Assignment No. 12.1

Attribute Information:

1. Sequence Name: Accession number for the SWISS-PROT database   
2. mcg: McGeoch's method for signal sequence recognition.   
3. gvh: von Heijne's method for signal sequence recognition.   
4. alm: Score of the ALOM membrane spanning region prediction program.   
5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.   
6. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.   
7. pox: Peroxisomal targeting signal in the C-terminus.   
8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.   
9. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

10. Class Distribution. The class is the localization site. Please see Nakai &

Kanehisa referenced above for more details.

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

The questions to be answered in this assignment are

1. Perform ANOVA test on the discriminant analysis scores of nuclear localization signals of both nuclear and non-nuclear proteins by class variables (Target).
2. Which class is significantly different from others?

Initially we test whether the data is intact and it conforms to the data definitions

> yeastprotien <- read\_excel("E:/kamagyana/Computing/DARET/Assignments/yeastprotien.xlsx")

> View(yeastprotien)

> str(yeastprotien)

Classes ‘tbl\_df’, ‘tbl’ and 'data.frame': 1485 obs. of 10 variables:

$ Sqname: chr "ADT1\_YEAST" "ADT2\_YEAST" "ADT3\_YEAST" "AAR2\_YEAST" ...

$ mcg : num 0.58 0.43 0.64 0.58 0.42 0.51 0.5 0.48 0.55 0.4 ...

$ gvh : num 0.61 0.67 0.62 0.44 0.44 0.4 0.54 0.45 0.5 0.39 ...

$ alm : num 0.47 0.48 0.49 0.57 0.48 0.56 0.48 0.59 0.66 0.6 ...

$ mit : num 0.13 0.27 0.15 0.13 0.54 0.17 0.65 0.2 0.36 0.15 ...

$ erl : num 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...

$ pox : num 0 0 0 0 0 0.5 0 0 0 0 ...

$ vac : num 0.48 0.53 0.53 0.54 0.48 0.49 0.53 0.58 0.49 0.58 ...

$ nuc : num 0.22 0.22 0.22 0.22 0.22 0.22 0.22 0.34 0.22 0.3 ...

$ class : chr "MIT" "MIT" "MIT" "NUC" ...

> nrow(yeastprotien)

[1] 1484

> table(yeastprotien$class)

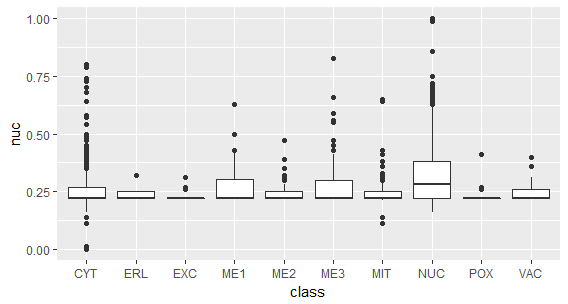
CYT ERL EXC ME1 ME2 ME3 MIT NUC POX VAC

463 5 35 44 51 163 244 429 20 30

> sum(is.na(yeastprotien))

[1] 0

> ggplot(yeastprotien,aes(class,nuc)) + geom\_boxplot()



The box plot clearly shows that the class NUC representing the nucleatic localization site is showing the highest median discriminant score of all the classes. More over the discriminant score is not showing any kind same variance across the various classes which is visible by the plotting of the minimum and maximum and the outliers.

> result <- aov(nuc~class,data=classnuc);

> summary(result)

Df Sum Sq Mean Sq F value Pr(>F)

class 9 1.993 0.22141 22.01 <2e-16 \*\*\*

Residuals 1474 14.825 0.01006

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The ANOVA test REJECTS the NULL hypothesis that there is no significant difference in the mean discriminant scores across the classes, since the F statistic’s significance level is much more lesser than 0%. So the discriminant scores differ across the various class variables. Since we cannot identify the specific class that displays a high score, we go on to perform the following test to identify the same.

> TukeyHSD(result)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = nuc ~ class, data = classnuc)

$`class`

diff lwr upr p adj

NUC-CYT 0.0744429005 0.05314812 0.095737681 0.0000000

NUC-EXC 0.1051068931 0.04924622 0.160967571 0.0000001

NUC-ME1 0.0634965035 0.01319448 0.113798528 0.0026743

NUC-ME2 0.0861881256 0.03912110 0.133255152 0.0000004

NUC-ME3 0.0616783217 0.03244028 0.090916364 0.0000000

NUC-MIT 0.0906947151 0.06521504 0.116174388 0.0000000

POX-NUC -0.0976783217 -0.17037079 -0.024985855 0.0009144

VAC-NUC -0.0810116550 -0.14102211 -0.021001196 0.0008397

The ANOVA results show that there is significant difference in the discriminant score of nuclear localization signals of both nuclear and non-nuclear proteins across class variables. This is due to the p-value being extremely lower than 0 and hence REJECTION of the NULL hypothesis, that there is no such difference.

Intially the box plot has shown larger median for NUC class, but to verify statistically we again undertook the Tukey Honest Significant Differences test and then confirmed that NUC appears to be among all the pairs significantly and its adjusted p-value also shows always a significance level of less than 0.05 or 5%.

HENCE the class NUC is significantly different and higher in its mean discriminant score than the other classes.