The genomes we used in this practical is listed below.

|  |  |  |  |
| --- | --- | --- | --- |
| **GenomeID** | **Taxa** | **NCBI taxaID** | **STRING taxaID** |
| 1 | E.coli 536 | 362663 | 362663 |
| 2 | S.coeli A3(2) | 100226 | 100226 |
| 3 | S.cerevisiae | 4932 | 4932 |
| 4 | R. xylan DSM 9941 | 266117 | 266117 |
| 5 | H.halaphila | 1053 | 349124 |

1.

Script: *connectivity\_plot.py*

Purpose: This script calculates the average connectivity of the genome and plots the distribution of the number of interactions for each node.

Usage: Python3 connectivity\_plot.py genomeFile

The average connectivity of genome1 = 220.134

, , ,Genome2 = 298.384

, , ,Genome3 = 325.411

, , ,Genome4 = 203.621

, , ,Genome5 = 159.817

2.

A close up of a map

Description generated with very high confidenceA close up of a map

Description generated with high confidenceA close up of a map

Description generated with high confidenceA close up of a map

Description generated with high confidenceA close up of a map

Description generated with high confidence

The yeast genome (genome 02) follows power-law distribution, containing many nodes (>100) with only a few links (<10) and a few hubs with a large number of links. The proteome of eukaryotics should have higher stability and more complex interaction compared to proyokarutos. Following the power-low distribution makes the random changes more likely happen in low-connectivity nodes, which would do much harm to the whole system.

3.

Script: *findOverlapGeneSet.py*

There are two gene sets that has the most overlap, each of them has five overlapping genes with the eukaryote genome.

Geneset1:

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

Geneset2:

FAS2 RPB8 HIS6 CEM1 SPE3 ILV5 ILV3 ARO8 IMP1 OAR1 MAE1 PNP1 ARO1 PRO2 BNA7 MET14 DHH1 RIB1 COX8 RNH1 FAS1 SNQ2 GUT1 TFB1 CAB4 MET8 HIS5 GLT1 ARO8 URM1 PHA2 HIS1 PNP1 STE6 PRP38 ARG4 PNP1 GLN1 EHD3 PMS1

Note: There are duplications in the *experiment.txt* file, which should be deleted.

4.

4.a

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **FunCoup** | | **STRING** | |
|  | Geneset1 | Geneset2 | Geneset1 | Geneset2 |
| Nodes | 70 | 67 | 40 | 37 |
| Links | 525 | 574 | 22 | 31 |
| Hubs | TRP5 (32)  PRE9 (32)  RPB8-(28) | ILV5 (41)  DHH1 (41)  ILV3 (41) | SSL1  TRP5  RPB8 | ARO1  OAR1  ILV5 |

The two gene sets have similar gene numbers and gene set 2 has higher average connectivity. The hubs in gene set 2 has higher connectivity and there are more hubs. In general, gene set 2 is more significantly connected than gene set 1.

Using the same confidence level STRING predictes less nodes and connections compared to FunCoup.

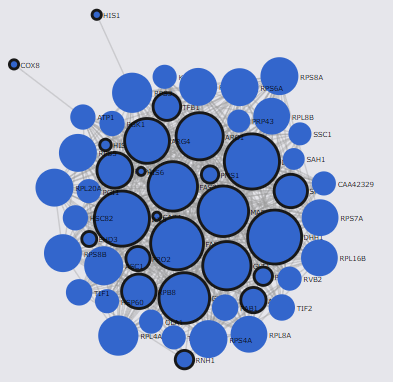
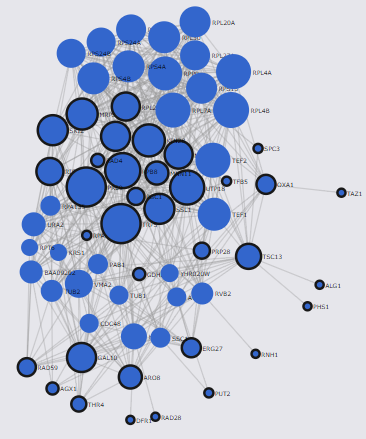


Figure 1 the interaction network of geneset1 (left) and geneset2 (right) with FunCoup (confidence = 0.800)

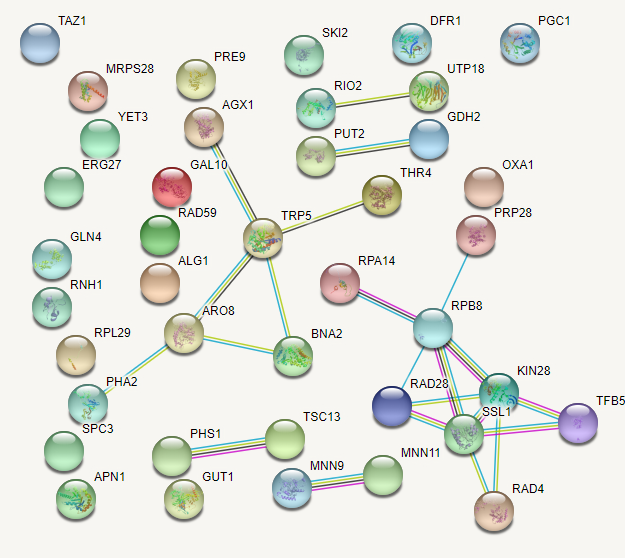
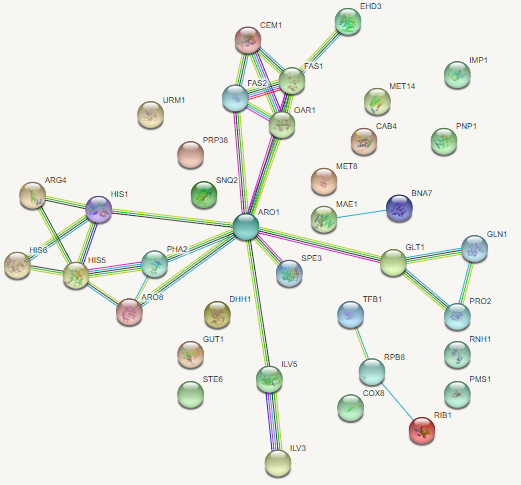
 

