**Project plan**

(To pass bold parts are not needed)

Created a git hub repository which has structure for my project

**Information about my dataset:-**

Data is of membrane aplpha helices. The data file is in 3 lines. With a protein id, protein sequence and the feature sequence.The features have 3 possblities Inside membrane or outside

**Steps:**

**Overview:**

I will parse the file and extract the features into a format on which I can input into sklearn.

I will need to divide the dataset into 3 or 5 different file on which to use for cross-validation in order to train and test a SVModel. Train a SVM on sequence info with sklearn.

to be continued………...

**In-depth steps:**

1. Extract the feature from your dataset

Open my data file

Sort through the text file. Separating the the different info into lists.

- id

- sequence

- feature

Index?

2. Create cross-validated sets

3. Train a SVM using single sequence information, using sklearn

4. Check different window sizes for the inputs

**\*5. Add evolutionary information by running psi-blast and extracting the information**

**\*6. Train a SVM using multiple sequence information**

**\*7. Optimize the performance of the SVM**

8. Analyze the results and compare it to previous work

**\*9. Use random forests and a simple decision tree and compare the performance with the SVM performance.**

**10. Extract the data from 50 other proteins and test the performance**

11. Review the state of art for your predictor

12. Write a report