

Milestone 1

Part 1:

Description

The dataset we chosen was “Transcriptomics in Yeast”. This dataset displays the condition in which the yeast cell was placed in, the cellular component involved, the molecular function and the biological processes of a gene exhibited by the yeast.. In simple terms, the cellular component is where the gene is located within a cell, molecular function is what the gene is doing in the cell and biological processes is the genes involved in the process. The author of this dataset was attempting to determine if there is a possibility for improvement in the genetic classification by modeling the data.

1. Conditions Annotations

- a. ID - Identification of condition
- b. Primary - primary condition
- c. Secondary - secondary condition
- d. Additional Information - any additional information that needed to be noted in conditioning.

2. Biological Processes

- a. GeneID - Identification of the gene
- b. Validation -Is the gene verified, uncharacterized or dubious
- c. Localization - where in the yeast cell is the gene being expressed in

3. Cellular Component

- a. GeneID - Identification of the gene
- b. Validation- Is the gene verified, uncharacterized or dubious
- c. Localization - Primary localization
- d. Loc2- other location
- e. Loc3- other location
- f. Loc4- other location
- g. Loc5-other location
- h. Loc6 -other location

4. Molecular Function

- a. GeneID - Identification of the gene
- b. Verified - Is the gene verified, uncharacterized or dubious
- c. Molecular Function - what the gene is doing in the yeast cell

5. Expression Values

- a. X-axis is the condition in which the yeast cell was placed in (92 columns)
- b. Y-axis is the ID of the gene (6071 genes) per yeast cell (6071 rows)
- c. Note: Evaluation was done on gene level. Each gene has its own row. Expression levels were normalized by TPM (transcripts per million), as a default normalization procedure. Raw counts would have been integers, normalized they are floats