1.1 Methods

1.1.1 Feature Description

To predict essentiality of the genes in yeast *Saccharomyces cerevisiae*, our analysis uses features associated with each gene from two sources: 1) fourteen sequence-derived features from the Seringhaus, et al., 2006 study, and 2) eight additional non-sequence-derived features from the Ensembl website (ref). Feature definitions are listed in Table 1 (ref). Five of 22 features (mitochondria, vacuole in\_how\_many\_of\_5\_proks\_blast, intron, er) were removed from the analysis due to low content (less than 20% of non-zero values). At lower training sizes, their low content resulted in deterministic models because the randomization employed multiple seeds until the design matrix had no columns of zeroes. Genes which had values for this low-content features would have been selected more frequently than genes without information in these features.

1.1.2 Cross-Validation Strategy

The cross-validation strategy incorporates an *unbalanced strategy* to the test set (Figure 1) along with a *contamination rate*. For an unbalanced design, test sets utilize the remaining genes not used in the training sets rather than a balanced strategy which matches training and testing set sizes. The unbalanced strategy was chosen because, in practice, an investigator would typically want to test all the remaining genes for essentiality rather than just a subset of genes. Another concept interwoven into the analysis is contamination. Contamination considers the dilemma from falsely assigned genes in the training sets. The semi-supervised and unsupervised methods do not consider negative labels in their computations and, thus, are unaffected by contamination. For supervised methods, positive labels in the training set are mixed with negative labeled genes for analysis when varying percentages of contamination are introduced.

In the unsupervised simulation, semi-supervised was compared against the unsupervised method described in section 2.4. We train our methods from one of two sets of essential genes (positive labels): 1) known essential genes (769) and 2) a subset of pre-2002 known genes (64). Additionally, we contrasted all 22 features against a subset of 14 sequenced-derived features as predictors of essentiality (see Section 1.1.1). Training set sizes were based on increments of 5 with minimum set sizes greater than the number of features to prevent rank deficiency in training sets.

In the supervised simulation, semi-supervised was compared against three supervised methods (LASSO, SVM, and Random Forest) at low training set sizes. AUC performance of these four methods was compared across training set sizes between 1% (n=##) and 10% (n=##) from all ### genes. Genes randomly chosen for the supervised training sets reflect the same ratio of positive and negative labels as seen in the full data set. Among the 3500 yeast genes, there are 769 essential genes resulting in a 21% ratio. Therefore, as an example, at 1% training size, 35 randomly chosen genes contained 7 positive labels (21% of 35) and 28 negative labels for supervised methods, while semi-supervised methods analyzed 35 positively labeled genes. In order to mimic contamination, negative labels were reassigned a positive label at rates of 0%, 20%, and 50%.

For all results, unique initial seeds were chosen based on the iteration number, training set size, and contamination (for supervised comparison only). Thirty random iterations were used to average the AUC or other metrics used to evaluate performance. Iterations were increased to 100 for lower training sets to better discriminate. Once the cross-validation data was generated by a seed, the same data was used to compare each method.

1.1.3 Algorithms

All simulations were performed in R version 3.3.3. The semi-supervised and unsupervised analysis utilized functions from the *lcmix* package. The *lcmix* package developed and implemented in the previous paper by Dvorkin, Biehl, and Kechris A Graphical Model Method for Integrating Multiple Sources of Genome-scale Data and can be downloaded from [http://r-forge.r-project.org/projects/lcmix/](http://r-forge.r-project.org/projects/lcmix/" \t "pmc_ext). LASSO was performed using the *glmnet* command in the *glmnet* package (ref Hastie and Qian). Using *cv.glmnet*, k-fold cross validation optimized the minimum lambda for the LASSO function. SVM analysis used the *svm* command under the *e1071* package (ref David Meyer). Various runs using different criteria revealed a radial kernel density and C-classification optimized AUC performance. Random Forest was performed with the *randomForest* command under the *randomForest* package (ref Breiman, L.) All supervised predictions used the *predict* command in the *stats* package.

1.1.4 Performance

The AUC mean, variance, and CV (median absolute deviation) of the three supervised methods were contrasted against semi-supervised method. Because LASSO outperformed the other supervised models in AUC across all training set sizes and contamination rates, a closer evaluation of its performance was compared with semi-supervised method. In order to fairly contract LASSO performance to semi-supervised, the prediction scores were rescaled to be between 0 and 1. Precision, recall, and f-measure further discriminated the two methods with four rescaled prediction score cutoffs including the median and prediction scores of 0.5, 0.8, and 0.95. The median cutoff is a relative measure based on the data while the other three cutoffs are absolute. The f-measure was calculated from the average precision and recall at each training set size from 1% to 5%. The number of predicted positive values at each training size are presented in Supplemental Table 1.

1.2 Results

We used cross-validation simulations to compare our hierarchical mixture model semi-supervised method with unsupervised and supervised methods. We hypothesize that semi-supervised method outperforms both unsupervised methods at any training set size and supervised methods at low training set sizes especially when positive labels are contaminated. Using the unsupervised comparison, we also explored the effect of different training sets or features. First, we describe each of the features and report the univariate performance in predicting essentiality for each one in Table 2 (ref). Based on the range of AUC, many of the features have low predictive value on their own but in the following comparisons we will explore their combined predictive power.

