Comparative Genomics 2018
Practical 8: Interaction Networks

Group 11
Fuqi Xu, Milda Valiukonyte, Shuhan Xu

Summary
In this practical, we extracted the interactomes of our five organisms from STRING and investigated the average connectivity and degree distributions of these networks. Then, we identified two sets of differentially expressed genes from the experiment.txt file which have the most number of overlaps with the genes in our eukaryote chromosome. Following this, we use FunCoup and STRING to visualize the subnetworks containing our two gene sets. Finally, we used PathwAX and DAVID to identify enriched pathways associated with our gene sets.

Comparative network analysis using STRING

We retrieved the STRING NCBI taxon-Id of the following organisms.

	NCBI taxid	**STRING NCBI taxon-Id**
Escherichia coli 536	362663	**362663**
Streptomyces coelicolor A3(2)	100226	**100226**
Saccharomyces cerevisiae	4932	**4932**
Rubrobacter xylanophilus DSM 9941	266117	**266117**
Halorhodospira halophila*	1053	**349124**

*since STRING does not have the species Spiribacter curvatus, we used Halorhodospira halophila as a replacement for this exercise

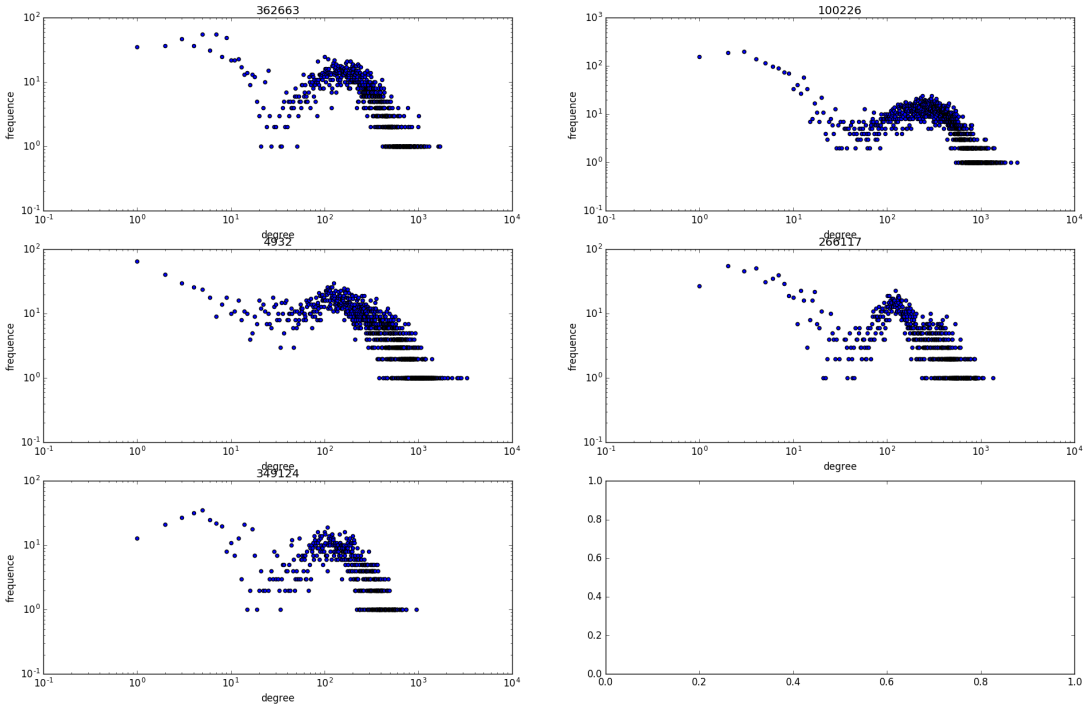
1.
Script for calculating average connectivity and plotting degree distribution:
See attachment *network.py*

Usage: python3 network.py <path to file protein.links.v10.5.txt.gz> <taxon-Id 1> <taxon-Id 2> <taxon-Id 3> <taxon-Id 4> <taxon-Id 5>

Description: The script takes the protein-protein interaction network file and taxon Ids as inputs and outputs the average connectivity of each interactome. In addition, it plots the degree distribution of each interactome in a log-log scaled scatter plot.

