5/24/2013

A Supervised Machine Learning Model for Gene Regulation during Differentiation of Embryonic Stem Cells

Matt Ziegler, advised by Mark Craven (Biostatistics and Medical Informatics, Computer Science,) Ron Stewart (Morgridge Institute,) and Scott Swanson (Morgridge Institute.)

Introduction

The identification of regulatory relationships among genes is an important and challenging part of genetic and developmental biology research, and can help us understand the mechanisms behind cell type differentiation. An important task in embryonic stem cell research is to understand how stem cells are "programmed" – how they are signaled to differentiate into different cell types.

To make inferences about which transcription factors (TF's) and µRNA's are involved in gene regulation during differentiation, we developed a supervised learning model to try to predict which genes have changes in expression during development. As evidence for the model, we used predicted binding sites of TF's and µRNA's in the promoters, enhancers, and 3' UTR's of each gene.

Methods

To learn about the regulatory mechanisms at work during differentiation, we developed a supervised machine learning method that tries to use information about which transcription factors and µRNA's might be binding to genes' regulatory regions to explain why some genes have changes in expression. The goal of the learning algorithm is to find a "rule" about transcription factor and µRNA binding which separates genes with expression changes from genes without expression changes. (An example of such a rule could be: "if TF A and TF B bind to a gene's enhancer, and µRNA C does not bind to the gene's 3' UTR, then the gene's expression will go up.") We focused primarily on a Multiple Instance Logistic Regression model as our supervised learner, but we also experimented with random forests, a logistic regression model, and a naïve Bayes classifier.

We built our model using gene expression data from the Bing Ren lab at University of California San Diego in 6 cell types: H1 embryonic stem cells, mesendoderm, trophoblasts, mesenchymal stem cells, neural progenitor cells, and IMR90 cells. Because all of the other cell types (except for IMR90) are derivatives of H1 cells, we can consider H1 cells to be before differentiation and the other cell types to be after differentiation. To set up our learning algorithm, we selected genes with more than 15fpkm expression in H1 cells, less than 15fpkm in all other cell types, and less than 5fpkm in at least one cell type to be our "positive" set of genes, and other genes that had more than 15fpkm expression in one cell type and less than 5fpkm expression in another cell type to be our "negative" set of genes.

Also from the Ren lab, we used computational predictions of enhancers from ChIP-seq data for histone marks. For transcription factor binding site evidence, we used ChIP-seq data for 61 TF's in H1 cells from the ENCODE project. For µRNA's, we used computational predictions of binding sited from the TargetScan database. In our model, we represent each gene as a vector of presence or absence of a binding site for each TF in the gene's promoter or enhancer, or for a µRNA in the 3' UTR.

To measure the accuracy of our model, we divided our genes up into a "training set" and a "test set." We train our model on the training set, use the model to predict expression changes in the test set, and compare the predictions to the actual expression changes. Because of the small number of genes in our training set (93,) we used a cross validation method to test our model – we divided the genes into 10 partitions and conducted 10 separate tests, holding back one partition from the training set and using it as the test set, and then averaging the accuracies across all of the partitions.

Results

By plotting our features (presence or absence of a TF or µRNA binding site) in ROC space (figure 1) and computing Fisher exact tests for over-representation of each feature in a particular class (positive vs. negative) of genes, we observed that no single feature can be used as an accurate predictor of a gene's expression change during differentiation.

We trained and tested several different machine learning models on our expression and binding data, and though all performed better than would be expected by guessing at random, none found a strong enough signal to make inferences about the regulatory mechanism at work (table 1 and figure 2).

Conclusions

There are several possible reasons why our models fail to find a signal in the data: key information might not be present in our data set (eg. a transcription factor that isn't measured,) the enhancer and promoter regions might not be identified accurately enough, or the learning method may be failing.

Possible next steps for the project are to adjust the parameters of our data models (eg. change how many bases wide we consider an enhancer to be when associating binding sites with enhancers, or changing which sets of genes we use as our positive and negative sets) to adjust the algorithm parameters, and to try additional algorithms.

