CSCI 5622 Project Update: Predicting Essential Genes in Yeast

Team: Nicolas Metts, Matthew Pennington, Rani Schwindt, Carter Tillquist

Progress so far:

- We have a working Python script that allows using multiple classifiers (Logistic Regression with Stochastic Gradient Descent, SVM, AdaBoost, KNN, Gaussian Naive Bayes) and any set of available features.
- The script is also configurable to allow using pre made test/train set files or using cross-validation to split the data into training and test sets.
- The metrics being used are accuracy, precision, and recall.
- The script also outputs classification results into a log file so we can keep track of our progress
- We are currently able to obtain results that are an improvement over the baseline.
- We curated a new, expanded data set (5,700 genes and 50 features) in order to make the results more generalizable and to explore features not previously used for this type of classification.
- The proposed timeline remains attainable, so no changes have been made.

Remaining to be done:

- Comparing results of various classifiers according to the three metrics (accuracy, precision, and recall), as well as what set of features give the best performance on these metrics.
- Deciding on the best set of features and classifiers to be used for each metric. All feature engineering will be completed by December 1st. The entire team agrees that this is a reasonable goal.
- Consider methods of combining classifier results to improve overall predictions and reduce overfitting.
- Complete error analysis, including any visualizations that may be informative
- Analysis of results will begin after December 1st. Analysis will be completed by December 8th, and the remaining week will be spent formally documenting the analysis in the form of a write up and presentation.