Practical 8: Interaction Networks

Group number: 2

Group members: Maximilian Senftleben, Zhong Hao Daryl Boey

Summary: In this practical, the functional association networks of our proteomes is investigated. First, the average connectivity is derived with a certain python code to observe the number of interactions vs. the number of genes. Second, a scatter-plot is derived from the node degrees and their frequencies, which gives a better insight into the degree distribution. Nevertheless, a power-distribution is not observed. Thirdly, the protein BLAST output was compared against an experiment dataset for overlapping genes. 2 gene sets with the highest number of overlaps were fed to FunCoup and STRING for comparative network analysis, as well as PathwAX and DAVID for gene enrichment studies, with the results compared.

Activities

Comparative network analysis using STRING

1.

Organism	TaxID NCBI	fileID
29.fa.txt: Saccharomyces cerevisiae	4932	1
44.fa.txt: Leuconostoc gelidum	927691	2
47.fa.txt: Neisseria meningitidis	122586	3
16.fa.txt: Rhodopirellula baltica	243090	4
20.fa.txt: Thermotoga maritima	243274	5

2. Extraction

gzip -cd prot… l grep “^nr\.” > file

for i in 927691 4932 122586 243090 243274; do gzip -cd ../../protlinks/protein.links.v10.5.txt.gz l grep "^$i\." > $i.txt; echo "$i done"; done

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

1.

makeblastdb ... l blastp …
blastp -db -outfmt 7 -max_target_seqs 1 > outblastp

Key Questions:

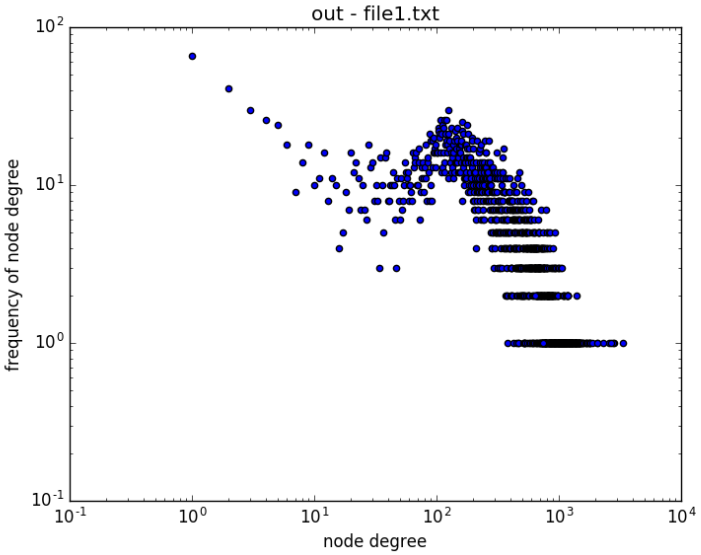
Comparative network analysis using STRING

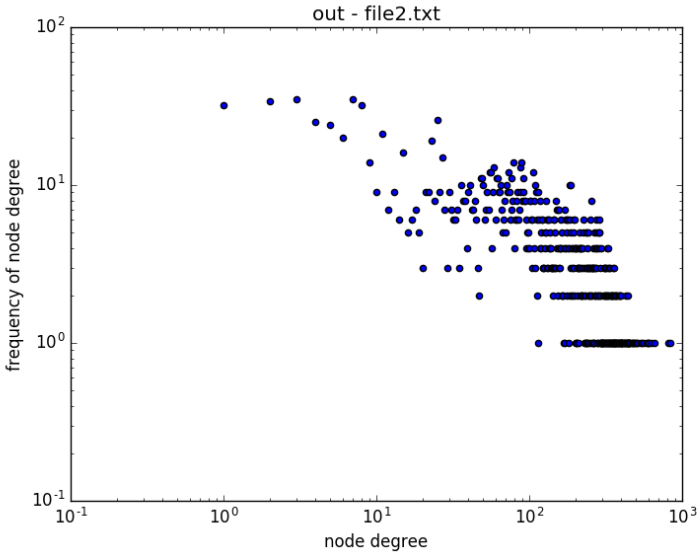
1.

