The highlight part is answers we are not sure about.

The right parts need better phrasing

Scripts that need changes.

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

No no straight line.

3.

Script: *findOverlapGeneSet.py (we need to use shuhans’s scripyt)*

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 | 70 | 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 |

4.b

FunCoup: Protein Interaction

STRING

4.c

STRING: Database and mining, prediction, experiment, itegrates all known knowledge.

FunCoup: Prediction and experiment,good at findng new.

5.

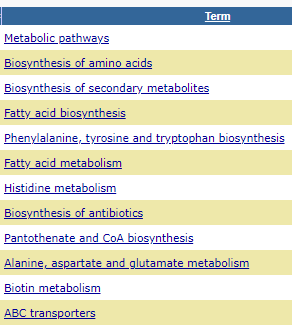
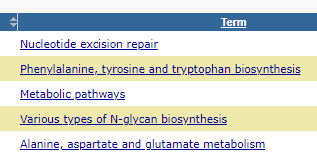
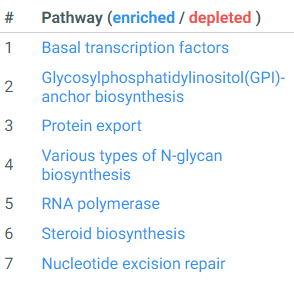


Figure 4 Enriched KEGG pathways in PathwAX of gene set 1(top left) and gene set 2 (bottom left), Enriched KEGG pathways in DAVID of gene set 1(top right) and gene set 2 (bottom right)

5.2

In pathwAX, genetic information processing and metabolic pathway. Also, in geneset 1, there is protein export.

In DAVID, it mainly analysed the metabolic pathway.

5.3

The size of the gene set is limited to 400 in PathwAX because of the limit of network connectivity matrix. And DAVID is good at processing large gene set.

DAVID is based on overlapping. In small gene sets, there are few overlapping which leads to false negative. PathwAX solves this problem by using network crosstalk and is good at processing small datasets.