100k+), consider filtering by recent 5/10 years
4. Go to 'Send to' -> select abstracts (text) and download

Step 3: Build a custom term list

Literature networks for gene interactions are an interesting overview of the groups of gene that are associated by their areas of study. This can be enhanced by including other non-gene words in the search too, such as field-specific terminology.

There are various ways to go about this. It might be worth using several of these together for best results!

The first way and most basic, it to simply go to the Wikipedia Page of your chosen disease or genetic area of study and look for the blue words! Most of the relevant associated areas of study will be linked to their own page on Wikipedia via their primary keyword/term. Gather as many of these (20+ key terms would be good), into a list. Then simply add this list onto the end of your gene list.

The script we will use to build the network files accepts both single words and phrases with spaces in them as entries in the terminology list.

If you were looking for a quicker way to pull out key terms from a very long article, https://tagcrowd.com/, select 'web page URL' and enter the Wikipedia page. This won't generally work as well as your manual curation though.

Another way is to use a similar frequency decomposition to tag-crowd, but one you can manually control. For this purpose, we have included a python script with full comments describing its function. [**kernelize.py**]

In Linux:

```
Python3 kernelise.py pubmed_results.txt 300 5
```

Where 'pubmed_results.txt' are the collection of downloaded abstracts. 300 is the minimum frequency of occurrence you wish to see, and 5 is the minimum word length you wish to see in the output.

In Windows or with a GUI in Linux:

Using your choice of python script editor, open the script, make sure your editor's file location is set to the place you have saved all your scripts/files for this exercise, and run the script from here. You may directly edit the file names inside the script if you wish to.

The method summary can be described like this:

1. Take a set of abstracts which you aren't interested in (a totally unrelated field of study). In this case we have included abstracts on paediatric anaesthesia.
2. Frequency-decompose the text. This means finding the occurrence rates of every word in the text.
3. Take your main set of abstracts, and repeat step 2.

4. Filter the words in the main abstract by the irrelevant words in the unrelated abstract frequency list which occur more than *x* times.
5. Print a sorted list of words by frequency which are unique to the abstracts of interest, which are longer than *p* characters, and which are higher than *x* in frequency.

Be sure to read through all the comments in the python script to gain a better understanding of what you are doing here.

You may then scan this list of field-specific words to see if there are any interesting terms to include in your network. It can also be quite common to get place names or university names in this list. Whilst we might consider excluding these, it might not be a bad idea to include them. For example, the study of a disease via a subset of genes or other terms might localise around certain institutions or countries – which is itself informative about the nature of the field of study. You may already see some gene names in this list too, however these should already be included in your symbol list.

Another thing to consider is the difference in frequency between the terms you include. Some very general terms might simply correlate with everything and not be very informative. Looking for very specific terms with medium frequency levels can result in a network with a less cluttered layout. Too many super-hubs can make interpretation more difficult!

Step 4: Add Discrete Annotations

This is quite a quick step, but an important one. In the final network we will need a way to distinguish between the systematic biology words (the genes), and the field specific terminology.

Before doing anything else, open your gene list (now with added terminology) in excel or equivalent. (There are a million other ways to do this also). And simply copy-paste a '1' in the cells adjacent to the genes, and a '2' into the cells adjacent to the terminology. If you have chosen a very large field of study, or simply have a lot of terminology in there, you could choose to add more categories to the terminology. For example, which could take all anatomy-related terms and give them a '3', and all cell biology related terms, give them a '4', and all place/university names and give them a '5' etc.

Make sure you save this file as a 'tsv' file. This is a flat text utf-8 tabular format with no column or row names and tabs as delimiters.

Step 5: Build the Network Files for Cytoscape

Here we are going to take the term list you have created in step 4, and use it search the main abstract list for network connections.

After this step we will have *TWO* files to import to Cytoscape.

One file will be a three-column list of node attributes. This will include the name of the gene symbol or term, the discrete annotation you gave it, and the frequency of its occurrence in the network.

The second file will be another three-column table; however, this will be the network file. The first column will be a node, the second column will be the number of times they were co-referenced, and the third column will be the node to which the first is connected. This is the file Cytoscape will use to build the network visualisation.

Here we will use a second (included) custom python script to perform this operation. [**build_network.py**].