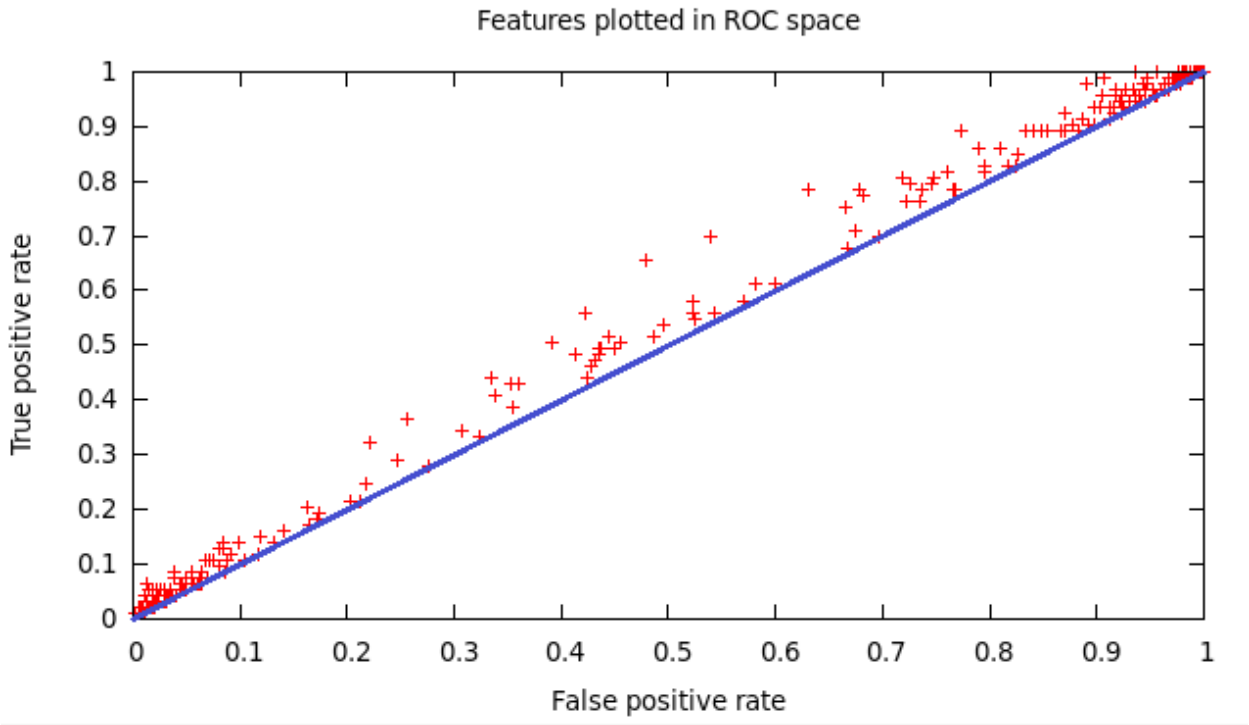

Figure 1: Features (transcription factor and μRNA binding site presence or absence) plotted in ROC space. The false positive rate and true positive rate are plotted for each feature, using each feature as a classifier of "positive" genes with H1-specific expression. The line from the origin to (1,1) shows the ROC curve expected from guessing randomly. None of the features can serve as an accurate predictor of a gene's expression change during differentiation.

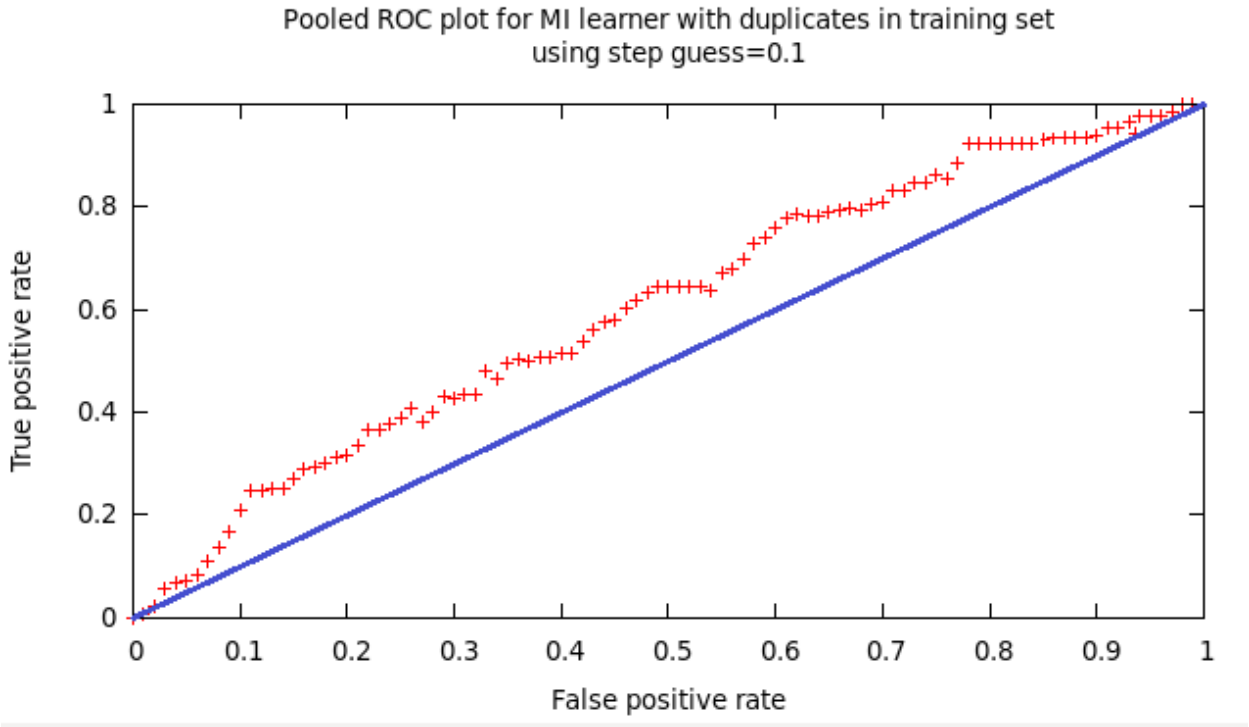

Figure 2: An ROC plot for the multiple instance learning model trained and tested on our data. The model performs only slightly better than random.

Model Type	**Area under Receiver Operating Characteristic**
Multiple Instance Logistic Regression	0.6091
Logistic Regression	0.55
Random Forest	0.596
Naïve Bayes Classifier	0.575
Random (baseline)	0.5

Table 1: Area under Receiver Operating Characteristic for 4 machine learning algorithms. An AUROC of 1 is a perfect classifier, and 0.5 is what we expect from a completely random classifier. All of the algorithms perform better than random, but none detect a strong enough signal to make inferences about which transcription factors are active during the differentiation.