Project Plan

Name: Youcheng ZHANG

Course: Project in molecular Life science (KB8024/KB8025)

Project: Signal Peptide in Gram-negative bacteria

Goal:	To develop a method to predict signal peptides in Gram-negative bacteria based on machine learning approaches
Background:	Evidences show that signal peptides uniquely function as guiders in various organisms such as Gram-negative bacteria (Chou, K. C., 2002), with features and properties on both sequences and structures (Costa, T. R. et al., 2015). Predicting and identifying signal peptides that utilize these intrinsic information by machine learning methods help understand the full complexity of bacteria.
Project design:	
week1	1) Manage template project folder structure in a new repository on GitHub
(Feb 15th - 18th)	2) Learn basic Bash, Git, Python, sklearn command
week2	1) Literatures search on project background about Protein secretion, signal peptide prediction
(Feb 19th - 25th)	2) List five papers and summary: List_of_papers_and_Summary and Write Project_Plan
	3) Extract features from raw dataset: gram-signal.3line.txt and Convert into array structure for sklearn.svm data input
	4) Run SVM and cross-validation (automatically) on dataset with sklearn
	5) Run cross-validation (manually) with self-selected data partition as training and validation set
	6) Compare different cross-validation method, and Evaluate the accuracy and performance
week3	1) Practice presentation and Peer review
(Feb 26th - 4th)	2) Extract features from raw dataset and Modify with different window sizes
	3) Run SVM and cross-validation to evaluate
	4) Change SVM parameters such as kernel types, etc. and Test model performance
	5) Write bash scripts for obtaining homologs of all the sequences in gram-signal.3line.txt by running PSI-BLAST locally, for constructing multiple sequence alignment (frequency matrix or scoring weight matrix), and for further extracting the features from the established MSA
	6) Run SVM and cross-validation on multiple sequence feature
week4	1) Paper presentation and Self-evaluation
(Mar 5th - 11th)	2) Change SVM parameters to evaluate model performance
	3) Compare model performance to existing prediction models
	4) Run random forests and decision tree on MSA data, and Evaluate the model performance
	5) Start write final essay: Introduction and Method section
week5	1) Find 50 other protein sequences to perform prediction with the model
(Mar 12th - 18th)	2) Process data and graphs
	3) Write essay: Result and Discussion section
week6	1) Submit final essay
(Mar 19th -)	2) Upload model GitHub