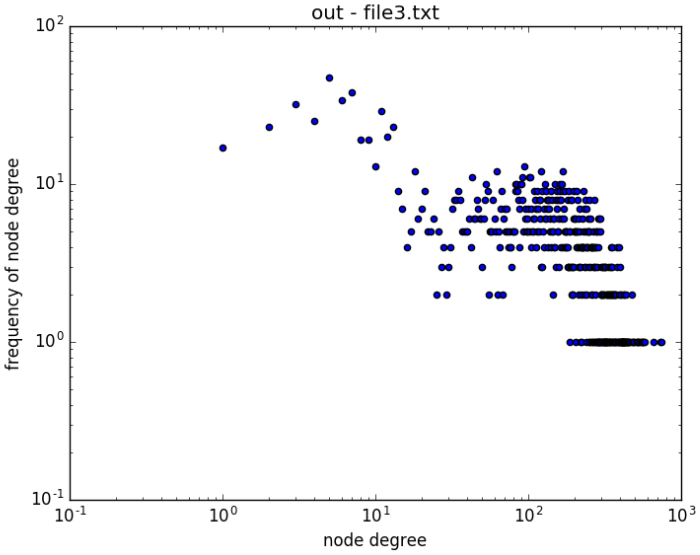
File: 1, Number of Genes: 6391, Number of Interactions: 2007134, Average connectivity: 314.0563
File: 2, Number of Genes: 1899, Number of Interactions: 238668, Average connectivity: 125.6809
File: 3, Number of Genes: 2049, Number of Interactions: 284938, Average connectivity: 139.0620
File: 4, Number of Genes: 7094, Number of Interactions: 951832, Average connectivity: 134.1742
File: 5, Number of Genes: 1850, Number of Interactions: 298816, Average connectivity: 161.5222

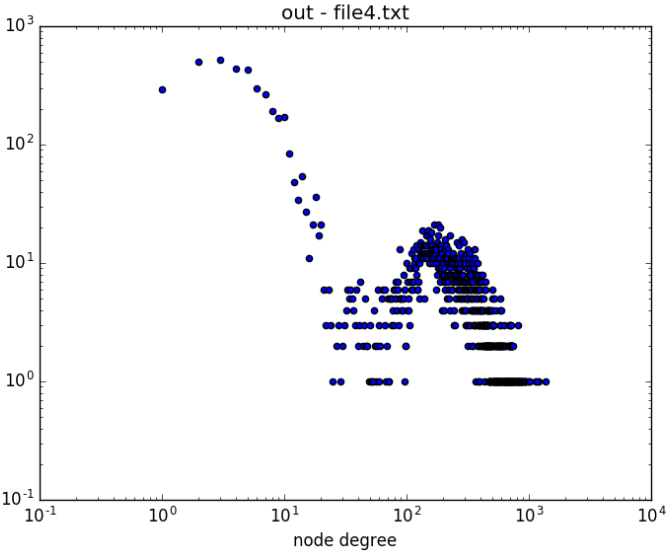
python script: key_questions_1and2.py – function quest1

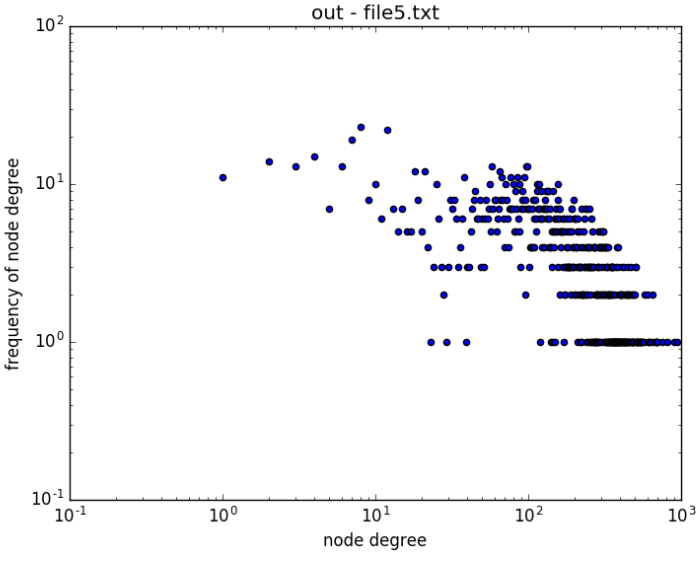
2.


out - file1.txt
frequency of node degree
node degree


out - file2.txt
frequency of node degree
node degree


out - file3.txt
frequency of node degree
node degree


out - file4.txt
node degree


out - file5.txt
frequency of node degree
node degree

A power-law distribution is when there is a quantitative relationship between two variables (in this case, the variables are the node degree and the frequency of the node degree). In all five interactomes, there is no power-law distribution existent.

python script: key_questions_1and2.py – function quest2

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 splP39529lYJU6_YEAST; sp l uniprot accession numberl uniprot entry name) for yoursearch of overlapping genes.

