STATISTICAL ANALYSIS OF PROTEIN-PROTEIN INTERACTION DATASET CONSTRUCTED BY CHECKING THE EXISTENCE OF PROTEINS IN INTERACTION IN DIFFERENT SPECIES

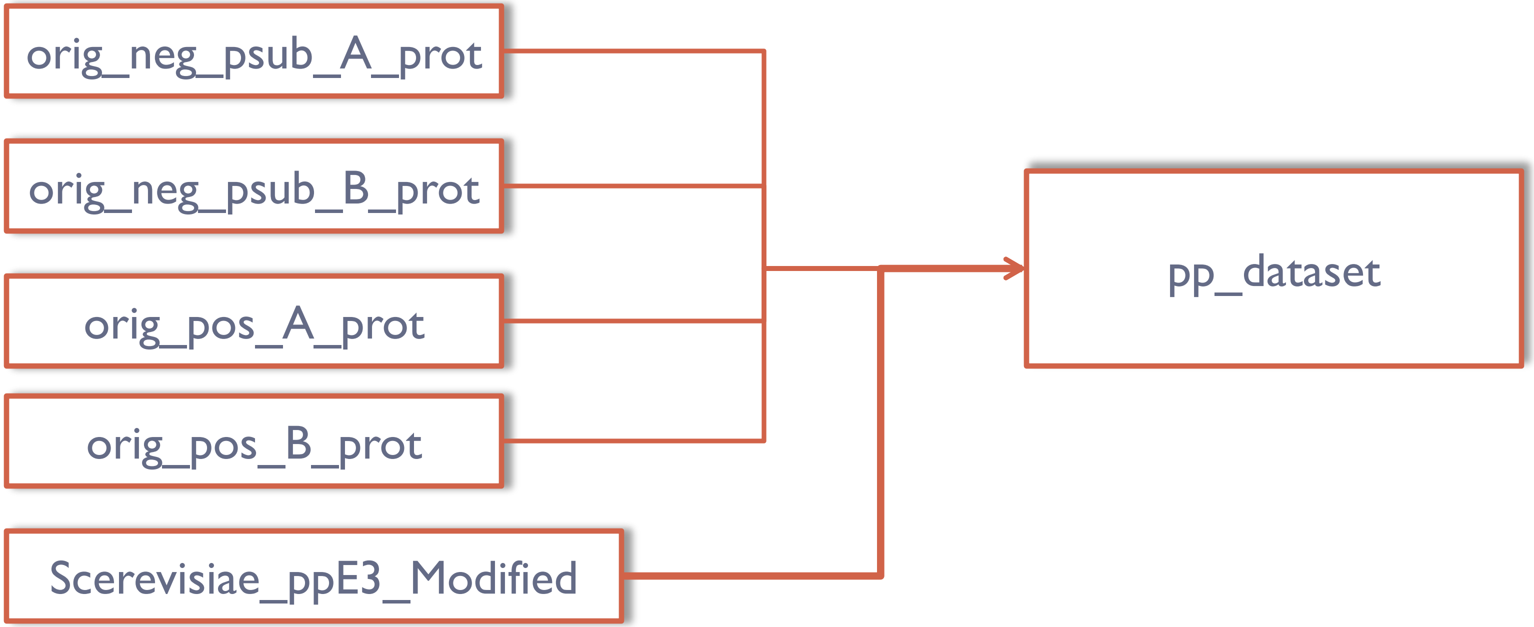
Ömer Nebil Yaveroğlu
METU – M. Sc. – Computer Engineering

The Construction Of Dataset

The organism used for extraction of the protein-protein interactions :

Saccharomyces cerevisiae (yeast)

FILE OPERATIONS



orig_neg_psub_A_prot

orig_neg_psub_B_prot

orig_pos_A_prot

orig_pos_B_prot

Scerevisiae_ppE3_Modified

pp_dataset

The Dataset

- 11698 protein-protein pairs extracted from yeast organism (samples)
- 450 different species (features)
- Constructed by a looking at the existence of proteins in the protein pair in different species
- Ranking :
  - If both proteins are included → 2
  - If none of the proteins included → 0
  - If one protein exists, other don't → -1

The Resulting Dataset Format

1	0	2	0	-1	-1	…
1	0	0	-1	0	0	…
1	0	2	0	0	0	…
0	-1	2	0	0	0	…
0	-1	-1	0	-1	0	…

The Aim of the Project

Finding an outlier between the positive and negative datasets by applying statistical methods such as:

- Principle Component Analysis (PCA)
- Multidimensional Scaling (MDS)
- K-means Clustering
- Support Vector Machines (SVM)

What have been done so far?

- Performed Principle Component Analysis
- Performed Multidimensional Scaling
- Performed K-means Clustering but features and samples switched

What to do next?

- Revision of the results obtained by previous method applications

- Checking for where the positive and negative datasets are placed after the application of the mentioned methods

- If needed, application of support vector machines in order to find an outlier