Comparative Genomics 2018

Practical 8: Interaction Networks

Assistants: Stefanie Friedrich, Miguel Castresana, Deniz Secilmis

All forms of plagiarism are forbidden, and if detected it will result in a lower grade.

PURPOSE

This practical is about functional association networks in your proteomes. You will identify and characterise networks, its topology, and compare the results about functional association gained with FunCoup and STRING. Further, you will compare the enriched pathways based on differentially expressed gene sets found in a yeast chromosome via two state of the art networks analysis tools (PathwAX, DAVID).

KEY QUESTIONS

Summarise shortly this practical.

Comparative network analysis using STRING

You will investigate the topology of your interactomes (i.e., the known and predicted networks of your genomes)

1. Write a Python script to calculate the average connectivity (i.e, number of links/interactions vs number of proteins) for each extracted interactome for your genomes, and present the results.
2. Further, plot the degree distribution as a scatter plot with log10-scaled axes for each interactome separately. The x-axis of each plot presents the node degree for each node protein in the network and the other axis the frequency of each node degree. Do you observe a power-law distribution in any of the interactomes? If yes, in which?

Experimental gene sets

3. Present the two gene sets with most overlap; for each, report how many genes of the set size overlap with the genes in your eukaryote.

Comparative network analysis using FunCoup and STRING

4. Compare your results gained with both tools based on the same input data (your two genes sets).
   a. How do these networks differ in terms of nodes, links, and hubs (the three nodes with the highest degree) for both of your gene sets?
   b. Which is the most common evidence type with a high confidence (>0.9)?
   c. What are the differences in terms of underlying data sources in the two databases? Explain them!

Enrichment analysis using PathwAX and DAVID

5. Which pathways are enriched ?
6. What differences do you observe in the results from PathwAX and DAVID?
7. Does the number of (input) genes matter for the results? If so, explain why!

MATERIAL & TOOLS

1. Your five proteomes, and experiments.txt in Comparative_Genomics/data/
2. Tools and databases
   a. Uniprot for downloading the Saccharomyces cerevisiae S228C yeast proteome webserver http://www.uniprot.org/proteomes/), and BLAST
   b. String (Mering et al., 2005) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539959/, webserver https://string-db.org for network analysis
   c. FunCoup (Ogris et al., 2017) https://academic.oup.com/nar/article/46/D1/D601/4641901, webserver http://funcoup.sbc.su.se/search/ for network analysis
   d. PathwAX (Ogris et al., 2016) https://academic.oup.com/nar/article/44/W1/W105/2499342, webserver http://pathwax.sbc.su.se/ for enriched pathways analysis
   e. DAVID (Huang et al., 2009) https://www.nature.com/articles/nprot.2008.211, webserver https://david.ncifcrf.gov/ for enriched pathways analysis

ACTIVITIES

Comparative network analysis using STRING

1. Download known and predicted protein-protein interaction (PPI) networks from STRING

   ```
   wget <link to file protein.links.v10.5.txt.gz>
   ```

   The entries of that file are *protein1* *protein2* *combined_score*
   Example *394.NGR_c00010* *394.NGR_c33930* *255*
   394 stands for the NCBI taxonomy identifier; *NGR_c00010 & NGR_c33930 are the gene symbols*
   Tip 1: The file is very large so do not copy or decompress it; however, you can check the first 100 lines in the terminal with

   ```
   gzip -cd protein.links.v10.5.txt.gz | head -100
   ```

2. You will need to extract the network belonging to all of your prokaryotic and eukaryotic organisms from this file `protein.links.v10.5.txt.gz`
   Tip 2: Python ETE3 package contains NCBI Taxa class which can be used to translate NCBI taxa identifiers to species names. Please double check the taxa id with https://string-db.org/ and https://www.ncbi.nlm.nih.gov/taxonomy if you extracted the correct species.

Tip 3: Use python gzip.open() in your Python script to open the file and to extract the links for your species

Creation of two experimental gene sets containing differentially expressed genes (DEGs)

Let us say experiments on *S.cerevisiae S228C* has been performed to investigate certain diseases, each comparing two conditions; significantly differentially expressed genes have been detected (one gene set per experiment). If any of these DEGs are present in your yeast chromosome, you want to extract and study the corresponding gene sets from *S.cerevisiae S228C* that overlap the most with genes in your eukaryotic chromosome.

1. Download the S.cerevisiae S228C yeast proteome from Uniprot (http://www.uniprot.org/proteomes/) and use BLAST to match your predicted genes against the S.cerevisiae S228C proteome.

```
module load blast
makeblastdb ... | blastp ...
```

2. Parse the blast results to extract the gene symbols (of only best hits) of genes present on your yeast chromosome
   Alternative: Use blastp with -outfmt 7 to get a list of predicted genes and gene symbols and -max_target_seqs 1 to get only the best hits.
3. Using the provided file *experiments.txt* containing gene sets (one set per row) of DEGs from the experiments in *S.cerevisiae S228C*, find two experimental gene sets that overlap the most with the genes in your eukaryotic chromosome. Save the two gene sets (from experiments.txt) with most overlap for further analysis.
   Tip 4: You will need the gene symbol, i.e. the first part of the third string (example: YJU6 in sp|P39529|YJU6_YEAST; sp | uniprot accession number| uniprot entry name) for your search of overlapping genes.

Comparative network analysis using FunCoup and STRING

1. Using your two experimental gene sets from the previous task, query FunCoup and STRING for sub-networks containing these genes.
   Tip 5: FunCoup works with a space-delimited list of gene names; STRING expects each gene name in new line for a query.
2. Use the same expansion depth or max number of interactors for searching both databases to get comparable results.

Pathway enrichment analysis using PathwAX and DAVID

Analyze your two experimental gene sets for an enrichment of pathways with both, PathwAX and DAVID (both using KEGG).
Tip 6: In DAVID you need to convert the gene symbols of your gene sets to *ENTREZ_GENE_ID* with the DAVID tool *GENE ID conversion*. If you do not obtain an enriched pathway, search on all organisms (not just *S.cerevisiae S228C*).