In Linux: `Python3 build_network.py pubmed_result.txt gene_list.txt`

This will output: `Attribute_file.txt` & `Network_file.txt`

Otherwise open in a GUI editor and play with the filenames and filters at the top of the script!

The method summary for this script is as follows:

1. Read and frequency decompose the abstracts file.
2. Read the gene list file
3. Iterate through each abstract in the abstracts file and decompose the frequencies individually
4. Filter each per-abstract frequency set for terms defined in the gene_list.txt document
5. Create network connections between all filtered terms remaining from each abstract
6. Add these network connections to a single network data structure
7. Print out the network data structure
8. Print out the node attributes, which includes all the terms in your input list which were successfully found in the abstracts, and the frequency of their occurrence.

Be sure to read through all the comments in the python script to gain a better understanding of what you are doing here.

Step 6: Import and use Cytoscape to Visualise your Results!

Open up Cytoscape.

Although primarily a visualisation tool, Cytoscape can be very useful for filtering networks and running standardised metrics on their structure.

The first thing to do is File -> Import -> Network from file.

If you have followed the previous instructions, there shouldn't need to be any changes to the default importing parameter at this stage.

You will hopefully see a big indistinct blob of nodes on your screen.

Before doing anything else to it, we want to import the attributes file.

Go to Import -> Table from file.

Again, Cytoscape should accurately identify which columns in this table are the keys for the nodes, and you shouldn't need to change any of the default loading options at this point.

There are now two core components to making a good network visualisation, these are **layout** and **visual attribute mapping.**

Layout

Layout simply means the way in which the nodes are arranged in the space, in relation to each other. It is quite easy to experiment with layouts by selecting the pre-sets in Cytoscape from the Layout drop-down menu. If there is a layout which seems to work for you, but the nodes are a little too far apart, or too close, you can also customise the parameters of these layout generation functions by going into Layout -> Settings and selecting the layout whose parameters you wish to alter.
WARNING: If you have ended up with 5000+ edges and 2000+ nodes in your network, some of the layouts will be exceptionally slow execute, and may even crash the software. It is recommended in this case that you filter your network first. (Step 7)

Visual Attribute Mapping

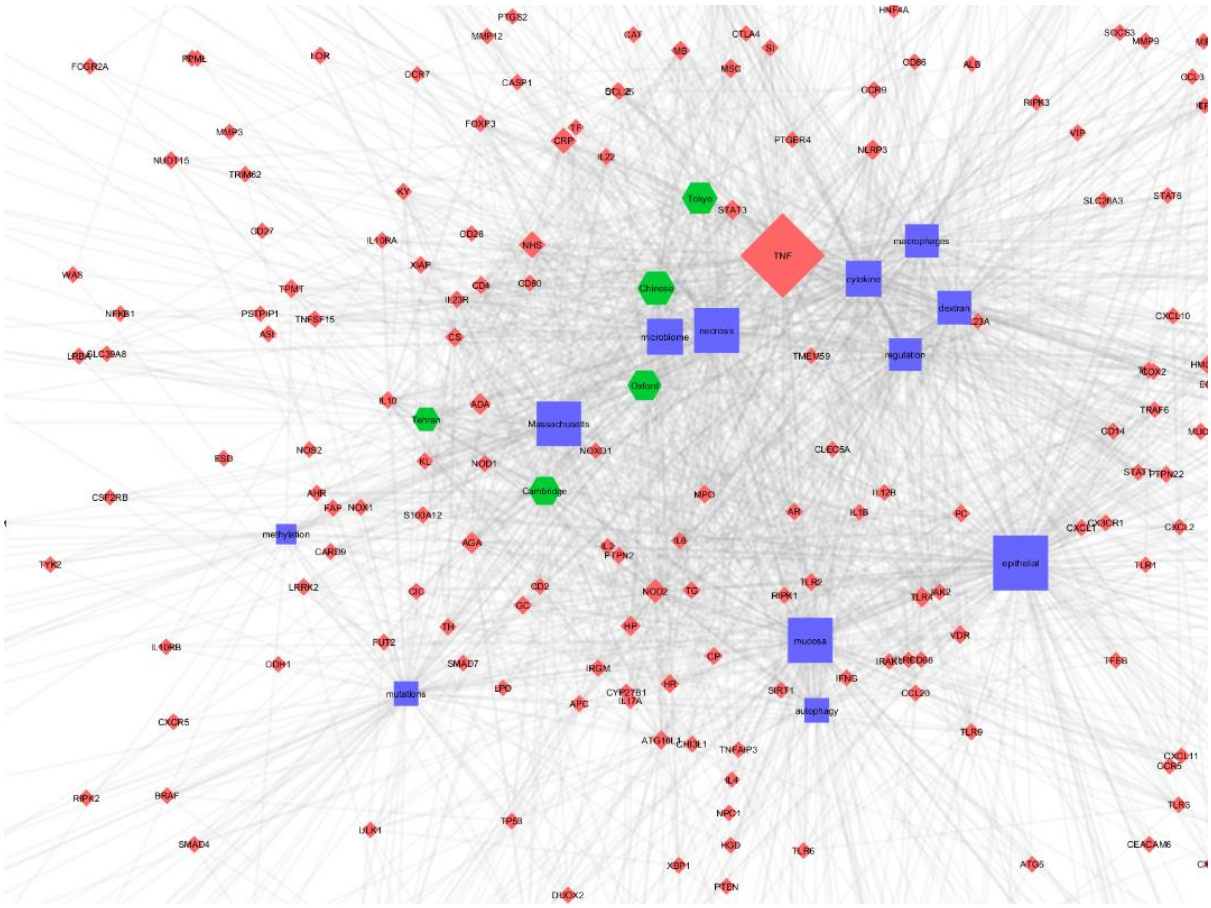
Near the top left of the screen, above your network name, you should see a few tabs. ‘Network’, ‘Style’ and ‘Select’. The ‘Style’ tab is where we will be editing the network visualisation.