Figure 2: The interaction network of geneset1(left) and geneset2(right) with STRING (confidence = 0.800)

4.b

Metabolic

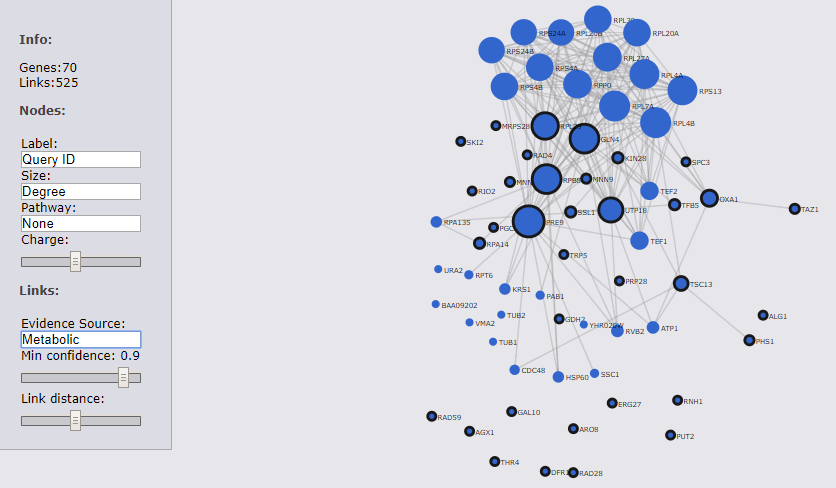


Figure 2 The interaction network based on metabolic evidences in FunCoup (confidence = 0.9)

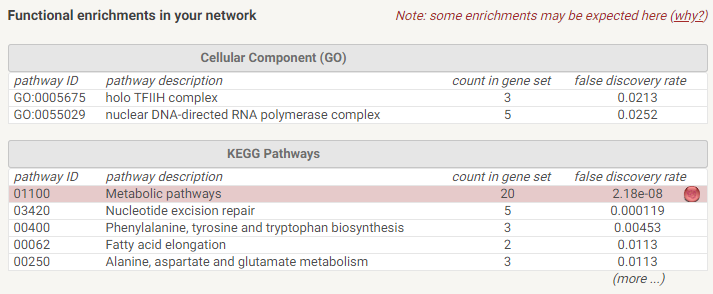


Figure 30 The interactions of geneset1 based on metabolic evidences in STRING (confidence = 0.9)

4.c

Data source in FunCoup:

PIN, MEX ,PEX, GIN,TFB, MIR,SCL,DOM,PHP,QMS

Data source in STRING:

[COG](http://www.ncbi.nlm.nih.gov/COG/),[Ensembl](http://www.ensembl.org/)       [Intact](http://www.ebi.ac.uk/Intact/)       [RefSeq](http://www.ncbi.nlm.nih.gov/RefSeq/)       [PubMed](http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed)       [Reactome](http://www.reactome.org/)       [DIP](http://dip.doe-mbi.ucla.edu/)       [BioGRID](http://www.thebiogrid.org/)       [MINT](http://mint.bio.uniroma2.it/)       [KEGG](http://www.genome.jp/kegg/)       [SGD](http://www.yeastgenome.org/)       [FlyBase](http://flybase.bio.indiana.edu/)       [SwissProt/UniProt](http://www.uniprot.org/)       [SwissModel](http://swissmodel.expasy.org/)       [HUGO](http://www.hugo-international.org/)       [OMIM](http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM)       [NCI/Nature PID](http://pid.nci.nih.gov/)       [PDB](http://www.rcsb.org/pdb/home/home.do)       [The Interactive Fly](http://www.sdbonline.org/fly/aimain/1aahome.htm)       [BioCyc](http://www.biocyc.org/)       [Gene Ontology](http://www.geneontology.org/)       [SIMAP](http://liferay.csb.univie.ac.at/portal/web/simap)

STRINGS contains protein structure and sequence data, and it used more pathway information.

FunCoup emphasice the relationship between orthologs and considered the co-ocurrence in the same operon.

5.

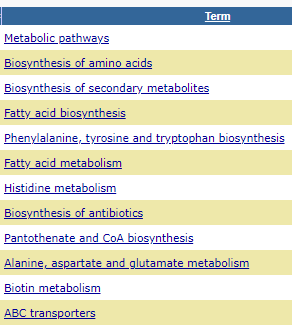
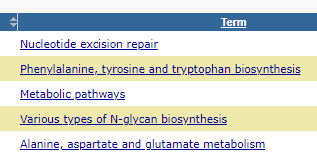
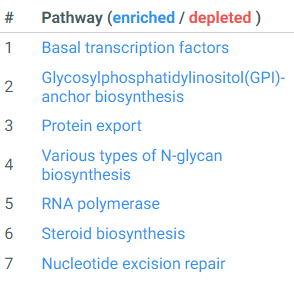


Figure 4 Enriched KEGG pathways in PathwAX of gene set 1(top left) and gene set 2 (bottom left), Enriched KEGG pathways in DAVID of gene set 1(top right) and gene set 2 (bottom right)

5.2

5.3