Wrote the script <exp_parse.py>, that produced output file <overlap_count>, which contains a count of the number of overlapping genes displayed with the gene set. There were a total of 3 sets which presented a total of 13 overlapping genes:

ID:26 count:13: ['ATP1', 'FAS2', 'FAS1', 'CEM1', 'BNA2', 'AMD2', 'LIP2', 'SEC59', 'ILV5', 'MTR3', 'FAS1', 'PRP28', 'RPB3', 'SNU13', 'COX13', 'HAM1', 'FAS2', 'PMS1', 'FAS2', 'ARO8', 'FAD1', 'OAR1', 'MRL1', 'PRP18', 'RRP4', 'FAS2', 'NTE1', 'URA1', 'DPB2', 'ALG2', 'DUT1', 'ADE1', 'HIS4', 'LIP2', 'LSM5', 'YMR1', 'ATP14', 'SUB2', 'PRI1', 'PMI40']

ID:38 count:13: ['DCD1', 'DUT1', 'YPC1', 'RRP4', 'OAR1', 'IFA38', 'DPB4', 'ARO8', 'HIS4', 'SGS1', 'PAN6', 'GAD1', 'ACO2', 'SPE2', 'BNA2', 'PGK1', 'SPC3', 'ARO8', 'SSS1', 'FCY1', 'VMA8', 'SUB2', 'HAM1', 'FAS2', 'PRO2', 'GDH2', 'GLN1', 'HIS5', 'LIP5', 'FAS1', 'LYS2', 'PAB1', 'FAS1', 'GAL10', 'NPY1', 'PFS2', 'PPX1', 'PRI1', 'ARO8', 'HAM1']

ID:52 count:13:['RAD4', 'PHA2', 'PRE9', 'DFR1', 'GDH2', 'ERG27', 'MNN9', 'RPL29', 'AGX1', 'PHS1', 'RAD28', 'GAL10', 'YET3', 'RAD59', 'SKI2', 'ALG1', 'RIO2', 'RPB8', 'GLN4', 'ARO8', 'PUT2', 'KIN28', 'TRP5', 'TAZ1', 'GUT1', 'OXA1', 'APN1', 'TSC13', 'BNA2', 'PGC1', 'THR4', 'SSL1', 'MRPS28', 'RPA14', 'SPC3', 'RNH1', 'UTP18', 'TFB5', 'MNN11', 'PRP28']

3 output files were also produced, one for each individual gene set (outID26, outID38, and outID52).

