INTRODUCTION

Spiro Stilianoudakis

August 9, 2018

INTRODUCTION

The advent of various genome-wide technologies, such as high-throughput conformation capture (Hi-C), have revealed how the spatial organization of the human genome may affect several epigenetic functions (Aiden et al.). Analyses have shown that the genome is tightly compacted into distinct compartments. There exist regions within these compartments that are highly conserved and self-interacting, and termed topologically associating domains (TADs) (Dixon et al.). Evidence suggests that regulatory elements and genes tend to interact more frequently within the same TAD (Symmons et al.). This suggests that the boundaries of TADs may play a role in restricting the function of elements such as enhancers, thereby impacting the transcription of genes.

More recently, it has been discovered that insulators have a primary role in orchestrating the topological arrangement of higher-order chromatin architecture (Phillips-Cremins et al). Insulators are multi-faceted regulatory sequences that moderate a variety of genomic processes including activation, repression, and enhancer blocking. Specifically, the insulator binding protein CTCF has been found to be enriched at the boundary sequences of topologically associating domains in human cells and may therefore act as a mediator of long range chromatin contacts (Zuin et al.). Likewise, it was found that DNase I-hypersensitive sites were enriched at the boundaries of domains and correlated with the CTCF signals (Hong et al). On the other hand, it is unclear how other types of regulatory elements such as histone modifications, which are associated with transcriptional initiation and open chromatin structure, are related to TAD boundaries.

The distinct patterns of some of these different proteins and functional elements point toward the opportunity of computational approaches in predicting the development of TAD boundaries. However, due to the size of Hi-C data and the abundance of available genomic features, few methods have been developed to study the role of specific sets of these features on the folding of chromosomes. Furthermore, many widely used methods ignore key characteristics of the data that may hinder the performance of certain parametric models. One such group have proposed a multiple logistic (MLR) model used to identify the most influential proteins

with regards to TAD boundaries (Mourad et al.). They also provide an MLR model with LASSO estimated coefficients that they believe is better suited for predictors that may be correlated. However, key aspects of the data were ignored with each of these models, such as the sparsity of domain borders throughout the genome. Likewise, data pre-processing techniques such as normalization, the elimination of low variance predictors, and variable selection techniques were not considered. Furthermore, due to the large number of genomic features that can be considered, accounting for the relationship among features with interaction effects becomes computationally infeasible.

Data pre-processing techniques such as normalization and the elimination of low variance predictors are an integral part of model fitting with regards to genomic data. Continuous predictors tend to be highly shewed. Therefore, a common practice is to perform a log base 2 transformation. It is also sometimes necessary to standardize continuous predictors, especially if they are on different scales. It is unclear, however, which combination of these two concepts improves prediction. Additionally, binary predictors may only be concordant with the outcome of interest in very rare cases. This contributes to predictors with near-zero variances. Eliminating these predictors prior to model fitting can reduce both the noise and the computational speed of the model. We use the nearZeroVar function provided in the caret package to determine which predictors have near-zero variance.

TADs can be up to a million base pairs in length and therefore the boundaries are sparse throughout the genome. Sparse data can create heavily imbalanced classes and may affect prediction performance of classification algorithms. There are many techniques that can be used to more evenly balance such data. These include oversampling the minority class, under sampling the majority class, and some combination of both. One such technique is referred to as SMOTE, which stands for Synthetic Minority Over-sampling TEchnique. SMOTE is a function in the DmWR package in R and incorporates both under-sampling and over-sampling. It has been shown that SMOTE out performs simple under-sampling of the majority class for some machine learning algorithms (Nitesh et al). We instead propose a method of taking multiple bootstrap under-samples from the majority class and then aggregating performance metrics by taking the average across all iterations. We then compare this method to SMOTE.

Modeling genomic data often involves a large number of predictors as well. As a result, different machine learning models can benefit from variable selection techniques to reduce the feature space, and thereby improve computational speed. There are several known selection techniques in the field of machine learning including forward, backward, and stepwise selection. These are known as wrapper methods because they measure the usefulness of a subset of features by actually training a model on it. Additionally, we incorporate recursive feature elimination as a method of variable reduction and compare it to the rest.

Furthermore, the model proposed by Mourad et al. falls short of addressing issues such as cross-validation and variable importance, which can be handled more efficiently through ensemble models like random forests and gradient boosting machines. Therefore, in addition to proposing a novel pipeline that addresses irregularities in the data mentioned above, we also apply a random forest model in order to find the key molecular drivers most associated with TAD boundaries. A random forest was chosen because of its ability to handle potentially correlated variables as well as its inability to overfit. We then compare the performance of the random forest model to the models proposed by Mourad et al.