Assignment Three: Supervised Machine Learning and Data Management Using Python

Transcriptomic analysis has given us unparalleled insights into how cells and organisms respond to changing growth and environmental conditions. Furthermore, as the number of sequenced genomes and gene annotations grow, we are better able to use pre-existing knowledge to hypothesize and explore the functions of important non-model organisms. Recently, the transcriptome of *Entamoeba histoytica*, the causative agent of amoebiasis, has been explored under varying growth conditions. This resulted in the identification of a number of short regulatory sequences upstream and downstream of the polypeptide start site which have been associated with enhancing virulence in this organism. However, it may also be possible to use machine learning, specifically supervised learning with Random Forests, to identify important regulatory motifs.

The original article can be found here:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-5570-z

Your task is to use supervised machine learning to identify which regulatory motifs are associated with gene up or down regulation within the two different treatment conditions (Serum Starved and Serum Replenishment). This project will focus on your ability to manage data and use a simple supervised machine learning model and introspection algorithm to analyze your results. I (Joe) will be making time next week and until the assignment is due for extra-help.

The input into the random forest classifier requires that your data have a specific format:

First, you need a matrix of genes (rows) and kmers (columns). The matrix should contain the counts of each kmer found within the gene. The columns of this matrix should refer to all possible kmers from all genes. An example using 2-mers can be found below. Second, you will need a list (below) that corresponds to the treatment condition and regulation status of each gene. Genes are found in only one condition. Finally, you will need a list containing the labels for each column in your count matrix, as below.

AA	AT	AC	AG	TA	TT	TC	TG	CA	CC	CC	CG	GA	GT	GC	GG

0	3	4	0	9		1
1	0	1	0	0		3
5	0	3	0	4		0
0	0	2	0	0		4
2	0	1	0	1	…	5
0	2	3	0	2		0
0	0	4	0	3		1
0	1	5	0	0		2

Ser S Up Reg
Ser S Down Reg
Ser R Up Reg
Ser S Down Reg
Ser S Up Reg
Ser R Down Reg
Ser R Down Reg
Ser S Up Reg

To make this matrix, I have provided you with three datasets: the genome, a list of differentially expressed genes under different conditions relative to a control, and the gff file. Using the data provided, you are to identify a 500bp region which flanks the AUG start codon of each differentially expressed gene. This region spans -400bp to +100bp with respect to the initiation codon of each gene.

Be careful how you treat DNA extracted from negative strand. To complete this task successfully you would need to do the following: For each gene's regulatory region, you will need to extract a list of all k-mers of length 12. We strongly suggest that you start with a smaller length because to make testing and finding errors in your code much easier.

Once the count matrix is constructed, you are to project it into two-dimensions using any of the visualization methods discussed in class on November 28th. Feel free to explore other projection and visualization methods.

Following this, you will need to train a supervised machine learning algorithm (A Random Forest or Extremely Randomized Trees Classifier) using the 'scikit-learn package'. Cross-validation using the Stratified Shuffle Split strategy should be used to gauge how well the classifier generalizes on unknown data.

Finally, we will be using the 'shap' package to extract the top 10 features (k-mers) which best predict the regulation status in each treatment.

The following links have information on these classifiers and approaches:

Cross-Validation:

https://scikit-learn.org/stable/modules/cross_validation.html

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score

Ensemble Methods:

https://scikit-learn.org/stable/modules/ensemble.html#ensemble

Projection Methods:

https://scikit-learn.org/stable/modules/decomposition.html#decompositions

https://scikit-learn.org/stable/modules/manifold.html#manifold

Questions (30 marks)

What was the mean cross-validation score and the standard deviation? Why is it important to cross-validate our models? What does the cross-validated performance tell us about our models (4 marks)?

What conclusions can you draw from your projection of this dataset (2 marks)?

How might your results change if you chose a different k-mer size (2 marks)?

Identify the top 10 putative regulatory motifs in each treatment condition. Were any of these found in the original study (Additional File 8)? What are the advantages and disadvantages of your method compared to the method presented in the paper (2 marks)?

The remaining 20 marks will be divided equally and depend on the quality of your code and figures, efficiency of your code, and commenting in your code.

Example Output With A K-mer Size of 4 (Figures):

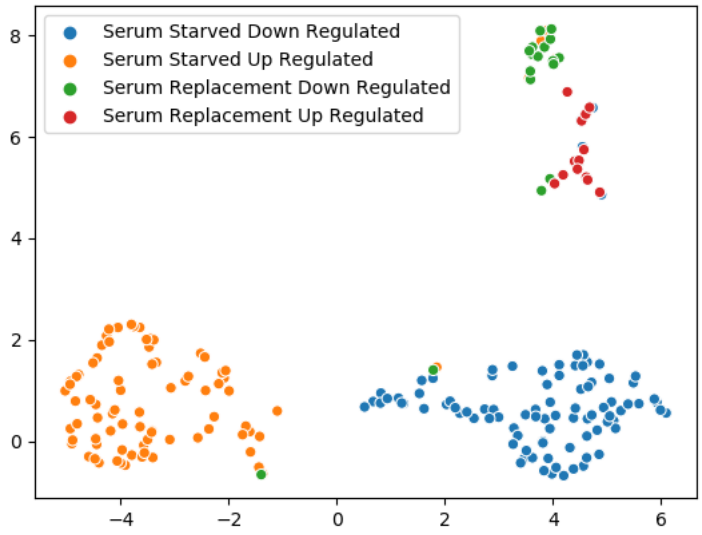

Figure 1: Projection of the 4-mer count matrix into 2D space.

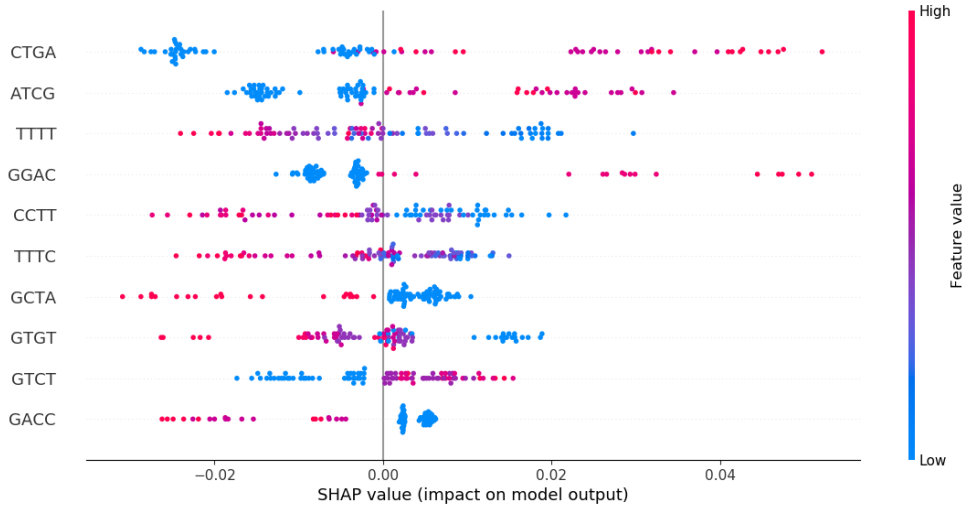

Figure 2: Important 4-mers identified after using the ‘shap’ package that best predict if a gene will be up or down regulated in the serum starved treatment. Note, positive SHAP values indicate that the presence of a particular 4-mer is predictive of the down-regulated state while negative SHAP values indicate that the presence of a particular 4-mer is predictive of an up-regulated state.