# CSCI 5622 Project Proposal: Predicting Essential Genes in Yeast

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

30 October 2015

## 1 Background & Motivation

DNA, the genetic "instructions" contained in each cell of a living organism, is made up of genes. Each gene is a functional sequence which encodes important information needed to make proteins. Some genes encode proteins which are required for the cell to survive. **Essential genes** are genes that, when deleted, cause a cell to die[1]. Predicting which genes are essential to a specific organism is important in many contexts, such as finding drug targets for pathogenic microbes or finding the genetic vulnerabilities of a cancer cell. For organisms with thousands of genes, essentiality prediction could bypass the need for expensive, genome-scale experimental screens and instead be used to inform validation experiments on a few predicted candidates. We propose that a gene can be classified as essential or non-essential by learning from several biological features of a gene, which we describe in Section 3.

For our analysis, we will be using the genes from *Saccharomyces cerevisiae* S288C, a well characterized strain of yeast. Each gene in the *S. cerevisiae* genome has been experimentally deleted in order to determine whether it is essential or non-essential, making this an ideal system for machine learning classification. Additionally, because *S. cerevisiae* genes have homologs in other organisms (including humans), discovering which features are important for predicting gene essentiality in yeast will provide valuable insight for performing this analysis in other organisms.

## 2 Baseline

In 2006, Seringhaus *et al*[1] used 14 biological features to train a classifier for predicting essential genes in *S. cerevisiae* and a related organism *S. miktae*. On 4,648 genes in *S. cerevisiae*, the classifier resulted in a precision $\frac{TP}{TP+FP} = 0.69$ and recall $\frac{TP}{TP+FN} = 0.091$. The classifier used was an average of 7 different classifiers, including logistic regression, Naive Bayes, and AdaBoost.

## 3 Data

We currently have the data from the 14 sequence-related features used in Seringhaus (2002)[1] for each gene, including features like subcellular localization, number of transmembrane helices, effective number of codons, % GC content, and length of protein. In addition, we have 7 other features already compiled, including scores for phylogenetic conservation and number of interaction partners.

## 4 Techniques

We plan to compile additional features, including Gene Ontology annotations, and perform feature engineering to find an efficient combination. Over the course of the semester we have a learned a number of techniques for approaching classification problems similar to this one. SVMs are powerful and relatively flexible, allowing the use of different kernels with minimal cost, and they have been shown to provide good results in practice. On the other hand, K-Nearest Neighbours is one of the most simple techniques that we have examined and its performance could be used as an additional benchmark against which we may compare our results. It might also be useful in uncovering relationships between subsets of features and suggesting ways of combining and manipulating them.

## 5 Timeline

**4 Nov** - all features compiled and ready to use
**6 Nov** - baseline classifier results obtained
**10 Nov** - feature engineering complete
**13 Nov** - improved results compiled

1. Seringhaus M1, Paccanaro A, Borneman A, Snyder M, Gerstein M. Predicting essential genes in fungal genomes. *Genome Res* 16(9):1126-35.(2006).