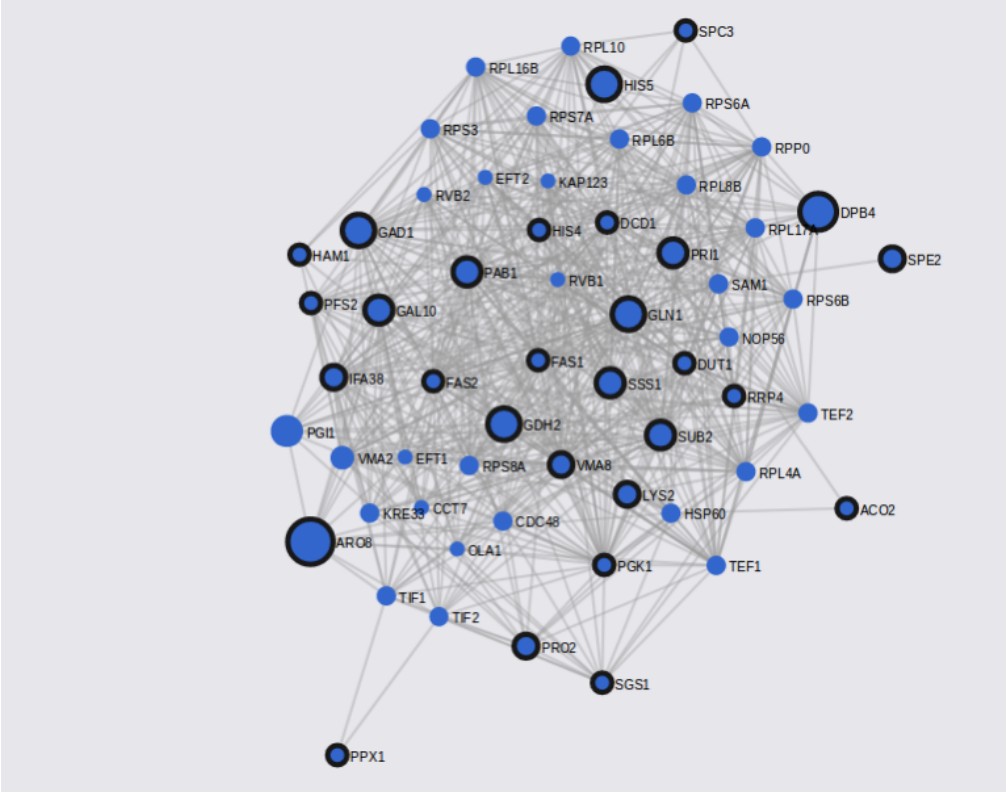
Comparative network analysis using FunCoup and STRING

4. Compare your results gained with both tools based on the same input data (your two genes sets).

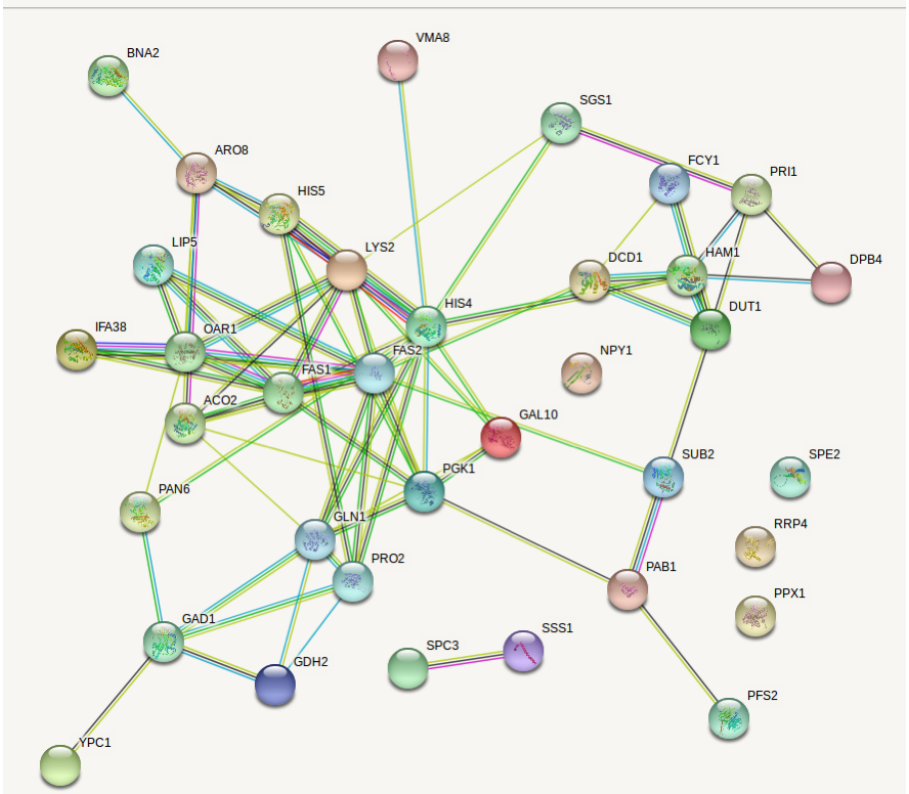
Decided to work on outID38 and outID52.

