The preprocessing procedure is implemented using InSyBio Dataset Preprocessing pipeline. In this procedure, columns that contain more than 70% missing values are removed. Then the rest are normalized using arithmetic sample- wise normalization and the missing values are filled using the kNN method of imputation.

The results are the following:

- Before preprocess:

Total number of calculated features: 61

- After preprocess:

Total number of calculated features: 41

Headers of removed columns: Homologous in Mouse, Homologous in Drosophila, Homologous in Yeast, Homologous in Ecoli, Subcellular Co-localization?, 0-14

-> See: DATASET\_PREPROCESSED\_70.csv (== FINAL\_DATASET\_PREPROCESSED.csv)