Result

To create a great dataset, changes need to be made to the downloaded dataset. The dataset has been gathered of 1484 yeast sequences from SWISS-PROT using the annotations from YPD. In this section, we are going to talk about the results we found from the dataset.

Codebook

As it is a large dataset, it is wise to look at the codebook first. However, there is no codebook. Nevertheless, there is a file with more information on all the attributes. This is where the information is extracted to make your own codebook.

Below a table with a description of all the attributes abbreviation, explanation and data types.

Table 1: *The codebook*

Name	Fullname	Datatypes
seq.name	Accession number for the SWISS-PROT database	str
mcg	McGeoch's method for signal sequence recognition	float
gvh	von Heijne's method for signal sequence recognition	float
alm	Score of the ALOM membrane spanning region prediction program	float
mit	Score of discriminant analysis of the amino acid content of the N-terminal region	float
erl	Presence of 'HDEL' substring	float
pox	Peroxisomal targeting signal in the C-terminus	logical
vac	Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins	float
nuc	Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins	float
loc.site	The class is the localization site	factor

The last column is the sequence's localization site. There are 10 different alternatives for this. Here are the 10 in question: 5

Table 2: *Sequence localization sites*

Abbreviation	Fullname	Amount
CYT	cytosolic or cytoskeletal	463
NUC	nuclear	429
MIT	mitochondrial	244
ME3	membrane protein, no N-terminal signal	163
ME2	membrane protein, uncleaved signal	51
ME1	membrane protein, cleaved signal	44
EXC	extracellular	37
VAC	vacuolar	30
POX	peroxisomal	20
ERL	endoplasmic reticulum lumen	5

There are alot of CYT and NUC localization. ERL localization is the least. There are only 5 of these in the dataset.

Transform data

The data has to be transformed before it can be worked with.