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?**

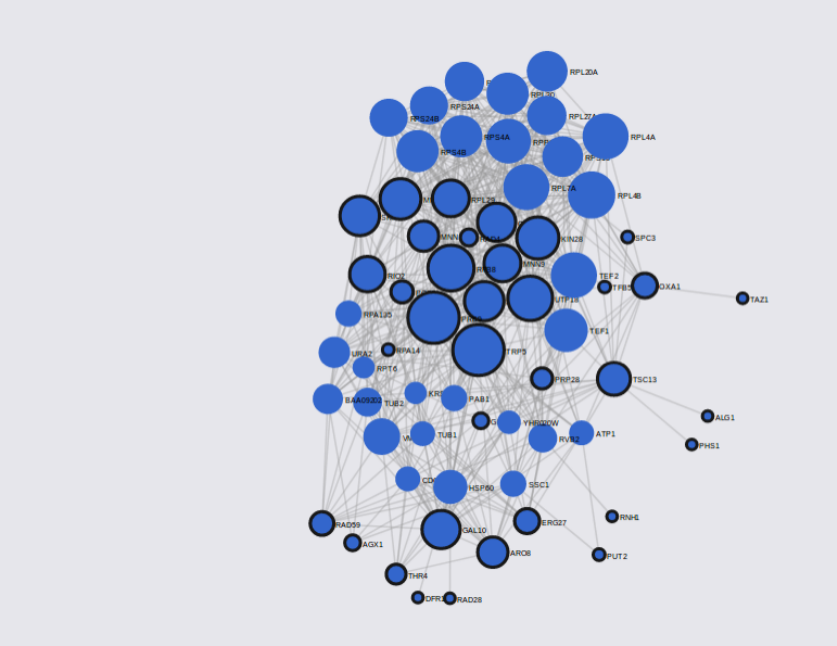
OutID38 FunCoup:



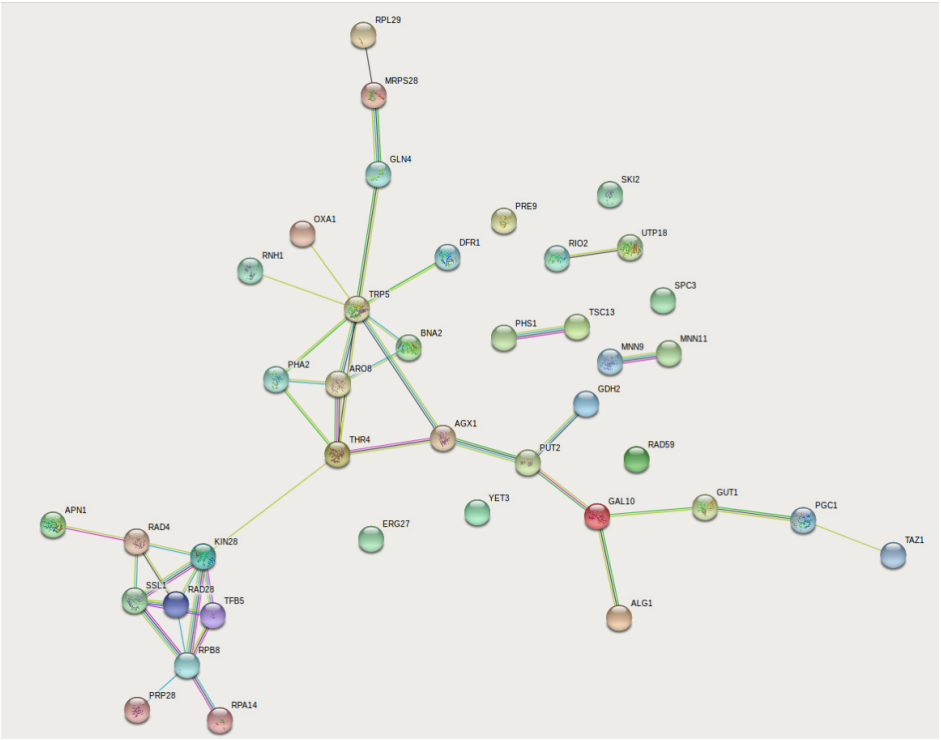
OutID38 STRING:



OutID52 Funcoup:



OutID52 STRING:



For OutID38, the 3 nodes of highest degrees were PAB1 (46), FAS2 (40), and VMA8 (40) in Funcoup, and FAS1, FAS2, HIS4 in STRING. For OutIDt52, the 3 nodes of highest degrees were TRP5 (32), PRE9 (32), RPB8 (28) in Funcoup, and TRP5, RPB8, and RAD28 in STRING. In general, all 4 networks produced vary quite significantly in terms of nodes, links, and hubs, with OutID38 producing a network with a denser network.

b. Which is the most common evidence type with a high confidence (>0.9)?