Taxon-Id	Average connectivity
362663	220.13376906318084
100226	298.3845057649955
4932	314.05632921295575
266117	203.62096516458934
349124	159.81704260651628

2.
Using the script from 1. we obtained the following degree distribution plots:



We do not observe power-law distribution in any of the interactomes as the dots in the scatter plots do not appear to fit a negatively-sloped straight line. Hence, these networks are not scale-free and may not have characteristics of scale-free networks such as resistance to random node removal and sensitivity to targeted hub removal.

Experimental Gene set

3.
Blast results of predicted proteome against S.cerevisiae S228C proteome:
See attachment *blastresults.txt*

Script for retrieving top two most overlapped gene sets:
See attachment parser.py

Usage: python3 <path to blast results (-outfmt 6)> <path to experiments file>

Description: The scripts find the top two experimental gene sets with the most number of overlaps with genes in our eukaryote chromosome

Output:

```
experiment no.: 53
number of overlaps: 5
gene set:
```

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

experiment no.: 38
number of overlaps: 5
gene set:
ARO8 BNA7 ARO1 GUT1 HIS5 PHA2 PRP38 COX8 FAS2 OAR1 MET8 EHD3 IMP1 DHH1 STE6
          PMS1 RPB8 RNH1 ILV5 PNP1 CAB4 MET14 GLN1 GLT1 ARG4 MAE1 RIB1
          HIS1 URM1 FAS1 SPE3 SNQ2 TFB1 PRO2 HIS6 CEM1 ILV3
```

4.
We tried to get comparable results by applying the following settings

FunCoup
Confidence threshold: 0.9
Expansion depth: 1
Nodes per expansion step: 30

STRING
Minimum required interaction score: 0.9
Max number of interactors to show: 1st Shell max interactors: 30

a.
Experiment number 53

	FunCoup	STRING
nodes	70	70
links	806	247
Hubs (top 3 nodes with highest degrees)	HSP60 (46 links) RPP0 (43 links) RVB2 (43 links) VMA2 (43 links) TEF1 (43 links) TUB2 (43 links)	PRE8 (16 links) SCL1 (16 links) PUP2 (16 links) PRE6 (16 links) PUP3 (16 links) PRE4 (16 links) RP026 (16 links) PRE3 (16 links) RPN11 (16 links) RPB5 (16 links) PRE1 (16 links) RPN5 (16 links) PRE7 (16 links) PRE2 (16 links) PRE10 (16 links) PRE5 (16 links) PRE9 (16 links) PUP1 (16 links) RPT6 (16 links)

Experiment number 38

	FunCoup	STRING
nodes	67	67
links	890	260
hubs	TIF1 (45 links) TIF2 (44 links) PAB1 (44 links) RPS3 (44 links)	RPB5 (22 links) RP026 (22 links) RPB4 (22 links) RPB8 (22 links) RPB10 (22 links)

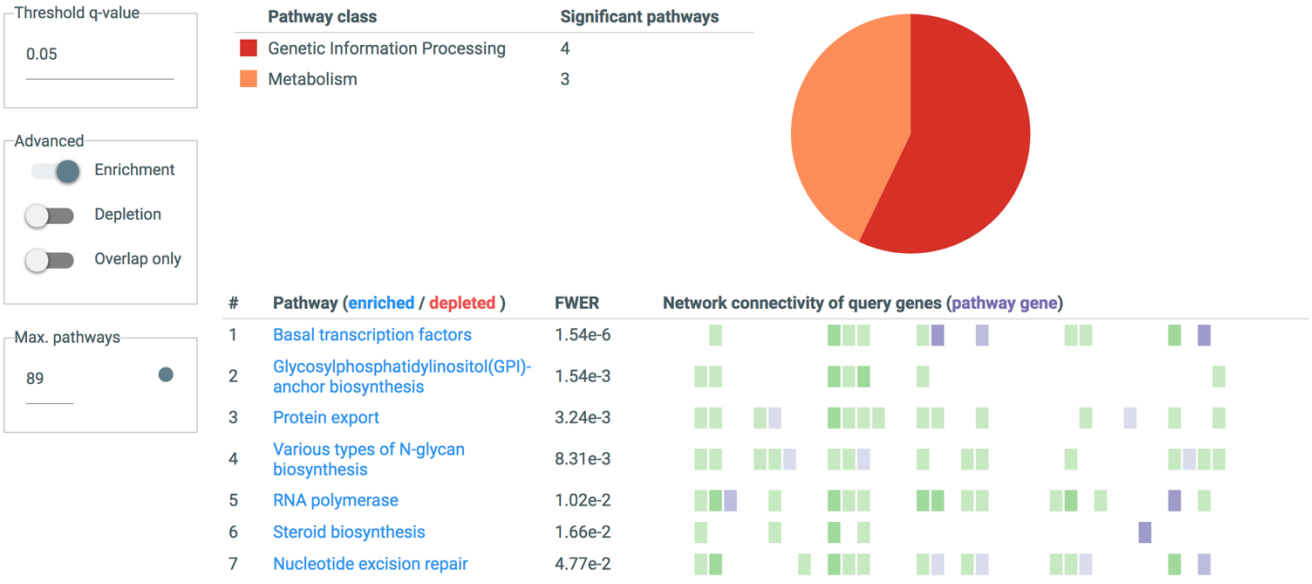
