Worksheet 2

Construction and analysis of a human protein interaction network

Introduction to Profile Areas in Bioinformatics: Protein Interaction Networks

Florian Klimm

Winterterm 2021/22

After the analysis of a small protein interaction network (PIN) of *Saccharomyces cerevisiae* in the seminar project, we will now investigate a larger PIN for *Homo sapiens*. Furthermore, we will study small-world networks.

Deliverables:

- **Report: 15.01.2022, 18:00**
- **Code on GitHub: 15.01.2022, 18:00**
- **Reviews: 18.0.1.2022, 18:00**

Exercise 2.1 Accessing Protein Interaction Data

There exist many different protein interaction databases that store information about the pairwise interaction between protein. For this project, we use the Biological General Repository for Interaction Datasets (BioGRID; see [1]). It is available under `https://thebiogrid.org/`.

BioGRID stores information about experimentally measured pairwise interactions between proteins in ~ 80 different species.

Subexercise 2.1.1 Downloading Protein Interaction Data

Download the file `BIOGRID-ORGANISM-4.4.205.tab3.zip` from the database and unpack the zip file. You now have 79 files, each contains the protein interaction information of a different organism. Sort the files by size; which organisms have the most interaction data available?

Subexercise 2.1.2 Investigate Protein Interaction Data

Before we create a program that reads these files to create a PPIN, we want to investigate the files and their structure. The files are in a tab-separated format, which is very common for data files. It is basically a large table with the first line as header that gives a brief description about the contents contained in each column. Each of the following lines contains one entry, i.e., an interaction between two proteins. Each field value of an entry is separated from the next by a tab character. Open the file `BIOGRID-ORGANISM-Human_Herpesvirus_6B-4.4.205.tab3.txt` with OpenOffice Calc and with a Text Editor.

Question: How many interactions are in this file?

Each of these files contains a lot of information on the interactions. Most of these informations are not important for our project. Have a look at the description of this file format under `https://wiki.thebiogrid.org/doku.php/biogrid_tab_version_3.0`.

Question: The information of which columns are necessary to create a PPIN? There are different choices possible. Read the column definitions to identify which columns are best suited.

Have a look at one of the larger files. Is it still feasible to open them in OpenOffice Calc or with a Text Editor?

Exercise 2.2 Construct PPINs from BioGRID data

Now we will create PPINs from these files. First, we create a small PPIN manually to get an understanding of what our program is supposed to do. Second, we write a program that constructs a PPIN from a file.

Subexercise 2.2.1 Create a PPIN manually

To get an understanding of what the program is supposed to do automated, we do it first manually ourselves. Open the file `BIOGRID-ORGANISM-Human_Herpesvirus_6B-4.4.205.tab3.txt`. Draw a PPIN with pen and paper.

Question: How many nodes and edges are in this network? What is the size of the largest connected component?

Subexercise 2.2.2 Write a program to Construct PPINs from a BioGRID file

Write a program that reads the file `BIOGRID-ORGANISM-Human_Herpesvirus_6B-4.4.205.tab3.txt` and creates a NetworkX graph from it.

Subexercise 2.2.3 Use NetworkX to Illustrate Graphs

Create a new program that draws the PPIN of Herpesvirus 6B by using NetworkX.

Question: Is this network the same as the one you drew? If not, check for errors!

Subexercise 2.2.4 Create PPINs for other Organisms

Thus far, we wrote a program that constructs a PPIN for a single organism. Create a new version of your program that takes a file name as an input and so is able to create a PIN of any of the 79 organisms.

Subexercise 2.2.5 Cleaning PPINs

Adapt you program such that it removes parallel edges (multiple edges connecting the same pair of nodes) and self-loops (edges connecting a node to itself).

Exercise 2.3 Analysis of the human protein interaction network

After defining a function that constructs protein interaction networks from BioGrid we will now analyse the network for *Homo Sapiens*. First, use your programme to construct the PIN for *Homo Sapiens*.

Subexercise 2.3.1 Largest component

Identify the largest component and what percentage of nodes are in it. For the remainder of this work, work only with the largest component.

Subexercise 2.3.2 Degree distribution

Compute the mean degree and illustrate the degree distribution. Does the degree distribution resemble the one of an Erdős–Renyi (ER) network?

Subexercise 2.3.3 Centrality measures

Identify the hubs in the network. For this, compute at least two different centrality measures. Give their proper definitions with references in the write-up. Discuss the top-ranked proteins for different centrality measures. Do you find the identical proteins with different measures? Illustrate the centralities of the proteins in a scatterplot and compute the correlation between the centralities. Discuss your findings.

Subexercise 2.3.4 Clustering coefficient

Compute the average clustering coefficient for the PIN. Compare the result with the values that you obtain for an ER network and a lattice network of same size (i.e., number of nodes) and similar density.

Exercise 2.4 Average path lengths

Next, we want to compute the average path lengths in the protein interaction network and compare it with ER and lattice networks. Fortunately, NETWORKX has a function called AVERAGE_SHORTEST_PATH_LENGTH for this. Unfortunately, computing the shortest paths between all pairs of nodes is very slow for large networks. To speed up the computation, we will use a common trick: subsampling, which means that we do not compute the average path length over all pairs of nodes but rather just a random sample (e.g., 500 pairs of nodes). Write a function that does this and compute the approximate average shortest path lengths for ER, Lattice, and PIN. Discuss your findings.

Exercise 2.5 Small world network

Subexercise 2.5.1 Reading Watts' and Strogatz' paper

Read the seminal paper by Duncan Watts and Stephen Strogatz [2]. Give a brief summary of their findings in your report.

Subexercise 2.5.2 Reproduce their results

Reproduce their findings from Fig. 2. For this, compute the characteristic path length $L(p)$ and the clustering coefficient $C(p)$ for the family of randomly rewired graphs described in their manuscript.

Subexercise 2.5.3 Optional: Reproduce their results for networks of the size of the PIN of *Homo Sapiens*

Reproduce their Fig. 2 results for a network that is of similar size and density as the human PIN (You may run this with less realisations to speed up the process). Compare the observed clustering and characteristic path length values to the one for the PIN. Are they comparable?

Exercise 2.6 Community detection in protein interaction networks

Subexercise 2.6.1 Compute community structure

NETWORKX has a modularity-maximisation for community detection available. Use this to compute community structure in the PIN of *Homo sapiens*. Illustrate the distribution of the community sizes.

Subexercise 2.6.2 Optional: Explore other community detection algorithms

It is known that standard greedy modularity maximisations are slow and can lead to suboptimal results. If you are curious, you may explore some other approaches (e.g., the LEIDEN algorithm). Furthermore, in many modularity-based community detection algorithms there exists a resolution parameter γ, which you can vary to obtain communities of different sizes, which might lead to biologically more interpretable results.

Subexercise 2.6.3 Gene ontology enrichment

We want to test the hypothesis that the obtained communities represent functional communities. For this, we may use tools for *Gene ontology enrichment*. One

widely-used tool is `http://geneontology.org/`. Explore whether the communities you computed have any ontology terms enriched.

Bibliography

[1] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. BIOGRID: A general repository for interaction datasets. *Nucleic Acids Research*, 34(suppl_1):D535–D539, 2006.

[2] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.