The protein interaction network (PIN) is the most common evidence type with high confidence.

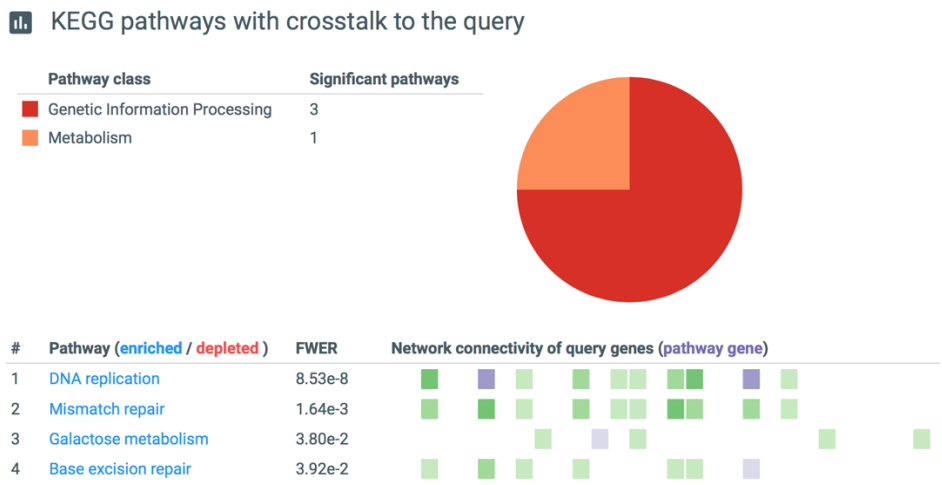
c. What are the differences in terms of underlying data sources in the two databases? Explain them!

Both STRING and FunCoup are integrated databases that serve to merge data from various sources, but differ in terms of of how the information is combined. FunCoup primarily uses the InParanoid database as a reference for orthologous information transfer and a large variety of datasets from various sources such as Gene Ontology (Ogris., et al. 2018), while STRING uses the KEGG database, and PubMed literature for protein association data (Mering., et al. 2005).

Enrichment analysis using PathwAX and DAVID

5. Which pathways are enriched ?

OutID38, PathwAX:



OutID38, DAVID:

Included DAVID output cluster as file outID38DAVID_Cluster. The most enriched pathways were metabolic pathways and catalytic activity.

OutID52, PathwAX:

Pathway class	Significant pathways
Genetic Information Processing	4
Metabolism	3

#	Pathway (enriched / depleted)	FWER	Network connectivity of query genes (pathway gene)
1	Basal transcription factors	1.54e-6	
2	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	1.54e-3	
3	Protein export	3.24e-3	
4	Various types of N-glycan biosynthesis	8.31e-3	
5	RNA polymerase	1.02e-2	
6	Steroid biosynthesis	1.66e-2	
7	Nucleotide excision repair	4.77e-2	

OutID52, DAVID:

Included DAVID output cluster as file outID52DAVID_Cluster. The most enriched pathways were metabolic pathways and biosynthesis.

6. What differences do you observe in the results from PathwAX and DAVID?

The enriched pathways given by the 2 programs differ significantly, for both gene sets. However both programs did state Metabolic Pathways as enriched.

7. Does the number of (input) genes matter for the results? If so, explain why!

Yes, depending on the number of links between the gene set and pathways (crosstalk) the pathway is determined to be enriched or depleted (Ogris., et al. 2016). Therefore a larger number of input genes improves the possibility of crosstalk, and such will affect the results.

References:

Christoph Ogris, Dimitri Guala, Mateusz Kaduk, Erik L L Sonnhammer; FunCoup 4: new species, data, and visualization, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D601–D607, https://doi.org/10.1093/nar/gkx1138

Von Mering C, Jensen LJ, Snel B, et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research*. 2005;33(Database Issue):D433-D437. doi:10.1093/nar/gki005.

Ogris C., Helleday T., Sonnhammer E.L.L.. PathwAX: a web server for network crosstalk based pathway annotation. Nucleic Acids Res. 2016;44:W105–W109.