b. FunCoup has protein interaction as the most common evidence type for both gene sets while STRING has database import as the most common evidence for both gene sets.

c.
STRING has three main types of data sources: experiments, predictions based on genomic context and data mining from databases and literatures. Its experimental data consists of high-throughput experiments and co-expression analysis. Its genomic context methods consist of gene fusion, gene order conservation and phylogenetic profiles.

FunCoup also has experiments and genomic context predictions as its data sources. They consist of domain interactions, co-expression patterns, phylogenetic profiles, quantitative mass spectrometry, protein-protein interactions, correlated genetic interactions, subcellular co-localisation, shared miRNA targeting and Shared transcription factor binding. However, FunCoup does not use text-mining evidences as it is error-prone. It also does not use curated databases as FunCoup is designed for detecting novel interactions instead of reproducing existing knowledge.

Enrichment analysis using PathwAX and DAVID

5.
Experiment number 53
PathwAX:



DAVID:

5 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Nucleotide excision repair	RT		5	12.5	3.7E-3	1.8E-1
	KEGG_PATHWAY	Phenylalanine, tyrosine and tryptophan biosynthesis	RT		3	7.5	3.5E-2	6.2E-1
	KEGG_PATHWAY	Metabolic pathways	RT		18	45.0	6.9E-2	7.1E-1
	KEGG_PATHWAY	Various types of N-glycan biosynthesis	RT		3	7.5	9.8E-2	7.5E-1
	KEGG_PATHWAY	Alanine, aspartate and glutamate metabolism	RT		3	7.5	9.8E-2	7.5E-1

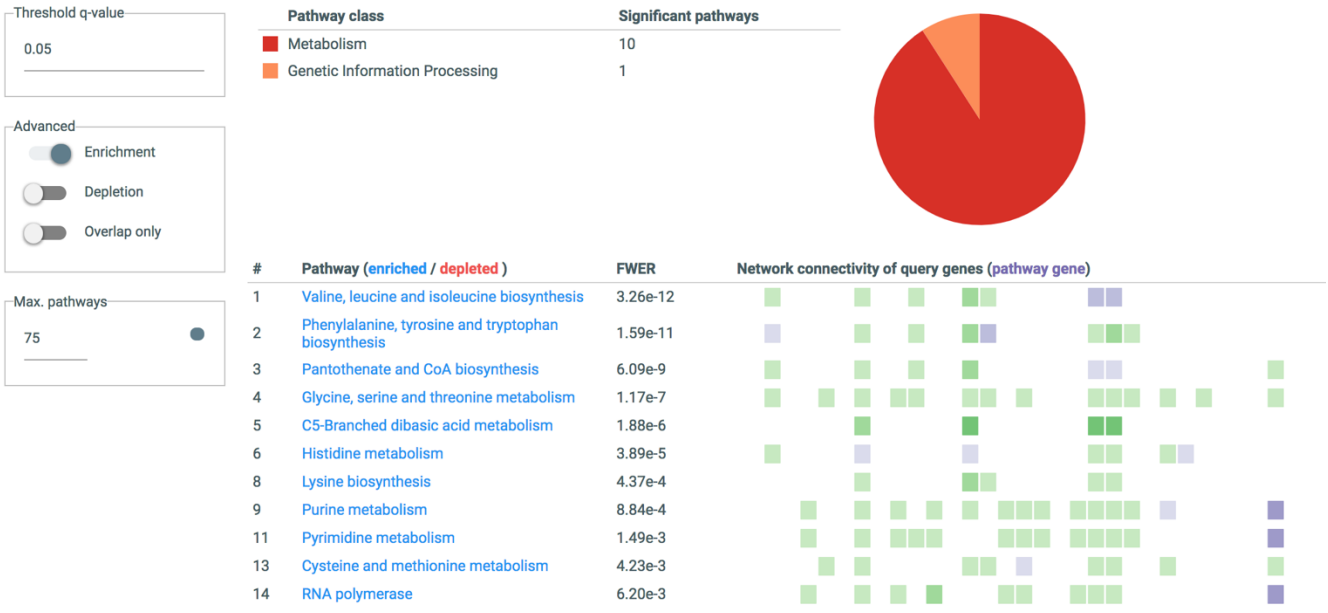
Pathways enriched in PathwAX:

- Basal transcription factors
- Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
- Protein export
- Various types of N-glycan biosynthesis
- RNA polymerase
- Steriod biosynthesis
- Nucleotide excision repair

Pathways enriched in DAVID

- Nucleotide excision repair
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Metabolic pathways
- Various types of N-glycan biosynthesis
- Alanine, aspartate and glutamate metabolism

Experiment number 38
PathwAX:



DAVID:

12 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Metabolic pathways	RT		27	73.0	4.1E-7	2.0E-5
	KEGG_PATHWAY	Biosynthesis of amino acids	RT		12	32.4	2.0E-6	4.9E-5
	KEGG_PATHWAY	Biosynthesis of secondary metabolites	RT		14	37.8	4.8E-4	7.7E-3
	KEGG_PATHWAY	Fatty acid biosynthesis	RT		4	10.8	6.1E-4	7.5E-3
	KEGG_PATHWAY	Phenylalanine, tyrosine and tryptophan biosynthesis	RT		4	10.8	2.4E-3	2.3E-2
	KEGG_PATHWAY	Fatty acid metabolism	RT		4	10.8	5.0E-3	4.0E-2