If you select the style tab, first note the new tabs which appear the bottom of this panel. These allow you to which between editing the styles for nodes or network edges.

There are two types of mapping you can use, discrete and continuous. Discrete mapping is good for categorical values, such as those you created for the gene/terminology list earlier. By selecting the shape option, for example, then column -> Annotation_type, and mapping type -> discrete, we can assign a specific node shape to the nodes based on their type. Repeat this step with the node colours, and you now have a convenient way to distinguish between the genes and the field terminology in your network.

Next, we could try mapping the frequency of a node to the size parameter, this way the most common nodes in the network show up as being more prominent. Click through all the options and see what you can map to what, and what the results are! Feel free to experiment and create interesting designs.

In this not-particularly-informative example image below, you can see that the place names have been mapped to green octagons, the biological terms have been mapped to blue squares, and the gene names have been mapped to red diamonds. The edges have been given a width and transparency based on the number of connections they represent, and the network has been organised by the ‘Edge Weighted Spring Embedded Layout’ -> using connections as an edge weight.



Step 7: Filtering the Network

Return to the three tabs where you selected the 'Style' option. Now select the 'Select' tab. Here you can define and execute filters on your network.

When you use a filter and click 'apply', all the nodes which pass the filter will be selected in the network. You can now choose either to delete them (hit delete), or to create a new network using just these nodes and edges. To create a new filtered network, after clicking apply, select the 'New network from selection' button from the toolbar. It looks like two overlapping page outlines. You will now have another network to work on separately, which might work better with your choice of layout.

There are two types of filter of interest here. The degree filter and the column filter. The degree filter allows you select nodes above/below a certain number of edges. The column filter allows you select nodes based on the attributes you provided when you created the network. For example, you might want to remove all nodes with two or fewer edges. Or you might want to remove all nodes with a frequency of 5 or fewer. Or perhaps you could use the clustering co-efficient metric to select only density clustered nodes in the network: this might make their underlying relationships clearer! Feel free to play around with the filters and find something that works, but make sure you understand what is going on if you use an interesting network metric. (Google it!)

Step 8: Continuing from here

The basic analysis loop it is recommended to follow from this point on is as follow:

1. Create a clear network visualisation
2. Assess whether your choice of terminology was wise (do any super-hubs need to be deleted, or are any terms redundant or unhelpful?)
3. Assess whether your choice of field of study was a good one for genetic network analysis
4. Are there any filtering options you can use in Cytoscape to improve the clarity of the network?
5. Re-run the network generation script with either
    a. New terminology
    b. Difference search results
    c. More, or fewer abstracts from the same search
    d. Etc.
6. Repeat!