1.2.1 Unsupervised Comparison

The complete essentiality data for Saccharomyces cerevisiae contains n=769 positive labeled genes [REF]. However, there is a smaller subset of genes that were identified before 2002 (n=64) [REF].. Therefore, we explored the use of the pre-2002 as a natural training set to predict essentiality on all other genes. We also used the complete set of positive labeled genes as a second set. Finally, to explore whether our conclusions were sensitive to the choice of features, we first used only 14 sequence-derived features from REF and then a larger set of 22 features, which included the 14 sequenced-derived features and additional features collected from Ensembl (see Methods).

Semi-supervised performs better than unsupervised for AUC regardless of training on pre-2002 or all essential genes, predicting with 14 sequence-derived or all 22 features, or training set size (Figure 2). The variance of the AUC for both methods increases as training size increases when training on all essential genes (Figure 2c,d). This is expected as the test set is relatively larger and more constant with the smaller training sets. The eight additional features added from Ensembl Biomart generally improves AUC performance and decreases variance for both methods, but with more dramatic improvement with the smaller pre-2002 training set (Figure 2a and 2b). Comparing the two training sets on the same size (e.g., n=25 and 50) for both sequence-derived (Figure 2a versus 2c), and all features (Figure 2b versus 2d), we see increased AUC using the larger training set.

Comparing training with pre-2002 known genes to the training sizes of 25 and 50 for all essential genes provides more resolution at low training sizes albeit with a specific subset of essential genes. Predicting with all features and training on pre-2002 essential genes results in the smallest variance across all training set sizes.

1.2.2 Supervised Comparison

Next, we compared the semi-supervised method with a supervised strategy using all essential genes for the training set and all 22 features. Supervised algorithms require both positive and negative labels. Therefore, we picked a random set of the non-essential genes to be the negative labels (see Methods), but also included some contamination (some essential genes labeled as negatives in the training set) since in practice, the complete set of negative labels will not be known in many situations.

LASSO and semi-supervised outperforms the other two supervised methods - SVM and Random Forest (Figure 3). At low training sizes (< 2%, n = 70), semi-supervised method has a higher mean AUC than the three supervised methods. LASSO does not match the stability (lower variance) of semi-supervised until around 5% (n=175) training set size. However, for larger training set sizes, the AUC variance of semi-supervised increases while variance from LASSO slightly decreases. As contamination increases, all three supervised methods decrease in performance. At 50% contamination, semi-supervised method bests all methods across all training set sizes (up to 10%). The CV (median absolute deviance/median) for semi-supervised is lower than LASSO across all contamination levels and training set sizes up to 5% (Figure 4).

1.2.3 Semi-supervised versus LASSO Performance

To compare the best performing supervised method, LASSO, to semi-supervised method, prediction scores were rescaled to be between 0 and 1. At 1% training level, LASSO kernel densities of prediction scores exhibit an unimodal distribution while semi-supervised methods exhibit bi- or multi-modal behaviors (Figure 5 Row 1).Uni-modal behavior makes it more difficult to find better separation of gene types (e.g., essential versus non-essential) At 5% training level, LASSO kernel densities of prediction scores continue to exhibit unimodal distributions while semi-supervised methods maintain their multimodal behaviors (Figure 5 Row 2).

Focusing on 0% contamination in Figure 6, the three absolute cutoffs (50%, 80%, and 90%) reveal a higher recall across all training set sizes for semi-supervised and the median cutoff shows semi-supervised outperforming LASSO up to 3% at which they become comparable. Also, up to 3%, semi-supervised outperforms LASSO in precision at the median cutoff. Precision generally increases as the absolute cutoff increases with LASSO besting semi-supervised as training set size increases. Contamination reduces all three performance measures (precision, recall, f-measure) for LASSO across training set sizes from 1% to 5% and all four cutoffs. Irrespective of contamination, the f-measure for semi-supervised outperforms LASSO for all training set sizes and cutoffs.

Potential Discussion Points

Due to the inherent capacity of supervised methods to utilize both positive and negative labels, they have natural advantages over semi-supervised methods which only handles positive labels.

The contamination is a strategy to emulate a real-world scenario that a researcher may know a certain number of positive labels for genes in their experiment but are unsure if the remaining genes are truly negative.

Lasso may have an advantage over the other methods because it can reduce the effect of poorly predicting variables by collapsing their betas to 0.

, when negative labels are unknown or tentative

Supervised methods such as LASSO with unimodal distributions do not intrinsically show a clear optimal cutoff compared to the multi-modality of semi-supervised predicted probabilities. The multimodality of semi-supervised prediction scores provides more natural cutoffs than the unimodal distribution from LASSO.

With posterior probabilities ranging from near 0 and 1, the mid-range for both methods is near 0.5. Because of the heavily, skewed low probabilities in the unimodal distribution from LASSO (Figure 4), the median would divide the set somewhere on the backside of the slope, greater than the maximum but less than 0.5. The median and mid-range for semi- supervise tend to fall near a local minimum, a useful indicator for separating distributional behaviors and cannot be evaluated in unimodal distributions. The most dynamic difference between the two methods is the behavior of the recall performance. Recall measures how many positive labels were predicted out of the true number of positive labels. In the mid-range cutoff for LASSO, recall would naturally be lower than the median cutoff due to the small area in the right tail greater than 0.5. The expected increase in precision wasn’t strong enough to outperform semi-supervise in the combined f-measure.

##Say something about how the 1-5% training set sizes correspond to realistic values of positive labels #s? (for example the Drosophila study and pre-2002 #s)