	KEGG_PATHWAY	Histidine metabolism	RT		3	8.1	1.8E-2	1.2E-1
	KEGG_PATHWAY	Biosynthesis of antibiotics	RT		9	24.3	2.2E-2	1.3E-1
	KEGG_PATHWAY	Pantothenate and CoA biosynthesis	RT		3	8.1	2.4E-2	1.2E-1
	KEGG_PATHWAY	Alanine, aspartate and glutamate metabolism	RT		3	8.1	8.5E-2	3.5E-1
	KEGG_PATHWAY	Biotin metabolism	RT		2	5.4	9.5E-2	3.6E-1
	KEGG_PATHWAY	ABC transporters	RT		2	5.4	9.5E-2	3.6E-1

8 gene(s) from your list are not in the output.

Pathways enriched in PathwAX:

- Valine, leucine and isoleucine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Pantothenate and CoA biosynthesis
- Glycine, serine and threonine metabolism
- G5-branched dibasic acid metabolism
- Histidine metabolism
- Lysine biosynthesis
- Purine metabolism
- Pyrimidine metabolism
- Cysteine and methionine metabolism
- RNA polymerase

Pathways enriched in DAVID

- Metabolic pathways
- Biosynthesis of amino acids
- Biosynthesis of secondary metabolites
- Fatty acid biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Fatty acid metabolism
- Histidine metabolism
- Biosynthesis of antibiotics
- Pantothenate and CoA biosynthesis
- Alanine, aspartate and glutamate metabolism
- Biotin metabolism
- ABC transporters

6.
For experiment number 53, PathwAX reported that the majority of enriched pathways are associated with genetic information processing. There are also some pathways associated with metabolism. DAVID only reported that the majority of the enriched pathways are

associated with metabolism. There is only one enriched pathway (nucleotide-excision repair) associated with genetic information processing.

For experiment number 38, PathwAX and DAVID have similar results in that majority of the enriched pathways are associated with metabolism, particularly biomolecule synthesis.

In contrast to DAVID, PathwAX is able to significant pathways which has no gene overlap with our gene sets. In addition, PathwAX can also identify depleted pathways which DAVID cannot.

7.
The number of input genes matters more for DAVID. DAVID uses gene overlap enrichment analysis which requires at least two overlaps between gene list and pathways to calculate EASE score. Many of the pathways identified by DAVID has two or three gene overlaps. If the gene list is shorter and there are less overlaps, DAVID will not be able to identify these pathways. PathwAX, on the other hand, does not have this limitation. It is able to identify significant pathways with one or even no gene overlaps.